THE USE OF VIDEO AS A RESOURCE FOR THE DEVELOPMENT OF L1 FOUNDATION PHASE LITERACY IN ISIXHOSA: A DIGITAL MULTIMODAL DISCOURSE APPROACH

A thesis submitted in
fulfilment of the
requirements for the
degree of
Doctor of Philosophy
of
RHODES
UNIVERSITY

Sasha-Lee Schafli
https://orcid.org/0000-0002-7874-0776

Student number: g15S3552
Supervisors: Dr. Ian Siebörger & Prof. Mark De Vos
Department: Linguistics and Applied Language Studies

Abstract

Videos as a resource for African language L1 literacy development are strikingly absent from the South African school curriculum for the Foundation Phase (Grades 1 to 3). Furthermore, an overall lack of isiXhosa same language subtitling (SLS) practices in South Africa for videos poses questions as to the benefits and challenges of SLS for L1 early grade readers of isiXhosa. The aim of this research was to test, describe and analyse the efficacy of five isiXhosa *YouTube* story videos with SLS for the development of literacy in isiXhosa at the Grade 2 and 3 L1 levels. In this mixed-methods study, the five videos were exposed to Grade 2 and 3 learners at a school in the Eastern Cape in a quantitative two-group experimental design in a three-month intervention. This was to determine whether video exposure resulted in significant difference for learners' literacy, focusing particularly on word recognition, Oral Reading Fluency (ORF) scores and broad semiotic awareness between the linguistic and visual modes. The five videos were also analysed with a Digital Multimodal Discourse Analysis (DMDA), illustrating potential areas in which these five videos could assist or pose challenges to literacy learning in this context. While learners' reading scores improved over the three-month intervention, nonparametric t-test results indicate SLS video exposure did not make a significant difference in learners' reading improvements. Results from both methods were triangulated with cognitive theories of multimodal literacy and Mayer's principles of learning with multimedia. The analysis highlights that while the videos' design may be conducive for learning, the subtitle rate in the videos is far greater than the learners' reading scores in this study or reading speed benchmarks expected of Grade 2 and 3 learners. This can result in these videos being ineffective as a resource to improve literacy in isiXhosa for this level. This research highlights the importance of the integration of multiple methods of analysis for multimodal resources, as well as the importance of subtitle rate as a system within a multimodal analysis for literacy research. Furthermore, Comparative Relations in Intersemiotic Texture is proposed as a useful system for examining learning resources. This research further suggests areas of focus for future video design in potential L1 Foundation Phase literacy resources for isiXhosa and other African languages.

List of Abbreviations

ANA – Annual National Assesment

CAPS – National Curriculum Assessment Policy Statements

DBE – Department of Basic Education

DMDA – Digital Multimodal Discourse Analysis

EGRA – Early Grade Reading Assessment

ePIRLS – electronic Progress in International Reading Literacy Study

L1 – home language/mother tongue

L2 – Second Language/First Additional Language

LoLT – Language of learning and teaching

NGO – Non-government Organisation

ORF – Oral Reading Fluency

PIRLS – Progress in International Reading Literacy Study

PRAESA – Project for the Study of Alternative Education in South Africa

SF-MDA – Systemic Functional Multimodal Discourse Analysis

SLS – Same Language Subtitled/Subtitles

SCPM – syllables correct per minute

SPM – syllables per minute

WCPM – correct words per minute

WPM – words per minute

Glossary

1 – 1st class

1a – class 1a

2 – 2nd class

2a – class 2a

3 – 3rd class

4 – 4th class

5 – 5th class

6 – 6th class

7 – 7th class

8 – 8th class

9 – 9th class

10 – 10th class

11 – 11th class

14 – 14th class

15 – 15th class

17 – 17th class

ANT – anterior

APPL – applicative verb extension

AUG – augment

CAUS – causative verb extension

DEM – demonstrative

DISJ – disjoint

FUT – Immediate Future

FV – final vowel

INF – infinitive

INSTR – instrumental case or instrumental prefix

LOC – locative case or locative prefix

NEG – negation

OM – object marker

PASS – passive verb extension

PERS – persistive

PL – plural

POSS – possessive pronoun or possessive prefix

PROX – proximal

PST –Recent Past or Past form of auxiliary verbs

REM– remote (referring to remote past or future tenses)

RM – relative marker

SG – singular

SM – subject marker

STAT – stative verb extension

SUBJ – subjunctive

Table of Contents

Figure List

Table List

Acknowledgements

It takes a village to raise a child and it takes a community to raise a PhD. This thesis would not be a reality if it were not for the support from the people on this page. My gratitude knows no bounds.

I am immensely grateful to Dr. Ian Siebörger, the most patient supervisor, who has never stopped believing in me from when I first approached him in my Honours year about postgraduate studies in multimodality. As the hymn goes, "We have an anchor, that keeps the soul, steadfast and sure as the billows roll," and our God, the anchor and rock to which we cling, was incredibly benevolent to send me such a kind and supportive supervisor.

I have been the luckiest student to not only have received excellent guidance from one supervisor, but two, and truly the best team with Prof. Mark De Vos, who has also watched me grow in the department. Thank you for your constant support throughout the years and our long discussions in the tearoom. I am beyond grateful for your guidance and mentorship. I thank both supervisors for their advocacy and support, which led to the presentation of this research in Spain, across the ocean in 2023, at the 32nd European Systemic Functional Linguistics Conference (ESFLC).

I would further like to extend my deepest gratitude to coordinators, delegates, and presenters at the ESFLC in 2023 for their guidance and feedback on this research. This research would also not be attainable without the National Research Foundation (NRF) for their funding. I am immensely grateful for their support.

A writer is nothing without their team of editors, and this thesis would not be possible to read without the assistance of Stephen Hunt and Erin Alexander. There are no words to describe how eternally thankful I am for your aide in the editing, as well as your unwavering emotional support, and to have such dedicated people in my life as you both. Erin, you have watched my career from day 1 and walked with me on this journey. I have no words to express my gratitude for your steadfast loyalty and invaluable friendship. While I submit this to receive a PhD, I already possess the highest honour in having you both in my life. I love you both.

For all their support up to this point, I would like to express my deepest gratitude to the Small Projects Foundation and Kistefos, with special thanks to Dr Paul Cromhout, Luke Baisley and Hilton Williams, for all their assistance and advice in this project. My heartfelt appreciation goes to the principal and teachers who allowed me to conduct the research at the school. To Nonkosi Vuyo and Velwano Velile Mpikashe, my sincerest thanks and respect for your invaluable hard work and cooperation in this project.

While most of this thesis was created during the pandemic and the tearoom was not yet open, technology afforded the possibility for the community to remain in contact and I am forever grateful to have been a part of such a community. I am truly grateful to Hlumela Mkabile and Onelisa Slater for your translation aide, Tracy Bowles for your research advice, Paige Cox for your statistics assistance, Jarryd Futcher for your methodology advice and Stefan Savić and Camilla Christie for your grammar advice. I am not only grateful to have had your aide, but even more grateful for your emotional support as friends and colleagues.

This thesis occurred during one of the hardest times of my life and I would like to thank both the staff at T Birch & Co (PTY) LTD and the staff at the Confucius Institute for their emotional support as I continued to reach for that red gown. Gisela Zipp and Hanyang Zou, I am forever appreciative of your friendship. Friends are like precious flowers that you cherish forever, and I am especially grateful to Zakeera Docrat for all of your support and belief in me throughout the years. Landing this plane and opening a new chapter of my life, I also wish to extend thanks to my colleague and friend Stuart Whittle, who continues to remind me of the power of human kindness.

Many started this race with me, but few ended with me. This is to thank those of you who started this journey with me and contributed in many ways to this thesis, both knowingly and unknowingly, but left before I had won the race. My undying gratitude also goes to my parents, Stanley and Yvonne Kriel, and to my sister, Jolene Kirton, for all their moral support in achieving the dream I have had since I was a small child. I am indebted to my Heavenly Father for giving me the strength to survive through this degree and produce this dissertation. All the glory goes to Him.

Chapter 1 – Introduction

A tradition of spreading knowledge through written texts and literacy began with Egyptian hieroglyphs in Africa and other ancient writing systems in Asia and South America at approximately the end of the fourth millennium BC and it continues today throughout the world (Wolf 2012). This research broadly examines the use of isiXhosa same-language subtitled (SLS) videos as a resource for first language or mother tongue (L1) literacy development in this language at an early stage of literacy development (second and third year of primary school education, within the Foundation Phase). It consequently focuses on opportunities and challenges in isiXhosa L1 literacy development. While the practice of reading seems commonplace in a Global North context, many of the early writing and reading practices developed in the Global South. It is therefore surprising that, in a digital age where reading and writing practices are commonplace and normalised, South Africa in the Global South has been left behind in providing adequate literacy for its children. Reading has become central to epistemological access of knowledge in the world today, but South Africa is currently facing a literacy crisis (Howie *et al.* 2006, Howie *et al.* 2016) in all 12 of its official languages, including isiXhosa (du Plessis 2023). IsiXhosa is an official language of the Eastern Cape Province in South Africa, which is where this research is situated, and has the second-highest population of L1 speakers in the country (Statistics South Africa 2011, Made 2010).

Due to a colonial history, and assisted by a later apartheid regime that ended in 1994, Standard English and Afrikaans were the languages that received the most research and attention, despite the majority of South Africans speaking African languages of the Bantu language family (Rees 2016).[1] In 1994, upon the commencement of democracy post-apartheid, South Africa recognised 11 official languages, which are listed according to the percentage of home language speakers recorded in the 2011 census. These are isiZulu (28.55%), isiXhosa (20.03%), Sepedi (11.41%), Setswana (9.85%), Sesotho (9.41%), Xitsonga (5.58%), SiSwati (3.38%), Tshivenda (2.97%), English (2.97%), isiNdebele (2.62%), Afrikaans (1.49%) (Statistics South Africa 2011). A 12th official language, South African Sign Language, was

[1] The use of the term 'Bantu'' is used internationally for a specific language group within the Niger-Congo language family and not the derogatory term that referred to people under the apartheid regime. Nevertheless, I acknowledge that the term has derogatory connotations within South Africa, but it is used in this thesis as a technical linguistic term purely to refer to the language family in question.

included in 2023 (0.5%) (du Plessis 2023). Yet the education of learners of colour was characterised by poor funding, a lack of teacher training, inadequate facilities, sparse learning materials in the Bantu languages and over-crowded classrooms influenced by the social context from colonial rule through to the apartheid regime (Webb *et al.* 2010). Despite the attempts to redress this by the current government, Howie *et al.* (2006; 2016)'s discussion of factors that influence poor literacy rates (discussed further in section 1.1.1) demonstrates that schools' face many of the same challenges in present day South Africa.

Digital technology has created a digital divide, which has led to social imbalances in skills, as well as political and economic power (Mossberger et al. 2003). Without technologisation, South Africa may fall behind in terms of global economics, a factor that as we have seen earlier in this section, has contributed to the growing literacy crisis (Howie *et al.* 2016). South Africa is also falling into an energy crisis which may further impede our ability to grow in the realm of technology (see Heinemann 2019). Despite the challenges in the environment, we have to rise to overcome these obstacles, creating new knowledge as the tough South Africans that we are in the face of adversity.

To understand how isiXhosa literacy and text relate to a broader set of systems in the world, how these systems relate to one another, and how this thesis is structured, this thesis understands literacy as a semiotic process of interpreting and re-creating meanings through a semiotic system (Jewitt *et al*. 2016). Semiotic meanings (often conflated with semantics in formal linguistics) arise from choices from different resources in various modes (O'Halloran 2023) for example, language and animation in the case of isiXhosa SLS videos used in this research. As O'Halloran (2023) describes, people understand their physical environment with biological systems (made up of physical matter and biological matter) using combinations of different senses (e.g. hearing and sight). They comprehend and construct meaning in the world using meaning-making resources. "Sensory input" from the world and environment is influenced by "social factors" that are comprised of life experiences, values and beliefs, culture and context (social systems) (O'Halloran 2023:176). Physical, biological, social and semiotic systems all contain matter (and are material) but can be ordered in terms of complexity. Physical systems consist of matter (O'Halloran 2023). Biological systems (biological and life systems) further include physical systems (O'Halloran 2023). Social systems include both biological and physical systems, but also social organisation (O'Halloran 2023). Semiotic systems have all of these former systems, including meaning (O'Halloran 2023). According to

O'Halloran (2023), semiotic frameworks have to consider the four systems and dimensions within these systems, as a change to one system could create a domino or ripple effect across the entire "meta-system" (O'Halloran 2023:174).

This thesis understands isiXhosa SLS video texts as a physically, biologically, and socially produced set of semiotic resources. It further investigates its impact on people as biological systems that use both physical, biological, and social systems to interpret semiotic resources for meaning. Thus, literacy in this thesis is understood as a semiotic process, which encompasses social, which include cultural and contextual aspects of literacy, as well as physical and biological domains, which include physical and biological, neurological and cognitive structures and processes of literacy. In this mixed-methods research, five videos were exposed to Grade 2 and 3 learners at a school in the Eastern Cape in a quantitative two-group experimental design in a three-month intervention. This was to determine whether video exposure resulted in significant difference for learners' literacy, focusing particularly on word recognition, Oral Reading Fluency scores and broad semiotic awareness between the linguistic and visual modes. The five videos were also analysed with a Digital Multimodal Discourse Analysis (DMDA), illustrating potential areas in which these five videos could assist or pose challenges to literacy development in this context.

The introduction focuses on the social systems aspect from the next paragraph until 1.2.2. This includes the literacy crisis context in South Africa (1.1.1), the further implications for the COVID-19 pandemic and digital learning (1.1.2), and SLS videos in a literacy context (1.1.3). By examining isiXhosa L1 SLS videos and their influence on literacy rates in this language, it may potentially assist in alleviating the literacy crisis highlighted above. The COVID-19 pandemic's influence on SA education (1.1.2) provides further insights, which are foundational to the understanding of this research context. SLS videos as potential tools for literacy improvement is described in other contexts, therefore this thesis examines literature on this semiotic system (1.1.3).

This thesis examines instances of isiXhosa L1 subtitles, based on written conventions of the language that follow the standard dialect(s). IsiXhosa is an official language of the Eastern Cape, the most spoken language of the province and spoken by the amaXhosa people (with amaGcaleka and amaRharhabe as standard dialects), as well as isiXhosa-speaking clans (Bhaca, Mpondo, Mpondomise, Thembu, Bhele, Hlubi and Zizi) (Made 2010; Mtsatse and van

Staden 2021). The ancestors of the amaXhosa migrated to the Eastern Cape during the 1200s and the isiXhosa-speaking groups came from Natal as war refugees, settling in the region of the amaXhosa in the late 17th century, assimilating and adopting their culture and language (Mtsatse and van Staden 2021). However, within the dialects of the isiXhosa-speaking groups, there remain slight differences from the original dialects (Mtsatse and van Staden 2021). Examining the factors that different dialects have on the literacy rates is beyond the scope of this thesis. It is important to note for this research, however, that the subtitles in these videos adhere to the isiXhosa standard dialect(s) writing conventions. This study seeks to contribute towards alleviating the crisis for this language, by examining how standardised isiXhosa SLS videos could potentially assist L1 literacy for this language. To discuss isiXhosa and literacy rates, it is important to understand the language in its socio-cultural context (social system) (1.2). Due to this being a linguistic study and analysis that examines language and visuals, an understanding of the nature of the isiXhosa language is integral to understanding the findings of this research. Thus, section 1.2.1 introduces the linguistic features of isiXhosa, as language is one of the many semiotic systems in this research, defined by O'Halloran (2023). Section 1.2.2 discusses relevant policies on literacy for this language and their connection to multimodality and literacy education, which is another aspect relating to the semiotic systems category. This research, as stated in its title, undertakes a Digital Multimodal Discourse Approach. Multimodal approaches are required to fully grasp the complexity of semiotic systems (O'Halloran 2023). This leads to an introduction of the concept of multimodality and other important concepts of this research in this section to understand why this approach has been undertaken (1.2.3). Section 1.3 introduces the research goals and questions of this thesis and finally, section 1.4 concludes this chapter.

1.1 Further motivations for study

As introduced by this chapter, the examination of isiXhosa SLS videos as a resource is to understand its influence on isiXhosa L1 literacy and observe how it would operate in a South African context, in which exist a literacy crisis and the occurrence and after-effects of the COVID-19 pandemic. The purpose of this section is to highlight the motivation and rationale for this study, and to highlight the contexts for the semiotic systems and social systems of isiXhosa as a language and in L1 education in South Africa. In section 1.2, Contextual information (social system information) is discussed in relation to the literacy crisis (1.1.1) and

the COVID-19 pandemic (1.1.2) in this section in more detail. This research seeks to potentially assist in alleviating the literacy crisis. The recent COVID-19 pandemic has further indicated the lack of digital literacy and resource access in many parts of the country when face-to-face contact had been limited. This research illustrates whether SLS videos in this context are useful or not in terms of L1 isiXhosa literacy education. This research could further highlight potential opportunities and challenges in the South African context for L1 isiXhosa literacy development with this resource post-pandemic at the Foundation Phase, a phase of early reading development (see Spaull *et al.* 2020; Probert 2019).

SLS videos as a potential tool for the improvement of literacy is also introduced in relation to previous research in other languages and contexts (other social systems) in 1.1.3, which consequently supports the undertaking of this research. In conclusion (1.1.4), I summarise the gaps in the research which are further addressed in the Literature Review in Chapter 2.

1.1.1 Literacy crisis

The Progress in International Reading Literacy Study (PIRLS) in 2006 called attention to a crisis in literacy in South Africa. South African learners in the Intermediate Phase were unable to read in their L1 and South Africa's literacy scores were the lowest in the 40 countries that participated in the PIRLS assessments (Howie *et al.* 2006). School samples in PIRLS came from majority peri-urban areas (62%), with 21% suburban schools and 17% urban schools that approximately total to 30 000 children from 400 schools (Howie *et al.* 2006). The findings of this study thus illuminate potential low literacy scores and literacy issues for the research site used in this thesis (discussed in 4.2.1.1). While the literacy crisis affects schools in all areas, whether urban or peri-urban, often access to resources and technology is more challenging for schools in peri-urban areas. The research site investigated in this thesis is a school in a countryside, but due to its proximity to a city, can be classified as peri-urban. For isiXhosa in particular, Grade 5 readers had the lowest literary reading score (approx. 225 points from the written test battery where the International Average was 500). This was in comparison to other languages, where isiXhosa scores had the second lowest after isiNdebele for Grade 4 readers (at approx. 190 points) (Howie *et al.* 2006). One percent of isiXhosa Grade 4 readers and four percent of isiXhosa Grade 5 readers could not meet the Low International Benchmark or 400 points on the scale of achievement in PIRLS, with no learners meeting higher benchmarks (Howie *et al.* 2006). Howie *et al.* (2006) correlate reading success in South Africa to

educational and financial advantages (the latter also including access to more literary resources) and parental engagement in reading activities individually and with children, particularly prior to schooling.

This is combined by the 2019 statistics revealing that 12% (4.4 million) of adults are illiterate in South Africa, which is not far ahead of the global average of 14% (Khuluvhe 2021). This figure remaining highest amongst Black South Africans in comparison to any other racial grouping (Khuluvhe 2021). Statistic figures like this, however, do not correlate with the literacy score results found in the PIRLS studies (Howie *et al.* 2006; 2016), as qualitative answers to whether a participant in a survey can read or not do not consider the extent of a participants' reading ability. The state of literacy or illiteracy is not necessarily a dichotomous state, as the PIRLS (Howie *et al.* 2006; 2016) study indicates that learners can read to some level, but this level is too low for full comprehension and understanding of what is being read. Thus, over 80% of the population, according to PIRLS (Howie *et al.* 2006; 2016), are literate, but functionally illiterate and the score of 12% may potentially be a larger number than participant answers indicate. The previous educational disadvantages for this language exacerbated by colonialism and the apartheid regime present one of the reasons why many parents, even if they wished to assist their children with reading, are simply unable (Department of Higher Education and Training 2021).

Webb *et al.* (2010) state that language is a determining aspect in learners' poor performance. English as a preferred language of learning and teaching (LoLT) by non-L1 English speakers, who often do not have adequate English language proficiency (Webb *et al.* 2010). This claim by Webb *et al.* (2010) is supported by the statistics of the South African Social Attitudes Study (2001-2018), early responses are discussed by wa Kivulu and Morrow (2006). According to wa Kivula and Morrow (2006), all except the Western and Northern Capes favoured the use of English for Grades 1-3 as LoLT, while 44% of the Eastern Cape would prefer to use isiXhosa. According to Webb *et al.* (2010), due to schools teaching African languages of the Bantu language family perceived as providing a lower standard of education, Bantu languages in education became associated with low quality education. Supported by a binary logistic regression of the data, wa Kivula and Morrow (2006) discuss that people of colour more than white South Africans are more likely to select English as the main language of instruction, as well as participants with higher income and who are younger. L1 Education was only selected by white people, people with high personal income or those with low level education (wa

Kivula and Morrow 2006). Webb *et al.* (2010) claims this stems from British colonisation until 1961 and former apartheid racial segregation of education facilities through the Bantu Education Act from 1954 to the end of apartheid. This historical context as shown by Howie *et al.* (2006; 2016) indicate the social and historical disadvantages of these languages having ripple effects on a general lack of resources and a further lack of reading culture in the home in the present day.

Another PIRLS study was conducted in 2016 and illustrated not much had changed from 2006 in terms of the crisis. South Africa's literacy score for Grade 4s (320 points, whereas the average was 500) still ranked the country at the bottom of participating countries (Howie *et al.* 2016). The Eastern Cape ranked second last of the provinces and 100% of Eastern Cape schools had participated in the study (Howie *et al.* 2016). The score for isiXhosa readers was 283 (whereas the mean score was 500), possibly due to a broader sample of Grade 4s, but ultimately it scored penultimate before Sepedi (Howie *et al.* 2016). In South Africa, 78% of all Grade 4 learners did not meet the Low International Benchmark and most learners of isiXhosa Grade 4 learners (88%) are unable to read for meaning, which indicates that learners have not achieved the Foundation Phase outcomes (Howie *et al.* 2016). Of learners attending schools in peri-urban areas, 89% were unable to meet the Low International Benchmark; along with 85% of learners living in small towns and villages and 83% of learners attending township schools. In contrast, 53% of learners attending urban schools did not achieve the Low International Benchmark and 48% of suburban schools did not achieve the Low International Benchmark, indicating geographical location and remoteness of the school also plays a role in literacy rates. Once more in the 2016 PIRLS study, a lack of school resources remains a factor in poor literacy scores (more on this discussed in 1.2.) and safety, school environment and climate were also reported as potential factors that influence literacy scores (Howie *et al.* 2016).

A summary report for the Nguni languages by the Department of Basic Education (DBE) has established Oral Reading Fluency (ORF) benchmarks for the Foundation Phase as follows: 40 letter-sounds correct per minute at the end of Grade 1, 20 words correct per minute (WCPM) at the end of Grade 2 and 35 words correct per minute (WCPM) at the end of Grade 3 (Ardington *et al.* 2020). These were established with the Nguni languages isiXhosa, isiZulu and SiSwati in mind, as there was not enough sample data to explore this with isiNdebele. While these benchmarks have been set quite recently, the report tells us meeting these benchmarks remains a concern (Ardington *et al.* 2020). By the end of Grade 3, 53% and 76%

of learners in their sample had only reached the minimum 20 WCPM threshold expected at the end of Grade 2 and only a quarter of learners reached the 35 WCPM benchmark by the end of Grade 3 (Ardington *et al.* 2020). Furthermore, the report highlights the complex issue of comprehension measurement, which caused comprehension thresholds to not be included within the set benchmarks that only measure Oral Reading Fluency (ORF). Ardington *et al.* (2020) state that comprehension levels of what learners were reading were low across the board. Thus, there is a danger of setting up learners to be able to read letters on a page, but not understanding what they are reading.

Other studies after the 2016 PIRLS have indicated that learners are still not meeting the benchmarks for reading set in 2020. Spaull's (2020) study indicate that isiZulu readers reached a median of 19 WCPM in the beginning of their Grade 3 year in 2017 from 65 schools in Gauteng, Kwazulu-Natal and Limpopo. However, Probert (2019) illustrates that isiXhosa readers in the Eastern Cape from 2 schools reached a median of 12 WCPM at the final quarter of Grade 3, showing they had not reached the minimum benchmark for reading set at the end of the Grade 2 year. While Probert's (2019) sample size is far less than Spaull's (2020) study, the schools in her study are from low socio-economic class communities situated in peri-urban townships in the Eastern Cape, which is the province where this study is based. This illustrates the potential outlook for isiXhosa and schools in the Eastern Cape.

Meiring (2020) conducted an intervention study in 2019 with the literacy NGO Funda Wande, and her data of thirty experimental group schools and twenty-nine control schools illustrates the reading norms in a typical school year without a literacy intervention. In the study, 50% of learners could not read a single word correctly at the end of Grade 1, with 16-24% of learners unable to read a single word correctly at the end of Grade 2 (Meiring 2020). When they could read words, the average WCPM was 10 for Grade 1 and 20 for Grade 2 (Meiring 2020). However, one should take into account that the no-fee schools in this control group were urban and peri-urban schools. Furthermore, schools were excluded on the bases of classroom overcrowding (not more than fifty per class), management difficulties, or less than twenty learners in a grade (Meiring 2020). Consequently, while these WCPM rates match the benchmarks in the 2020 report, these schools may not entirely reflect conditions in low socio-economic class communities and conditions in peri-urban townships in the Eastern Cape. According to the Eastern Cape Provincial Government, more than 1500 schools of the 5451 schools in the Eastern Cape are overcrowded, with 1598 schools having inadequate sanitation

during the 2020 academic year and these factors also play a role in literacy rates in the province (Phandle 2021). Nevertheless, for Grade 1 learners, Meiring (2020)'s findings with isiXhosa are similar to Schaefer and Kotze (2019)'s findings with isiZulu. Schaefer and Kotze (2019) found that only 45.7% of learners could read a word in isiZulu correctly at the end of Grade 1 in Mpumalanga. This illustrates that the crisis of literacy remains an issue, in which a vast number of learners in the Eastern Cape and South Africa remain pre-literate after Grade 1. Learners cannot reach the 20 WCPM threshold at the end of Grade 2 in isiXhosa and languages with a similar orthography, such as isiZulu. All of these studies took place before the COVID-19 pandemic, which has further disrupted literacy education in South Africa. Foundation Phase learners' inability to read this early further indicates that measures to address the literacy crisis need to start in early reading phases of education.

1.1.2 The COVID-19 pandemic and online/digital learning

COVID-19 has further resulted in learners in South Africa experiencing disruptions in their schooling, which has naturally worsened the current literacy crisis. The COVID-19 virus, which appeared in South Africa in the beginning of 2020, caused schools to experience disruptions, due to multiple lockdowns and lack of face-to-face contact **(**Ramrathan 2020). For example, the implementation and assessment of literacy with many paper-based assessments, including Early Grade Reading Assessments (EGRA), was disrupted. This has had broader socio-economic impacts on schools as the pandemic also caused funding to schools and literacy non-government organisations to be reduced or suspended.

The Early Grade Reading Assessment (EGRA) "is a tool used to measure students' progress toward learning to read" (Gove and Wetterberg 2011:2) and an adapted Familiar Word Reading section and an Oral Reading Fluency section for the isiXhosa assessment are used as tests in this thesis to measure intervention results. This is discussed in Chapter 4 of this thesis. The entire assessment is a spoken test, administered to one learner at a time, and assesses basic pre-reading and reading skills in about 15 minutes in total. The EGRA framework was developed by RTI International with a range of education experts to implement them in more than 50 countries and 70 languages (Gove and Wetterberg 2011). These assessments have caused many legislators and teachers to seek resolutions for challenges in literacy, which the EGRAS have illustrated for many countries, including South Africa (Gove and Wetterberg 2011). For example, the school (research site) in this thesis was part of a literacy intervention that had

begun to record EGRA results from 2019, but the entire programme had to be suspended, primarily due to COVID-19. Thus, EGRA 2020 results were not collected for that year.

Most face-to-face school activities were forced to try and move online during the COVID-19 lockdowns. In 2016, South Africa was involved in the 2016 electronic adaptation of the Progress in International Reading Literacy Study (ePIRLS), but due to numerous schools having obsolete or malfunctioning equipment and unused computer laboratories, there were too many challenges for ePIRLS implementation (Combrinck and Mtsatse 2019). Consequently, the ePIRLS requirements were not met and South Africa was not included in the report (Combrinck and Mtsatse 2019). This illustrates the access challenges to online resources and technology that persist in South Africa that was highlighted particularly during the COVID-19 pandemic (Ramrathan 2020). This also projects an impact for PIRLS in future years. The Department of Basic Education (DBE) wants to open online schools, and private education institutions like Curro online have opened an online school from Grades 4-10 (Bolowana 2021). However, data and access remain a difficulty for online learning to occur in South Africa, unless they are on zero-rated[2] platforms (Ramrathan 2020). The Covid-19 pandemic has presented opportunities for new online learning resources to be developed, and for new ways to be found to use these resources.

Furthermore, online learning initiatives seem to be scarce for isiXhosa, and there is a lack of research on their effectiveness. While the DBE has an interactive workbook that runs on Adobe AIR for isiXhosa Grade 1, Adobe AIR has been discontinued (Adobe Corporate Communications 2019). The DBE website has a small selection of books in isiXhosa for Foundation Phase learners (DBE 2011). Electronic readers have been made available for learners in partnership with Vodacom, MTN, Telkom and Cell C (Malinga 2020), with Vodacom having its own online e-school of resources, but one has to have teacher access, mainly from Intermediate to Senior Phase. Same language subtitles (SLS) or intralingual subtitled videos are not available on these service provider websites or on the DBE website, yet they have been shown in other languages and contexts to be helpful for children in language learning and learning to read. Therefore, it is imperative that we explore resources such as subtitles that could potentially assist in alleviating South Africa's literacy crisis. This is discussed in the next section.

[2] Zero rated means free or without charge to access and use the platform (Ramrathan 2020)

1.1.3 SLS videos as a potential learning resource for literacy

Videos as tools for learning are not new in the classroom. While videos have been shown to be successful resources in foreign language learning contexts, particularly with English as a foreign language (e.g., McConachy 2018), video as an L1 learning and literacy resource seems to be relatively under-researched in South Africa. Diab *et al.* (2016) found that video was beneficial for recall, pronunciation, communication, as well as vocabulary learning with adult medical students, who were isiZulu L2 speakers. Learners who interact with SLS videos can better define content words and pronounce novel words, infer happenings in the videos and recognise lexical items (Gernsbacher 2015).

Kothari and Bandyopadhyay (2014) built on two earlier studies. A controlled experiment examining decoding abilities of children after 3-month exposure to SLS Bollywood videos in Hindi (Kothari *et al*. 2002), and a second six-month study which assessed the decoding skills of youth and adults (Kothari *et al*. 2004). Decoding skills refer to the set of skills for mapping written text to sound without necessarily comprehending meaning (Pretorius 2010). In both of these earlier studies, exposure to SLS videos contributed positively to decoding skills (Kothari and Bandyopadhyay 2014). The researcher doubled weekly exposure and extended the overall period of exposure, reporting the following results: "Where schools are failing to get nearly 45% to decode a single letter after five years of schooling, the SLS complement at home brought this shocking figure down to 13%" (Kothari and Bandyopadhyay 2014:44). Due to the success of SLS video exposure in India in improving literacy, I draw upon this research in India when conducting the intervention in this thesis. These studies are discussed in more detail in section 2.2.6.

Subtitled videos are not discussed as literacy resources by the DBE. One reason for this in the case of isiXhosa may be due to the lack of videos that exist. Two channels on *YouTube* contain SLS isiXhosa videos that are on the public domain: *Xhosa Kids* and *MPBXhosa*. While the channel *Xhosa Kids* contains isiXhosa songs, the videos come from English nursery rhymes and songs and were dubbed and translated over to isiXhosa, with many inconsistent word boundaries, capitalisations, and spelling errors, for example, "Igqirha lendlela ngu qongqotwane" instead of "IGqirha lendlela nguqongqothwane" (the diviner of the road is the beetle) (Xhosa Kids 2019; Simelane-Kalumba 2014:43). This was the reason videos from this channel could not be used in this study. *MPBXhosa*, the channel selected for this study, has

translated stories from English to isiXhosa. These translations are linguistically accurate. They were designed from a collaboration between the creators of animated children's book Moving Picture Books and an NGO, The Project for the Study of Alternative Education in South Africa (PRAESA), based in Cape Town. The *YouTube* channel *MPBXhosa* has a series of SLS videos of stories, five of which are analysed in this thesis.

Potential challenges with interlingual subtitles have been reported that may also occur with SLS. Lavaur and Bairstow (2011:457) list various potential difficulties with learning from subtitles. These include information loss from condensed subtitles, subtitle distraction from visuals or visual distraction from subtitles, limited reading time and challenges with the spatial positioning of subtitles (Lavaur and Bairstow 2011:457). In South African research specifically on subtitles, difficulties were found in L1 Sesotho and L2 English subtitling of a French spoken language video (Hefer 2013a). Adult participants were shown to spend too much time on subtitles and while reading was slightly faster for the English subtitles than Sesotho subtitles, reading rates and comprehension of subtitles were considerably low in both languages. In other words, the viewers read too slowly to fully follow and comprehend the text. This is in comparison to adult Afrikaans-English bilingual speakers in which L1 Afrikaans subtitles were read faster and more efficiently than L2 English subtitles (Hefer 2013b). In addition, within a different educational context, Matthew (2020) reported that subtitling English lectures made very little difference in comparison to when students with English as L2 viewed non-subtitled lectures. This indicates that challenges and opportunities vary according to participants' literacy levels and the languages involved. He further highlights that more studies are needed in the other African languages and also in earlier levels of education (Matthew 2020), which this study explores with isiXhosa L1 subtitle reading in the Foundation Phase. This literature is discussed in further detail in Chapter 2.

1.1.4 Conclusion

Based upon the discussions in this section, the following reasons for this thesis can be summarised as follows:

1. South Africa is undergoing a literacy crisis where most of the learners cannot reach lower literacy benchmarks in their L1, especially in township and peri-urban schools, according to the PIRLS studies, and isiXhosa is one of the languages that rank at the

bottom in terms of literacy scores. The PIRLS studies highlighted a lack of resources, both in terms of books and digital resources, that correlated with poor literacy scores.

2. The literacy crisis has been further exacerbated by the disruption of schooling, assessments and funding for literacy interventions due to the COVID-19 pandemic. The pandemic has highlighted the lack of access to online and digital resources for learning and literacy in South Africa.
3. Studies exist in other contexts and other languages that show SLS videos as useful tools for the improvement of literacy, however the DBE has not recommended SLS videos or intralingual subtitled videos as literacy resources, and in the case of isiXhosa this may be due to the lack of resources that exist. This is further highlighted in the literature review in Chapter 2.
4. Current literature on African language subtitling with adults indicates that viewers may read too slowly to comprehend subtitles (Hefer 2013a). This is a potential challenge that this research investigates.

This study seeks to analyse five SLS story videos on the *YouTube* channel *MPBXhosa* and conduct an intervention to ascertain whether SLS videos can assist in improving literacy for isiXhosa in South Africa. As established earlier, we need more research in the areas of early literacy and technology for the transformation for isiXhosa in terms of the use of this language in digital subtitled media and the empowerment of isiXhosa L1 speakers from this transformation. Furthermore, we need more research on literacy resources and their development to help alleviate the literacy crisis in the country for this this particular area of isiXhosa language transformation. By focusing on isiXhosa L1 literacy at the Foundation Phase level, this research contributes knowledge on subtitles in this context and specifically for this language. It provides knowledge on isiXhosa L1 literacy teaching and learning practices in Foundation Phase for educational bodies. This includes staff and learners in schools, as well as all educational practitioners and policy makers in the private and public sector.

This thesis examines isiXhosa and due to the history of colonial oppression, my positionality in this research is extremely vital to contextualise the observational lens of this research. My position is that of a white woman researcher, with privilege as a white researcher. However, I have less privilege as a woman, further situated in a lower social class level within my racial group, living in a rural context within Makhanda, in the Eastern Cape. Makhanda is a small town with a deteriorating infrastructure and water shortages (Pamla *et al.* 2021). I grew up in

a multilingual home with a father that could speak English, Afrikaans and isiXhosa fluently, many of my family members have spoken fluency in isiXhosa, and it was a language that I grew up hearing at home. I was born in 1994, after apartheid had ended, and have been educated in isiXhosa since primary school, a benefit of being born free of the regime. As I have grown up and studied isiXhosa and African linguistics in my undergraduate degree, I have learned to recognise shared and differing aspects between my South African culture and the various South African cultures. While learning applied and formal linguistic methods through my undergraduate studies, I have been working towards applying the knowledge I have gained in other language teaching and learning contexts and investigating it in an African context with isiXhosa. This is to empower and build knowledge for a community that did not get many of the advantages I gained.

My approach to this research is not to enter a community, as many white researchers have done in the past, with suggestions of what to do and repeat the past of a relationship of racial power. Rather, in this research, I examine this research area by using methods formulated by researchers from the Global South and using my privilege to raise awareness of what is happening with the language and in communities that speak the language.

At every stage of this research, there has been participation from L1 isiXhosa speakers both inside and outside of the community. During the research stage, I collaborated with L1 isiXhosa linguists and translators on research methods. During the implementation of the intervention and pre- and post-tests, I collaborated with L1 isiXhosa speakers from the community of the research site in terms of data collection, as well as societal impact of this research. In collaboration with the community on this project, we have learned from one another. This research and conducting an intervention at a school in this community has taught me how SLS videos influence the L1 isiXhosa Foundation Phase literacy context. This was the primary outcome for this research, but I have learned many secondary outcomes from engaging with the community and conducting the intervention at the school. I hope that readers of this thesis can gain similar understandings and insights in this area. Through this intervention, the school has had an opportunity to reflect on current teaching and learning practices in the school in this research. The recommendations that come from this research seek to expand on knowledge that is still unknown, with future gaps in the research that provide L1 isiXhosa researchers further opportunities to expand in terrain and areas where I could not reach, due to my positionality. This is a crisis, and we all have a role to play as South Africans in doing what we can, while

knowing our limits. The next section (1.2.) examines the research context of isiXhosa and its use in classrooms in South Africa.

1.2 IsiXhosa L1 literacy in the Foundation Phase

This section examines linguistic features of isiXhosa (1.2.1), in order to inform linguistic analyses in Chapter 5 and 6 and to comprehend linguistic issues regarding literacy and literacy policies later in this section (1.2.2). It also discusses Foundation Phase education policies for L1 isiXhosa (1.2.2) and the gap between policy and reality, in order to highlight current issues in the context to take into account when examining videos as a potential resource for literacy. Videos as multimodal resources and key concepts of multimodality for this research are examined in 1.2.3, as this thesis uses a Digital Multimodal Discourse Approach. This section is concluded in 1.2.4.

1.2.1 Linguistic features of isiXhosa

As mentioned earlier in this chapter, isiXhosa (S41[3]) is a language of the Bantu language family, within the sub-group classified as Nguni languages (Probert 2019). Bantu languages were only introduced as a LoLT in missionary schools in the 19th century (Webb *et al.* 2010). In terms of isiXhosa specifically, the main reason isiXhosa was written down was for missionaries to teach to isiXhosa L1 learners effectively at schools (Maseko 2017). A missionary from the Glasgow Missionary Society, John Bennie, was the first to codify isiXhosa, producing the first printed text in the language in 1823 and his spelling system was devised as the standard (Oosthuysen 2015). He compiled the complete orthography, pronunciation, and spelling lists, as well as religious texts in isiXhosa with other missionaries (Maseko 2017).

The orthography for isiXhosa is relatively transparent or shallow (Probert 2016). This means that there is a regular grapheme-to-phoneme correspondence (Probert 2016). This is in comparison to a language like English with an opaque orthography, in which the letter /a/ can

[3] Guthrie (1971) code to identify languages.

represent many sounds, such as [æ] in “rat”, [ɑ] in “car” (Rao 2018) to name a few. To understand the orthography of isiXhosa better, this section examines first the phonology within the orthography (1.2.1.1) and then the morphology of isiXhosa (1.2.1.2). Ideophones are discussed as a distinct feature of the language in 1.2.1.3).

1.2.1.1 Phonology and orthography

IsiXhosa uses the Roman alphabet system (Pasch 2008). Phonologically, it has seven vowel sounds [e, i, a, u, o, ɔ, ɛ] depicted by five letters *“e, i, a, u, o”* (Jordan 1966:6). There are at least 58 consonant sounds of isiXhosa (Saul 2013). There are also complex graphemes in the language, e.g. [ɬ] written as /hl/ in *umhlali* (community member). There are limited consonant clusters that are only found in loan words, e.g. [kl] in *ikliniki* (clinic) (Saul 2013). Probert (2016) argues that there are no consonant clusters in isiXhosa, as they rather appear as single consonants and phonemes such as [ŋ] “ng” in “kungekudala” (soon). Due to this argument, this thesis attempts to keep the syllable as focused on one vowel and consonant clusters are treated as one single syllable. An example of consonants that are counted as its own syllable without a vowel in this research is the object marker circumfix -m-, as a syllabic consonant (see Mowrer and Burger 1991).

IsiXhosa has two basic tones and five surface level tones (Le Doeuff 2020). Tone can distinguish between various words, yet it is not marked in the orthography (Le Doeuff 2020). Other further complexities in orthography relate to the letter /h/, which represents three phonemes, aspiration [ʰ], the voiced glottal fricative [ɦ], and its voiceless counterpart [h] (Saul 2013). Some examples include ihashe [íɦə̰ːʃe] (horse), *utata* [uˈtʔaːtʔa] (father) and *uthatha* [uˈtʰ aːtʰ a] (to take). In addition, there are three basic click types categorised according to three places of articulation in isiXhosa -- the dental [ǀ] written as “c”, the alveolar-lateral [ǁ] written as “x” and the alveolar [ǃ] (or palataoalveolar [ǂ], see Christie 2019) written as “q” respectively (Gxilishe 2004). These three basic clicks, when occurring with neighbouring consonants that can represent nasalisation and aspiration, e.g. “isiXhosa” [isikǁʰóːsa] (the xhosa language) can represent around 15 phonological clicks (Gxilishe 2004).

IsiXhosa has a consonant-vowel (CV) syllable structure (Mkabile 2019) e.g. “i-ko-fu” (coffee). It is further a syllable-timed language, as other Bantu languages, which mean syllables occur at approximately regular intervals (Neubert 2020).

Furthermore, due to the consonant-vowel syllabic structure of isiXhosa, it appears decoding is grasped easier in isiXhosa at the level of the syllable over the level of the phoneme (Probert 2019). This and the grain reading size of isiXhosa is discussed further in 2.1.2. IsiXhosa usually has an SVO word order, however due to the rich morphology within the language, often with intricate verb complexes, the language allows for flexibility in the word order (Mkabile 2019).

1.2.1.2 Morphology

IsiXhosa has complex verbs because it is an agglutinative language like most Bantu languages, in which all the bound morphemes are affixed to a root (Mkabile 2019**)**. Other agglutinative languages include Finnish and Turkish (Hakkani-Tür *et al.* 2002). These vast number of bound morphemes, whether affixing to an adjective, noun, or verb, each typically represent a unit of meaning (Mzambo *et al.* 2019). IsiXhosa is written conjunctively with no space in between these morphemes (Moropa 2007). This results in fewer spaces between orthographic words than in English, the verb complex in isiXhosa is lengthy, with morphologically complex words and nine available positions for morphemes (Gxilishe *et al.* 2009).

The conjunctive writing system of isiXhosa resulting in long, morphologically complex words also creates difficulty in understanding the concept of a word in isiXhosa. In order to understand words in isiXhosa and this thesis’ use of the term ‘word’ in isiXhosa, one must understand the distinction between an orthographic word and a linguistic word (Louwrens and Poulos 2006). The label ‘word’ has been used for a lexeme (underlying or root lexical form of a word), such as *fund-* (read), as well as a grammatical word, which is an inflected form of the lexeme, such as *ndiyafunda* (I am reading) (Louwrens and Poulos 2006). The grammatical word *ndiyafunda* is also an orthographic word (a word written between spaces). However the future tense expression *ndiza kufunda* (I will read) is two orthographic words, but is technically one grammatical word and thus one linguistic word. Linguistic words are thus an autonomous group of linguistic elements and their status is usually proven through the application of one or more of the four-word tests: replaceability, isolatability, transposability, and separability (Louwrens and Poulos 2006). The example *ndiza kufunda* (I will read) which is two

orthographic words, but one linguistic word, is due to the affix *-za ku-* (future tense marker) which results in an orthographic space between the subject of the verb and the verb itself. Languages such as English usually do not have such a disparity between linguistic word and orthographic word (Louwrens and Poulos 2006) as isiXhosa does, and this leads to difficulties in establishing the unit for counting words in isiXhosa, as simply counting an orthographic word in isiXhosa may be misleading. This complexity of word separation in isiXhosa is further compounded by morphological complexity, such as the distinction between disjoint (long) and conjoint (short) form.

The following examples illustrate morphologically complex verbs in isiXhosa and the difference between disjoint (long) form (1) and conjoint (short) form (2), described as syntactic and semantic (Savić 2020). Syntactically, the conjoint form (2) needs to be followed by an object noun for it to not be in the disjoint form, and not include the disjoint marker -ya- (Savić 2020). Furthermore, the disjoint form can include an object marker, but the conjoint form cannot, implying a semantic focus on the verb in the disjoint form, but a semantic focus on the objects of the verb in the conjoint form. Further note the number of words in English in comparison to these single words and the resulting long lengths of orthographic units or words in isiXhosa:

(1) Ndiyayifunda
Ndi-ya-yi-fund-a
SM.1SG-DISJ-OM-read-FV
I am reading it

(2) Ndifunda incwadi
ndi-fund-a i-ncwadi
SM.1SG.-'read'FV 9-'book'
I am reading the book

While verbs have a root, to which bound morphemes affixate, nouns in isiXhosa consist of a stem and prefixes or suffixes (Mkabile 2019). Prefixes provide information related to grammar, such as negation, subject and object agreement, modality, tense or aspect and relative clause markers (Mkabile 2019). Suffixes or verb extensions change the argument structure, for example, individual suffixes that indicate individual passive, stative, reciprocal and causative

constructions (Mkabile 2019)[4]. Examples from (3) onwards are taken from the video in this thesis, for easier reference to specific examples referenced in later chapters:

(3) Ayisabonakali

A-yi- sa-bon-akal-i

NEG-SM.9-PERS-see-STAT-NEG

It (The fly) cannot be seen

The affix -sa- indicates an event that persists or progresses (Savić 2020) and is thus glossed as persistive aspect (see Nurse 2008:24). The suffix -akal- here is a form of the general stative -ek- in isiXhosa (see Khumalo 2009 for this form in Ndebele). Satyo (1985) provides a discussion on the differences between stative and passive constructions, in which the focus of stative constructions is on a state or condition, and its potential of achievement, rather than passivity of the agent elided. An example of an applicative and passive verb complex is in (4).

(4) ziyafuna ukukhathalelwa nazo

zi-ya-funa uku-khathal-el-w-a nazo

SM.10-DISJ-'want' INF--'care for'-APPL-PASS-FV 'also'

the cats also needed (lit: wanted)[5] to be cared for

Savić (2020) provides a detailed description of tense and aspect in isiXhosa and how this is realised in the language in the Indicative (seen as Halliday's Declarative and Interrogative Moods combined), Subjunctive and Imperative Moods, which include a variety of verb forms. Savić (2020) does not recognise modality, however, in the same way as Halliday and Matthiesen (2014), thus what is included as representing modality in isiXhosa is slightly more limited, yet nevertheless diverse. Savić (2020) alludes to the modal semantic categories of probability, obligation and inclination, however he does not recognise usuality, which limits his categorisation of what constitutes as modality (the systems of Mood and Modality in the SFL framework in Halliday and Matthiesen 2014 are outlined in 3.1.1.2). Nevertheless, Savić's (2020) realisations of modality in isiXhosa that are relevant to this thesis in isiXhosa include the infix *-si-/-s-* (3), which is realised in the negative as *-nga-* (in the Present Participial, titled Circumstantial, see Creissels 2021), as well as other realisations of the affix *-nga-* (7). Further

[4] Stative in this context refers to the stative verbal extension use -ek-, -k- or -akal- and not relating to other stative aspect markers of verbs (see Mkabile 2019; Savić 2020).
[5] Savić (2020:334)

realisations include clauses in the Imperative Mood (commands) (5), future auxiliary verbs *-za* (5) and *-ya* (6) with accompanying *ku-* prefix on the lexical verb, subordinate clauses introduced by the conjunction *ukuba* (that) (7), verbs *-za/-ze*, meaning and then do/did (9) and *-zange*, meaning never did (10). This thesis further identifies the following realisations of Modality: *-vumela*, which means to allow (11) and subordinate clauses introduced by the conjunction *xa* (if/when) (9). This thesis also recognises temporal verb forms as realisations of modality, including *-soloko* (always) (6), *-mana* (do continuously) (15), and *-qhele* (to be usual/normal) (8). It also recognises a Circumstance of Time, *ngamanye amaxesha* (sometimes) (12). Examples of these realisations are discussed along with Modality according to Halliday and Matthiesen (2014) in 3.1.1.2.

(5) Uza kubona!
U-za=ku-bon-a!
SM.2SG-FUT-'see'-FV
You will see!

(6) Kum uyakusoloko eqaqamba.
Ku-m u-yaku-soloko e-qaqamb-a.
'to'-'me' SM..3-REMFUT-'always' SM.3-'shine bright'-FV
To me, he will always shine bright.

(7) Amachaphaza engwe ayayinceda ukuba ingabonakali.
A-ma-chaphaza e-ngwe a-ya-yi-nced-a ukuba i-nga-bon-akal-i.
AUG-6-'spots' LOC.9-'leopard' SM.6-DISJ-OM.9-'help'-FV 'to be' 9-NEG.SUBJ-'see'-STAT-NEG
The spots of a leopard help it to not be seen.

(8) Ibiyinto eqhelekileyo,
ibi-yi-nto e-qhelek-ile-yo
9.PST-OM.9-'thing' RM.9-be=usual-ANT-DISJ
It was a usual thing.

(9) Zonke izinja zazisuke zize kuye zibaleka ukuza kumbulisa xa zimbona.
Zonke izi-nja za-zisuke zize ku-ye zi-balek-a u-ku-za=ku-m-bulis-a xa zi-m-bon-a.
SM.10-'every' 10.'dog' SM.10.REMPST-'go' and 'to'-'her' SM.10-'run' -FV AUG-15-FUT-OM.1-'greet'-FV 'when' SM.10-OM.1-'see'-FV
All the dogs would just run to greet her when they saw her.

(10) Akazange alahlekane nam

Aka-zange a-lahlekan-e na-m

NEG.SM.1-‘never=did’ SM.1-‘lose’-SUBJ ‘also’-me

He (the friend) has never lost me

(11) izilwanyana zalo, naluphi na uhlobo zazingavumelekanga kweso sakhiwo sakhe.

i-zi-lwanyana za-lo, na-lu-phi na u-hlobo za-zi-nga-vumel-ek-anga

AUG.SM.8-‘ánimals’ 8.DEM.PROX.3, ‘also’-11-’where Q 11-‘type’ SM.8.go.REMPST-‘allow’-STAT-ANT.NEG

all kinds of animals were not allowed

(12) Ngamanye amaxesha ndizoba imifanekiso yosapho lwam.

Nga-ma-nye a-ma-xesha ndi-zob-a i-mi-fanekiso yo-sapho lwa-m.

INSTR-SM.6-‘some’ AUG-6.’time’ SM.SG1.’draw’-FV AUG-4-‘picture’ ‘of’-11-‘family’ POSS.11-SG

Sometimes I draw pictures of my family.[6]

There are 15 noun classes in isiXhosa (Mkabile 2019). For example, in “ikati” (cat), the prefix *i-* represents class 9, of singular words while the stem of the noun is *-kati*, meaning cat. Some noun prefixes have an additional initial vowel, or pre-prefix, known as the augment (11), which can be dropped in certain tenses and negation (Mkabile 2019). Noun prefixes is the foundational structure for agreement with other grammatical constituents (Mkabile 2019). For example, in (1), the agreement between the verb and the object “incwadi” (noun class 9) with prefix “i-” is reflected as an object marker in the verb complex “-yi-.” This would also occur for the subject agreement markers that would agree with the noun class prefix of the subject noun (Mkabile 2019).

Due to the rich morphology of this language, prefixes can often act as pronouns, which can result in argument deletion (Oosthuysen 2016). This is illustrated in example (1), the word for I in isiXhosa is “mna” and is dropped, as the subject marker (ndi-) acts as a pronoun. Furthermore, the word “incwadi” (book) as direct object of the verb can be deleted, as it is represented in the verb complex by the object marker “-yi-” that agrees to the noun “incwadi” (book). There is also argument removed for sentential topics (Ma 2017), as in (1) if the book

[6] These glosses illustrate the complexity and agglutination of isiXhosa. Many of these examples will be referred to in Chapter 6. The rest of the examples of isiXhosa are not glossed, as particular morphemes are described where necessary.

was the topic of the sentence and introduced before, a particular book already known, the presence of the object marker allows for it to be removed.

IsiXhosa has thus rich inflectional morphology. There is also derivational morphology in isiXhosa (Demuth 2003), but as it was not pertinent to the DMDA of the subtitles, it is not elaborated on in this section and is beyond the scope of this thesis. A further linguistic feature of isiXhosa are ideophones.

1.2.1.3 Ideophones

Due to the presence of specific ideophones in the videos, this lexical class in the language is briefly discussed here. For a full understanding of the complexity of various phonological, morphological and syntactic features of ideophones in isiXhosa, please see Andrason (2020). Ideophones have extra-systematic properties that cause them to act contrarily to the typical standards of phonology, morphology and syntax of isiXhosa, and this extra-systematicity is largest in the morphology (Andrason 2020). Ideophones in languages like isiXhosa tend to have predicate or modifier properties, and the large number of variations of ideophones across languages have often led to their classification as types of verbs, adverbs and adjectives (Childs 2003). Ideophones convey meanings related to perceptions, emotions, and events, while imitating components of these meanings (Voeltz and Kilian-Haltz 2001). In isiXhosa, the ideophone itself cannot be inflected, but requires a verb that is almost auxiliary in nature, such as "-thi" (to say) for overt inflections (Andrason 2020). This results in bound morphemes, such as person and tense, to be affixated to the verb "-thi," for example "cwaka" (silent) and "ethe cwaka" (who remained silent) in (13). Note that this auxiliary verb is removed in the imperative form in (14).

(13) kwakuthe ngcu usele-sele oluhlaza, ethe cwaka.[7]
kwa-ku-th-e ngcu u-sele-sele olu-hlaza, e-th-e cwaka

[7] Andrason (2020) translates "cwaka" as quiet or calm. Due to the presence of -thula (quiet) and in order to distinguish subtle differences in meaning between these verbs, I translate "cwaka" as "silent" and "-thula" as "quiet." The predominant reason for this choice is the presence of "ethe cwaka" (he remained quiet) in a Relative past form with "-e", which in the context of the sentence, indicates that the state of silence started in the past and has not changed or repeated.

at-SM.15-‘say’-ANT perched=on SM.1a-Sele-sele RM-green, RM.9-say-ANT silent

perched[8] green Sele-sele, who remained silent.

(14) Cwaka usele-sele!

Cwaka u-sele-sele!

‘silent’ SM.1a-Sele-sele!

Silent Sele-sele!

(15) Waman’ ukuyithi krwaqu le mpukane, ibhabhela kufuphi.

Wa-man’	uku-yi-thi krwaqu	le	mpukane,	i-bhabhel-a	kufuphi.
SM.3REMPST-keep=doing	INF-OM.9-‘say’ ‘glance’	DEM.PROX.9	‘fly’	SM.9-‘fly’-FV	‘nearby’

He kept glancing at this fly, flying nearby.

During this process, the verb “-thi” loses its original meaning, i.e. “say” and becomes semantically bleached (see examples in Andrason 2020). One can also observe this in (15), in which the frog keeps glancing at the fly, with the ideophone “krwaqu” (glance) and the fly represented as an object marker (-yi-) in “ukuyithi,” the semantically-bleached verb. Classifications of these instances in the DMDA are discussed in Chapter 6. Glossing of isiXhosa in further chapters is not conducted and only the sentences and translation are provided, as specific morphemes and associated meanings are referenced when required within the body of the text. Examples of various DMDA realisations are also provided in 3.1.1.

This section has provided a broad overview of isiXhosa linguistic features. Thus, while a more transparent letter-to-sound correspondence than English should assist in reading, the length of words with a complex morphological structure may pose challenges for learners as they read (Schaefer *et al.* 2020). Findings from literacy studies that examine isiXhosa are discussed in 2.1. In the next section, policies of isiXhosa L1 literacy are discussed that further highlight reading challenges in the Foundation Phase.

1.2.2 Foundation Phase policies and assessments

The Language in Education Policy only requires that the LoLT in government schools is in an “official language” (DBE 1997: 2). In practice, however, the Bantu languages are only

[8] Andrason (2017), ngcu is also a type of ideophone, with contrary and extra-systematic phonological properties.

implemented as LoLT until the end of Grade 3 (Webb *et al*. 2010). The governing bodies of individual schools decide languages taught, and this decision is informed by multiple factors including available materials, predominant languages in the region, preferred language and SA policy (Rees 2016). This section discusses a fundamental policy for L1 isiXhosa education at the Foundation Phase, as well as assessments that are based on these, namely the National Curriculum Assessment Policy Statements and the Annual National Assessments.

The National Curriculum Assessment Policy Statements (CAPS) are, according to the DBE, "a single, comprehe[n]sive, and concise policy document [...] for learning and teaching in South African schools [...]" (DBE 2021). They are organised according to subject and grades (DBE 2021). According to the document for isiXhosa as a home language and for Foundation Phase, 7 or 8 hours per week are set for L1 instruction, with suggestions on how to teach it to learners (DBE 2011).

From the CAPS document, it is expected that 26 letters of the alphabet should already be taught in Grade R, including the ability to break a sentence up into words and morphemes, but in this sense the word refers to a written word divided by spaces in its natural context with prefixes, suffixes and roots (DBE 2011). Grade 1 learners are expected to be able to pronounce, construct and write words with all the letters (DBE 2011:27). Once this skill is mastered, learners should be able to master trigraphs, e.g. "ntl" (e.g. *intlanzi,* meaning fish, which occurs in *The Icky Sticky Frog*, Video 1 of this thesis) and complex consonant sequences, e.g. "ngcw" (e.g. (*ingcwele*, meaning holy) (DBE 2011:27), and be able to produce short sentences that could potentially be five syllables in length (which are seen as short sentences because they are only comprised of two words). For example, "(Imfama) iyakhokelwa[9]" ("A blind person is led") (DBE 2011:27). Word length is also not taken into account by CAPS as a potential challenge in literacy. These are complex tasks and CAPS are failing to recognise their complexity (De Vos *et al.* 2015). This illustrates CAPS not taking linguistic aspects of learning to read into account (De Vos *et al.* 2015). As De Vos *et al.* (2015) discuss, there remains a lack of knowledge in the linguistic interplay between the orthography, morphology, syntax and lexis in languages from the Bantu language family, such as isiXhosa. They also discuss how these translate into the standards or literacy thresholds for each grade (De Vos *et al.* 2015). Although

[9] The verb complex here is translated to the farmer is guided, with the i- prefix representing imfama (farmer) as subject.

research is growing in this area since the publication of their paper, the CAPS have not been revised to include these findings yet.

Most of the CAPS for isiXhosa as an L1 is a direct translation from the English L1 CAPS and this further illustrates how CAPS for isiXhosa as an L1 has not taken many linguistic aspects of learning to read in isiXhosa into account when compiling this policy. Rees (2016) highlights an example translation issue in CAPS (DBE 2011), which originated from the English CAPS and was translated unchanged for the isiXhosa CAPS without considering orthographical differences. In the isiXhosa CAPS (DBE 2011:16), instruction that teachers could teach "c", a letter that represents a simple consonant in English but a click in isiXhosa, before the letters "l" and "a". Yet, "l" and "a" in isiXhosa relates primarily to single simple sounds, due to isiXhosa's shallow, transparent orthography, and would be far quicker to teach and for learners to grasp than the isiXhosa representation of "c" or the English representation of "a', which depending on the word, can represent a plethora of vowel sounds (see 1.2.1.1). As highlighted in 1.2.1, the structure of isiXhosa with its agglutinative nature and transparent, shallow orthography makes it a vastly different language to English. There is a need for more research to inform how learners learn to read in isiXhosa, focusing on the linguistic aspects of the language (De Vos *et al.* 2015). This in turn should inform educational policies and the isiXhosa CAPS should be revised to focus specifically on isiXhosa as an individual language and not drawing on the English CAPS, when the languages share very little in common. The inadequacy of the CAPS may be contributing towards the literacy crisis, as an uninformed curriculum could hamper learners reaching literacy benchmarks in their L1.

According to ANAs (2014), a summary of Grade 1 L1 results indicate weaknesses in writing words, sentences and punctuation. The Annual National Assessments (ANAs) are national assessments conducted annually by the DBE and are administered country wide, focusing on Language and Mathematics with learners in Grades 1-6 and 9 (ANA 2014). The last diagnostic report compiled on ANA results was for 2014. Learners had to number written sentences of a story according to the correct order of narrative, which were out of sequence. This was also a weakness for them, indicating their inability perhaps to decode and comprehend sentences and words (ANA 2014). There is no Oral Reading Fluency (ORF) at the word and sentence levels, therefore one cannot conclude from these tests whether the problem is in a metalinguistic skill used for decoding or whether they can decode the words but cannot understand them without the options represented visually. A similar picture and sentence match task to the Picture-Word

Match test in this research was further used (see 4.2.2). The ANAs for Grade 1 include a picture-sentence match task, and I include a similar test in this study, but at the word-picture level and a simple reading test rather than asking the learner to write as in the ANA (see Appendix B). In the ANAs, this was reported as a learner strength, indicating that learners can associate words with their meanings, yet there were spelling errors in their writing (ANA 2014). The way the tests are constructed, however, causes it to be a challenge to determine where the specific problems that learners face lie in the reading process or in the spelling and writing process.

The ANAs present a national average score for all home languages in these assessments without a specific breakdown per language. This makes it difficult to navigate which languages need the most attention. ANAs reported improvements between 2012 and 2013 (3%) and 2013 and 2014 (2%) with an above 50% average for all tested home languages. The ANAs, however, do not indicate whether specific languages tested may have fallen below this average or why. While Grade 1 learners were found to struggle with the recognition of vowel blends or coalescence, e.g. (Associative formative na-, meaning and/with, + iikati = neekati, as a+ii = ee) (see Sibanda 2009). Grade 2 learners were found to struggle with vowel and consonant digraphs, such as /hl/ mentioned in section 1.2.1.1, indicating that this was not achieved in Grade 1 either (ANA 2014). Yet as stipulated earlier, this is an expectation of CAPS by the end of Grade 1, further highlighting the problems in CAPS and the wide gap between policy expectation and the reality of teaching and learning. Furthermore, while multimodal resources such as illustrated books are alluded to in CAPS, CAPS does not integrate or discuss multimodality in its policies or mention subtitled videos. The next section discusses this in more detail and describes videos as multimodal resources and introduces key concepts of multimodality that are used in this research.

While subtitled videos are not included as a potential literacy resource in CAPS, the idea that visuals paired with the words assist in word comprehension, thus a multimodal understanding of reading, is present in the five-finger strategy of CAPS. The "five finger strategy" in CAPS provides five reading strategies for new and challenging words based on the fingers of a child's hand (DBE 2011). They are defined as follows:

- Ubhontsi: Lishiye igama usifunde sonke isivakalisi. (Thumb: skip the word and read the whole sentence)
- Umnwe wokuqala: Jonga emfanekisweni (First finger: Look at the illustration)

- Umnwe wesibini: Jonga igama ubone ukuba akho na amalungu owaziyo egameni. (Second finger: Look at the word to see if there are any parts of the word that you recognise)
- Umnwe wesithathu: Funda izandi ozaziyo zegama elo. (Third finger: Pronounce the word aloud)
- Umnwe wesine: Cela uncedo lokufunda igama okanye ukwazi intsingiselo yalo. (Fourth finger: Ask for assistance for the reading of the word or understanding the meaning of the word)

(DBE 2011)

The first finger calls to look at an image that may provide clues in deciphering the word, thus highlighting the importance of illustrations for early grade readers to be used with the text. This also highlights the importance of multimodal reading strategies in using semiotic resources in one or more modes to assist understanding meanings conveyed from another mode(s). Subtitled videos could also assist reading new and challenging words due to their audio-visual nature and the combination of language and image require a multimodal outlook to reading and literacy. Due to the suggestions by the DBE (2011) of a multimodal view of reading, this thesis now examines key concepts to understanding multimodality.

1.2.3 Key concepts of multimodality

Videos are a multimodal resource (Iedema 2001). A mode "refer[s] to a socially organised set of semiotic resources for making meaning" (Jewitt *et al.* 2016:221). Examples of modes include visual and linguistic, the latter including spoken or written sub-modes (Stöckl 2004). It is not only videos, however, that are multimodal, rather the society at large in the digital age is multimodal (O'Halloran 2023). For example, even as I read this thesis on a visual computer screen looking at and processing written language, which has a spoken and auditory component, I then type words in this thesis. I type on visual keys to create the meanings that I see and comprehend. Readers will comprehend these meanings after the process has finished, in a similar way to how I read this thesis but will bring their own understandings and experience to the text to comprehend it further or in alternative ways.

Jewitt *et al.* (2016:221) describes modes as a collection of socio-cultural semiotic resources. A semiotic resource is described as "the actions and artefacts we use to communicate, whether they are produced physiologically [...] or by means of technologies" (van Leeuwen 2005:3). O'Halloran (2023:176) provides examples of semiotic resources as "graphs and diagrams" within the visual mode and visual systems of analysis that understand their meaning potentials. Semiotic resources are arranged within the design of a text by a designer or sign-maker for a

specific purpose within the social context (Jewitt *et al.* 2016). Thus, this thesis recognises four semiotic resources that are analysed in the Digital Multimodal Discourse Approach in this thesis: (1) SLS isiXhosa story videos, which are multimodal texts, (2) subtitles as a realisation of the written linguistic mode, (3) spoken language as a realisation of the spoken linguistic mode and (4) dynamic images as a realisation of the visual mode. Semiotic resources of the videos that are not examined in this research include the gestures, music and sounds within the videos. This thesis highlights a Digital Multimodal Discourse Analysis (DMDA)[10] (Tan *et al.* 2015:559; Schafli 2020) as a useful theoretical framework for analysing SLS videos and the interaction between written text and educational videos as audio-visual (Tan *et al.* 2015:559). Multimodal analysis explains how "people draw on distinctly different sets of resources for meaning making (e.g., gaze, speech and gesture)" (Jewitt *et al.* 2016:218). The use of computer software, *Multimodal Analysis Video* in DMDA enables "the micro-level of discourse" to be investigated along with "the macro-level" (Tan *et al.* 2015:560). The different inter-disciplinary frameworks through the software accommodates a holistic examination of the multiple modes, and still affords the description of individual modes and a graphical representation (Jewitt *et al.* 2016:88). Thus, a micro-textual examination affords in-depth insight of integral features for conducive video design for L1 literacy development in this context. In keeping with standard practice in DMDA, for the visual and linguistic modes, Systemic Functional Multimodal Discourse Analysis (SF-MDA) is used. The theory on DMDA in this study is described in Chapter 3 and the implementation of the DMDA is described in Chapter 4 of this thesis.

Multimodal research has been used in analyses in Africa ranging from discourse of university bathroom graffiti and website pages in South Africa to Kenyan communities and discourse on AIDS (Banda and Oketch 2011; Mafofo and Banda 2014; Ferris and Banda 2015). Furthermore, multimodal research in pedagogy from South Africa exists, but it only focuses on English as an L2 in classroom learning and literacy (Archer and Newfield, eds. 2014; Stein and Newfield, eds. 2006). None of these studies examined video, but rather written texts with images, from which this research seeks to deviate with its DMDA of video. This literature is discussed further in 2.3.

[10] For uniformity, the DMDA mostly refers to Digital Multimodal Discourse Analysis, but is used throughout the thesis referring to DMDA as analysis, as well as approach and method.

Semiotic resources are thus material realisations of different potentials for making meaning (Jewitt *et al.* 2016:182). The various potentials for making meaning realised by semiotic resources from different modes are defined as modal affordances (Jewitt *et al.* 2016:182). Modal affordances influence how a mode is most effectively used within a specific social context to communicate meanings, i.e. modal affordances are different ways of modal usage to convey potential meanings (Jewitt *et al.* 2016:182). The knowledge of modal affordances, meaning potentials and design features of different modes and their semiotic resources are an aspect of a broader proficiency skill set, designated as semiotic awareness in multimodal literacy research (Lim and Toh 2020)

The term semiotic awareness is defined in various ways in literature and is addressed in the research questions of this thesis, described in 1.3 (Abrams 2016; Lim and Toh 2020). To fully grasp semiotic awareness in children, I briefly examine Piaget's (1969) semiotic (symbolic) function of typical child cognitive development. The preoperational stage of typical child cognitive development occurs from the ages of two to seven and is based upon a development of two skills, intuition, and for our interest, semiotic representation, also known as the semiotic function, which includes the linguistic symbolic (Piaget 1969). This is the ability of a child to construct mental representations between an object that is not present,[11] usually understood as an idea or even as a visual representation of a real-world concept in the preoperational stage (Piaget 1969). These mental representations are associated with situational models in 3.2 and mental representations of words in 2.1. Mental representations are thus constructed by the child's perceptions (Piaget 1969).

Furthermore, the child can relate this mental representation to another object that is present and is usually understood as spoken words (labels) for an object or another visual object (Piaget 1969)[12]. The semiotic (symbolic) function has roots in early infancy (Piaget 1969) but manifests greatly between 2 and 4 years of age (Piaget 1969). Activities in further research associated in an awareness between material forms of semiotic resources and their meaning potentials include viewing photographs, computer screens, language production, drawing, and repeatedly requesting adults to identify and name objects to construct mental associations (Development Teaching and Learning Group 2020; Scott and Cogburn 2023). In the latter activity in particular, this can extend to frequent requests to be read to or watch the same video

[11] associated with the term 'signified' in 3.1 (Piaget 1969)
[12] associated with the term 'signifier' in 3.1 (Piaget 1969)

repeatedly, until children can repeat every word and discuss the story (Development Teaching and Learning Group 2020).

Piaget's (1969) discussion of a semiotic function, while not taking developmental differences into account, elaborates on an almost critical period for early multimodal literacy development (and by proxy, early linguistic literacy development). This critical period seems to be rooted in early infancy and intersects periods of childhood development between pre-school ages (pre-literate stage of reading or Early Childhood Development in a South African context) and early reading development (the Foundation Phase in South African schools). Within this critical period for early multimodal and early literacy development, children manifest a level of engagement with meanings between different visual modes and between spoken linguistic and visual modes (associated with Intersemiotic Texture in 3.1.3). Within this period, it is further highlighted that repeated exposure to visual and both spoken and written linguistic stimuli (in the form of reading stories with adults or even words on screens) can initiate comprehension of meaning relations between modes (associated with Intersemiotic Texture in 3.1.3). Children shift from private and self-focused symbol/semiotic construction to eventually mastering socially shared signs in later stages of development as an active process (Lenninger 2006).

The multiple abilities within Piaget's (1969) semiotic function indicate the complexity of skills and abilities for semiotic representation, perception and inter-relation of meanings to occur. These abilities and skills, I believe, fall within semiotic awareness, which is generally relegated to a students' knowledge of design features, affordances and meaning potentials of different modes and the knowledge of how these modes are incorporated together to communicate meaning (Lim and Toh 2020). This definition, while accurate and used across multimodal literature (e.g. Jewitt 2006; Abrams 2016), does not fully encompass the scale of intramodal awarenesses.

Intramodal in this sense is referring to a mode such as the written linguistic mode, which is an amalgamation of two separate modes, the visual and linguistic, and is thus a singular sub-mode of these two broader modes, yet multimodal. Examples of intramodal awarenesses within semiotic awareness include metalinguistic awarenesses for literacy, such as phonological awareness and word recognition, see Pretorius (2010), or spatial awareness for visual literacy, see Wilmot (1999). I distinguish this from intermodal, which integrates meaning potentials, affordances and design features from different modes that do not form a sub-mode. However, intermodal still acknowledges meaning potentials, modal affordances and design features at

the level of a single mode of written language, e.g., linguistic literacy. Intermodal literacy skills refer to broader multimodal meaning relations of spoken language, visuals, written language and other modes. I use the term multimodal literacy for intermodal literacy and literacy for intramodal literacy of the written language in this thesis.

The definition of semiotic awareness by Lim and Toh (2020) further leans toward complex knowledge and skill processes that may be expected of early adults, rather than younger children. Towndrow *et al.* (2013:326) provide a broader definition of the complexity of semiotic awareness:

> Semiotic awareness may be understood as an alertness to the representational possibilities that any form can afford, in which contexts and for whom and how and why. It generally follows, too, that the extent to which a new media designer, like Jeremy, is semiotically aware will predict the degree to which his design choices will articulate across different modes and media, befit purposes at hand, and speak to the experiential bases and interests of given audiences, who must similarly engage their own semiotic awareness capacities toward actualizing the intended meanings (Towndrow *et al.* 2013:326)

Towndrow *et al.* (2013:326) understand semiotic awareness to be a set of skills employed by designers, and in teaching and learning contexts, teachers/facilitators and learners that do not necessarily always align and correspond. Furthermore, in order to understand meanings conveyed from multiple modes, one should naturally have sufficient ability to process meanings from individual modes and their semiotic resources. Thus, this thesis focuses on the following strata toward a definition of semiotic awareness, which contains intramodal and intermodal knowledge:

(1) a literacy of the written linguistic mode (an example of intramodal literacy) contains a set of skills and awareness which are necessary for advanced intermodal semiotic awareness and multimodal literacy. These foundational intramodal literacies relate to a simple stratum of semiotic awareness.

(2) A more advanced stratum of intermodal semiotic awareness includes the knowledge and incorporation of meaning potentials of different modes, meaning affordances, inter-relational knowledge, features of design and meaning transformation based on that knowledge.

Thus, this thesis examines the videos with a DMDA in order to present an advanced stratum of meaning potentials, affordances, inter-modal meaning relations and design features that can potentially inform L1 isiXhosa literacy development in early reading. The DMDA is the

primary approach of this thesis, however, this thesis further investigates opportunities and challenges for specific skills that are situated within the simple and advanced strata of semiotic awareness. These skills are word recognition and ORF (tests mentioned in 1.1.2 and discussed in 4.2.2), which are both metalinguistic and semiotic awareness skills for intramodal literacy of the written linguistic mode. Thus, these tests and intervention measure and infer aspects of learners' capacity for semiotic awareness, i.e., word recognition and ORF. A Picture-Word Match test is also conducted to investigate semiotic awareness for intermodal literacy (multimodal literacy) between written language words and static images (test mentioned in 1.2.2 and discussed in 4.2.2). Theories of multimodal literacy that support this research definition is discussed in 3.2.

1.3 Research goals and questions

The main aim of this research is to test, describe and analyse the efficacy of five isiXhosa story *YouTube* videos for the development of isiXhosa literacy at the Grade 2 and 3 L1 levels. This aim is elaborated on in the following research questions:

1. How is meaning construed in terms of the systems and sub-systems of the mode of spoken language in the videos?
2. How is meaning construed in terms of the systems and sub-systems of the mode of written language (subtitles) in the videos?
3. How is meaning construed in terms of the systems and sub-systems of the visual mode in the videos?
4. What is the relationship between the various modes of language and image within these videos according to the DMDA systems and sub-systems of these modes?
5. How does the combination of the modes influence learner literacy levels?
6. What effect does repeated exposure to the videos have on literacy scores?
7. What are the opportunities and challenges associated with learning word recognition, oral reading fluency and semiotic awareness with these videos?
8. What recommendations can be made for the design and use of future educational videos in line with the policies of the DBE?
9. How can DMDA be used in combination with psycholinguistic approaches to literacy to investigate and foster literacy development in schools?

1.4 Chapter summary

This chapter has introduced this research and highlighted motivations for this thesis. These can be summarised as stemming from the social system, which is influenced by the current literacy crisis in South Africa, the impact of the COVID-19 pandemic and the possible potential of SLS video as a tool for literacy development in other social systems. These sections have resulted in Chapter 2, the literature review, being divided into literacy research for Bantu languages including isiXhosa as a semiotic system (2.1) and subtitled video research (another semiotic system) (2.2). It also examines the research context further by focusing on the linguistic features of the isiXhosa language and relevant Foundation Phase policies and assessments, which provide necessary background for section 2.1.

This chapter introduces key definitions and concepts of multimodality and the semiotic framework of analysis that is used in this thesis, as well as the research goals and questions. Section 2.3. discusses an overview of multimodal research relevant to this study, another social system. Chapter 3 describes and elaborates on the theory of DMDA and theories relating to multimodal literacy. Chapter 4 outlines the mixed-method methodology in this research. Chapter 5 discusses the statistical analysis of the findings from the intervention. Chapter 6 discusses DMDA, with a triangulation of findings from both methods (two-group experimental design method and statistical analysis and DMDA, see Chapter 4 for elaboration on these methods) concluding this chapter. Chapter 7 concludes this thesis with answers to research questions and proposed areas for further research.

Chapter 2 – Literature review

This chapter reviews the literature to ascertain the area of research that this thesis addresses and highlights the gaps in the literature that are investigated in this research.

Since this thesis is focused on the area of literacy and Grade 2 and 3 learners of isiXhosa (and all systems of semiosis included in this process), it examines the available literature on this language and children in South African schools. This is to ascertain how this study differs and is similar to previous studies (2.1). This area of literature provides context and highlights the gaps in the literature, on which research questions 5 and 6 of this thesis are focused (see 1.3 for research questions). This further motivates the areas of triangulation with DMDA in research questions 7-9 (see 1.3 for research questions). In examining the literature on literacy, this chapter also describes methods and tests for literacy measurement in other studies, to assist in test and research design in this study, which is discussed in 4.2.2 in Chapter 4. Furthermore, the videos examined are same-language subtitled (SLS) and since the learners may be influenced by the written linguistic mode in the video, research on subtitles themselves are addressed (2.2). This is to foreground the mode of the written language discussed in research question 2, as well as the potential language learning and literacy development opportunities and challenges in research questions 5-7 (see 1.3 for research questions). Furthermore, it describes the various methodologies used to investigate subtitles, particularly SLS in various contexts, and highlights similarities and differences between these methodologies and the methodology within this thesis.

Due to the DMDA that is being undertaken as part of research questions 1-4 and 7-9 (see 1.3 for research questions), the literature in multimodality (2.3) is described to highlight the gaps in this area in the contexts of literacy research and subtitle research for literacy development (particularly in isiXhosa). In addition, this section describes the various other types of multimodal analyses to ascertain why DMDA and its systems were chosen and how this thesis contributes to this area. To conclude the chapter, its contents is summarised in 2.4.

2.1 Literacy

Literacy in the written linguistic mode is a complex product from an array of abilities and knowledge that makes meaning for a specific purpose (Pretorius 2010). I understand that reading is not only an individually cognitive-linguistic achievement, but also a socially constructed practice (Pretorius 2010). Literacy as a multimodal semiotic process and practice is realised from social, biological and physical systems (O'Halloran 2023), introduced in the beginning of Chapter 1 of this thesis. Thus, it is examined at the level of the social semiotic, and a Digital Multimodal Discourse Approach (see 3.1), and psycholinguistic, including cognitive, theories of multimodal literacy, which are discussed in more detail in section 3.2 in Chapter 3. Thus, this thesis treats bottom-up reading models (form to meaning) and top-down reading models (meaning to form) as a set of skills that are not necessarily sequential (Frehan 1999). This is aligned to Stanovich's (1980) interactive-compensatory model of reading, in which a reader (particularly in a broader multimodal context) may draw on stronger reading skills to compensate for weaker skills. This distinction between literacy of the written linguistic mode and multimodal literacy between the written linguistic mode and other modes, such as visuals and spoken language, is understood as requiring a set of stratified knowledge, termed semiotic awareness (see 1.2.3). Multimodal literacy between the written linguistic and other modes and advanced skills of semiotic awareness are discussed in 1.2.3 and Chapter 3.

However, literacy of the written linguistic mode is also a multimodal and semiotic practice (Prinsloo and Stein 2004). It uses a visually based print system and maps these to spoken language (Probert 2019). Recognising that literacy of the written linguistic mode is a further set of metalinguistic skills (see Wilsenach 2013), situated within a broader skill set of semiotic awareness skills, this thesis focuses on the metalinguistic skills of word recognition and Oral Reading Fluency (ORF). These are discussed in 2.1.1 in relation to automatic reading.

This section further examines literacy studies in South Africa on isiXhosa and languages that are similar to isiXhosa. Section 2.1.2 examines Grain Size and its relation to reading fluency to understand research on the early reading strategies of a complex language such as isiXhosa to understand ORF in relation to early reading strategies and word recognition (2.1.3).

2.1.1 Literacy, word recognition, oral reading fluency and automaticity

This thesis focuses on word recognition and Oral Reading Fluency (ORF) aspects of the three building blocks mentioned by Pretorius (2010) (the other is phonological awareness, which this thesis categorises as an aspect of word recognition and ORF). Word recognition and ORF encompass literacy skills and thus they are described in more depth before studies on them are examined.

Word recognition is "the ability to recognise words rapidly and accurately" (Pretorius and Mokhwesana 2009). It "involves interactions between language structure, orthography and metalinguistic skills" (Probert 2016). Metalinguistic skills are "abilit[ies] to identify, analyse and manipulate language forms" (Koda 2007:2). For example, this section discusses studies focused on the metalinguistic skills of phonological awareness. Orthographies are "a set of linguistic and social compromises developed in particular social contexts" (Probert and de Vos 2016). For example, as mentioned in 1.2.1 in Chapter 1, isiXhosa has a more transparent/shallow orthography with a conjunctive writing system (Probert 2016). In addition, learners have their own cognitive tools such as memory and lexical access strategies that they use to recognise unfamiliar words and these are developed the more they are exposed to reading (Probert 2016, Probert and de Vos 2016).

Pretorius (2010) describes word recognition as an important foundational building block of decoding. Decoding is the process of mapping and recoding orthographic words from the visual symbolic writing system to phonological and morphological units and involves reconstituting these units into spoken words and/or mental words which can lead to accessing representations of the words (Pretorius 2010). Thus, linguistic information from the text is extracted without necessarily comprehending the text (Pretorius 2010). Probert (2016:43-44) states that word recognition also "includes the ways in which people recognise words and access the corresponding word representations stored in their mental lexicon." Therefore, decoding and the comprehension of written text when reading are not mutually exclusive acts (Pretorius and Mokhwesana 2009).

Comprehension refers to the process of assigning text information with new or prior knowledge and meaning (Pretorius 2010). In terms of comprehension, Mkhize (2013) finds evidence of transfer of skills across Zulu and English. The same is found in Grade 7 with Northern Sotho,

however proficiency in their L1 was not a predictor of reading ability, rather L2 (English) predicted L1 reading ability, probably due to learners' exposure to more English reading. Generally, in this study, reading comprehension scores were low and at the end of the year, they were slightly higher for English than Northern Sotho. Pretorius (2012) found the same with Northern Sotho in terms of low comprehension levels and that early L1 instruction did not create advantages for academic literacy skills for Grade 6 learners. The same was found in 88 high school learners (Grade 8) in a township school in Gauteng with Setswana (Matjila and Pretorius 2004). This illustrates that if comprehension scores are low in other languages in intermediate and senior phase, they may also be low in foundation phase and in isiXhosa.

Oral Reading Fluency (ORF) allows for decoding to become comprehension, as the act of reading becomes more automatic the more children decode (Pretorius 2010). Automaticity in reading relates to cognitive capacity and memory, as the less cognitive processing a familiar resource demands in decoding, results in a greater cognitive and memory capacity for high-level processes related to "meaning-making" (Pretorius 2010:345). Pretorius (2010) provides an example of a child developing improved skills for both decoding and comprehension from increased exposure in combination with social context motivation, which facilitates more automatic reading and processing via neural pathways. Increased exposure can be understood as duration and frequency of occurrence.

Duration and frequency of occurrence are in every semiotic resource in every mode, however there are distinct difference between duration and frequency (Lantz *et al.* 2020). Information reporting frequency of occurrence results is represented in discontinuous instances whereas information reporting duration is sustained over time (Lantz *et al.* 2020). Duration and frequency of occurrence may lead to stronger recognition for both visuals and words, as well as other modalities (Lantz *et al.* 2020). These temporal aspects are discussed with reference to the methodology in 4.1.2.1, 4.2.3 and 4.3.2.

According to Spaull *et al.* (2020:6), ORF "reflects the speed, accuracy and naturalness that readers display when reading a text aloud, following the intonation and rhythm of spoken language." Due to prosody being a challenge to accurately assess, ORF tests focus on the components of speed and accuracy (Ardington *et al.* 2020). Accuracy reflects the total number of words that are correctly read out loud (Spaull *et al.* 2020). The speed of reading addresses the number of correct words read within a particular time frame (Spaull *et al.* 2020). It is this aspect of ORF that Adams (1990) described as a primary aspect of skillful reading. With the

study conducted by Spaull *et al.* (2020), readers are typically given texts to read with a time-limit of a minute and errors are deducted from total number of words that were read by the learner, which give a total word correct per minute (WCPM) score. An accurate ORF score is ideally arrived from the mean WCPM of a group (Spaull *et al.* 2020) As decoding and comprehension are separate processes, a separate comprehension test is usually also conducted (Spaull *et al.* 2020). ORF scores can be affected by a variety of factors, which include familiarity with topic and genre and the difficulty level of the text (Spaull *et al.* 2020). ORF growth occurs in the Foundation Phase level of reading (from Grades 1-4) (Spaull *et al.* 2020). The suggestion by Spaull *et al.* (2020) is that once reading is fast and accurate, only then do other factor indicate differences in reading comprehension, such as skills of inference, vocabulary, and knowledge of texts and their backgrounds and genres. A measure used by Spaull *et al.* (2020) and others in this section include the EGRAs for ORF, described in 1.1.2. Adaptations of these tests were used in this thesis and are described in 4.2.2.

Fuchs *et al.* (2001) discuss oral reading fluency as a better indicator of reading comprehension than silent reading fluency. This is primarily due to relying on the reader to self-report the last word that was read (Fuchs *et al.* 2001). Thus, despite subtitles usually not being read aloud, but being read silently, ORF tests to provide a more accurate estimate of reading rates are used in this thesis. The next sub-section discusses studies that examine grain size and studies that use oral reading fluency.

2.1.2 Grain size and ORF studies

Grain sizes are the size of a segment at which words are decoded (Goswami 2020). Grain sizes manifest as levels of visual forms of a word (e.g., graphemes and morphemes), levels of phonological forms of words (e.g., phonemes, syllables) and the linking of the visual and their phonological forms (Goswami 2020). Grain sizes stem from the psycholinguistic grain size theory by Ziegler and Goswami (2005). While the syllable is the grain size for isiZulu, it is at the level of phonemic awareness that word reading, and oral reading fluency is improved (Pretorius 2015). This could be the same for isiXhosa and its conjunctive writing system. Pretorius (2015) examined Grade 4 learners for decoding skills in isiZulu and found syllable identification in isiZulu did not display a correlation to any other decoding skills, yet phonemic awareness correlated with word reading and oral reading fluency.

Syllables are the “dominant grain size” over the phoneme in Setswana and isiXhosa and the morpheme is a “secondary grain” in isiXhosa (Probert 2019: i). Probert (2019) examined different orthographies for Grade 3 isiXhosa learners and Grade 4 Setswana learners by grain size and how this influences their reading. Probert (2019) examined 74 primary school children in total from four different schools situated in semi-rural townships and they were tested with four linguistic tasks, which included an ORF task. Her findings indicate isiXhosa learners unpack words at the syllable rather than the phoneme. Thus, this thesis measures words correct per minute (WCPM), as well as syllables per minute (SCPM), to determine a rate more relatable to the grain size of the language.

Meiring’s (2020) study reflected that 50% of learners could not read in isiXhosa at the end of their Grade 1 year, with 16-24% of learners unable to read isiXhosa at the end of their Grade 2 year. When words were read, 10 was the average WCPM for Grade 1 and 20 for Grade 2 (Meiring 2020). However, the no-fee schools in this control group were urban and peri-urban schools, and schools were excluded on the bases of classroom overcrowding (not more than fifty per class), management difficulties, or less than twenty learners in a grade (Meiring 2020). Thus, while these WCPM rates match the benchmarks in the 2020 report (Ardington *et al.* 2020), these schools may not entirely reflect conditions in low socio-economic, peri-urban communities and conditions in the Eastern Cape. This thesis examines WCPM and syllable per minute (SCPM) rates at a peri-urban school, and therefore contributes to this area of research.

Schaefer and Kotze (2019) examined Grade 1 isiZulu and Siswati learners' L1 and English literacy skills at 80 schools. They examined literacy in similar ways to this thesis (an EGRA-based test), however they also included a vocabulary task, listening comprehension, word span task, non-word repetition task and tasks for phonological awareness and word-reading fluency. They found a variation regarding oral language proficiency, listening comprehension and phonological working memory. There was no variation for vocabulary skills, possibly due to the task not being sensitive enough. Ultimately, at the end of Grade 1, learners could recognise on average seven letter-sound correspondences in their L1 per minute and read aloud only 5.51 L1 words per minute. They could read slightly more words in their L1 than English (their L2), but their reading scores were low at the beginning of Grade 1, where the word recognition task got floor effects. This was the reason for this thesis to examine Grade 2 and 3s only, as it was assumed the majority of children in similar schooling conditions to their study may be in a

preliterate stage of reading. To ensure learners had a full Grade 1 year of literacy development at the school, learners in this thesis were tested at the beginning of Grade 2.

Spaull *et al.* (2020) and Schaefer and Kotze (2019) indicate that vocabulary knowledge does not factor into early grade reading as much as decoding does. However, a study by Wilsenach (2015) indicates measurements of vocabulary are early literacy predictors and can further call attention to children that have challenges with phonological awareness and basic literacy skills (Wilsenach 2015). She describes that when learners have the vocabulary knowledge of a word, they can read that word aloud faster and easier, which indicates vocabulary knowledge may support decoding and ORF. This study on 99 Grade 1 learners of L1 Northern Sotho illustrated low vocabulary knowledge and that Northern Sotho receptive vocabulary predicted phoneme-grapheme mapping, phonological awareness and early writing outcomes (Wilsenach 2015). As learners in this thesis are exposed to vocabulary from the video, it seeks to gauge whether automaticity is improved from words from the video in comparison to words not in the video.

Veii and Everatt (2005), in their study of Herero, showed L1 ORF comprehension and L2 (English) phonological awareness were predictors of L1 literacy skills. Transparent orthography caused literacy development to be faster in Herero than in English. Makaure (2017) found that phonological processing and reading skills were more developed in their L1 Northern Sotho in which children received their literacy instruction. IsiXhosa has a transparent orthography, similar to Herero and Northern Sotho, indicating if isiXhosa is taught in their L1, learners' reading speeds may be greater in their L1. Despite the study by Schaefer *et al*. (2020) that illustrated phonological awareness is a better predictor in later grades. Wilsenach (2013) found this to be an earlier predictor in Grade 3 children with Northern Sotho as their L1. She found that children with Northern Sotho as their L1 who are taught in their L1 read their L1 and L2 better than Northern Sotho children who are not taught in their L1. This highlights the importance of a solid foundation in a child's L1 education that further assists in bilingual and multilingual literacy (Wilsenach 2013). She suggests phonological awareness, lexical ability, working memory and automaticity are not properly developed in children who have reading issues (Wilsenach 2013). Malda *et al*. (2014) with Setswana also showed that phonological processing is integral in more transparent orthographies and vocabulary and working memory were stronger skills needed for English reading. Yet the overall connections between reading and cognitive skills were not significantly different across English, Afrikaans and Setswana.

This illustrates that phonological awareness and processing within word recognition is an important metalinguistic skill for isiXhosa readers.

Piper and Zuilkowski (2015) illustrate ORF is a good predictor of comprehension and for monitoring literacy outcomes in the Foundation Phase. Kiswahili L1 children in Grade 2 had slow reading rates and low comprehension levels, with 24 WCPM (Piper and Zuilkowski 2015). Makalela and Fakude (2014) examined ORF amongst Northern Sotho readers in Grades 4-7 and found slow reading rates, with Grade 7 readers on average reading 65, 56 and 66 WCPM. No statistical differences were found in reading rates between the other grades with means of 54, 49 and 56 WCPM. While these are good numbers to bear in mind for this thesis, Northern Sotho has a disjunctive orthography and thus shorter word length than the conjunctive writing system of isiXhosa. Furthermore, the Northern Sotho readers were at a higher grade than the Kiswahili L1 children, which are at the same grade level as isiXhosa learners in this thesis. Thus, it may be expected that the Northern Sotho reading scores are faster than the averages for isiXhosa. The WCPM scores for Kiswahili Grade 2 children with its conjunctive orthography would be more accurate in comparison to isiXhosa WCPM scores for Grade 2. The next sub-section examines studies in the areas of word recognition strategies, orthography and writing.

2.1.3 Word recognition strategies and conjunctive orthography

Probert and de Vos (2016) examine strategies for word recognition in isiXhosa and their transfer to the L2, English. Data was from 47 Grade 4 learners collected at two urban schools differing in language of teaching and learning (LoLT), English and isiXhosa, with their L1 being isiXhosa. Learners were given word-reading and pseudo-reading tasks and errors were analysed in order to ascertain the types of transfer errors. They found strategies for decoding real and non-real words are used by isiXhosa learners, whereas lexical whole word recognition and phonological strategies are used by English LoLT learners to decode real words, and mainly phonological for non-words. There is a phonological and restricted transfer of these strategies from the transparent orthography (isiXhosa to the deep orthography (English) (Probert and de Vos 2016). Literacy skills can be transferred to a language that is not the LoLT when isiXhosa learners decode English words incorrectly, e.g., [lalamaɹi:] instead of [kalamaɹi:] or when they do not use lexical reading strategies for whole words enough for the

English language, e.g. saying “well” instead of “wheel.” Ultimately, an isiXhosa as LoLT for L1 learners in the Foundation Phase aids in developing skills to English and L2 literacy, however, isiXhosa literacy and English literacy have to be taught separately and distinctly (Probert and de Vos 2016). L2 literacy has to be introduced carefully, to ensure isiXhosa learners can learn whole word recognition skills and transfer these to English. (Probert and de Vos 2016). This suggests that, while isiXhosa learners in Grade 4 employ both lexical whole word recognition and phonological reading strategies, they predominantly use phonological reading strategies for isiXhosa, particularly for unfamiliar words. This thesis employs word-level and Long Passage reading ORF tests, and this study indicates that learners may use phonological processes to read these tests more than relying on lexical access reading strategies.

Thus far, ORF tests have been a method to measure word reading speeds. Another method is eye-tracking. Van Rooy and Pretorius (2014) examine eye-tracking and isiZulu in Grade 4s and found the eye movements for the conjunctive orthography were slower and this is attributed to longer word units in the conjunctive writing system of isiZulu, which is similar to the writing system of isiXhosa in this study. Land (2015) finds in adult readers from ages 16-61 that isiZulu texts take longer to read, at a mean score of 815 letters a minute, with shorter success, longer fixation durations and more frequent regressions in eye movement than English. Land (2015) further finds an inverse relation between sentence length and reading rate, with slower reading in longer sentences in the same age group. She also finds that when similar letter strings recurred, it slowed down reading rate. Less than 1% of words were skipped, 24% of words may have been immediately recognised and 23% of words had few regressions or fixations. This may illustrate how quickly or slowly reading is in isiXhosa for this thesis, bearing in mind this thesis deals with a younger age group.

While I focus on subtitles and the relationship between visuals and the language and their influence in particular to literacy, my shift of focus in this thesis is not in how the learners can read the subtitles and break them down, but whether they’re able to read them at an adequate speed to follow the subtitles or not. Furthermore, whether learners can make connections between the image and language meanings. This section reviewed the literature in the area of literacy, highlighting word recognition and ORF as early predictors of skilled reading and important decoding skills to lead to comprehension skills. This section has thus illustrated the adoption of ORF tests in this thesis to gauge influence on learner literacy skills in L1 isiXhosa

Foundation Phase. It also highlights the lack of literature in L1 isiXhosa literacy development in the Foundation Phase in terms of subtitle reading and recognition of meanings between the visual and the linguistic modes.

The grain size at the level of the syllable, with the morpheme as a subordinate grain size in isiXhosa, motivated this study to also examine syllables correct per minute (SCPM). This is to ensure that the word reading rate was calculated as accurately as possible, as word lengths in isiXhosa are longer than for the English language due to isiXhosa's conjunctive system of writing. Schaefer and Kotze (2019)'s and Meiring's (2020) results indicate that isiZulu, Siswati and isiXhosa learners could not read for meaning at the beginning of Grade 1 and relatively little at the end of Grade 1. This motivated this thesis to only examine literacy scores from Grade 2 and 3 for fear of floor effects in Grade 1 learners. This next section examines research in subtitles in particular in order to understand literacy with subtitles and SLS.

2.2 Subtitles

This study seeks to examine how same language subtitled (SLS) videos assist in literacy; therefore, it is integral to examine the literature on subtitles. This section examines the types of subtitles according to the literature in 2.2.1, followed by a description of SLS in 2.2.2. The next section, 2.2.3 focuses on SLS conventions, particularly drawing subtitle speeds and agglutinative language research (2.2.3.1). The trends in the research of these subtitles alongside SLS in comparative literature in terms of L2 language/foreign language (FL) learning are discussed in 2.2.4. There is a section on subtitle research and research in South Africa in 2.2.5 and 2.2.6 examines SLS and literacy research.

2.2.1 Types of subtitles

Zanon (2007) separates subtitles into three different types: foreign language (FL) spoken language with L1 subtitles (interlingual subtitles), reversed subtitles and FL subtitles with FL spoken language (intralingual or bimodal subtitles). FL audio with L1 subtitles is termed interlingual subtitles. An example of interlingual subtitles include an Afrikaans series such as *7de laan* subtitled in English for L1 English viewers. This is found more typically in South African television, discussed further in 2.2.5.

Reversed subtitles occur when the video has L1 audio and L2 subtitles (Zanon 2007). An example of this would be if the South African series *7de laan* was dubbed in English for English L1 viewers and subtitled in Afrikaans. This type of subtitle might be useful in foreign language (FL) classrooms, but it is not typically found in a broadcast fashion in South Africa (see 2.2.5). Platforms like Netflix and Showmax, however, often list a variety of subtitles for shows. So, the viewer could have the option to set reversed subtitling on the show if the show has been subtitled in multiple languages.

Foreign language subtitles with foreign language audio are known as bimodal or intralingual subtitles (Zanon 2007). This would be if the series *7de laan* was subtitled in Afrikaans while its original audio is in Afrikaans for an English-speaking audience. Other names for intralingual subtitling include captions, particularly in deaf and hard-of-hearing research, or Same Language Subtitling (SLS). This type of subtitle where the spoken and written language are the same language is the type of subtitle examined in this thesis. Closed captions are different as they not only show the written form of spoken language, but often include descriptions of noises and non-spoken language for deaf or hard-of-hearing audiences (Kuosmanen 2020).

For uniformity, this thesis refers to any subtitle that differs in language from the spoken language as interlingual subtitles. This thesis also adheres to the term Same language Subtitles or SLS to avoid confusion with Zanon (2007)'s limitation for intralingual subtitles to only describe foreign language subtitles paired with foreign language audio and captions. This is also to separate this research from the use of the term captions for the deaf and hard-of-hearing (e.g., research such as Tiitula *et al*. 2018).

SLS "is a form of screen translation which involves the transfer from oral language into written language" (Caimi 2006). The study by Kothari *et al*. (2002), on which this research method is based, has a similar definition, but restricts their definition to motion media programmes of television and film. This definition is not broad enough to include video clips from other sources, such as *YouTube*. Thus, this research uses the definition taken from intralingual subtitles from Caimi (2006). SLS are now described in more detail in the next sub-section.

2.2.2 Same Language Subtitles (SLS)

According to Caimi (2006), SLS can be categorised according to two modern purposes: an accessibility aid for the deaf or hard-of-hearing and a teaching and learning resource for those who are unfamiliar with the spoken language in the video. Subtitles in general first appeared as intertitles, which took up the entire frame, in silent films (Fong and Au 2009). Thus, the first subtitles were, in essence, Same Language Subtitles, except the original audio was absent.

The categorisation by Caimi (2006) illustrates the issue of intralingual subtitles considered only as a foreign language or L2 aid (Zanon 2007) and not as an L1 literacy aid or considering the ways they came to be used. Subtitles began to take over from intertitles still during the era of silent films, however, production found it cheaper to edit text within a picture rather than separate text frames (Fong and Au 2009). The Nordic countries began to take the lead in subtitling techniques. Stamping printed titles onto moist frames was invented in Norway in 1930 and later, these titles were inserted using heat in Hungary, but despite often poorly defined letters due to dirt and erosion, this technique is still used in parts of the world in Eastern Europe and South America (Fong and Au 2009). Chemicals were added to refine this process in between, which is still used today in older film studios, except the plates have been modernised (Fong and Au 2009). Due to many patents, France and Norway dominated the subtitling market from 1933 until the mid-50s.

Based on the history of subtitles, we can note that initially they were created with SLS in mind in silent films, and that Nordic countries were prominent leaders in the subtitling field. Hence, as we look at the agglutinative language Finnish to seek a language similar in structure to isiXhosa, it clarifies the large amount of research on subtitles for Finnish as a Nordic language in comparison to other agglutinative languages such as Basque or Estonian. Furthermore, subtitling developed as its own entity that was at first quite separate from the visual and audio modalities, and as time has passed, slowly became more incorporated in the overall video production process.

While studies in section 2.2.4 show that SLS can be used as a didactic aid, I disagree with Caimi's (2006) description of SLS, as it is not limited to an unfamiliarity of spoken language (Kothari and Bandyopadhyay 2014). Office of communications (2006) already estimated the

current curve facing video-viewing today -- of the 7.5 million TV viewers in the United Kingdom watching with SLS, only around 1.5 million of them were hearing impaired. Studies also illustrate that SLS may be beneficial to assist in focus for people that have ADHD, e.g. Lewis and Brown (2012). In terms of another agglutinative language like isiXhosa, Finnish, this trend extends. Kuosmanen (2020) in a Finnish academic context illustrates an SLS trend of general videos and films viewed regularly that seems to be reported in various popular media, such as the Guardian (Davies 2019) and the Wired (Kehe 2018). Using Reception Theory, she examines subtitle preferences in Finland on Netflix amongst over 154 participants across two surveys in 2014 and 2015, primarily with young adults ranging from 20-30 years of age. Her participants were L1 Finnish speakers and L2 English speakers (Kuosmanen 2020). She found 73% of participants watched with Finnish subtitles in comparison to 20% with other subtitles and 6.3% without any subtitles at all. With no correlation between language proficiency and subtitle usage, Kuosmanen (2020) claims this high volume of subtitle users could be due to the prevalent exposure of participants to subtitles on Finnish television, indicating that a culture of subtitle usage can be established. The high volume of L1 subtitle usage seems to correlate more with frequent exposure to subtitles. This aside from passivity accounted for her findings in participant responses that she categorises using subtitles as a "force of habit" (Kuosmanen 2020:78).

Most participant answers in the study by Kuosmanen (2020), however, described the role of subtitles as aides, finding it easier to comprehend with subtitles. This was very rarely associated with language proficiency issues entirely, such as vocabulary, but those skilled in the languages still found difficulties in terms of accents, unclear speech and incoherent mumbling (Kuosmanen 2020). Volume also played a major role. If headphones are not in use, keeping loud noises under control could be difficult, particularly in the evenings near others, as well as stemming from unbalanced dynamic audio where voices are softer than louder noises (Kuosmanen 2020). Subtitles were thus used as an aide to control the volume, in contrast to constantly having to adjust the volume (Kuosmanen 2020). A few responses also show that if they were not used for themselves, subtitles were used for the benefit of the other people watching with the participant (Kuosmanen 2020). This indicates the social value of subtitles, perhaps in circumstances where it may be difficult to hear the audio when talking amongst each other (Kuosmanen 2020).

Often, issues that caused participants to use less subtitles were that they could conversely be a distraction or that they could be poor quality translations (Kuosmanen 2020). While contexts differ and use of subtitles and perhaps subtitle culture may differ geographically and individually, this study illustrates these as factors for subtitle usage. Furthermore, it also highlights that a lack of studies in subtitle effectiveness in South Africa in Bantu languages, particularly for children, may be simply due to the lack of usage of these language subtitles in South Africa. While the South African Broadcasting Commission (SABC) does have subtitle conventions for English in their Request for Proposals (RFP) released in 2014, this is only for English. Consequently, there is not a standard for subtitles in the Bantu language family. This thesis examines subtitle conventions and efficacy in Finnish, as it is another agglutinative language with a similar orthography to isiXhosa. This is particularly discussed in relation to subtitle speed (2.2.3.1) within the next section. 2.2.3, same language subtitle (SLS) conventions.

2.2.3 Same language subtitle (SLS) conventions

SLS subtitlers have a different role to subtitlers of interlingual subtitling, which have to change more content based on subtitle size limitations and the languages involved. SLS subtitlers focus primarily on the transfer of what is said into a concise message that retains the kernel semantic content and eliminates redundancies if necessary (Caimi 2006).

Effective SLS must be highly legible, with a clear layout including an easily read typeface and a spelling mistake in a subtitle, as well as too many integral changes between subtitle and spoken language, can be distracting for the viewer (Caimi 2006). This thesis particularly examines the conventions of subtitle speed, as the rate of the subtitles can be analysed considering learners' reading speeds. Readers that are familiar with subtitles can absorb the information automatically while non-accustomed viewers may have reduced reading speed as they only spend a section of their viewing time reading, when presented with visuals and audio (Caimi 2006).

2.2.3.1 Same language subtitle (SLS) speed

Kruger *et al.* (2022) highlight subtitle speed as the most misconstrued topic in audio-visual translation research, and an area of a lack of research. (Kruger *et al.* 2022). Kruger *et al.* (2022)

describes that research often contains a lack of comparability between experimental designs, not controlling for genre differences and editing, and a lack of natural speech, by editing and adjusting (see, for example, Szarkowska and Gerber-Morón 2018 and Szarkowska and Bogucka 2019). This thesis controls for semantic, syntactic, and genre differences, as well as the authenticity aspect discussed by Kruger *et al.* (2022). Within this thesis, five isiXhosa-only videos were selected from the same producers from the same *YouTube* channel and included the same narrator's voice for each video (see 4.2.1.2). The videos were further sampled within a range of playback duration (see 4.2.1.2).

When only examining average subtitle speed, Kruger *et al* (2022:215) explains that "[t]his inevitably obfuscates the impact of speed, especially when reporting only global measures." Another area of research has been on the duration of time viewers use to read subtitles (Kruger *et al.* 2022). They state that the duration of reading can only be examined in relation to the efficiency of processing by analysing language in lexical units (words) and understanding the extent in which subtitles are scanned cursorily or extensively processed (Kruger *et al.* 2022). Thus, there is still a lack of knowledge whether subtitles and their presence influence reading speeds outside of the context of reading subtitles, such as learners' L1 literacy skills in every-day reading. The average subtitle reading speed in this study was calculated for length of time which only contains subtitles, and not overall video length, which further controls for these variables. This thesis examines the influence of speed, as well as comparing the rate learners can read when constrained for time and the subtitle rate to determine if these were compatible. This thesis examines this by using reading tests in an intervention in comparison to the average word and syllable rates within the videos.

Before discussing further insights from Kruger *et al.* (2022), these discussions are integral to identifying gaps in research that examine subtitle reading and subtitle reading speed. As illustrated in 2.2.4, 2.2.5 and 2.2.6, many of the research in subtitles and SLS are in the context of L2 language learning, particularly adults, with few focused on early L1 literacy, which is the context of research in this thesis. Many of the studies in the sections that follow (2.2.4, 2.2.5 and 2.2.6) use eye-tracking software, which could have indicated whether learners in this thesis gazed at the subtitles and the length of time in which they did. However, eye-tracking software was not seen as the best research tool for this investigation and purpose of this study, as this software, aside from the finances involved, is incredibly difficult to use in a peri-urban

context in terms of mobility and electronic access. If this could be achieved in further research, this would add further layers of investigation to what this research achieves.

Furthermore, an issue in the context of measuring average word speed by character per minute in eye-tracking research (e.g. Kruger *et al.* 2022), particularly with English, assumes that the decoding of a word takes place at the level of the phoneme (see 1.2.1.1). As discussed for isiXhosa in 1.2.1.1, this is at the syllable, which includes at least two characters. However, due to the complexity of orthographical features (i.e. many characters representing particular sounds), this thesis calculates average word and syllable per minute rates, rather than examines characters. Learners do not decode words at a grapheme level of characters for isiXhosa for this context.

Kruger *et al.* (2022) explains that if subtitles are too fast, readers may not be able to resolve linguistic (decoding-level) issues, as they may not have time to re-read or re-scan areas of the subtitle previously read. They also point out that unlike in static reading, on which this thesis particularly focuses, subtitle reading involves a process of horizontal reading (subtitle only) and vertical reading (between subtitle and visuals) (Kruger *et al.* 2022). They explain they serve different purposes, as horizontal reading may indicate decoding and comprehension processes at the sentence level to make sense of the written subtitles (Kruger *et al.* 2022). This connects to linguistic literacy, skills for metalinguistic awareness and an intramodal literacy stratum of semiotic awareness (see 1.2.3). Vertical reading (i.e. integrating visual images with the text) may allow for a more advanced-level comprehension, which is found in longer passage reading (Kruger *et al.* 2022). This connects to multimodal literacy and an intermodal literacy stratum of semiotic awareness (see 1.2.3). Kruger *et al.* (2022) claim that a reduction in either types of these reading for multimodal texts would hamper or at least reduce linguistic processing in subtitles and their integration with visuals for better comprehension of the video. This is also described by a multimodal-integrated language framework for multimodal reading (Liao *et al.* 2021), which is discussed as theoretical framework in this research in 3.2.

The integration of semiotic resources in a multimodal text depend on accurate word recognition, object recognition and the processing of sentences (Kruger *et al.* 2022). If information is missed by readers or there is ambiguity, more cognitive processing of an aspect is required, which could result in the missing of other aspects (Kruger *et al.* 2022). Less time available for text processing could thus result in vital aspects from the subtitles being missed

or misinterpreted (Kruger *et al.* 2022). The frequency and length of words (the complexity of the linguistic mode) could slow down the reading process, increasing the likelihood that a sentence cannot be processed before the subtitle fades from the screen, else if the reading increases to the speed of the subtitle, linguistic areas of complexity could not be processed adequately (Kruger *et al.* 2022).

While Kruger *et al.* (2022) conducted a study with adult L1 English participants and videos from a nature documentary, their findings reveal that fast subtitle speeds influence depth of subtitle processing. As the speed rate rises, the reading speed of viewers increases, which leads to fewer fixated words, more skipped words and potentially difficulties in sentence processing (Kruger *et al.* 2022). Subtitle rates that are too fast may likely negatively influence both the language processing, as well as integration of the written linguistic mode with the other modes (Kruger *et al.* 2022). More complex visuals and more complex visual interactions may result in less cognitive capacity by viewers to engage in subtitle processing, either resulting in a lack of object or word recognition (Kruger *et al.* 2022). Kruger *et al.* (2022) state that fast subtitle speeds do not imply that viewers may have the inability to change their cognitive literacy strategies to follow the subtitles, but it will influence the depth of that processing.

Furthermore, other agglutinative languages, such as Turkish, Hungarian, Estonian, Basque and Korean (Chai 2007; Mihajlik *et al*. 2007), are also lacking in SLS research. Nevertheless, for the purpose of another agglutinative subtitle rate to use as a baseline to compare the type of rate of the videos in this thesis, I turn to discussions of Finnish subtitle norms and rates. Lappalainen (2021) states subtitles should be divided into no more than two lines and words belonging together (e.g. nominal groups) should appear in the same line. The maximum duration of a subtitle is seven seconds, and the minimum duration is 1.8 seconds (Szarkowska 2016). There is a 12–14 character per second (CPS) maximum speed suggested for reading (Szarkowska 2016). Szarkowska (2016) states the maximum character per second cap for Finnish is 16. Countries that use words per minute (WPM) are the UK, Australia, Spain, Portugal and Iran, while the Nordic countries tend to adhere to CPS (Szarkowska 2016).

Kuosmanen (2020) provides an idea as to a WPM average reading speed for Finnish, stating that sources differ whether it should be 150-180 WPM or 200-250 WPM. Yet Sandford (2016, cited in Kuosmanen 2020) found 9% of viewers felt subtitles too fast at 170 WPM, while the percentage rose to 22% at 200 WPM, indicating the latter range may not be the most ideal.

However, it is important to note that the Finnish context is radically different to the South African context in that most of the Finnish population are exposed early to subtitles in various ways, and these figures examine adult viewers/readers proficient in reading in their L1. While this is a start, this does not tell us how early readers of the language would fare with subtitles in South Africa, as WPM reading speeds in South Africa are very different. In a study by Land (2015), adult readers of isiZulu had mean scores of 122 and 114 WCPM on an easier set of texts, and 91 and 96 WCPM on a more challenging set of texts. The early reading benchmarks for isiXhosa are 20 WCPM for Grade 2 and 35 for Grade 3, thus for early readers, their reading rates are far lower than adults reading rate abilities (Ardington *et al.* 2020).

A study by Pihlajakoski (2008) examined children's preferences in Finland between a dubbed version of the movie *Shrek* versus a Finnish subbed version, and children seemed to prefer the dub. She indicated that subtitling conventions in Finland were so that dubbing was done with children and subtitles seemed to be more catered towards adult viewing. While interest remains a factor, not only in terms of the video, but also towards the subtitles themselves, this also may indicate the interest a particular child may have for reading, as well as their literacy level. This may also factor into their use and reading of the subtitles on-screen. If there is a lack of interest by the learner, literacy cannot be improved. The next sub-section provides a broad overview in trends in the literature that discuss interlingual and SLS subtitles together in terms of L2/FL learning.

2.2.4 Subtitles as a didactic aide for L2 language/FL learning

When examining the literature on the two types of subtitles, some major trends in the research can be established. In both interlingual and SLS research, the focus has often been on L2 language learning (Garza 1991; Neuman and Koskinen 1992; Huang and Eskey 1999; Caimi 2006; Kruger *et al*. 2013) or foreign language (FL) learning (d'Ydewalle and van de Poel 1999; Caimi 2006; Karakaş and Sariçoban 2012) with mostly adult learners (Markham 1999; Bird and Williams 2002; Perez *et al*. 2013). Caimi (2006) argues that for these videos to assist learners in the areas of language learning, they need to be supported and complemented by appropriate teaching materials, which allows listening comprehension and information retrieval to be facilitated and enhanced. Kothari and Bandyopadhyay (2014), however, illustrate that simple viewing of SLS without any additional materials can still improve learning, and even literacy. Most studies in this section examine English or another language

Indo-European language family, unless specified otherwise. None of these studies except for the South African literature in section 2.2.5 that examine Sesotho investigate a language from the Bantu language family as examined in this thesis.

This trend of focusing on FL and L2 language learning dates to 1983 with the first FL subtitle study on English (Price 1983). Comparative language studies between interlingual subtitles (L1 subtitles) and SLS (EFL or ESL) of English videos report two main findings: either it was discussed that L1 subtitles of English videos are potentially better than SLS English videos (Markham *et al*. 2001; Markham and Peter 2003; Bianchi and Ciabattoni 2008) or vice versa (Stewart and Pertusa 2004). In Taylor's (2005) and Specker's (2008) studies, SLS of English videos were not found to assist adult students at all. All these studies focused on Spanish and English and on intermediate language learners, except Taylor (2005) who focused on beginner language learners and Bianchi and Ciabattoni (2008) who focused on Italian language learners. They also included beginner and advanced students in their studies. Aside from the language and literacy skills, SLS has been shown to assist in overall motivation and they can be used as learning and teaching tools to test the memory's capacity to retain information (Caimi 2006).

Based on the varied results of these studies, Matielo *et al*. (2015) suggest that literate adult students, who are at an intermediate level of their FL, would have improved listening comprehension if introduced first to interlingual (their L1) subtitles. They are able to understand SLS videos with less difficulty than videos without subtitles, as their FL literacy skills progress (Matielo *et al.* 2015). While these studies focusing on L2 language or FL learning differ from the literacy lens this thesis takes, one can observe from their findings that adult students who are more unfamiliar with a language tend to do better with their L1 subtitles. As 3.1 in Chapter 3 discusses, there is potential that when L1 subtitles are combined with L1 spoken languages in video, there is opportunity for reinforcement of those subtitles for easier cognitive processing by the viewer.

There are also major trends in the methods of this type of study. Most of the studies examining interlingual subtitles and SLS tested students in terms of vocabulary and comprehension (Vanderplank 1990; d'Ydewalle and van de Poel 1999; Markham 1999; Markham, *et al*. 2001; Markham and Peter 2003; Stewart and Pertusa 2004; Danan 2004; Bianchi and Ciabattoni 2008; Guichon and McLornan 2008), with some examining both skills simultaneously (Huang and Eskey 1999; Winke *et al*. 2010; Perez *et al*. 2013). Less studies focused on word

recognition in subtitles (Neuman and Koskinen 1992). Most studies used control group experiment designs similar to this thesis (Markham *et al*. 2001; Markham and Peter 2003; Stewart and Pertusa 2004; Taylor 2005; Bianchi and Ciabattoni 2008; Winke *et al*. 2010; Perez *et al*. 2013; Matthew 2020). Further psycholinguistic studies also used eye tracking methods (d'Ydewalle and De Bruycker 2007; Specker 2008; Perego at al. 2010; Kruger *et al*. 2013; Winke *et al*. 2013; Kruger *et al*. 2014). These studies illustrate that more listening comprehension and vocabulary skills have been examined in terms of L2 language and FL learning with subtitles, and that control group experiment design and eye tracking seem to be common methods of investigation when examining subtitles. Consequently, more attention should be given to the skill of word recognition, which this thesis examines, and this confirms the use of the control group design component of this thesis with former studies in this field.

A study by Winke *et al*. (2010) suggests that the effectiveness of SLS depends on the language in question. They examined SLS with EFL university students, whose L1 included Arabic, Chinese, Spanish and Russian and it was shown that SLS were more effective aides when they completed vocabulary tests than if they completed the tests without SLS. They found SLS to be "attention depleting" (Winke *et al*. 2010), which highlights the potential that SLS has to reduce cognitive load. Yet another aspect that made this study unique was its examination of various L1 languages simultaneously. Winke *et al*. (2010) found a trend that captions were more beneficial for participants with Russian and Spanish as their L1 in their first viewing of the videos than their second and the opposite for the L1 Chinese and Arabic speakers. This illustrates that former research made by d'Ydewalle and Gielen (1992) on the parsing of subtitles being automatic (regardless of familiarity with subtitles) might not necessarily be the case for all languages. This was particularly noteworthy in Specker's (2008) thesis, where Arabic L1 children had longer eye movements patterns in reading captions while English L1 children had faster eye movements. The effort in cognition that is required to parse SLS may influence potential benefits utilised by the learner (Winke *et al*. 2013). While this study uses SLS, it is important to remember that the viewers were all EFL learners and therefore the SLS were not in their L1. This observation made by Winke *et al*. (2010) and (2013) that different languages may influence subtitle parsing can be found in South African subtitle research, which is discussed in the next section.

2.2.5 Subtitle research and exposure in South Africa

Difficulties were found in L1 Sesotho and L2 English subtitling of a French spoken language video, i.e. interlingual subtitling (Hefer 2013a). Adult participants were shown to spend too much time on subtitles and while reading was slightly faster for the English subtitles than Sesotho subtitles, reading rates and comprehension of subtitles were considerably lower. In other words, the viewers read too slowly to fully follow and comprehend the text. While subtitles with these languages do not seem to result in a cognitive overload (Kruger *et al*. 2013), Kruger *et al*. (2014), found, when examining participant eye tracking, that the Sesotho subtitles were read much less than the English ones, often even avoided. This is in comparison to adult Afrikaans-English bilingual speakers in which L1 Afrikaans subtitles were read faster and more efficiently than L2 English subtitles (Hefer 2013b). In addition, within a different educational context, Matthew (2020) reported that subtitling English lectures made very little difference in comparison to when students with English as L2 viewed non-subtitled lectures. This indicates that challenges and opportunities vary according to participants' literacy levels, subtitle familiarity and the languages involved, but also that the study of subtitles in South Africa seems to also focus on adult comprehension of subtitles and not in early L1 literacy.

South Africa, as mentioned in Chapter 1, has 12 official languages, not including unofficial languages and dialects used in the country. However, the language of subtitles, when they occur at all, are all in English (Olivier 2011), which as mentioned earlier is only the fourth most spoken language in the country. Examples of this include the soap operas *Isidingo* which aired up until 2020, *7de laan*, *Muvhango* and *Generations* that air/aired on the South African Broadcasting Commission (SABC) (Kruger 2012). In 2012, while 37 programmes were reported as subtitled on SABC, the Independent Communications Authority of South Africa (ICASA) stipulates that all broadcasting channels (including SABC and e.TV) should provide English subtitles, when the spoken language is not English (Kruger 2012; Commissioning Protocol 2010). Olivier (2011) argues against this, particularly in incidents of SLS, opting for bilingual or pivot subtitling approaches to cater for more languages simultaneously.

Furthermore, in 2012, 76% of programmes on SABC were in English, thus further illustrating an Anglophone trend in South African media, both due to sourcing media from the US and Europe, but also due to the political and historical context of English in South Africa and its global status as a lingua franca (Kruger 2012). The annual report from SABC for 2020 does

not go into detail as to the number of English programmes versus non-English, rather how successful these targets were when compared to ICASA quotas (SABC Annual Report 2020). Nevertheless in 2019, SABCs submission on the ICASA draft indicated that English was still the default language for subtitling and subtitle translations into the 11 official languages at the time would be "cautiously considered"[13] (SABC 2019). SABC had further stated that they had acquired specialised software for subtitling. According to DSTV, subtitling is done either by a third party or Sky, and all their apps are also like this which include BBC Player and Showmax (DSTV Now 2022). None of their subtitles are for languages within the Bantu language family.

Reasons for the lack of subtitle research in Africa in particular may be due to economic and technological constraints (Mubenga 2020). Furthermore, since African countries tend to not produce their own subtitles, and media outside of Africa often comes already subtitled (Mubenga 2020). This may also be why research in subtitling is lacking, as there is no motivation to better a practise that is not localised across the continent as it may be in other continents (Mubenga 2020). Subtitling software is expensive and thus not a priority for many African countries with large socio-economic issues (Mubenga 2020). The lack of language subtitles from the Bantu language family illustrated in this sub-section further highlight the lack of subtitle standards and conventions, including for SLS, for these languages.

The reviewed literature illustrates the focus of this research on L2 and FL learning in adults and specific languages in the Indo-European family, with little research on Bantu languages and word recognition. Nevertheless, insights can be gained from these studies in that SLS seem to mostly benefit language learning overall, particularly in reading comprehension and vocabulary. However, this may differ according to the language in question. South African research illustrates that for Sesotho L1 adult speakers, SLS may be more of a challenge than a benefit. Within comparative studies, SLS seems to assist in the improvement of comprehension and vocabulary in children, like they do with adult students (Neuman and Koskinen 1992). Van Lommel *et al*. (2006) found that unlike vocabulary, grammar may be too complicated for children to acquire from short SLS videos. The next section solely focuses on SLS used in the literacy context.

[13] This is not including South African Sign Language, as it was adopted as the 12th official language in 2023, see du Plessis (2023)

2.2.6 SLS and literacy

Unlike the plethora of literature on reading comprehension in L2 language and FL learning with SLS, the use of subtitles in the area of L1 literacy is lacking research (Black 2021). Black (2021) states that of the ten SLS studies that research SLS efficacy in terms of literacy (nine for L1 literacy, one for L2), nine of the SLS studies showed that SLS improves literacy, in terms of vocabulary, word recognition and comprehension. She also notes that the majority of the research is on English (Linebarger and Walker 2005) and was conducted over a longer time period. For example, Linebarger and Piotrowski (2010) examined 7-year-old children's literacy improvement in the US with English and found SLS to improve word recognition skills, comprehension and the reading of pseudo-words, yet there was no improvement in ORF. In New Zealand, Parkhill, Johnson and Bates (2011) examined subtitles in literacy improvement with 10–13-year-olds and found improvements in comprehension, vocabulary, ORF and engagement with reading.

Suzanne and Tiokou (2015) conducted a study in Cameroon with university students, asking their preferences of subtitling. While the study acknowledged the potential benefits of subtitles, Suzanne and Tiokou (2015) states that students reported experiencing difficulties with the speed of the subtitles. This again highlights the aspect of subtitle rate as an area of research. Minucci and Cárnio's (2010) results were inconclusive, and they suggested that since 2nd graders are at the decoding level and 4th graders are at comprehension level in reading, the way they are taught at school will influence their reading of subtitles.

This thesis draws heavily on literacy studies on SLS in India. In 2001, 273 million illiterate people and 389 million officially literate people who cannot read simple texts were identified (Kothari 2014). Kothari *et al*. state:

> "Is there such a butterfly for literacy? A butterfly that could flutter ever so gently to transform the literacy forecast of a nation that is one billion strong and home to one-third the world's illiterates? Possibly, even if it sounds somewhat overstated. And that simplest of creatures is 'same language subtitling' (SLS)" (2002:55)

Their intervention was undertaken, like the one in this study, within three months at a school representative of most government schools, with a pre- and post-test study design (Kothari *et al*. 2002). These were conducted with three groups of Grade 4 and 5 learners, with 46 children in each group (Kothari *et al*. 2002). One group was shown five subtitled Hindi songs in a 30-

minute session, the other was shown the songs without subtitles and the third group was a control group (Kothari *et al*. 2002).

The test comprised of four sections of separate words (Kothari *et al*. 2002). There were 38 mono-syllabic words in the first section and there were 20 words ranging between two and four syllables in each of the second, third, and fourth sections (Kothari *et al*. 2002).

In the post-test scores in the monosyllabic section, an additional 1.57 syllables in the subtitle group were read (Kothari *et al*. 2002). In the post-test scores in the higher syllabic section, an additional 2.93 syllables and 1.51 words were read (Kothari *et al*. 2002). The mean scores were highest in the subtitle group, second highest in the non-subtitle group and lowest in the control group (Kothari *et al*. 2002). Thus, the improvement of learners' ORF scores were due to subtitle exposure and not video exposure in this context (Kothari *et al*. 2002).

In their 2004 study, Kothari *et al*. examined whether an improvement was also visible for adults and youth who were exposed to Hindi film songs. 1500 participants from two villages and two slums in two districts were exposed to 25 episodes at 30 minutes each for 6 months, thus 8 hours in total of SLS exposure in 6 months.[14] It also followed a control and experimental group design, but only separated participants into two groups. Results indicated a difference, but more discreet than with the children. Post-experiment, there was found to be an increase in reading for the experimental group at 4.4 syllables a minute in exercise 1, 3.5. syllables per minute/2.4 words per minute for exercise 2 and 4.1. syllables/2.6. words per minute for exercise 3.

In the same year, Rangoli was also examined with the same test group for SLS (Kothari 2008). They found that 56% of learners could read proficiently after five years with exposure to videos at 30 minutes a week, whereas 24% of children became good readers without this exposure (Kothari 2008). Furthermore, in the no-subtitle group, 25% of learners could not read after five years, while 12% could not read in the subtitle group. In other words, the percentage of improved readers was doubled by SLS and the number of illiterate children was halved (Kothari 2008).

[14] See 4.1.2. The longitudinal nature of these studies, which highlight better results, would have been ideal to emulate in this thesis. However due to the financial constraints of this project and it having to be completed within a timeframe of three years, this thesis could only model the three-month period Kothari *et al.* (2002).

A difference was also seen after five years in adults, as 12% could read proficiently in the subtitle group, while only 3% illiterate adults could read proficiently in the no-subtitle group. There were 83% of adults who could read in the no-subtitle group after five years, while a less adults (68%) remained illiterate in the SLS group. In both earlier studies, exposure to captioned videos of film songs contributed positively to decoding skills (Kothari 2014). Kothari (2014:44) doubled weekly exposure and extended the overall period of exposure, reporting the following results: "Where schools are failing to get nearly 45% [of learners] to decode a single letter after five years of schooling, the SLS complement at home brought this shocking figure down to 13%."

This (2014) study by Kothari built on the two earlier studies. Since a similar crisis is occurring in South Africa, similar interventions such as this one for Hindi may assist learners in a South African context for L1 literacy, such as with isiXhosa.

This section has examined the types of subtitles, including SLS and situating this type of subtitle within the broader contexts of L2 language and FL learning, as well as literacy. Finnish SLS conventions were examined to gain insight into potential isiXhosa SLS conventions. Challenges of reading Sesotho L1 subtitles in adults indicate challenges isiXhosa learners may face when reading in their L1, as not only do isiXhosa subtitles rarely occur in everyday media, thus there is no subtitle exposure, but also, (see 1.1.1), there is a literacy crisis for this language. L2 language/FL learning research indicates that SLS may provide opportunities to improve vocabulary, listening comprehension and word recognition, amongst skills such as memory and motivation, which can easily be utilised when reading.

As this section has discussed the results for L1 speakers' exposure to subtitles are generally positive; however, a longer continuous intervention seems to be the most ideal. On the other hand, for L2 and FL learners, the results are mixed. While subtitles can enhance certain aspects, they can also have counterproductive effects in some cases. Subtitled videos contain many meanings that are communicated across multiple modes. The next section examines multimodality and while there are very little studies that examine subtitled videos, literacy and multimodality altogether, the next section highlights important studies to understanding the field and contributing towards this research.

2.3. Multimodality

This study examines subtitled videos as a literacy resource with a Digital Multimodal Discourse Approach. In order to situate this thesis and its research questions in relation to other multimodal studies, the literature on multimodality and video is reviewed and discussed, with particular emphasis on the areas of language learning and literacy. Within this section, the literature is organised according to the different types of multimodal analyses stemming from different disciplines. This provides a brief review of these types of analytical frameworks and explanation as to the choices that I made to select an analytical framework for this thesis. Section 2.3.1.1 discusses relevant studies on the analysis taken for this thesis, namely DMDA and SF-MDA. Following this, 2.3.1.2 discusses geo-semiotic and multimodal corpus analysis studies. Section 2.3.1.3 then discusses social semiotic analysis studies, because this is the analyses that led to the creation of DMDA. Consequently, 2.3.1.4 elaborates on multimodal ethnographic analyses that took place, particularly in a South African context in schools, which is similar to the context of this study. Section 2.3.1.5 examines a relevant study that uses multimodal interactional analysis as it looks at the triangulation of student feedback and text that closely resembles this thesis' methodology. Finally, 2.1.1.6 briefly describes multimodal translation and multimodal reception analysis to introduce the way subtitles are examined multimodally in the literature. The actual DMDA systems used in this thesis are discussed in Chapter 3 (section 3.1).

2.3.1 Multimodality and video

As mentioned in 1.2.3, a mode "refer[s] to a socially organised set of semiotic resources for making meaning" (Jewitt *et al*. 2016:221). Language and image are examples of modes (Stöckl 2004). The study of multimodality or how multiple modes work together to create meaning has stemmed from multiple disciplines (Kress 2015) such as linguistics (Kress and Van Leeuwen 2006), anthropology (Scollon and Scollon 2003), and fine art (O'Toole 1994). Jewitt *et al*. (2014) outline eight potential types of multimodal analysis:

- multimodal social semiotic analysis
- multimodal interactional analysis
- systemic functional discourse analysis (SF-MDA)
- multimodal reception analysis

- multimodal ethnographic analysis
- multimodal corpus analysis
- geo-semiotics
- multimodal conversation analysis

The current analysis DMDA, stemming from SF-MDA and multimodal social semiotic analysis, integrates specific systems of analysis. These systems not only emphasise the visual like the former social semiotic analysis (Jewitt 2016 *et al*.; Stein 2008), but rather have a more balanced relationship of interpretation between the linguistic and the visual. While the multimodal social semiotic approach was an initial and preferred analysis for multimodality, the later SF-MDA is systemic and functional, a framework that works on building the shared systems under the metafunctions first discussed in Kress and Van Leeuwen (2006). The nature of SF-MDA highlights meaning at the micro-level whereas a multimodal social semiotic analysis focuses on the macro-level (Jewitt 2009), thus is less fine-grained. However, there are more studies in multimodal social semiotic analysis and other analyses than SF-MDA and DMDA, and consequently this sub-section discusses relevant findings of studies of SF-MDA and DMDA (2.3.1.1). The definition and description of DMDA and SF-MDA, due to its complexity and for ease of reading, are described separately in 3.1.

2.3.1.1 SF-MDA and DMDA

SF-MDA "understand[s] the ways in which meaning systems are organised and used to fulfil a range of social functions" (Jewitt *et al*. 2016). Due to the integration of various digital systems and the inclusion of other systems that may not be easily organised according to Systemic Functional Theory categories, Tan *et al*. (2015:559) uses the term "digital multimodal discourse approach" to which this thesis adheres. The majority of video studies with SF-MDA and Digital Multimodal Discourse Analysis (DMDA) have examined classroom interaction and classroom discourse that included recorded video (Tan *et al*. 2016; Fei 2011; Amundrud 2017). *Multimodal Analysis Video* has been used to understand 3-D virtual spaces used by university students in English classrooms (Tan *et al.* 2016). Fei (2011), on examining classroom interaction, illustrated that the teacher did not treat the video as an integral learning resource, further not elaborating relevant information on them to students, although most of the classroom time was taken up by the videos (Fei 2011:238-239). He states that video "present[s] semiotic selections for the teacher in the construction of the lesson experience for students"

(Fei 2011:56). This illustrates the potential of video, but also challenges when it is not used effectively. *Multimodal Analysis Video* has also been used to understand interaction between university students in English classrooms (Amundrud 2017: xiii). These studies are not described in detail in this thesis as they examined recorded natural discourse in the classrooms, which is a different context to the research context in this thesis.

Tan *et al*. (2015) discuss DMDA used in this analysis, which combines social semiotic and critical discourse analysis using software to analyse divergent leadership styles in higher education videos. While they examine intonational systems, they do not include subtitle rate, such as this thesis does for pedagogic purposes. Since O'Halloran and Fei (2014) examine television advertisements, the findings from these studies of the genre of the texts analysed do not really add to this thesis' research topic. Therefore, this literature review sub-section focuses on other forms of multimodal analysis, namely multimodal social semiotic analysis, multimodal ethnographic analysis and multimodal interactional analysis. It explains why geo-semiotics, multimodal conversation analysis and multimodal corpus analysis are not examined for the purpose of literature review in 2.3.1.2.

2.3.1.2 Geo-semiotics, MCA and multimodal corpus analysis

Geo-semiotics, popularised by Scollon and Scollon (2003), is a "descriptive analysis of the semiotics of sign, their emplacement and the participants' embodied interaction in the space being investigated" (Jewitt *et al*. 2016). Since the emphasis in this type of analysis is on space and signs, i.e., a focus on geographic location, and video and subtitles have not been looked at in regard to this type of analysis, this thesis does not examine literature on this multimodal analysis type. Furthermore, multimodal conversation analysis (MCA) focuses on conversation and multimodal corpus analysis on a large collection of texts. This thesis seeks to examine subtitled videos as an L1 literacy resource for learners. Multimodal corpus analysis could have had the potential to assist with this aim, but there are not enough isiXhosa subtitled videos to establish a corpus of videos as of this thesis' publication. This thesis, unlike a focus on only the text, or the space and culture in which the text is situated, is concerned with the triangulation and relationship between the meanings conveyed from the multimodal video text and the reception of that meaning to build literacy skills for the readers. Prior work on children's literature with systemic functional linguistics-based analyses focus on the literature and ideological assumptions (e.g. Smith and Adendorff 2021) rather than assumptions that can be made for learners' literacy improvement. Since multimodal analysis in terms of children's

literacy is rare in the literature, this section now focuses on literature relevant to this thesis topic that use multimodal social semiotic analysis with videos. It also broadens the areas to language learning and not limiting the field of research to literacy due to the lack of literature on this topic in a multimodal area of research.

2.3.1.3 Social semiotic analysis

Due to social semiotic analysis stemming from Hodge and Kress (1988) and Kress and Van Leeuwen (2006), video appears to be examined with this analysis the most, ranging from advertisements (O'Halloran and Fei 2014) and comedic videos (Adetomokun 2018) to story-telling (Gachago *et al*. 2014), music videos (Brady 2015) and software (Kaltenbacher 2004). A social semiotic analysis typically entails "detailed analysis of texts, sometimes a few, sometimes a larger collection and sometimes involving historical comparisons" (Jewitt *et al*. 2016). Eventually, a metafunctional framework stemming from Systemic Functional Linguistics was included by Kress and Van Leeuwen (2006) and Van Leeuwen (2005) to eventually lead to what O'Halloran (2008) refers to as SF-MDA. In terms of language learning, Videos incorporated in English FL software "often deliver 'multimodality' which actually disrupts learning" (Kaltenbacher 2004:119). Flat-toned speech and non-synchronised movements of linguistic articulators, i.e., unnatural gestures, that may lead to FL learners misunderstanding information due to paralinguistic cues. It could also lead to a misinterpretation of conveyed meanings that could negatively impact learning (Kaltenbacher 2004). Due to the text-centric focus of this study, Kaltenbacher (2004) does not include findings from the viewers' perspective, and whether English FL learners would notice and be affected by these modal discrepancies. This highlights one of the disadvantages of this kind of analysis by Iedema (2001), in that due to its primary concern with textual structures, social semiotic analysis does not examine viewers' readings of the text.

Jewitt (2006:2), however, illustrates how this kind of multimodal analysis can be used to better understand learning, examining how technology impacts teaching and learning in classrooms in schools. Her analysis of an interactive English novel *Of Mice and Men,* which had different modes such as video, spoken language narration and image, to allow for different engagement with the story, illustrates the potential for learner engagement at the level of the narrative and mode. "Modal learning" (Jewitt 2006:2) allows for particular moments, objects and characters in a story to be emphasised, allowing for various readings of the same novel via video, image and text, as some learners read the story by only watching the integrated videos (Jewitt 2006).

Her study shows that children do not always read a text as it was designed to be read initially in static text and visuals in a format from beginning to end (Jewitt 2006). They prefer to use different modes such as video to 'read' the story, and they prefer visuals and colour over writing and movement (Jewitt 2006). Thus, the visual mode is the preferred one, which may cause images to be a distraction to children in their learning from the novel and improvements in their English language use. Jewitt (2006) illustrates how they are simply another potential mode for learning to occur, and if more effectively in this mode than it does not necessarily act as a hindrance. She emphasises that learning may be negatively impacted if student modal preferences and principles of reading differ from principles of the curriculum designer and teacher (Jewitt 2006). She also states that interest and engagement are particularly important for learning (Jewitt 2006). Thus, learning is defined as "internalising the representational and communicative means of the subject discourse. Put more directly, learning involves taking on the identity and discourse of the [school] subject" (Jewitt 2006:25). However, this does not consider external learning practices and she describes that learners can assume a specialist identity, which may not always occur. There is often not enough of a balanced approach in her analysis, with less of a focus on linguistic content, which is the hallmark of a social semiotic approach (Stein 2008; Jewitt *et al* 2016). This thesis seeks to avoid this by placing a balanced focus on all modes involved. Nevertheless, it is also important to note that the creation of the study of semiotics included language as a mode (Hodge and Kress 1988).

In South Africa, multimodal social semiotic research has been conducted on a variety of different discourses and contexts (Banda and Oketch 2011; Mafofo and Banda 2014; Ferris and Banda 2015). Archer's (1997) social semiotic research in South Africa allowed her to plan an English lesson with university students, teaching stereotypes with a video from *Mind Your Language*, which is a British sitcom. The analysis only discusses visuals (Archer 1997:210-216). Student results are not triangulated with a micro-level video analysis triangulated, rather Archer (1997) explains it was dictated by the teacher, with the semiotic analysis conducted by the researcher and teacher informing curriculum (Archer 1997:170-171). However, to understand different multimodal approaches to pedagogy, Archer (2004:41) examines writing in higher years of study in tertiary education in South Africa with multimodal social semiotic approaches such as talking, writing online and mind-mapping (Archer 2017).

Multimodal pedagogy and discourse in the classroom space are prominent contexts in South African multimodal research. Stein (2008) also observes classroom practices and resources in

a social semiotic approach that includes drawings, video-taped recordings of children performances and diary and story writing. Due to her focus on how children create meaning, these resources are examined with a multimodal social semiotic analysis and not focused on how these improved their literacy scores in the subjects taken. Yet in the manner that Stein (2008) examines opportunities and challenges for pedagogy through practises, this thesis seeks to do the same with video. However, SF-MDA is a more balanced, systematic analytic approach in comparison to the social semiotic, which can often overemphasise the visual over the linguistic. SF-MDA is described in detail in 3.1. Stein (2008) did find that teachers "translate" knowledge into different modes to promote understanding from learners and that all learners potentially have preferred semiotic modes, where in the classroom space writing is seen as the dominant mode. Learners are then assessed as not doing well in school when they do not perform well in the written mode, yet this simply may not be their preferred mode of expression. Stein (2008) states that the written mode offers use of action and narrative, for learners to concentrate more on happenings, while visuals like drawings depict detailed moments of the action which may not be present in the written mode.

Guijarro and Sanz (2008) conduct a multimodal analysis of a picture book to determine the extent of meaning created by the linguistic and visual modes. They state that animals are frequently utilised in children's literature (Guijarro and Sanz 2008). Animals are usually anthropomorphic with human attributes, which allow producers of text to eliminate social category biases that human characters could provide, such as gender and social status (Guijarro and Sanz 2008). The following systems within children's narrative patterns are described and defined in 3.1.1. According to Guijarro and Sanz (2008), children's narrative patterns include primarily Material and Mental Processes more than Verbal, Relational and Behavioural Processes. This is due to children's books showing actions and feelings and thoughts of characters, and Circumstances of Location relating to the setting of the stories (Guijarro and Sanz 2008). There are mostly "offers of information" with very little demands or requests (Guijarro and Sanz 2008:1604). Long shots are primarily used which imply objectivity and distance, however, Front and Medium shots create involvement with the viewer, as they suggest an equal power relationship between Participants and viewers (Guijarro and Sanz 2008). Placing Participants visually Centre allows for Salience, and the main Participants are usually located in the initial slot of the clause as Themes so that the young child, still learning to read, does not struggle to comprehend the story (Guijarro and Sanz 2008). However, it may

run the risk of making a text somewhat tedious, despite being easy to follow (Guijarro and Sanz 2008).

Zohrabi *et al.* (2019) in their analysis, however, illustrate how if these systems deviate from the aforementioned patterns, particularly when both human and animal Participants are present, there may not be enough involvement between viewers and animal Participants for better comprehension and reading experience. Thus, interactions between viewer and human Participants may be better for learning for the stories in their analysis than the interactions between viewer and animal Participants. This primarily relates to deviations of Front and Eye-level shots. This thesis takes the importance of these shots into account when determining whether the learners find Participants in the videos engaging. However, further research on differences between animal and human Participants with quantitative tasks is needed, as this was not the focus of this study.

In their article on time in children's narratives, Tseng and Djonov (2023) state that the co-patterning of temporal features is an integral variable to consider when using video for language comprehension. Their multimodal study examined how 28 children aged 7-10 years old comprehended time from two Disney animations (Tseng and Djonov 2023). They state that by 4-5 years old, children have basic causal reasoning of chronological events, can distinguish between past and future activities, and list activities of a day (Tseng and Djonov 2023). The children in this thesis are around the age of the children in this study. It is assumed, therefore, that they have a similar temporal awareness in terms of early childhood development at 4-5 years old. Their study highlighted that a co-patterning of temporal features resulted in better comprehension of a non-sequential scene, but frequency of exposure did not make a significant difference on comprehension (Tseng and Djonov 2023). However, this thesis seeks to examine videos as a literacy resource more broadly, and while Circumstances of Location: Time appear in the videos frequently, comprehension of time was not specifically tested in the research. Nevertheless, this study indicates that co-patterning with clear temporal cues to ensure a film sequence is cohesive and coherent is more influential on comprehension than the frequency of exposure to a particular video. Further research is needed to investigate time, temporal features and Modality in the context of this thesis.

2.3.1.4 Multimodal ethnographic analysis

Prinsloo and Stein (2004) use a multimodal ethnographic analysis, which "makes visible the cultural and social practices of a particular community" (Jewitt 2016 *et al*.) to examine various multimodal classroom literacy practises in preschool classrooms. They focus on activities and the way teachers teach learners how to read, whether it be in isiXhosa or in English. They argue that literacy is not a neutral absorptive process by children, but rather that it is a social and cultural practice in which learners are active in their own literacy development. This is also observed in the multiple multimodal ethnographic case studies undertaken by Stein (2008). This thesis does not include activities with the videos, which make it a more passive activity, to measure the influence of the videos themselves and the tests used in this study are to measure whether they were effective or not. Ethnographic multimodal research does, however, suggest that by making children active participants in learning, more effective learning occurs. Due to the goals of ethnographic studies differing to the triangulation study this thesis undertakes, it was decided not to include this framework in the methodology. However, ethnographic observation measures were used in describing and understanding the context of the school site used in this research. This approach deepens ethnographic understanding with DMDA and psycholinguistic approaches. Multimodal ethnographic findings, however, do highlight the importance of social interaction in learning and how videos as passive activities may not be the most beneficial for learning and literacy development.

2.3.1.5 Multimodal interactional analysis and multimodal reception analysis

Abrams (2016:343) examines German as a foreign language (FL) and German beginner university students learning with tasks designed for *Rosenheim-Cops,* a L1 German video series. She employed a multimodal interactional analysis (MIA), which "explore[s] how a variety of semiotic resources are brought into and constitutive of social interaction, identities and relations" (Jewitt *et al*. 2016). Her results show that "the tasks [which accompany the video] encouraged semiotic awareness (see 1.2.3), helped activate referential knowledge useful for accessing multimodal resources, and elicited a positive response to authentic L2 use in context" (Abrams 2016:343). There were, however, difficulties in that some students were hindered in learning from lack of linguistic knowledge (Abrams 2016:343). The reason this thesis does not employ an MIA is due to this method potentially shifting away from the

linguistic mode (Jewitt *et al.* 2016), which is vital in a literacy study. Abrams' (2016) study also does not look at media as having potential meaning affordances, which this study does not disregard (Jewitt *et al.* 2016). Thus, for the understanding of the choice of SF-MDA and a triangulation of methods and theories in this thesis, Figure 1 illustrates that these various multimodal analyses focus on culture and text, and they can focus on the receiver of the text's discourse. This is usually restricted, however, to an analysis of spoken or written language as classroom discourse for specific ideologies or meanings, rather than measuring cognitive and perceptual processes or skills (see Figure 1). Multimodal reception analysis allows for the examination of eye-tracking to observe the focus of receivers of the text (Jewitt *et al*. 2016), but this does not in itself tell us that focus has transmitted to learned information. Eye-tracking software was not seen as the best research tool for this investigation and purpose of this study, as this software, aside from the finances involved, is extremely difficult to use in a peri-urban context in terms of mobility and electronic access. Multimodal reception analysis also emphasises the visual and cinematic aspects of film and media, and in the context of this thesis it was integral to analyse the visuals in relation to language and literacy development, with the subtitles and the linguistic aspects of it as the primary focus of this research. Thus, this study integrates data findings usually discussed in multimodal reception and interactional analyses, as well as psycholinguistic studies, with findings analysed in multimodal social semiotic and SF-MDA contexts (discussed further in Chapter 4). Multimodal reception studies are examined alongside other psycholinguistic studies in 2.2. This sub-section concludes with a particular method of analysing subtitles in a multimodal manner and that is multimodal translation studies.

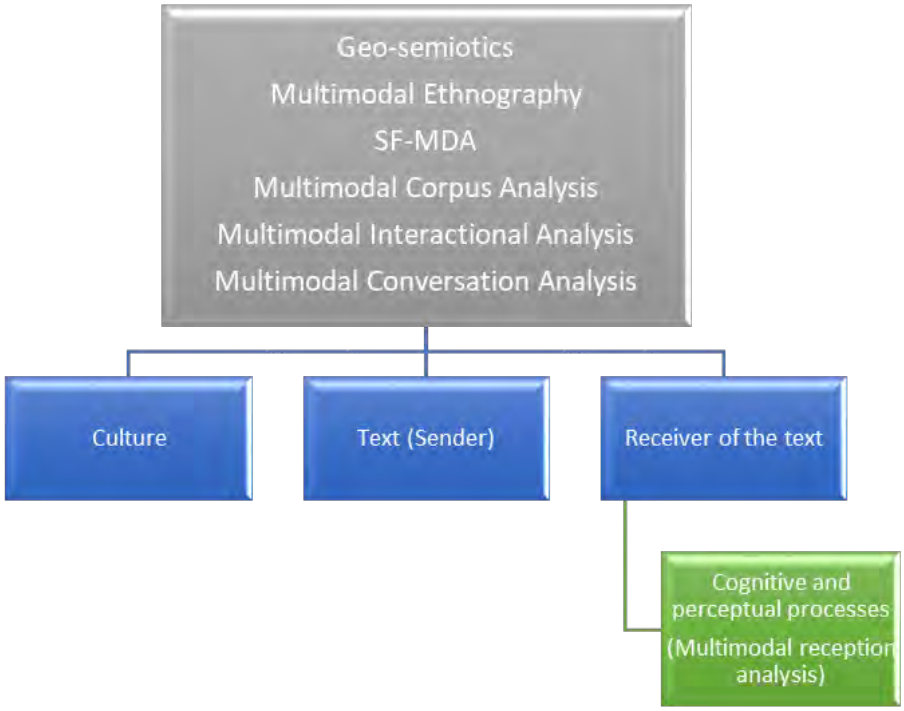


Figure 1 Multimodal analysis types and their focus

2.3.1.6 Multimodal translation

Subtitles in particular have been examined multimodally with two approaches – a multimodal reception analysis (discussed in 2.3.1.5) and a multimodal translation approach. A multimodal translation approach examines the interplay between visual and linguistic mode, as well as linguistic translation (Jewitt *et al.* 2016). This area of multimodal research, multimodal translation, was initiated by Taylor (2003)'s multimodal transcription of subtitles that starts to examine subtitles in films and the production of subtitles from a different spoken language alongside visual features of film. Transcriptions methods are derived from Thibault (2000) and Baldry (2000). Thus, the visual and the linguistic modes in a manner of translation, and their interaction with one another, are examined. Bączkowska (2011) examines the systems of omission (what is removed from the spoken language in the subtitle after interlingual translation, but also between the linguistic and verbal mode) and reduction of meaning. These systems are included in the DMDA employed in Chapter 3, but these studies examined videos or films that had subtitles differing from the spoken language, thus not SLS. SLS with subtitles based on the spoken language creates a kind of reduplication of meaning not examined in the study of multimodal translation analysis, which is understandable, as they focus on linguistic as well as multimodal translation of meaning. The translation of meaning between modes and interplay between visual and linguistic meanings is analysed in this thesis with the system of

Intersemiotic Texture and Comparative Relation from SF-MDA, described in Chapter 3. This system was selected due to its systematic nature. This chapter concludes with a summary in 2.4.

2.4 Chapter summary

In this chapter I have argued that literacy is a social and multimodal activity (2.1.1) and is a semiotic system and process (see Chapter 1 and O'Halloran 2023). As such, while quantitative studies of literacy skills in isolation shed light on reading strategies (section 2.1), a fuller, more holistic understanding can be obtained through balancing these with a DMDA analysis (section 2.3.1.1, see 3.1).

This approach is important because research shows that effective subtitling involves a combination of technical measures (e.g., font size, speed of subtitles, etc. in section 2.2.3), literacy fluency (section 2.1.1) as well as multimodal interpretation of subtitles in the context of the visual semiotic resources of the video (section 2.3). This raises important questions about how meaning is construed in the interplay between the spoken language of the videos, the written text of the subtitles and the visual display of the videos themselves.

Section 2.3 highlighted that most multimodal analyses mentioned here favour visual over linguistic and this thesis seeks to balance the modes. Classroom discourse and intonation were examined with DMDA and SF-MDA for ideology but not in relation to learner performance. Systems of SF-MDA and a DMDA were thus chosen due to their examination of context at the micro-level, a balanced approach to modes and their fine-grained nature. In terms of subtitles in 2.2, Finnish conventions are compared to the isiXhosa rate and reading rates as a baseline. The reading rate from the Department of Education benchmarks and the ORF reading rates from the learners in the intervention (see Chapter 5), as well as the subtitle rates in the video (see Chapter 6), are triangulated in 6.6. L1 Sesotho subtitle challenges for adult readers illustrate potential difficulties learners may face reading in isiXhosa. However, SLS has the potential to improve literacy, including in the areas of vocabulary, listening comprehension, word recognition, as well as assist in memory and motivation in other contexts. The lack of isiXhosa subtitle exposure may also pose a challenge to learners in this study, which may make reading them a challenge. Nevertheless, this study accounts for this with viewing of the videos twice per session and exposure in 15 sessions before post-tests were completed. Section 2.2.

also highlighted the motivations for this study in terms of methodology and control group design, but also for the way the study was conducted as per Kothari *et al*. (2002). Section 2.1 illustrates the grain size of isiXhosa is the syllable and how word recognition and ORF are good predictors of reading. Multimodal literacy theories are discussed in 3.2, which is a continued discussion from 2.2.3.1.

This chapter has illustrated a trend in the literature in 2.1, in which studies barely examine the text that is read, rather focus on reading scores. In 2.3, studies examine the text, along with participant feedback either in qualitative form or from the observer's perspective, such as the analyst. This thesis seeks to triangulate these aspects with subtitled videos for a holistic view of both the subtitled videos as text and learners reading abilities for the subtitled videos in question. In order to do this, a DMDA is conducted as well as a two-group experimental design method, discussed in Chapter 4. The next chapter examines first what DMDA and its systems are in this thesis, before examining multimodal theories of literacy in greater detail.

Chapter 3 – Theory

Two sections form this chapter, which discusses the main theoretical frameworks used to analyse the data in this thesis that are not outlined in Chapter 2 (the Literature Review) or Chapter 4 (The Methodology). The first section (3.1) introduces DMDA (3.1.1), and outlines the separate systems used in this thesis. Section 3.1.1 discusses the systems used for analysing spoken language and the visual mode, organising them into the ideational metafunction (3.1.1.1), the interpersonal metafunction (3.1.1.2) and the textual metafunction (3.1.1.3). Section 3.1.2 discusses written language (subtitles), and in 3.1.3, Intersemiotic Texture is discussed. These descriptions are given to aid in understanding the analysis reported in Chapter 6.

The second section briefly departs from DMDA to consider theories related to multimodal literacy (3.2) and Mayer's cognitive theory of multimedia learning (3.2.1) that are pertinent to the DMDA analysis in this thesis. These theories can be used to triangulate and postulate potential benefits and challenges to choices in video design for literacy development with SLS videos in this research context.

3.1 Multimodality and DMDA

To understand the selection of the main framework of analysis used in this thesis (DMDA), one has to examine the development of multimodality and DMDA. The foundations of multimodality began with work from De Saussure (1974, first published in 1916) in semiotics and Halliday (1976), who founded Systemic Functional Linguistics.

Social semiotics derived from "De Saussure's rubbish bin" (Hodge and Kress 1988:17), which is composed of other semiotic systems aside from language. The focus began on speaking as an act (*parole*), "extra-semiotic phenomena" (1988:18), such as culture and society and the sign, which consists of a form (signifier) that represented patterns of meaning (signified).

Kress and van Leeuwen (2006) started designing a visual grammar of design modelled on the linguistic metafunctional system. This system of grammar for the visual mode was connected to an already established system of analysis for linguistics - namely, a grammar that had

categories inspired from the linguistic metafunctional system of Halliday (1976). In other words, they realised that Halliday's (1976) ideas of metafunctions to interpret meanings from the linguistic mode could be extended to include systems alongside language found in the semiotic realm, such as gestural and visual modes. Kress and van Leeuwen (2006) focused on the concept of interest that led to particular ways of text design. In this context, interest is a "transformative, productive stance towards sign-making [which] is at the same time a transformation of the sign-makers' subjectivity" (Kress and van Leeuwen 2006:12). If a child was asked to draw a person, and they drew only the head as a circle, the circle represents a person, thus "person" as a concept to the sign-maker, in this case, the child. They would focus on the person's characteristic of having a head and construct meaning (signified) out of the form (signifier) based on the child's interest. The structures of the visual grammar introduced from Kress and van Leeuwen were incorporated into SF-MDA and DMDA and are still used. The framework for DMDA is discussed in 3.1.1, 3.1.2 and 3.1.3.

The idea of combining the semiotic and linguistic into one framework led to the development of the term multimodality, first used in Kress and van Leeuwen (2001). Modes and their relationship with sign-maker interest had been referenced early in this development as "in the interplay between spoken and written ways of meaning, and in their relation to other modalities of communication e.g. image, sound, activity)" (Martin and Rose 2007:256). The study of multimodality has thus sought to shift towards a perspective of multimodality in which universal semiotic principles, including language as a set of signs, could function in and across different modes and move away from different modes in multimodal texts having strict specialist tasks (Kress and van Leeuwen 2001). This was due to academic disciplines that have used "one language to speak about language (linguistics), another to speak about art (art history), yet another to speak about music (musicology), and so forth, each with its own methods, its own assumptions, its own technical vocabulary, its own strengths and its own blind spots" (Kress and van Leeuwen 2001:1). These blind spots also often arose from lack of space in thesis and research articles to include everything (Kress 2015). As the editing process for digital multimodal texts shifted from independent specialists in these modes to a producer or single entity of producers with multiple skills (Kress and van Leeuwen 2001), it created tension stemming from a mono-disciplinary perspective across disciplines when examining multimodal texts with a mono-disciplinary lens.

Thus far, the focus of a grammar or framework for multimodal analysis was primarily on static visuals when first establishing a visual grammar, but Kress and van Leeuwen begin to mention other modes such as gestures and music in film (2001). Multimodal approaches were not categorised within linguistics in nature, due to the "extra-linguistic" emphasis (Kress 2015:33). Texts have always been multimodal to some degree (Kress and van Leeuwen 2001), for example written language on a page is both linguistic and visual, thus the modes themselves do not exist in a vacuum. An unbalanced analysis can derive from not including meanings from multiple modes and only focusing on one. Thus, a balanced analysis of more than one mode in a context requires fair integration of knowledge from these disciplines.

Kress (2015) states that multimodality is a domain, not a theory, yet there is a need for an integrated theory that fully accounts for all the entities in the multimodal domain, which is important when reflecting on connection between multimodality and applied linguistics. Linguistics, in particular, tends to view the semiotic as non-linguistic, rather than containing frameworks that seek to encompass linguistics and other modes, their semiotic resources and affordances equally (Bezemer and Jewitt 2010).

As the need developed for an integrative multimodal framework, the linguistic influences on multimodal semiotic analysis had begun to cause developments in visual grammar. Van Leeuwen (2005) focused less on interest and semiotic thought, describing his later grammar with Kress as a "systemic-functional grammar of visual design" (van Leeuwen 2005:77). The use of ideational, the interpersonal and textual metafunction were a result of understanding former categories of representation, interaction and composition developed by Kress and van Leeuwen (2006, first published in 1996) in a broader framework for understanding meaning from multiple modes. Van Leeuwen states, "In the 1990s I was influenced by my work with members of the critical discourse analysis group" (2005: xii). This shift from interest to discourse became a viewpoint from which linguistic and visual modes could be more equally examined and initiated the establishment of Systemic Functional Multimodal Discourse Analysis (SF-MDA).

Stemming from systemic functional theory, "SF-MDA provides a transdisciplinary bridge across traditionally distinct fields of study" (O'Halloran 2008:444-445). It thus uses adapted frameworks of analyses for language (Halliday's 1976) in Systemic Functional Linguistic theory (Halliday 1976), as well as for visuals in Fine Art (O'Toole 1994) and Systemic Functional Visual theory (Van Leeuwen 2005).

Both visual and linguistic modes are organised in Content and Expression strata and an SF-MDA examines distinct modes separately as conducted in earlier approaches, but also examines them as parts of a larger entity, (O'Halloran 2008:444). The way this is done with both visual and linguistic semiotic resources is discussed in 3.1.1. As the framework has expanded to analyse the potential of various semiotic resources, other frameworks have been added, for example, in Tan *et al.* (2015).

Jewitt (2014) distinguishes SF-MDA from former multimodal social semiotic analysis. SF-MDA, described by O'Halloran (2008), focuses on examining texts in units of shots and clauses, i.e., the micro-textual level, as opposed to the macro-textual level or whole-work approach taken by former social semiotic analysis (Jewitt 2009:31). This causes SF-MDA to be more fine-grained. There is also a shift of focus towards the social and cultural use and meaning production (Jewitt 2009:31). In SF-MDA, the metafunctions are emphasised as the underlying semiotic resources, succeeding the former exclusive examination on the social semiotic (Van Leeuwen 2005, Jewitt 2009:32).

The uses of the term "semiotic" can be confused with former multimodal social semiotic approaches that mainly focused on visual analysis. Tan *et al.* (2015:559) describe "a Digital Multimodal Discourse Approach" in their analysis of systems including intonational systems. While most systems used in this thesis can be categorised using a systemic functional model, the use of digital resources to analyse them have caused the term Digital Multimodal Discourse Analysis to be used (Tan *et al*. 2018). The inclusivity of this term to include other forms of analysis that are not restricted by the systemic functional model is the reason why the term Digital Multimodal Discourse Analysis is used in this thesis and not SF-MDA.

State transition machines (STMs) highlight the duration of system choice occurrences, as well as co-occurrence of meanings, across the modes from the software *Multimodal Analysis Video* after analysis has taken place (see 2.1.1 and 4.3.2.5.). States may also represent couplings from general systemic functional linguistics. These are when a pair or more meaning choices co-occur between multiple systems (Zappavigna 2011; Zappavigna *et al*. 2008:170). A system choice or a combination of them can be represented by a state in an STM, whereas a coupling can only represent multiple system combinations and not a single occurrence of one system. For example, co-occurrence of the Circumstance of Location: Time, "ngobu busuku" (tonight) and Extent Circumstance, "Ngamanye amaxesha" (sometimes) in the Relational Process within the clause "Ngamanye amaxesha inyanga yam inkulu, njengokuba injalo ngobu busuku"

(Sometimes my moon is big, as it is tonight). Multiple instantiations of motivated couplings which formulate patterns of meanings are known as syndromes (Knox *et al*. 2011:101). An example of syndromes found in the videos are the wide range of Material or Mental Processes, Action Processes and Circumstances of Location that co-occur, which is typical of children's narratives discussed by Guijarro and Sanz (2008) DMDA is now discussed and described according to its metafunctional categories. System choices in DMDA are capitalised for ease of reading and easier identification in the thesis, along with capitalisation of Mayer's (2009) principles in 3.2.1.

3.1.1 DMDA: Spoken linguistic and visual mode systems of analysis

Figure 2 Halliday's (1985 and 1994) model of Systemic Functional Linguistics

Pertinent to our current discussion of multimodality is Halliday's (1985 and 1994) linguistic model in order to situate the discourse semantics in its context and to describe the formation of SF-MDA and the current application of DMDA. The model devised by Halliday (1985 and 1994) examines linguistic grammar and realisations of meanings in terms of differences between concrete realisations of language (i.e., phonology and/or graphology) and more abstract levels of structures that organise these concrete realisations (e.g., the genre of a text).

A representation of this model (based on Halliday and Matthiessen 2004) is provided in Figure 2. The metafunctions are often examined at a number of levels: phonological/graphological, the lexico-grammar (the structures within clauses), discourse semantics (meaning in context), register (the situational context of the production of the text) and the genre (context of the culture). The "expression plane" is composed of phonology and graphology, while the "content plane" is composed of the lexico-grammar, discourse semantics and the "context plane" of the register and genre (Halliday and Matthiessen 2004: 25). Field comprises of text content and topic, tenor comprises of the involved parties and their relationship and mode comprises of how a text is constructed and communicated, whether spoken or written (Halliday and Matthiesen 2004). The discourse semantics level with ideational, interpersonal and textual metafunctions in Figure 2 is used in the SFL and SF-MDA component of this analysis and is described further below in 3.1.1.1, 3.1.1.2 and 3.1.1.3.

When examining the visual mode with DMDA, as well as the relationship between visual and linguistic modes with DMDA, O'Halloran (2008) refers to the level of discourse semantics containing discourse from the linguistic mode and Work from the visual image mode. Work is the top layer of O'Toole's (1994) scale in his grammar of visual art and refers to a work of art as a whole (O'Toole 1994:225). Liu and O'Halloran (2009) also add Intersemiotic Texture, which explores meaning relations between modes. This does ensure that the visual image as a whole and the linguistic discourse as a whole are kept separate while being analysed in terms of their relationship.

However, "Big 'D'" Discourse in comparison to "discourse" with a non-capitalised letter "d", where Discourse is defined as "the ways in which people enact and recognise socially and historically significant identities or "kinds of people" through well-integrated combinations of language, actions, interactions, objects, tools, technologies, beliefs, and values" (Gee 2015:1). The "discourse" with a small "d" is defined as restricted to the language itself. Thus, while the visual image is separate from "discourse", both work and discourse form the Discourse in discourse semantics, sharing the same metafunctions, despite different lexicogrammatical categories. Therefore, this thesis uses the capitalised "Discourse" when referring to both discourse and Work, as well as capitalising the "D" in "Discourse Semantics" to ensure it is understood as a level encompassing both linguistic and visual modes. Thus, Work can be a part of a broader understanding of Discourse.

The word, word group/phrase, clause and complex for the linguistic components and scene, episode, figure and part for the visual components are at the level of lexico-grammar. A scene connects more to the idea of a scene in a film, comprising multiple episodes (visual happenings, actions and relations), which in turn consist of many figures (humans, animals or objects) (O'Toole 1994:221). A further rank within Figure is Part such as a button on a telephone or skin colour (O'Toole 1994:222). The next sub-sections describe each metafunction, first in terms of the linguistic mode and then in terms of the visual mode.

3.1.1.1 Ideational metafunction for language and visual modes

Within the level of Discourse Semantics for linguistic and visual modes, the ideational metafunctional meanings conveyed refer to ideas and knowledge of the world in the experiential metafunction and the relationship between knowledge and ideas in the logical metafunction (Bloor and Bloor 2004). The logical metafunction is not as easily identified in visuals as it is in clauses and is rather described in terms of Intersemiotic Texture (3.1.3). In terms of the experiential metafunction, the systems within the ideational metafunction for both modes include Participants, Processes and Circumstances. Participants are abstract or non-abstract objects, Processes are acts, and examples of Circumstances provide information about place, time, manner, and associated participants (with whom? with what?); including time and place (Bloor and Bloor 2004:53, 111). Circumstances are separated into Accompaniment, Angle, Extent, Manner, Locatoin, Contingency, Matter, Role and Cause, but only Extent, Location, Manner, Matter and Accompaniment appear in the videos. Bloor and Bloor (2004) only state Location and this category includes Location: Place and Location: Time, for example, in *The Icky Sticky Frog*, "Kwelinye ichityana elihle elizuba" (in some little blue lake) and "kufuphi" (nearby). This thesis chooses to analyse this in the same way, as distinction on Time and Place was not a focus of the thesis due to Circumstances not being as frequent as other clausal structures in the videos. Examples of Circumstances of Manner in the videos include "ngendlela efanayo" (in the same way as) in "Usele-sele walinginya ibhungane ngendlela efanayo naginya ngayo impukane" (Sele-sele swallowed the beetle in the same way he swallowed the fly). An example of Circumstance of Extent includes "kwisithuba seeveki" (within/after a few weeks) and an example of Circumstance of Accompaniment is "ngoKram" (Kram! [sound]) in "imilomo yabo ivala "ngoKram!"" (their mouths close with a Kram sound). An example of Circumstance of Matter includes "ngemo yezulu, nokuba loluphi usuku esikulo evekini." (about the weather and in which day of the week we are) in the sentence "Senza

nezinye izinto ezonwabisayo ezifana… nokuthetha ngemo yezulu, nokuba loluphi usuku esikulo evekini" (We do other fun things… such as talking about the weather and which day of the week we are in).

In terms of language, the Processes can refer to Actions (Material processes), thoughts and perception (Mental processes), possessing an attribute (Relational processes), speaking (Verbal processes), behaviour (Behavioural processes) and simple existence (Existential processes).

The categories of Participant for the linguistic mode are as follows (taken from Bloor and Bloor 2004:111-126):

- The Material Process may have the Participant roles of Actor and/or Goal. An example of Actor and Goal is in "inyanga indilandele ukuya ekhaya" (The moon followed me home) from *The Moon Followed Me Home* with "inyanga" (the moon) as the Actor and "-ndi-" (me) as the Goal of the Material Process. Thus, in isiXhosa, Participants can be realised by prefixes on verb complexes, as this example illustrates. The Beneficiary is a Participant that benefits from the Process (Bloor and Bloor 2004), e.g., "umama wam" in "ndifika ndibonise umama wam omnye umsebenzi endiwenzileyo" (I always get to take my mother and show her some of the work I did) in Video 5.
- The Mental Process may have Participant roles of either or both Sensor and Phenomenon. For example, "USue wayezithanda izilwanyana" (Sue loved animals), "USue" (Sue) is the Sensor and izilwanyana (the animals) is the Phenomenon.
- The Relational Process may have the Participant roles of either or both Carrier and Attribute or either or both Identifier and Identified Participants. An example of a Relational Process with Carrier and Attribute Participants is "Ndinomhlobo wam" (I have my friend) where the Carrier is "Ndi-" (I) and the Attribute is "umhlobo wam" (my friend). A Relational Process with Identifier and Identified Participants is "asisodwa" (we are not alone). The prefix "si-" (we) is the Identifier and "-sodwa" is a Relational Process that also has the role of Identified (alone).
- The Verbal Process may have Participant roles of Sayer, Verbiage Reported and Quoted, however only Verbiage, Sayer and Quoted are in the videos in this thesis. Examples include "sicula iingoma" (we sing songs) and ""Ube nemini emnandi" atsho utata" ("Have a nice day", father says). The Sayers in these clauses are "si-" (we) and "utata" (father), the Quoted is ""Ube nemini emnandi"" ("Have a nice day") and the Verbiage is "iingoma" (songs).

- A Behaviour Process may have Participant roles of either or both Behaver and Behaviour. Only Behaver appears as a Participant in the videos, for example "ndingalala emthini otofotofo" (I sleep on the soft mat). "Ndi-" is the Behaver.
- Existential linguistic Processes emphasise the existence of Participants (Bloor and Bloor 2004), e.g., "nalo" (there is/this) in "Nalo ulwimi lukasele-sele, luncangathi kwaye lulude" ([There is] that tongue of Sele-sele, sticky and long).

In the visual mode, Participants, Processes and Circumstances are included in the representation category from the visual grammar Kress and van Leeuwen (2006). The Processes are basically separated into Conceptual and Narrative Processes (Kress and van Leeuwen 2006)

Kress and van Leeuwen (2006:59) identify the following narrative structures for the visual mode: Action, Mental, Verbal, Conceptual, Existential and Behavioural Processes. Action Processes can be identified in different forms: unidirectional transactional actions, non-transactional actions, transactional reactions, non-transactional reactions and bidirectional transactional actions. These processes are all centred around vectors, which are lines emanating from a Participant, visible or invisible, that imply relationships between the Participants. Unidirectional transactional actions usually connect an Actor and a Goal. Actors here are defined as Participants where the vector is initiated and involved in an Action process, in comparison to a Goal which is a passive Participant and where the vector ends in an Action process. For example, 1 minute 29 seconds into the video *They All Loved Sue,* there is a visual representation of Sue stroking a dog, which includes Sue as Actor and the dog as Goal. Non-transactional actions are an action type where the Action Process from the Actor is not directed at any other Participant. For example, there is a child in the video *I Can Go To School* that shares a story but the listeners of the story are not in the image.

Another component of actions is Reaction processes that are either transactional (where a Reactor reacts to a Phenomenon) and non-transactional (i.e Reactor reacts, but not directed at anything visible in the camera shot). Their look creates the vector, rather than their involvement in a particular action. Bidirectional transactional actions involve potentially two Actors where there is action from both sides, which then causes the Actors to become Interactors as they have dual roles of both Reactors and Actors. For example, in *The Icky Sticky Frog*, *Usele-sele* (Sele-sele, the frog) is the Actor and attacks a fly with his tongue. The fly buzzing on his tongue is also an Actor but is further a Reactor to the frog in this instance. Reaction and Action Processes

are discussed together in this thesis when they co-occur with Material Processes in the language in order to simplify comparison between meanings for Similarity and Contrast.

A Mental Process is a vector that connects a Sensor to a Phenomenon with a vector device like a "thought bubble" (Kress and van Leeuwen 2006:74). This is seen in *The Icky Sticky Frog* where the frog, *Usele-sele*, thinks of all the insects he has eaten, which are the Phenomenon in this case and appear in his thought bubbles. These are more than just related to Participant gaze and vectors, rather these visual Processes depict thoughts and nuanced perceptions of Sensors and Phenomenon.

A Verbal process uses a dialogue or speech bubble, in which a Sayer produces an Utterance enclosed in a dialogue or speech bubble. There are no examples of this in the videos selected, however the action of children singing, like the linguistic process, has been classified as a Verbal Process in the visual Mode when it appears in *I Can Go To School*.

Conceptual Processes are usually in the form of arrow representations and gestures, but there are no conceptual processes in these videos, therefore, they are not discussed in detail here.

Existential Processes containing an Existent may be Participants that are not involved in any action and thus are just existing, for example there are shots of Sue just existing and not engaging in any activities. Particular actions relating to Behaviour, including the girl falling asleep like in *I Can Go To School* are classified as Behavioural Processes.

The Circumstances in the visual mode relate to a setting, which is usually background visuals painted in less detail or partially obscured by participants, such as the jungle in the video *Jungle Tumble*. Means are type of Circumstance that is initiated by a tool, such as the child telling a story in *I Can Go To School*, as they tell a story with a drawing, which is the tool they are using. An Accompaniment is a Participant which has no vector relationship with other Participants, e.g. the dog with Sue in *They All loved Sue* at the timestamp 3:55.

3.1.1.2 The interpersonal metafunction in language and visuals

Meaning as social engagement between the perceiver and producer of texts is realised as the interpersonal metafunction (Bloor and Bloor 2004:11). Components of the interpersonal metafunction in the linguistic mode include Mood and Modality.

Mood in the clause illustrates if services or goods are requested or provided (Bloor and Bloor 2004:283). There are three types of Mood: Imperative (commands), Interrogative (questions) and Declarative (statements). As mentioned in 1.2.1.2, isiXhosa's Subjunctive and Indicative Moods are realised across Declarative and Interrogative Moods, while the Imperative Mood in isiXhosa is recognised as its own unique Mood. An example Declarative in the video *Jungle Tumble* is "Izikhwenene zizifihla phezulu emithini" (Parrots hide themselves high up in the trees). An example Interrogative is "Ingaba yinyoka emdaka le ihleka emthini?" (Is that a brown snake laughing in the tree?). An example Imperative in the video is "Ube nemini emnandi" (have a nice day!).

According to Halliday and Matthiessen (2014), Modality is the semantic area, which lies between degrees of polarity of four different types: probability (epistemic modality), usuality, obligation (deontic modality) or inclination (dynamic modality). While they had referred to the former two types restricted to propositions and the latter two types restricted to proposals (Halliday and Matthiessen 2004). This thesis restricts its analysis of modality to these four types, values within each type as high, median or low, and polarity as positive or negative. Instances of modality, according to Halliday and Matthiessen (2014) are often realised in two ways, by a finite modal operator, such as modal verbs in English, or an expansion of the Predicator, for example, a passivised verb or adjective. This indicates that passive-like constructions, such as stative, may carry modal properties. Realisations of modality in isiXhosa are found in examples (3)-(12) and (15) in 1.2.1.2. For example, in *The Icky Sticky Frog*, the presence of the affix *-s-* and the stative verbal extension *-akal-* in the negative polarity modal construction "Ayisabonakali" (the fly cannot be seen), realises high probability (3 in 1.2.1.2). A positive polarity modal construction is found in *Jungle Tumble* in "Uza kubona!" (You will see!), and immediate future markers *-za* and *-ku* indicate median probability (5 in 1.2.1.2).

An instance of modality illustrating high usuality with positive polarity is in "Kum uyakusoloko eqaqamba" (To me, he will always shine bright) in *the Moon Followed Me Home* (6 in 1.2.1.2). This is from the temporal verb –"soloko" (always). An instance of modality illustrating high usuality with negative polarity is in "Akazange alahlekane nam" (He has never lost me) in *the Moon Followed Me Home* (10 in 1.2.1.2). This is realised by the verb form "-zange" (never did) An instance of modality illustrating low usuality with positive polarity is in "Ngamanye amaxesha ndizoba imifanekiso yosapho lwam" (Sometimes I draw pictures of my family) in *I Can Go To School* (12 in 1.2.1.2). This is realised from the words "ngamanye

amaxesha" (sometimes). A positive instance of high obligation and low inclination is in "uSue uyewaqonda ukuba neekati ziyafuna ukukhathalelwa nazo" (Sue felt that the cats also needed (lit: wanted) to be cared for) in *They All Loved Sue* (4 in 1.2.1.2). In this compound verb construction, Sue felt that the cats needed to be looked after, thus there is a high value of obligation by Sue, but a low inclination from the cats (represented by *zi-* in the verb) realised in the verb *-funa* (want). A further positive example of high probability, in isiXhosa with the "-nga-" affix indicates "can or is able to", e.g. in the title of *I Can Go To School* (*Ndingaya esikolweni*). While Modality is discussed in 6.6.5 in relation to emphasis, this system was not fully investigated as it was not the sole focus of this thesis. Furthermore, Modality is not analysed with STMs in Chapter 6, as there are various smaller constituents that trigger various aspects (modal type, value and polarity) that permeate across the clause and multiple clauses. Furthermore, there are not many cases of modality in all of the videos, thus they are merely highlighted to indicate the complexity of meanings and grammatical constructions in the videos.

In the visual mode, the systems are organised along the ranks of Figure, Episode and Work, as discussed in 3.1.2, based on the category of interaction by Kress and van Leeuwen (2006). Low and high angles, lighting and shot distance all provide information on the position of the text creator in relation to the viewer and the power relations between them (O'Halloran *et al*. 2018). Perspective is examined and categorised as Horizontal (as either Front or Angled) and Vertical (as either Low, Eye Level or High angle). Focus is categorised in terms of Camera Movement and Zoom [the vertical Tilt of the camera, a horizontal Pan of the camera, a Dolly shot following particular Participants and Processes and Pedestal raising or lowering of the camera vertically in relation to the subject]. Another system of the DMDA to analyse interpersonal meanings in the visual mode is Gaze. Gaze can be direct or indirect, depending on whether the Participant is gazing directly at the viewer or away from the viewer (indirect) and elicits deeper interaction between Participants and the audience, as well as producer of the text and receiver (Kress and van Leeuwen 2006). For example, if Sue is looking directly at the viewer in the video *They All Loved Sue*, there is a greater direct engagement with the audience, but if the frog is looking away from the audience in *The Icky Sticky Frog*, there is less viewer engagement. In addition, another category is visual Salience, which is an extra layer of focus based on lighting and contrast on specific Processes or Participants (Kress and van Leeuwen 2006). Examples of Salience include Colour Contrast, Focus Sharpness, Lighting, Foreground

and Background Participants and Processes (O'Halloran *et al*. 2018; Kress and Van Leeuwen 2006).

An example of Salience in *The Icky Sticky Frog* includes when Sele-sele appears through a Sharpness of Focus from a blurred image, as well as the locust and beetle foregrounded to place emphasis on them. While Kress and van Leeuwen (2006) used to categorise Salience within the textual metafunction, SF-MDA categorises it as interpersonal. Due to Salience illustrating meanings more within the interpersonal metafunctiona than the textual metafunction, Salience is allocated to this metafunction in this analysis.

3.1.1.3 The textual metafunction in language and visuals

Structure, coherence and organisation of the text is realised by the textual metafunction (Bloor and Bloor 2004:11). Theme (information fronted as priority) and Rheme (rest of the clause) are systems within the textual metafunction to examine structure (Bloor and Bloor 2004). An instance of Rheme and Theme is found in the sentence "USue wayonwabe kakhulu, kakhulu!" (Sue was very very happy) in the video *They All Loved Sue*. "USue" (Sue) is the Theme and the remainder of the clause is Rheme. The Theme and Rheme may either represent Given or New information. Given information is shared and potentially older information that is usually at the beginning of a sentence and New information is added information, on which the speaker is focused (Bloor and Bloor 2004:65).

Visual Framing devices including vectors and their presence or absence are the realisation of the textual metafunction at the visual mode, which was known as the composition category in Kress and van Leeuwen's (2006:59) visual grammar. Vectors are in essence lines, such as the line of the tongue of the frog (Sele-sele) leading to the fly in *The Icky Sticky Frog* or even invisible lines between Participants from Gaze. While the tongue is more than one line in essence, it connects one Participant to another. Framing devices in terms of the textual metafunction separate visual components from each other, usually by a visible vector (Kress and Van Leeuwen 2006). For example, in *The Icky Sticky Frog*, the boundary line (the edge of the log) that separates the beetle crawling on the log from the rest of the frame is a framing device. Another system that realises textual meaning in the visual mode is in the placement of visual elements (Kress and van Leeuwen 2006). This is known as Information Value, which examines this in terms of left, right, top, bottom, or centre (Kress and van Leeuwen 2006)

The next section discusses another system of DMDA used in this thesis, namely speech and subtitle rate.

3.1.2 Subtitle rate

As mentioned in 2.2, subtitles can be categorised according to whether they are interlingual or intralingual (SLS). The type used in the *MPBXhosa* videos are the same-language subtitles that deal with dialogue only, thus intralingual. According to Liu (2014), there are two parameters that one can examine when looking at subtitles: the linguistic and the technical. The type of subtitle fits in the linguistic and the technical includes whether subtitles are open (cannot be removed from the film, alternatively termed burnt-on subtitles) or closed (broadcast separately). However, this does not include the visual nature of subtitles or include how viewers experience them in terms of the rate they move across the screen. Due to the SF-MDA model of examining the linguistic and visual modes as previously outlined, a way to examine this type of subtitle and classify it according to both modes is necessary.

Subtitle rate "is a measure of the length of time the subtitle stays on screen, as a factor of the amount of text that has to be read during its display" (Kruger *et al*. 2022:212). It is also referred to as "presentation rate" or "subtitle speed" (Kruger *et al*. 2022:212). It seems to vary and is not always due to the language, but also due to the pace of video components, such as the visuals, specific editing requirements, as well as the speech rate (Kruger *et al*. 2022). Higher presentation speeds might result from budget constraints in editing (Kruger *et al*. 2022). Yet what may be considered too high by specific viewers could be due to their differences in terms of language and reading proficiency, hearing status, fatigue and viewing environment, which may cause difficulty for viewers to follow the subtitles (Kruger *et al*. 2022). This description of subtitle rate indicates that the perception and cognition of subtitles by the viewer influences an ideal subtitle rate for the specific context.

Despite the adequacy of the subtitle rate depending so much on the cognition and perception of the viewer, it is such an important part of the language for comprehension and communication that is often ignored (Kruger *et al*. 2022). Even if this thesis examines subtitle rate from an area of cognition and perception, it still seeks a rate from the video itself and has to analyse the video for that subtitle rate. This thesis thus takes the stance that subtitle rate is an affordance of the subtitle semiotic resource, along with other subtitle affordances like

colour, size, spatial presentation and segmentation. As discussed in semiotics mentioned in 3.1, signs include a physical form and meaning. Without the physical manifestation of that sign, there could not be a conveyed meaning (Hodge and Kress 1988). Even though this thesis examines the form nature of subtitle rate over any meanings the subtitle rate conveys, it is nevertheless a system choice by the video editor(s). The choices by the video editor(s) are influenced by their own social context, such as specific editing guidelines and their experience, but also the other semiotic resources in the video. Furthermore, the nature of this rate influences the number of meanings conveyed by the language and other semiotic resources to an audience and what is perceived.

Speech rate may influence subtitle rate (Kruger *et al.* 2022), but it is not analysed in this thesis, as the rate of speech would not necessarily influence learners' literacy levels as directly as the subtitle rate. If the subtitles are too fast or slow due to the speech rate, it automatically implies the speech rate should be adjusted to match the subtitle rate.

However, since the subtitle rate and speech rate are both linguistic presentation rates that are simply different in modes (between visual and spoken), in order to discuss subtitle rate within Halliday's (1985 and 1994 model), I briefly describe how speech rate is situated in the model. Speech rate (often discussed as speech tempo in the literature) has often related to the rhythm of speech and intonation patterns as prosody (Fletcher 2010). Phonetic analysis and sociolinguistics may examine fluctuations in speech rate and possible reasons for fluctuations, as they would for tone variation (Fletcher 2010). Fletcher (2010) further states that certain styles of Discourse are associated with slower or faster tempo, as well as a speaker's emotional state, where tempo varies in spontaneous speech and slower interaction, but may be slower for listener comprehension, or when the speaker is sad, grieving or bored. In areas of discourse analysis, van Leeuwen (1999:47) further emphasises the context and genre. Furthermore, tempo differences have been related to age, sex, sociolinguistic and cultural differences, with older people speaking slower than younger adults (Fletcher 2010). In terms of speech rate, Fletcher (2010) describes of speech rate, the difference between objective tempo (measured tempo) and subjective perceived tempo (the tempo that is perceived by the listener). The subjective perceived subtitle tempo is not measured in this study. Rather the objective subtitle tempo from the video is compared to the ORF reading scores of children to determine how probable it is that learners can follow them.

Therefore, if speech rate was used to examine speaker choice behind fluctuations in an SF-MDA manner, it may relate to the mode category in context of the situation of Halliday's (1985 and 1994) model, with connections to tenor. Alternatively, a micro-level analysis would propose it as phonology (see Figure 2). This is similar for subtitle rate, as if the speech rate influences subtitle rate, and speech rate can be realised at the levels of context of situation and/or phonology, then subtitle rate can be realised at the levels of context of situation (specific editing requirements and pace of visuals and speech) and/or graphology (language-specific). This thesis examines the subtitle rate at the level of phonology/graphology in Halliday's (1985 and 1994) model, by examining subtitle rate. The DMDA further suggests choices for the context of culture (comprised of genre), context of situation (comprised of field, tenor and mode) and Discourse-semantics (the choices within the lexicogrammar for the linguistic and visual modes). The DMDA suggests choices which may maximise the benefits and minimise the challenges for L1 isiXhosa literacy development in this research context. Therefore, the system of subtitle rate includes an analysis at the micro-level to inform the macro-level and recognises subtitle rate as an affordance of subtitles, which are a semiotic resource. This is at the level of design, while the complementary intervention and tests measure specific aspects of subtitle perception, which could be influenced by their reading proficiency, fatigue, and viewing environment (social context) (Kruger *et al.* 2022).

As discussed in 2.2.3.1, if the subtitle rate is too fast for viewers, it could negatively influence the processing of meanings in the written linguistic mode and the integration of meanings between the written linguistic mode and other modes (Kruger *et al.* 2022). Subtitle rate is a focus for literacy development in the context of this research. Due to the emphasis on the written linguistic mode in literacy, this mode is the priority and primary mode in this research context and the subtitles are the priority and primary semiotic resource, while the visual and the spoken linguistic modes are ancillary. The visuals and the spoken language are thus aides to the subtitles in this context. Since the spoken and written linguistic modes convey identical language content, if there is complexity in one of the linguistic modes, it will influence the other linguistic mode. If the visuals as well as both linguistic modes, contain too much complexity, this could lead to a lack of focus on the subtitles (Kruger *et al.* 2022). In terms of literacy influence, it could lead to a lack of depth of processing of the subtitles (Kruger *et al.* 2022), a lack of word recognition (Kruger *et al.* 2022) and an overall lack of significant influence on early L1 isiXhosa literacy skills by the SLS videos (and subtitles).

As discussed in 2.2.3.1, subtitles are read in a multimodal text in two different ways, horizontal reading (referred to in 1.2.3 as intramodal) and vertical reading between the subtitles and other visual components (referred to in 1.2.3 as intermodal or multimodal). Even if the entire subtitle on the screen cannot be read, parts of the subtitle, if understood, can be combined with visual elements and spoken language to facilitate comprehension of meanings that are conveyed from multiple modes (Kruger *et al.* 2022). This could assist or hinder processing of the subtitles, however, which are a semiotic resource of the written linguistic mode that is investigated as a potential resource for literacy development within L1 isiXhosa SLS story videos.

Based on the research aims of this thesis, the DMDA is not used in the traditional sense of a deeper understanding of subtitle rate choice by the designer, rather the focus shifts considerably more to how the affordance (rate) of the semiotic resource (subtitle) influences learners' literacy skills in this research context. The mixed-method approach examines whether the choice of subtitle rate may be appropriate for the social context of this research (advanced stratum of semiotic awareness of the designer in 1.2.3). It further examines how the semiotic resource of subtitle, and its affordances relate to other semiotic resources in the video, namely spoken language and dynamic visuals. This is to present principles of design (based on Mayer's Cognitive Theory of Multimedia Learning, see sub-section 3.2.1), which relates to the arrangement of semiotic resources (subtitles, spoken language and dynamic visuals) in the video. This type of approach creates the potential for a DMDA, in which a text alone is not only examined from an analyst's perspective, but also the reception of the text in the environment it is used, which provides a holistic understanding to the comprehension and interpretation of meanings from a text.

The subtitle rate is calculated according to De Linde and Kay (2014). The average duration of subtitles were downloaded from Excel and the annotated duration is used as the total subtitle presentation rate. According to De Linde and Kay (2014:45), subtitle presentation rate is "a measure of how quickly a subtitle appears and disappears from screen" and is measured in words per minute. While isiXhosa is an agglutinative language and the words are considerably longer than English words, the language with which this formula was first used, the syllable per minute rate was also counted to ensure that adequate averages were calculated and accounted for these discrepancies in linguistic differences.

Research on languages like English often use the word per minute (WPM) rate of measurement, and while this unit of meaning was also used in this study, isiXhosa has a conjunctive

orthography, and this causes small morphemes that are built together to create a longer word than what would usually be found in English. Due to the syllable being the phonological grain size and most used unit of literacy processing for isiXhosa, a syllable per minute (SPM) count was also taken from the video subtitles. Words and syllables correctly read per minute (WCPM and SCPM) were also measured from learners' mean reading scores of learners in the intervention. Solving the challenge of measuring subtitle rate uniformly in psycholinguistic research is not within the scope of this thesis. However, this thesis desires to present as an addition to a trans-disciplinary, multimodal analysis, the potential for examining the affordance of subtitle rate to examine potential opportunities and challenges for word recognition and ORF for L1 isiXhosa learners. This affordance is also examined along with multimodal affordances and the interplay of meanings between visuals and language in order to examine the potential opportunities and challenges for L1 isiXhosa learners' literacy levels with regard to semiotic awareness. In terms of the interplay of meanings between modes, semiotic awareness in multimodal literacy and multimodal affordances for L1 isiXhosa learners' literacy levels is analysed in the system of Intersemiotic Texture in the DMDA. This is later triangulated with learners' mean scores for the Picture-Word Match test and multimodal literacy and learning theories (3.2). The system of Intersemiotic Texture is described in the next sub-section.

3.1.3 Intersemiotic Texture

Intersemiotic Texture refers to "semantic relations between different modalities realised through the Intersemiotic Cohesive Devices in multimodal discourse" (Liu and O'Halloran 2009:369). Intersemiotic Texture in this section is based largely on Liu and O'Halloran (2009). This is also why this area is often referred to as Inter-Relations, as it deals with semantic relations between semiotic modes (associated in this thesis with the advanced stratum of semiotic awareness described in 1.2.3). Texture consists of register and cohesion that differentiate a connected text from sporadic pieces of information completely separate from each other (Liu and O'Halloran 2009). Intersemiotic Cohesive Devices are found mainly within the textual, logical and experiential metafunctions (Liu and O'Halloran 2009). Meanings in the visual mode and linguistic mode are not identical, but rather converge and are similar in meaning in an exponential manner or diverge and differ in meaning completely (Liu and O'Halloran 2009). Comparative Relations as a type of Intersemiotic Texture is a focus of this analysis, which is discussed in this section.

Comparative Relations is an important sub-system of Intersemiotic Texture used in this thesis to identify how modes conform to design principles for Mayer's Cognitive Theory of Multimedia Learning discussed later in this chapter (3.2.1). Meanings in language and visuals either converge or diverge in Similarity in the system of Comparative Relations – thus, these meanings realise co-contextualisation or re-contextualisation (Liu and O'Halloran 2009). When meanings between semiotic modes converge, this convergence is called co-contextualisation, whereas when meanings between semiotic modes contrast with one another, this is known as re-contextualisation (Liu and O'Halloran 2009). An example of co-contextualisation from the videos is in the subtitle "Ngamanye amaxesha ndizoba imifanekiso yosapho lwam" (Sometimes I draw pictures of my family)[15], where the visual shows the child drawing a picture of her parents in front of a house. The meanings conveyed in the visual and linguistic modes in this example are similar. An example of re-contextualisation from the video *I Can Go To School* appears with the visual portraying soft toys sleeping and the subtitle stating "Emva kwemini siba nexesha lokusebenza ngokuzolileyo" (In the afternoon, we get/have time to work calmly). While the visual emphasises the calm and quietness, it shows the opposite meaning to the Material Process "work." According to Liu and O'Halloran (2009), these are mutually exclusive and do not appear simultaneously. However, there are instances in the subtitles (the written mode), where due to shot transitions occurring, Similarity and Contrast do occur simultaneously. This is discussed in Chapter 6, when this occurs, in more detail. Comparative Relations are discussed in the experiential and logical metafunction in the ideational metafunction, as well as the interpersonal and textual metafunction (Liu and O'Halloran 2009). While Intersemiotic Comparative Relations in the experiential, logical and textual metafunctions have been discussed to an extent in literature (Liu and O'Halloran 2009), I discuss an interpretation of meanings in Comparative Relations within the interpersonal metafunction that best suits the videos analysed in this thesis. Other types of Intersemiotic Texture are mentioned in the analysis in Chapter 6, and these along with Comparative Relations are discussed according the metafunctions in which they manifest in 3.1.3.1, 3.1.3.2 and 3.1.3.3.

[15]see (12) in 1.2.1.2

3.1.3.1 Experiential metafunction

In the experiential metafunction, Comparative Relations are realised in terms of Parallel structures. An example of a parallel structure is the utterance “Ukudlala iphazile yeyona nto ndiyathandayo” (Building a puzzle is the thing I love the most) where the visual depicts the child building a puzzle. In addition, intersemiotic cohesive devices include Intersemiotic Hyponymy, Meronymy, Polysemy, Antonymy, Correspondence and Collocation, with Polysemy and Correspondence realising Co-contextualisation and Antonymy as an instance of a cohesive device realising re-contextualisation (Liu and O’Halloran 2009:385). Hyponymy, Collocation, Polysemy and Meronymy are used as they are in the linguistic mode in this case but are expanded to describe relationships between modes (Liu and O’Halloran 2009:372; 375-377). An example of Intersemiotic Hyponymy from the videos is the phrase “USue wayezithanda izilwanyana. Wayezithanda zonke iintlobo zezilwanyana” (Sue loved animals. She loved all kinds of animals.). The visuals in the example simply illustrate Sue with a squirrel and a bird, which are hyponyms of the word “izilwanyana” (animals). An example of Intersemiotic Polysemy includes the sun in a thought bubble of a child in *I Can Go To School* with the sentence “nokuba loluphi usuku esikulo evekini” (and which day of the week we’re in). The sun is associated with the word “usuku" (day), and generally an image of a sun can represent many different word meanings related to days (and day of the week, particularly Sunday). By contrast, an example of an Intersemiotic Collocation is visuals of a big and small lion in *The Moon Followed Me Home*, when the language only refers to “mna” (I) and “utata” (father). The visual relationship between the lions is connected to the words and the semantic relationship of father and son. Thus, the image of these two lions has more than one meaning in this video, based on the language used, whereas the image of the lions on their own would have been limited to a visual representation of lions and a narrower range of meanings. Therefore, the relationship between the big and small lions are collocated to the relationship between father and son. No examples of Intersemiotic Antonymy and Meronymy at this level have been found in the data analysed in this thesis, and, therefore, are not discussed here.

Intersemiotic Correspondence is taken from Jones (2007:371) as an umbrella term to describe Intersemiotic Relations of Repetition and Synonymy. An example of Intersemiotic Correspondence from the video *I Can Go To School* in this metafunction include “Usoloko

ezingca ngam" (She is always proud of me). The visuals simply show the mother with an arm around the child, which supports the subtitle, but they are not really identical in meaning.

3.1.3.2 Logical metafunction

Within the logical metafunction, there are implication sequences which illustrate logical processes that formulate a solution to a problem (O'Halloran 2005). Implication sequences are divided into comparative, additive, consequential and temporal sequences (Liu and O'Halloran 2009).

Comparative Relations in the logical metafunction map out the Similarity and Contrast between clauses and scenes and illustrate co-contextualisation. For example, in the video *They All Loved Sue*, "USue wayezithanda izilwanyana" (Sue loved animals) and "Wayezithanda zonke iintlobo zezilwanyana" (She loved all kinds of animals) are two different subtitles that appear simultaneously with the same visual of Sue, the squirrel and the bird. This illustrates a Similarity in content between the two subtitles and the visual representing them. In Intersemiotic Additive relations, different meanings are conveyed, but the semiotic resources of one mode add information to another (Liu and O'Halloran 2009). For example, in the subtitle "Ndinomhlobo wam oyedwa (omnye) ophuma ebusuku kuphela" (I have one friend of mine who comes out only at night), the image shows the father lion and his son and the moon rising, thus the visual provides information on the friend's identity (i.e., the moon).

Intersemiotic Consequential Relations portray a relationship of enablement and are further divided into Intersemiotic Consequence and Contingency (Liu and O'Halloran 2009). Intersemiotic Consequence occurs when semiotic resources of one mode are a direct consequence of another, such as an image portraying the consequence of the meaning represented by language (Liu and O'Halloran 2009). There are no examples found in the videos, therefore, this is not illustrated here. Intersemiotic Contingency is similar, except it determines a possibility, with the after-effect not a guarantee (Liu and O'Halloran 2009). For example, a subtitle in *Jungle Tumble* says, "Tsala isebe lomthi…Uza kubona!" (Pull a tree branch… You will see!)[16]. From the previous information, it is implied that parrots hide in tree

[16] see (5) in 1.2.1.2

branches, and after this subtitle, a monkey pulls at the tree branch and a parrot is revealed and flies in the sky.

Intersemiotic Temporal Relations portray a relationship of succession between different modes' semiotic resources, e.g., linguistic instructions, whereby each linguistic step is followed by multiple images for elaboration (Liu and O'Halloran 2009). This is illustrated in the unfolding of the actions and subtitles in the video *The Icky Sticky Frog*, as when the subtitle includes an onomatopoeic word or description of the frog's tongue coming out and the disappearance of the insect, the visual mode depicts the frog either catching an insect with its tongue or swallowing it sequentially after the subtitle has announced it.

3.1.3.3 Interpersonal metafunction and textual metafunction

In the interpersonal metafunction, visuals can create co-contextualisation with Mood and Modality. Intersemiotic texture in the interpersonal metafunction is established with the presence or absence of visual techniques that create co-contextualisation with the linguistic speech functions (Royce and Bowcher 2006: 71-72). This includes the presence of Gaze or facial expressions from the Participant to the viewer that may clarify a question or a sentence. It could also include Horizontal camera angle which illustrates levels of involvement. For example, the subtitle in the Interrogative mood "Ingaba yinyoka emdaka le ihleka emthini?" (Is that a brown snake laughing in the tree?) could have had Direct Gaze, potential arrows pointing at the Participant in the video and Front and Angled views of the Participants to emphasise and converge with the meaning implied by the mood. However, instead in this situation, there are no Participants aside from an unknown, moving object that seems like a brown snake suggested by the language. Modality of obligation in the language can be highlighted using specific camera angles (such as Vertical angles) and Social distances. As discussed in 3.1.1.2, Modality in isiXhosa is quite complex and types of modal realisations in relation to Participants and their related Processes are restricted to the linguistic mode without an intersemiotic relational analysis of systems of Modality between modes. The videos were made in English before being translated to isiXhosa and thus, the elements of Modality within the isiXhosa language might manifest differently to their relation to the visual realisations of Modality. The English original videos had instances in the language that did not carry Modality, but in the isiXhosa translations, they do. For example, in *The Icky Sticky Frog*, the original

phrase in the video was “The fly was gone,” which was translated to “Ayisabonakali” (The fly cannot be seen).

Within the textual metafunction the Intersemiotic Cohesive Devices are Theme-Rheme and Given-New information organisation, with Given information usually presented on the left and New information on the right (Kress and van Leeuwen (2006). For example, in the video *Jungle Tumble*, the visual of a parrot on the left-hand side of the screen represents Given information that corresponds with the Theme in the simultaneous utterance and subtitle “Izikhwenene zizifihla phezulu emithini” (Parrots hide themselves high up in the trees).

Comparative Relations, Additive Relations, Consequential Relations and Temporal Relations are systems of Intersemiotic Texture that navigate meaning relations between modes. In particular, this thesis examines the types of Comparative Relations, which include Hyponymy, Meronymy, Polysemy, Antonymy, Correspondence and Collocation, in order to understand where meaning relations are similar or contrast with each other between modes. This relates to Mayer’s (2009) Principles of Coherence, Spatial and Temporal Contiguity discussed in 3.2.1, which suggests that similar meanings should co-occur or occur close to one another in a shot for better learning, as well as extraneous information be removed and occur less in the video. The next section discusses multimodal literacy and learning theories.

3.2 Multimodal literacy and learning theories

According to Walsh (2010), multimodal literacy is making meaning of digital and multimedia texts through reading, interaction, viewing, comprehension and producing or reacting. Screen reading is less linear than print, and there are different ways of navigation and processing used when reading on a screen (Walsh 2010:241). This refers to the distinction between horizontal reading and vertical reading in multimodal texts (Kruger *et al.* 2022), which relates to intramodal and intermodal (multimodal) semiotic awareness discussed in 1.2.3. The Picture-Word Match tests in this video are often combined in learning with video in foreign language contexts (see Schafli 2020) to illustrate whether viewers are aware of connections between different semiotic systems, in this case words and visual representations. (see 1.2.3 and 4.2.2). This thesis was interested in whether the learners had semiotic awareness skills in the advanced stratum, with which they could make connections between different multimodal semiotic resources and the meanings they convey. However, further analysis of specificities in learner

semiotic awareness would require a vast array of tasks for learners to complete and was beyond the scope of this thesis. This is an area for future research.

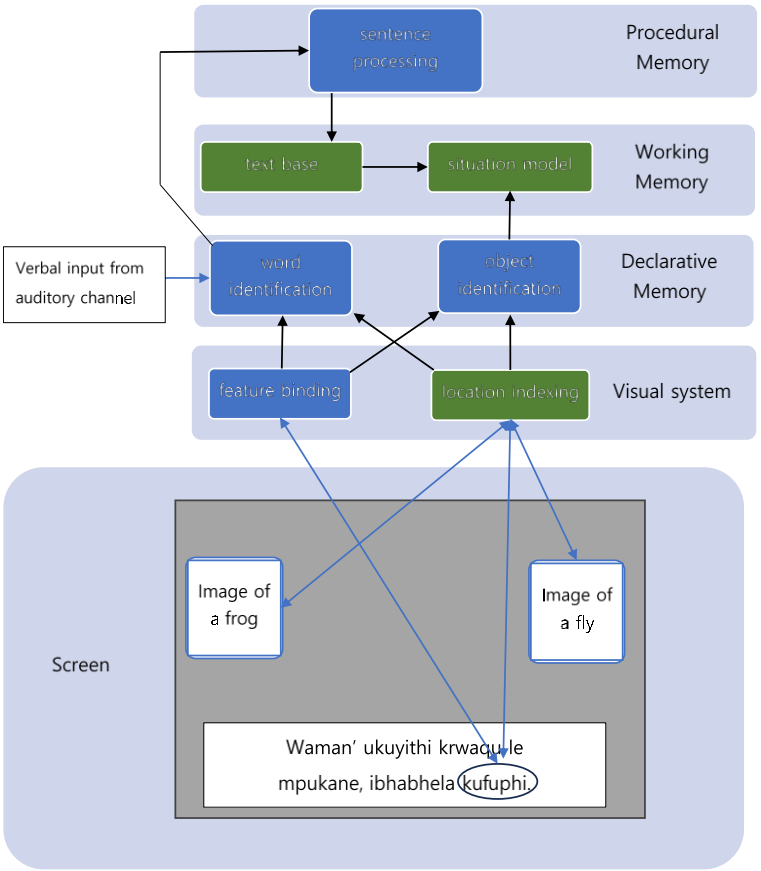


Figure 3 Adapted figure from Liao et al. (2021) of a multimodal-integrated language framework for multimodal reading

In Liao *et al*. (2021), a multimodal-integrated language framework for multimodal reading is proposed. In an example from the video in this thesis *The Icky Sticky Frog*, the shot contains the image of a frog (left) repeatedly glancing at a hovering fly (right) and the subtitle has a corresponding translation, he kept glancing at the fly, flying nearby (see Figure 3.[17] This would correspond with the Given and New Information systems proposed by Kress and Van Leeuwen (2006) in the textual metafunction. Due to the location tracking of visual objects and linguistic words on the screen in relation to one another (location indexing), and the visual identification of these objects in relation to their meaning, a model of the situation may be conveyed in the visual mode before the linguistic mode (see mental representations and semiotic awareness in

[17] see (15) in 1.2.1.2

1.2.3) (Liao *et al.* 2021). As depicted in Figure 3, the word “kufuphi” (nearby) is relating the location of the visual image of the fly to the visual image of the frog. However, the word “kufuphi” may need to undergo further sentence-level processing in procedural memory, as an additional step to word and object simultaneous identification in the declarative memory (see Figure 3). This is provided that the visuals and language convey diverging or contrastive meanings (Liao *et al.* 2021). If words in the written linguistic mode are identified and correspond with objects in the visual mode by their location (similar to Given and New Information in clauses in DMDA), words may not need to undergo further sentence-level processing. This is discussed as vertical reading of multimodal texts by Kruger *et al.* (2022) and related to an advanced stratum of multimodal semiotic awareness see 1.2.3, 2.2.3.1 and 3.1.2.

Another way text could be identified, according to Liao *et al.* (2021), is in selecting features of new objects serially on-screen, for example, letter identification, in the case of the written linguistic mode or selecting and identifying visual traits to identify visual objects. These features are then bound as a.) an identified word, b.) other words and grammar of a sentence. The information based on the text is then integrated into a model of the situation depicted and in working memory (Liao *et al.* 2021). This is discussed as horizontal reading of subtitles by Kruger *et al.* (2022) and related to a simple stratum of intramodal semiotic awareness see 1.2.3, 2.2.3.1 and 3.1.2.

Spatially focused attention is required for localisation and binding of object features to an area of focused attention. However, features that are salient and simple can be perceived without focused attention on object locations. Cognitive processes that allow simultaneous preservation of multiple representations are filled with the colour green in Figure 3. According to Liao *et al.* (2021), the locations of four to five objects may be visually tracked at the same time. Text comprehension is thus supported by concurrent object identification, which can assist construction or preservation of an appropriate model of the situation. For example, in Figure 3, if both the frog and the fly have been identified previously, skimming the accompanying subtitle would allow for a complete understanding of the situation, either in the written linguistic mode or the visual mode. A further possible entry point for spoken input from the auditory channel is depicted in Figure 3. Based on this model, it also suggests that spoken language is processed faster than written language, but slower than visual objects (see 1.2.3).

This model proposed by Liao *et al.* (2021) implies some foundations to take into account for this research. The comprehension of subtitled video may not only be reliant on one mode, but identified components in all the modes may allow for a creation of a model of the situation that creates a context, in which subtitles are read (Kruger *et al.* 2022). This could then ease the cognitive load on processing all information in the subtitles (Kruger *et al.* 2022). However, if a word or visual object is misunderstood or remains unidentified, it could jeopardise the processing of the sentence and negatively impact the model of the situation (Kruger *et al.* 2022). The inability of comprehending subtitles correctly from lack of word identification or incorrect feature binding could potentially impact on a misinterpretation of information for learners using the videos, thus impacting their comprehension skills.

However, this discussion that visual processing reduces the cognitive load of written language processing (further assisted by the presence of spoken language) raises the question of whether learners would focus on the subtitles at all, when they could potentially process the information from the spoken linguistic mode and the visual mode. Not reading subtitles, one could still comprehend a video. But one can only potentially develop literacy with SLS videos, which relies on the written language, by reading and engaging with subtitles. This proposed theory highlights the importance for educators to ensure learners focus their attention on the subtitles without focusing on them too much or ignoring them completely in order to learn from them for literacy purposes.

The focus of attention on subtitles, is however, impacted by the subtitle rate (Kruger *et al.* 2022). If the rate is too fast, subtitles could not only be completely ignored for other favourable modes, but they could also be the only mode on which a learner focuses, in order to try and read it all before it disappears (Kruger *et al.* 2022). If they cannot process it accurately, also due to a fast subtitle rate, a sole focus on this mode in such a context would then result in cues from the other modes being missed that could have aided in comprehension (Kruger *et al.* 2022). It also indicates that a Similarity of meaning between modes may reduce cognitive load, allowing for learners to focus on new and unidentified features for longer, therefore assisting in literacy development. Furthermore, this model, along with Mayer's (2009) Cognitive Theory of Multimedia Learning, follows the dual-coding theory that suggests that information is processed through two different channels; spoken linguistic, and visual (Paivio 1969). This is further in alignment with dual-route models of literacy and reading (see Coltheart *et al.* 1993), which is a motivation for why Mayer's (2009) Cognitive Theory of Multimedia Learning was

chosen as the primary theory of triangulation for the results from the DMDA and literacy intervention. While this thesis supports dual-route models for reading in relation to limited capacity in routes, due to the additional factor of agglutinative morphology in isiXhosa, reading of this orthography is described in relation to Grain Size Theory (see 2.1.2) to account for this processing grain (see Probert 2016). There is a lack of specific multimodal literacy theories in relation to intramodal literacy learning with video and principles of video design, particularly in isiXhosa. Mayer's (2009) Cognitive Theory of Multimedia Learning, although from a learning context and not a literacy learning context, contains broader principles of how semiotic resources conveyed in particular modes should be arranged for optimal cognition and perception. This theory relates directly to video design, which is also a focus of DMDA, but Mayer's (2009) Cognitive Theory of Multimedia Learning further relates this to the perception and cognition of the learner. The next section discusses Mayer's Cognitive Theory of Multimedia Learning.

3.2.1 Mayer's Cognitive Theory of Multimedia Learning

Mayer (2009)'s Cognitive Theory of Multimedia Learning focuses on learning as transfer from visuals and words to a comprehension of meanings in relation to multimedia design. Visuals can be static or dynamic and words in this context are spoken or written (Mayer 2009).

There are three assumptions regarding multimedia design: the dual-channel assumption, limited capacity assumption and active processing assumption. The dual-channel assumption presupposes that a person's cognition comprises of two separate auditory and visual channels for processing information. Mayer (2009) states that if the reader has no experience, text on the screen is initially processed in the visual channel. During this process, visual words are mapped onto sounds, thus requiring further processing in the auditory channel. A similar process occurs in writing, according to the dual-route approach to spelling (Rapp 2002).

The limited capacity assumption presupposes that each channel has a specific volume of storage space for information. The cognitive theory of multimedia learning describes three types of memory for learning: long-term, working and sensory memory (Mayer 2005). In sensory memory, the processing of images and words from presented media occurs (Mayer 2005). The transformation of sensory information into abstract knowledge occurs in the working memory and these processes include word and image selection, organisation and integration (Mayer 2005). The long-term memory is where lengthy storage of information

occurs after being processed in the working memory (Mayer 2005). The limited capacity of the working memory, in particular, to hold information before storage, is a founding assumption for the principles outlined in this theory. Thus, the organisation and presentation of information for learning has to be conducive with the limited capacity of the working memory and the dual processing channels that are used to take learned information into that storage.

The active processing assumption states that people use active cognitive processes such as information integration, information organisation and paying attention to create mental constructions of experiences, see 1.2.3 (Mayer 2005). This assumption presumes that learners are not passive, but active processors and that material is structured clearly with a guiding message.

Mayer (2009) proposes ten principles to successful learning with multimedia in the classroom and I discuss these in relation to DMDA. The Coherence, Signaling, Redundancy and Spatial and Temporal Contiguity Principles relate to the avoidance of irrelevant processing, which does not assist the construction of adequate models of a situation that aide learning (Mayer 2009). The Coherence Principle states the more irrelevant material (unnecessary information for the instructional goal) present in multimedia, the less cognitive capacity a learner has for integral information. This can relate to structure and content Similarity between modes and metafunctional systems, as content that is additional and contrasts with construed meaning within a particular mode may pose a greater cognitive load for the student. According to Mayer (2009), learning is best when it creates interest, but extraneous visuals, words, music and sounds are removed from multimedia, as well as when unnecessary words and symbols to learning instruction are eliminated. Interesting material is defined as such by learners; however, this is incredibly subjective and entirely subject to the learners, their preferences, and the context (Mayer 2009). While interesting content to the learner may arouse emotions and this may also allow for longer retention of this information, if this information is extraneous material, it can divert attention from the important material in working memory, and prime learners to integrate material under inappropriate topics (Mayer 2009).

The Signaling Principle states that highlighting essential words guides learners when processing knowledge and leads to more efficient learning (Mayer 2009). Highlighting can be in the form of vocal emphasis, textual headings and pointing words in the beginning of sentences such as “first, second and third” (Mayer 2009:109-110). Visual highlights can be in the form of cues such as distinctive colours, arrows, pointing gestures, flashing, and to grey out

nonessential areas (Mayer 2009:109-110). None of the visual or verbal highlights were found to occur in a way that highlights information in the videos as outlined by Mayer (2009). However, this thesis proposes that there are potentially more ways of highlighting information than outlined in Mayer (2009). In terms of DMDA, this may be visible in the systems of Salience, Social Distance, Camera Movements, Gaze, Framing and Information Value which focus on highlighting specific information in the visual mode. Theme may also illustrate a form of linguistic highlighting in the linguistic mode. Thus, this thesis takes an expanded definition of Mayer's (2009) Signaling Principle, that not only vocal emphasis and textual headings in the linguistic mode are a manner of highlighting essential words, but also placing words as Themes in clauses. In the visual mode, colours, arrows, pointing gestures, flashing and greying out can highlight essential information and words, but this can also be done by Camera movements and Shots, Participant Gaze, Framing devices, Salience and Information Value. This thesis proposes how this could be achieved from the video design, and further testing is required as to whether more or less highlighting in both modes does improve literacy learning in this context.

The Redundancy Principle implies that even if on-screen text and narration are the same, people learn better without on-screen text with narration and visuals, see 1.2.3 and 3.2 (Mayer 2009). This is an integral principle for this research that seeks to determine the opportunities and challenges with subtitles that have the same linguistic content as spoken language and that are presented with visuals, as this principle implies subtitles along with spoken narration and visuals, is redundant. Mayer's (2009) theory makes an opposite prediction from the research that shows the efficacy of SLS for learning. Mayer (2009) claims the presence of visuals and spoken and written text simultaneously cause redundancy, which creates extraneous processing as, by visually scanning pictures and on-screen text, the visual channel can become overloaded.

Furthermore, learners could expend more of their cognitive capacity in trying to compare information from both the visual and spoken linguistic mode than necessary (Mayer 2009). The Redundancy Principle may be less suitable for learning when the written language is concise and placed next to the visuals they describe, the spoken language is presented before the written language, rather than simultaneously and if there are no visuals, with short linguistic sequences (Mayer 2009). It is also argued that both written and spoken language as options for the learner to select their preferred mode to process information aligns with their preferences for processing information, however Mayer (2009) disagrees with this perspective, arguing for the

term redundancy in a restricted and narrow sense and aligning with the limited capacity assumption. This thesis reveals insights as to whether Mayer (2009) is correct for isiXhosa Foundation Phase learners in South Africa, or whether SLS videos influence literacy learning among these learners, and under what conditions within the context is this realised. While subtitle research discusses opportunities with subtitles, this principle argues that subtitles would not be beneficial for learning. However, in a literacy leaning context, where the written language is required in order for learning, and the spoken mode assists in understanding the written language, this thesis seeks to investigate possible instances of when the three modes presented simultaneously cause redundancy or when they assist in literacy learning.

The Spatial and Temporal Contiguity Principles discuss that if complementary spoken words and visuals are presented simultaneously or close to one another, learning is improved (Mayer 2009). The Spatial Contiguity Principle indicates that necessary words for learning (in this case, literacy development), should appear beside corresponding images, while the Temporal Contiguity Principle indicates they could be presented together at the same time (Mayer 2009). This relates to Similarity within Comparative Relations, because if the visuals and words present Similarity in meaning, they reduce cognitive load in comparison to if both channels present information differently at different times. This is argued as there is a greater chance for learners to simultaneously maintain visual and spoken linguistic information in their working memories (Mayer 2009). Similarly, there is less of a chance that learners can simultaneously maintain information from these two modes in working memory if images and pictures are located far apart on a screen or page, as they have to spend more cognitive capacity while visually identifying and binding information (Mayer 2009). The Spatial Contiguity Principle is most suitable if a learner is unfamiliar with the material, and thus this is proposed for literacy contexts and unfamiliar words or complex material (Mayer 2009). If the visuals are not fully understandable without words, the Spatial Contiguity Principle is also recommended (Mayer 2009). The Temporal Contiguity Principle is also recommended when the lesson is at a pace too fast for the learner and when they cannot control the pace or order of presented information (Mayer 2009). This further relates to similar meanings and words integration for multimodal literacy (Mayer 2009, see 3.2).

The Segmenting, Modality and Pre-Training Principles relate to successful management of learning processing. The Segmenting Principle stipulates that animation that is narrated should be at a pace the learner can follow (Mayer 2009). According to Mayer (2009), the principle is

most suitable when the lesson is at a pace too fast for learners, and the learner cannot alter the order of information presented or the pace. Mayer (2009:187) suggests segmenting information, such as providing learners with "Pause/Continue" buttons to control the information, in order to manage their cognitive processing and not miss information. I argue that this principle is not restricted to narrated animation, but also applies to the subtitle rate, which should also be at the pace of the learner in order to be comprehended (Kruger *et al.* 2022). This is this thesis' expanded definition of Mayer's (2009) Segmenting Principle. Since subtitles cannot be learner-controlled in video, unless they have access to pause and rewind, the videos should be played numerous times, but also the subtitles cannot be too fast for learner reading rates.

The Pre-Training Principle stipulates that if learners have prior knowledge of content, they will learn more easily (Mayer 2009). In fast-paced narrated animation, learners have to cognitively create a model of situation[18] (Mayer 2009). Pre-training on particular components, words and topics, in order for them to be preidentified, can help manage and distribute cognitive demands on processing (Mayer 2009). This is particularly advised when material is unfamiliar, fast-paced and complex (Mayer 2009). Prior exposure may assist, however, explicit training on words in order for it to be prior knowledge implies that the videos in this thesis would be more useful in lessons if learners had pre-training on particular concepts. This thesis investigates this by examining the significance of these videos alone on literacy learning improvement during the intervention conducted.

The Modality Principle postulates that spoken language and visuals used together are more suitable for learning of information than written language and visuals (Mayer 2009). This again relates to the visual channel potentially being overloaded, rather than using the two channels for information (Mayer 2009). However, in a literacy context, written language is integral, and visuals cannot be presented in the auditory channel, thus there is more visual information presented (Mayer 2009). This thesis investigates potentially how the presence of written language and visuals could be redundant or how they could potentially be optimal for literacy learning in the research context. This is also an important hypothesis for this research to gauge as to whether this is truly the case and whether a combination of all three modes (written language, spoken language and visual images) creates more difficulty for learners (Mayer 2009). This principle may be applied when material is unfamiliar, fast-paced and complex,

[18] See 3.2

whereas written language may be more suitable when symbols and challenging words are included, and when the learner is hearing-impaired or a non-native speaker of the language[19] (Mayer 2009). These conditions outlined by Mayer (2009) indicate how context can vary whether these principles are applicable, similar to his Redundancy Principle. If printed language is applicable for non-native speakers, in a literacy context where written language is the required learning material, it is a condition where the Modality Principle may not be applicable (Mayer 2009). As Mayer (2009:217) states, "design principles – such as the modality principle – are not immutable laws that must apply in all situations." They should be referenced as guidelines and consider how information is processed by learners cognitively (Mayer 2009).

The last four principles relate to general cognitive processes and they are the Multimedia, Personalization, Image and Voice Principles. The Multimedia Principle stipulates that the combination of pictures and spoken language causes easier learning than through spoken words alone, which is another integral stance this research seeks to investigate. Mayer (2009) explains that information from language and visuals have different meaning potentials, that they cannot be entirely the same. A presentation of both language and visuals allows for learners to construct linguistic and visual models in the mind, building connections between modes, which is also more realistic of the real-world environment (Mayer 2009). In this principle, Mayer (2009) organises visuals in four broad functional categories. Decorative illustrations may entertain or interest learners, but do not amply the meaning of information present (Mayer 2009). An example from the videos analysed in this thesis include the stuffed rabbit visuals that move in *I Can Go To School*. Representational illustrations may present an element in the visual mode that corresponds in the linguistic mode (Mayer 2009), such as the image of the fly for the word "impukane" (fly) in *The Icky Sticky Frog*. Organisational illustrations depict elemental relations in sequences, e.g., flow-charts, and no examples are truly illustrated in the videos (Mayer 2009). An exception might be the illustration of Sele-sele thinking of the fly, the beetle and locust in succession as the insects he had eaten. Explanative illustrations illustrate how a system works (Mayer 2009). There are few specific examples of information presented in the visual mode that illustrate further details into specific essential concepts in the linguistic mode from the videos in this thesis. A type of explanative illustration features lion Participants in the visual mode corresponding to "mna" (I) and "utata" (father) in the language,

[19] In a foreign language (FL) classroom, for example

but the visuals are not necessarily integral for the comprehension of the linguistic text. Mayer (2009) argues that Organisational and Explanative illustrations are more conducive for learning than Decorative and Representational, in terms of adding integral knowledge distinct to that mode, yet Representational illustrations could assist in knowledge integration from both modes (Mayer 2009). Mayer (2009) thus suggests that two modes separated via the visual and auditory channel are optimal for learning, and other modes such as sound and music in the auditory channel and written language and visuals in the visual channel could cause cognitive overload and a lack of focus of one of the modes being processed through that channel.

The Personalization Principle states that if a text is less formal, learning is easier (Mayer 2008). An example given by Mayer (2009:242) is the use of "you" and "your" in the narration instead of articles like "the." Mayer (2009) states that when learners perceive that they are directly addressed by the text, this primes the social response in the learner to prepare for conversation, making sense of what the speaker is saying in preparation for a social response. This may increase active cognitive processing by learners and reduce passivity in learning. Mayer (2009) suggests this principle is more applicable for beginner learners, which means it may be a principle useful for beginner literacy learners at Foundation Phase, such as in our context. The videos in this thesis are translations of storybooks and arguably are more formal than informal, with only instances of first-person and second-person pronouns being used in *The Moon Followed Me Home, Jungle Tumble* and *I Can Go To School.* The videos, however, are not entirely formal or academic. However, Mayer (2009) has limited examples of Personalization, and this causes a very narrow definition of "formal" and "informal," as with the term "highlighting" in the Signaling Principle. This thesis proposes an expanded definition of Mayer's (2009) Personalization Principle, in which direct Gaze in the visual mode, as well as Social Distance, Horizontal and Vertical angles may assist in establishing an informal relationship between viewer and Participant. This may result in more active engagement. This may also be achieved, according to this thesis, with more culturally appropriate and relevant material that would create a more informal and personal dynamic between the learner and the videos.

Mayer's (2009) Voice Principle proposes that learning is improved when a human voice is used instead of a machine voice for spoken narration and Mayer's (2009) Image Principle proposes that people may not necessarily improve learning with the image of the speaker present on screen. Since none of the videos include a visual of the speaker, and the narrations all include

a human voice for better learning, these principles are not discussed in greater detail in Chapter 6.

While the theory has been used in understanding how information is processed in learning with multimedia, a problem with Mayer's (2009) Cognitive Theory of Multimedia Learning is that it does not take context into account and how this influences Mayer's (2009) principles. Rudolph (2017) states how, in many of Mayer's own experiments, his research was limited to the subject area of scientific multimedia and further research is required on different types of instructional media, non-scientific subjects and learners under the age of 18, as this thesis seeks to address. We cannot observe cognitive processes and interactions on their own, but rather indirectly through the biological and physical systems involved in creation and reception of meaning and the semiotic and social systems of meaning (O'Halloran 2023), Thus, learning and literacy in this context can only be understood in relation to the multimodal artefacts in order to understand how semiotic meaning is transmitted and consequently interpreted, particularly in a L1 isiXhosa literacy learning context.

3.3 Chapter summary

This chapter examines multimodality, social semiotics and the theoretical framework of DMDA (3.1.) and unpacks the different systems of DMDA that are used in the analysis of the videos and their design reported on in Chapter 6. In order to analyse meanings in the linguistic mode, types of Participants, Processes and Circumstances are examined in the ideational metafunction. Mood and Modality are examined in the interpersonal metafunction and Themes are examined in the textual metafunction. In order to analyse meanings in the visual mode, types of Participants, Processes and Circumstances are examined in the ideational metafunction. Social (Shot) Distance and Zoom, Camera movement, Gaze, Horizontal and Vertical Perspective and Salience are examined in the interpersonal metafunction and Framing and Information Value are examined in the textual metafunction. The duration and frequency of usage of system choices in these systems could indicate a stronger or weaker adherence to Mayer's (2009) principles for improved learning potential, as well as the expanded definitions of Mayer's (2009) principles for the context of this thesis. Subtitle rate is calculated in order to triangulate with learners ORF reading scores. Types of Intersemiotic Texture and Meaning Relations are further examined, particularly focusing on Comparative Relations across the metafunctions in order to determine Similarity and Contrast between meanings between modes.

This is triangulated with the semiotic awareness skills measured in the Picture-Word Match tests in the intervention, as well as Mayer's (2009) principles of Coherence, Spatial and Temporal Contiguity for improved learning potential.

Furthermore, multimodal literacy and learning theories are discussed in 3.2, including the multimodal reading model by Liao *et al*. (2021). This model illustrates some foundations to take into account for this research, as decoding and comprehension are multimodal processes. Pre-identification of components allow for an ease of cognitive load in processing, and if a word or visual objects are not identified or misunderstood, it could compromise processing and comprehension. Mayer's (2009) Cognitive Theory of Multimedia Learning calls into question how to present information in multiple modes for ease of processing, and the potential redundancy and longer processing of written text when co-occurring with spoken text and visuals. Mayer's (2009) principles of multimedia design, which are further used in Chapter 6, particularly to triangulate the data between DMDA and the intervention results in 6.6, are examined in relation to how well they were adhered within the design of the videos analysed in this thesis.

Drawing on the literature from Chapter 2 and the theories discussed in Chapter 3, Chapter 4 examines the research design and methodology, including the literacy tests in the intervention and how DMDA was used to analyse the videos in this thesis.

Chapter 4 – Research methods and methodology

The past few chapters have indicated that a multidisciplinary research approach, which can examine a complex semiotic system is crucial, as changes in one system, whether physical, biological, social or semiotic, could cause a domino or ripple effect across the entire "meta-system" (see reference to O'Halloran 2023:174 in the introduction to Chapter 1). The main aim of this research is to test, describe and analyse the efficacy of five isiXhosa story *YouTube* videos for the development of L1 isiXhosa literacy at the Grade 2 and 3 levels. Videos as a semiotic resource and their perception is a complex research phenomenon, further complicated by complex research approaches in the literature stemming from various disciplines (see sections 1.1.3, 2.2 and 2.3). The complexity of combining disciplinary approaches of psycholinguistic, semiotic, and multimodality to research L1 isiXhosa literacy for the Foundation Phase, requires a research approach that can integrate these appropriately. Furthermore, the motivations for the research are complex, stemming from the current literacy crisis in South Africa, the impact of the COVID-19 pandemic and the possible potential of SLS videos as a tool for literacy development in other social systems (see 1.1).

While a complex and vast "meta-system" (O'Halloran 2023:174) cannot fully be grasped in a single study, the multidisciplinary mixed-method design of this research seeks to triangulate data from different sources. This is to provide an extensive insight as to whether SLS videos in the context of this research are a potential resource for L1 isiXhosa literacy development in the Foundation Phase. Within this thesis, a concurrent nested mixed-methods approach (Creswell and Plano Clark 2003) is undertaken. This research undertakes a statistical t-test analysis of quantitative pre-test and post-test data from a two-group experimental design of Grade 2 and 3 learners at a school. This research further undertakes a qualitative DMDA (see theory in Chapter 3) of the five subtitled isiXhosa videos used in the treatment, henceforth discussed as the intervention.

Despite the nested design where the qualitative data explains findings from the quantitative data, their findings are further triangulated in the analysis stages with multimodal literacy and learning theories (discussed in 3.2). This is to explain the results in terms of recommendations for video design in the context of L1 isiXhosa literacy development. This further illustrates that

there is a nested mixed-method design, however triangulation is an aspect within the analysis of the data and is a term used throughout the thesis when discussing the integration of data at the stage of analysis. The statistical analysis and DMDA are discussed as separate entities in 4.3, and each discussion of analyses is provided in a separate chapter (Chapters 5 for the intervention data and Chapter 6 for the DMDA of videos). The integrated analysis is embedded within the DMDA in 6.6 within Chapter 6, as this research examines quantitative findings in Chapter 5 within results from the DMDA. This research approach is discussed in greater detail in 4.1, which includes the rationale for the research approach and research paradigms. The sampling and collection of data and how this was conducted is discussed in 4.2, and finally the analysis of data is discussed in 4.3.

4.1 Mixed-method research approach overview

A research approach is comprised of the procedures and plans involved that describe and elaborate on every step of the research process from overarching assumptions to details of data collection, analysis, and interpretation methods (Cresswell 2014). In order to understand the concurrent nested mixed-methods approach (Creswell and Plano Clark 2003) undertaken in this thesis, this section discusses the rationale for this mixed-method approach in this section. The critical realist paradigm, which explains how both these types of analysis can be integrated into an account of the videos' influence on literacy, are discussed in 4.1.1. The research design and overview of two methods is concretised and described in 4.1.2. Research procedures are discussed in greater detail in 4.1.3.

There are multiple rationales for a mixed-method design in this research. I have defined my rationales according to Bryman's (2006) classifications of credibility, confirm and discover, explanation and illustration. A mixed-method design strengthens the credibility and integrity of findings when examining a single phenomenon with multiple methods (Bryman 2006). This is to gain a triangulated understanding into how SLS videos as a resource influence literacy, as the videos and their influence on learners' L1 literacy skills are examined, the former with DMDA and the latter with test scores. In terms of the confirm and discover rationale, the qualitative research of the DMDA provides potential areas of L1 isiXhosa literacy development in analysing the videos' design, while the quantitative findings examine these in a single project (research site). In terms of the explanation rationale (Bryman 2006), the quantitative results explain whether there was significant improvement in L1 literacy scores in terms of ORF and

broad semiotic awareness, while the DMDA explains potential reasons for the quantitative findings from the literacy intervention. This is similar to the rationale for illustration, where Bryman (2006:106) uses the metaphor of "qualitative "meat" findings placed on the "dry" bones of quantitative results.

There is also a rationale, which includes diversity of views (Bryman 2006). The types of this rationale that underpin the choice of a mixed-method approach include the production of interdisciplinary theory and comparing multiple data (videos and test scores) regarding the phenomenon of L1 isiXhosa literacy development. In this research, this includes explaining the interaction between human systems of video production and video viewing and elaborating on the complexity of videos and their potential use as resources specifically for L1 isiXhosa literacy development.

The literacy tests in this study outlined in 4.2. measure whether there was any difference before and after video exposure in a literacy development context. The DMDA seeks to examine meaning potential within the videos and the opportunities and challenges the videos may provide. The two-group experimental research design was based on a similar study conducted by Kothari *et al*. (2002) mentioned in chapters 1 and 2, as they had a similar research design to examine SLS in their context. The systems of analysis for the DMDA are outlined in 3.1.

In order to further substantiate the rationale for decisions made in the method choices for data collection and analysis, the critical realist paradigm of this research is discussed in 4.1.1. A research paradigm consists of a set of ideas, theories and assumptions that illustrates the lens through which a researcher observes and understands the world (Mackenzie and Knipe 2006).

4.1.1 Research paradigms

This thesis takes a stance in critical realism, combining a critical realist ontology with a relativistic epistemology (Stutchbury 2021). Critical realist ontology identifies systems, operations and interactions that are emergent, stratified and can transform (Fleetwood 2019). Epistemologically, this research paradigm sees the social world and natural world as one reality (Fleetwoord 2019. It believes reality can hold extra-discursive systems and multiple

Discourses[20] (Fleetwood 2019). One's social context influences one's Discourse, as well as formulating and making assertions about concepts (Fleetwood 2019). Critical realism acknowledges that there are different levels (strata) of reality (Stutchbury 2021). One can understand these stratified levels of reality as an iceberg, whereby the surface tip of an iceberg is the visible reality that can be observed and the bulk of the iceberg under the surface is most of reality (Stutchbury 2021). Thus, within this invisible, buried surface of reality, exist "causal mechanisms" or "explanations," which cause events and observed experiences to occur in the visible, observable reality (Stutchbury 2021:114). The causal mechanisms which determine events and experiences may not be observed, but the effects of causal mechanisms are (Stutchbury 2021). Thus, causal mechanisms can only be identified from inference, which draws upon data analysis, theoretical analysis and the research context (Stutchbury 2021).

Bhaskar (1975) posits reality in terms of three autonomous levels of depth, one domain for the invisible level and two domains for the visible levels. The invisible level is the real domain, which includes the invisible, deeper-lying structures and causal mechanisms. The two visible and observable levels are the empirical domain (visible/measurable experiences and perceptions of knowing subjects) and the actual domain (qualitatively investigated events and objects that occur in the real world). Critical realism seeks these "causal mechanisms" by focusing on the agency of people in the social context (observable structures in the empirical domain) in which they operate (Stutchbury 2021). Thus, we would examine the videos in a DMDA and use pre- and post-intervention test results in the empirical domain on visible events. These events occur in the actual domain, such as the learners' literacy skills and the development of this in their interactions with videos in the intervention within the surface tip of an iceberg (see Figure 4). This is to infer the invisible mechanisms that influence learners' literacy developments.

[20] See 3.1.1 for the definition of Discourse with a big "D" used in this thesis.

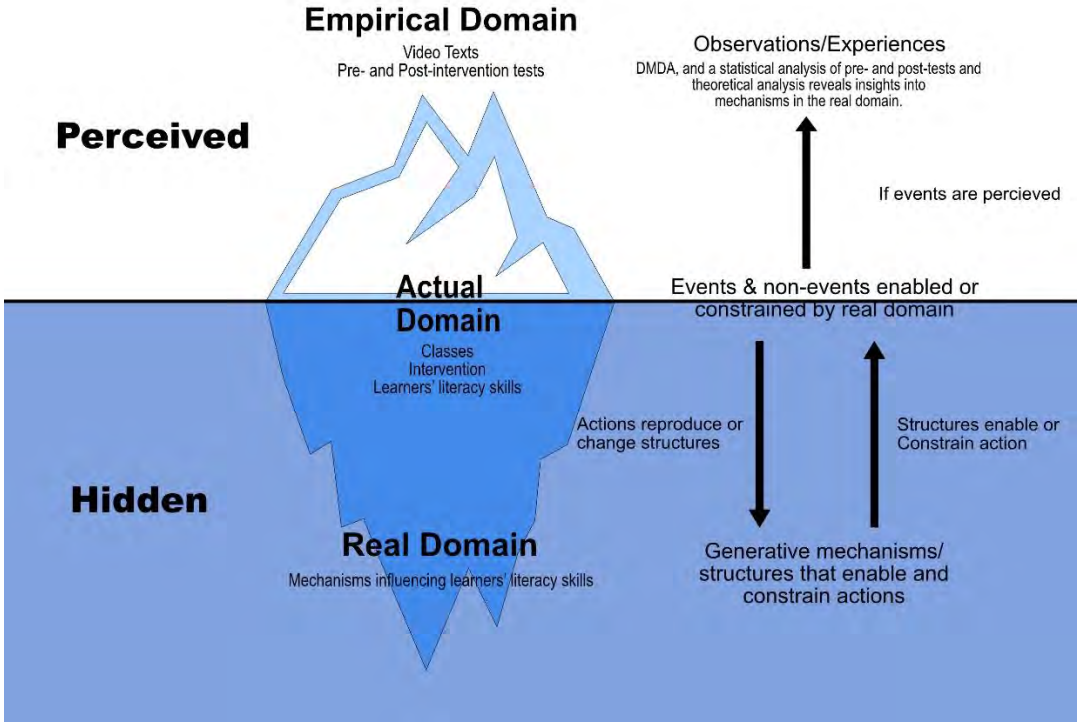


Figure 4 Critical realist paradigm of concurrent-mixed method design and this research methodology adapted from Bhaskar (1975)

The complementary use of statistical analysis of literacy scores and DMDA creates a critical realist paradigm. A critical realist paradigm provides an alternative paradigm to the opposing positivist paradigm of statistical analysis and interpretivist paradigm of text analysis, in which the natural and social world are treated as independent from each other (Fleetwood 2019). The events in the actual domain are the intervention, in which learners were exposed to the videos (thus, a Discourse was established) and participated in the tests. Another event within the actual domain included the production of the videos, which were later used in the intervention. Due to the presence of the videos production and integration in the intervention in the real world and actual domain, the videos as resources and data, along with the pre- and post-intervention test results, are both within the empirical domain. Both the videos and the pre- and post-intervention tests are empirical experiences generated by the events of intervention and production. The DMDA of videos, as with the analysis of pre- and post-intervention tests, and any triangulation of multimodal literacy and learning theories from Chapter 3 reveal insights into one Discourse (in this case, interpretations of the video by learners) and causal mechanisms that influences learners' isiXhosa literacy skills. The Discourse relates to the design of the videos, which influences the learners' experiences in the intervention and may influence their L1 literacy development, which is the aspect of investigation of study.

While there are potentially many causal mechanisms this study hopes to discover, two inferences of causal mechanisms for this study were identified in the conceptualisation stages of research design. These two inferences of causal mechanism are (1) the ability for learners to follow subtitles adequately to influence literacy development and (2) the nature of learners integrating meanings between visual and linguistic modes to assist in literacy development, i.e., effectively combining meanings from both modes to assist in the development of reading. Within the conceptualisation stages of research design, these were identified as the points of interface within the concurrent nested mixed-method model. The points of interface or points of integration is the area at which the core data (DMDA data) and supplementary data (statistical analysis data) are integrated and triangulated with theory (see Morse and Niehaus 2009).

4.1.2 Research design outline

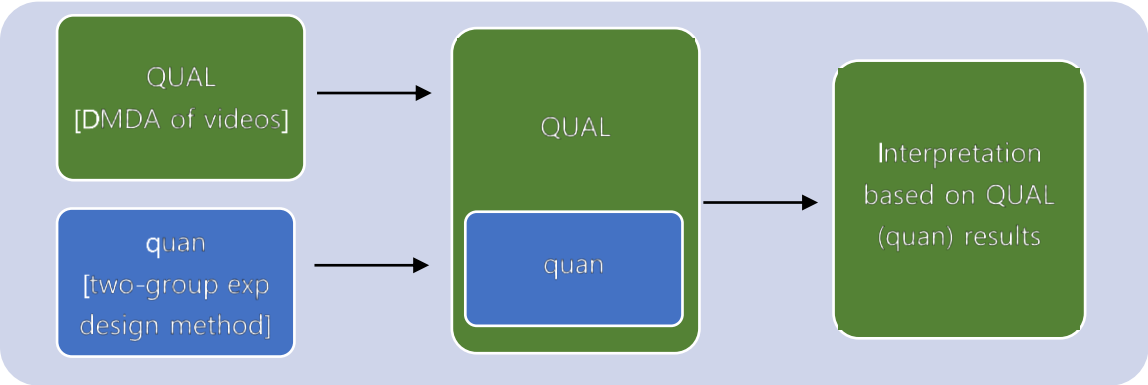


Figure 5 Concurrent nested mixed-method research design model (adapted from Creswell and Plano Clark 2003)

Within the conceptualisation stages of research design, these were identified as the points of integration within the concurrent nested mixed-method model. Timing was cross-sectional as the period of the PhD project from proposal to completion was set for three years, in which proposal and planning was conducted in year one, data collection and analyses conducted in year two and thesis completion in year three. The three-year deadline was constrained by external funding for the project and 3 years being the minimum requirement for the completion of the thesis.

The two inferences of causal mechanisms for this study are identified as the ability for learners to follow subtitles adequately to influence literacy development and the nature of learners integrating meanings between visual and linguistic modes to assist in literacy development. Thus, the points of integration in the analysis stage would be at the subtitle rate and the relationship between the visual and linguistic modes. In the DMDA, this was realised in the systems of subtitle rate and Intersemiotic Texture, more specifically the sub-system of Comparative Relations (see 3.1.3). Thus, tests were designed to explore perceptions of these two areas of video design from the learners in an experimental research method, specifically two-group experimental design. This was within a small-scale intervention. An elaboration on the small-scale intervention and decisions on sampling are discussed in 4.2, the ORF tests included Word Reading lists (Word List 1 including words in the videos and Word List 2 including words not in the videos) and Long Passage Reading, while another test included a Picture-Word Match. Data from Word Reading lists and Long Passage Reading were to be compared with the subtitle rate in the DMDA, and the Picture-Word Match results were to be understood in relation to the system of Comparative Relations. A more detailed overview of the two-group experimental method is discussed in 4.1.2.1.1.

This illustrates the nature of this mixed-method design model being an embedded or nested model. The quantitative pre- and post-test results were supporting data found primarily in the DMDA of the videos (see 4.3.2), which is the nature of embedded or nested mixed-method designs (Creswell and Plano Clark 2003). These two pathways of investigation were necessary for a comprehensive picture of how subtitled isiXhosa videos influence L1 literacy. A DMDA examines meaning and learning potential (Jewitt *et al.* 2016; Tan *et al.* 2015). In this particular case, not only do we see how meaning may be interpreted but how exposure to writing in the form of subtitles with visual aids influence learner literacy. The DMDA was selected as a method and analysis to answer the research questions 1-4. Research questions 1-4 are:

1. How is meaning construed in terms of the systems and sub-systems of the mode of spoken language in the videos?
2. How is meaning construed in terms of the systems and sub-systems of the mode of written language (subtitles) in the videos?
3. How is meaning construed in terms of the systems and sub-systems of the visual mode in the videos?
4. What is the relationship between the various modes of language and image within these videos according to the DMDA systems and sub-systems of these modes?

This is also illustrated by the two-group experimental design method and the statistical analysis of its resulting data being restricted to research questions 5-6:

5. How does the combination of the modes influence learner literacy levels?)
6. What effect does repeated exposure to the videos have on literacy scores?).

This illustrates the embedded/nature of the model, with less research questions being answered by the statistical analysis than the DMDA. The data from both analyses are triangulated at the research stage of triangulation for research questions 7-8:
The data from the statistical analysis and DMDA are triangulated with the theory outlined in 3.2 for research questions 7-9:

7. What are the opportunities and challenges associated with learning word recognition, oral reading fluency and semiotic awareness with these videos?
8. What recommendations can be made for the design and use of future educational videos in line with the policies of the DBE?
9. How can DMDA be used in combination with psycholinguistic approaches to literacy to investigate and foster literacy development in schools?

Priority is given to the DMDA results, which explains the nested mixed-method model (Creswell and Plano Clark 2003). Thus, both methods in the analysis are not given equal weight, rather the quantitative two-group experimental design is a small-scale intervention, embedded in the DMDA. This is an analytic procedure described as comparing results by Creswell and Plano Clark (2003), which involves a comparison between quantitative and qualitative results, in which the qualitative themes are supported by quantitative statistical trends.

4.1.2.1 Two methods overview

The subtitles as a semiotic resource within the written linguistic mode are seen to contain both form and meaning, which are chosen, presented and perceived simultaneously and not in isolation from one another as a motivated sign (Jewitt *et al.* 2016). However, in young children, understanding the relationships between form and meaning, as well as applying this understanding to other modes, such as from multimodal videos with dynamic subtitles to sole static text reading (semiotic awareness, see 1.2.3), are a set of skills stratified in terms of

complexity. Certain aspects of semiotic awareness, particularly in terms of written language, may develop earlier and others may develop later (see 1.2.3). In order to further understand semiotic awareness between modes, a Picture-Word Match test was used before and after video exposure to understand inter-modal relational abilities of young children at this level of literacy. The words from this test were from the videos, in order to understand how the frequent exposure to meaning occurrences and their duration influenced this aspect of semiotic awareness.

As discussed in 2.1.1, increased exposure facilitates development of literacy skills (an aspect of semiotic awareness) for automatic reading, which can be measured by ORF (Pretorius 2010). ORF is a skill that connects distinct but connected processes of decoding and comprehension (Pretorius 2010). As highlighted by the research questions (see 4.1.2), ORF at the word (word recognition) and passage reading level are tested before and after the repeated SLS video exposure to learners. This repeated exposure had a duration of 15 sessions and frequency of occurrence of three sessions per video, which were played twice in one session, thus learners were exposed to each video six times. ORF was measured by controlling the duration of oral reading in the tests for a minute. The mean scores of both pre- and post-tests were not only compared to the mean scores of subtitle rate in the video in the DMDA, but t-tests were also conducted to determine whether SLS video repeated exposure (the exposure of semiotic resources and their meanings from multiple modes) influenced learners' L1 literacy skills. Word Reading lists separating basic lexical units are examined to determine whether a frequency of meaning occurrences from semiotic resources influenced automaticity in ORF.

These aspects were then compared to both the rate of subtitles in the video (a measurement of subtitle duration) and the duration and frequency of meaning occurrences in the DMDA, which were analysed using separate meaning system choices. This was to relate DMDA findings to theories of multimodal literacy and learning regarding features of design (3.2) and, thus, elaborate on the potential influence of SLS videos on L1 isiXhosa literacy skills. The duration and frequency of meaning occurrences are represented and analysed in the DMDA by State Transition Machine Diagrams (STMs), with states representing meaning(s) duration in the videos and transitions representing meaning(s) frequency of occurrence (see 2.1.1 for the distinction between duration and frequency of occurrence).

This section introduces the methods in summary, first the two-group experimental method design (4.1.2.1.1) and secondly the DMDA (4.1.2.2.2) before describing steps of research procedure (4.1.3). This is to provide a brief overview of each method, before discussing the steps of research procedure, and further elaborating on choices and motivations for data collection and analysis of the data collected from each method.

4.1.2.1.1 Two-group experimental method design overview

This quantitative methodological process takes a two-group experimental research design. This was to ensure pre-test and post-test results could be compared to understand the efficacy of video treatment (an aim of this thesis). The decision was made for the research to be conducted in groups and follow a two-group design to truly ascertain the impact of the SLS videos on learners and to try and exclude any other influencing factors in testing. In this design, a control or "untreated" group is contrasted with a treatment or experimental group (Lavrakas 2008). A similar two-group experimental group design was conducted by Kothari *et al*. (2002), although they had two experimental groups, one with subtitled Hindi videos and one with Hindi videos containing no subtitles. The control group had no video exposure. Since this research was focused on the integration of subtitles within the videos, the viewing of videos without subtitles was deemed unnecessary for this research. It was also easier on the research team and learners to conduct video viewings in groups, rather than with individual learners, as this would spend more of the research assistants and learners' time than was necessary.

The pre-tests were identical to post-tests (see a-d of step 3 in 4.1.3). This includes the ORF tests for decoding word recognition skills in relation to Oral Reading Fluency (ORF) in isolation (Word List 1 and 2), longer passage reading (Long Passage Reading), as well as the Picture-Word Match test for word recognition in relation to semiotic awareness. These were the materials for this method, along with the videos, a laptop for projecting videos to children, recording devices (cell phones and another laptop), projector and projecting screen. Children had pencils to draw lines for Picture-Word Match tests and research assistants also had pens equipped to count words and write and record results, which were later compared with results in recordings.

The videos are the independent variables (and how often the videos are exposed to the learner) and the dependent variable is the learners' literacy levels. The hypothesis therefore is that with

regular exposure to the videos (independent variable), the learners' literacy levels (dependent variable) should improve. Four classes (two Grade 2 classes and two Grade 3 classes) of 70 children in total were selected as discussed in 4.2.1. The small sample size is further indication that this explorative study was embedded within the broader DMDA method. Participants were randomly placed in experimental group and control group according to classes within grades, so of the two classes per grade, one class in Grade 2 became the experimental group and another class in Grade 2 became the control group. This was initially implemented in order to conduct Wilcoxon Signed-Rank tests and Mann-Whitney U tests on smaller groups, as Dwivedi *et al*. (2016) indicate a sample size for nonparametric t-tests can be <16. Nevertheless, to ensure a robust sample, statistical analyses of experimental (N=37) and combined control groups (N = 33) were undertaken, as well as ones for Grade 2 experimental (N = 18) and control (N = 19) and Grade 3 experimental (N= 19) and control (N = 14). It was decided, since both combined and grade separated nonparametric t-tests illustrated the same results, combined results would be reported. This was also due to the Grade 3 control group being two participants under the recommended number for Wilcoxon Signed-Rank tests. This could further allow for parametric t-tests to be used if paired sample data was normally distributed. The separation by grade and groups is still conducted in the presentation of descriptive statistics in Chapter 5. However, the overall average reading rates (mean scores) by grade is discussed in Chapters 5 and 6 at the time of pre-test data collection in March 2022, and again at the post-test collection in July 2022 to compare with the subtitle rates in the videos.

A fifteen-day intervention in the experimental groups was undertaken between test phases where the five videos were screened to Grade 2 and 3 learners of the school on a data projector from a laptop, which also consisted of a brief discussion of the videos (see 4.2.2). The discussions were not recorded, only summarised from observation, and they were restricted to two questions. For each video, research assistants asked learners two questions in isiXhosa orally after the screening: "What was the story in the video?" and "how did the video make you feel?". The purpose was to encourage learner engagement with the videos and not to elicit data for the study, however. Each time a new video was introduced, discussion occurred. Some observations from the discussion from the researcher's perspective are discussed in 4.2.2.

The quantitative methodological process, pre- and post-tests (described in steps 1, 3, 5 and 6 of 4.1.3 and intervention described in step 6 of 4.1.3) allow for research questions 5, 6 and 7 to be answered. The answers to the first seven research questions in turn then assist in answering

research questions 8 and 9. Motivations for the type of data collected are outlined in 4.2. The reasons for ordering the steps of analysis are outlined in 4.3.

4.1.2.1.2 DMDA overview

The video analysis with DMDA (step 2 of research outline) was the primary method and analysis of this thesis. The theory and diagrams related to DMDA have already been discussed in Chapter 3. The choices of videos are outlined in 4.2.1 had to occur early in the research procedures (see steps in 4.1.3), in order to design the ORF word tests and the Picture-Word Match test for the intervention and for the DMDA to be undertaken. The process of video choice or selection was separate to the DMDA, which occurred concurrently to the pre-test phase and intervention as separate methods, as the DMDA did not inform the test design in a sequential manner. The DMDA as analysis and discussion of how it was undertaken is separated to 4.3.2.

As mentioned earlier, this method allows for research questions 1-3 to be answered and the data was triangulated with statistical analysis data and theories for research questions 7-8. These questions relate to how meaning is construed in terms of the systems and sub-systems of the modes of spoken language, written language (subtitles), and the visual mode. All of these are answered by means of State Transition Machines (STMs) that result from the DMDA described in 4.3.2.5, as well as the formula calculations conducted manually for the subtitle rate. Materials used were the videos, a laptop for analysing the data. The *YouTube* videos were downloaded for offline use for this study, which was permitted by PRAESA on behalf of the *YouTube* channel *MPBXhosa*. *Multimodal Analysis Video* was used to annotate the videos and produce the STMs and produce subtitle presentation rate figures to calculate subtitle rate.

4.1.3 Research procedures

The following steps were followed in conducting the research:

1. Sampling occurred as research site and classes were separated into experimental and control groups for two-group experimental design method. This is discussed in 4.2.1.1.

2. Five video clips from the *YouTube* channel *MPBXhosa* have been selected and downloaded. This is discussed in 4.2.1.2.
3. Pre- and post-tests were designed to measure word recognition, Oral Reading Fluency (ORF) and these literacy skills in relation to semiotic awareness. Word recognition and ORF are discussed in 2.3.1 and Semiotic Awareness in 1.2.3 and 3.2. Reasons and motivations for design choices are discussed in greater depth in 4.2.1:
    a. EGRAS adapted to include words from videos selected in an isolated word-level Oral Reading Fluency (ORF) test, titled Word List 1.
    b. Ensured that words in a second ORF word-level test, titled Word List 2, were not in the videos.
    c. A longer ORF connected passage text from the EGRAS was taken without changes, titled Longer Passage Reading tests. The total words and syllables were chosen in order for later counting scores of words and syllables correct per minute (WCPM) for this test and ORF tests described in a-c.
    d. Still images and words from the five videos were used for Picture-Word Match test design. They were drafted according to answers required for research questions (details of this are discussed in 4.2.1.1).
4. System and system categories established in the *Multimodal Analysis Video* software package. Videos uploaded into software. Details of this are discussed in 4.3.2.
5. A Digital Multimodal Discourse Analysis (DMDA) was initiated on the videos, which includes annotating the videos (discussed in 4.3.2). At the initial phase of the DMDA, at the same time, pre-tests in intervention were conducted (see 4.2.3). The specific tests formed from step 2 influenced the system and system categories for the DMDA in 3, rather than the DMDA informing the pre-tests and implementation, and test design. Rather, the two points of interface as discussed earlier at the conceptualisation phase were what informed both test design and DMDA system choices initially. This included the annotation of the videos in the systems outlined in 3.1.
6. A fifteen-day intervention in the experimental groups was undertaken where videos were screened to Grade 2 and 3 learners of the school on a data projector from a laptop, which also consisted of a brief discussion of the videos (see 4.2.2).
7. A post-test identical to the pre-test was conducted, to examine to what extent the students' literacy skills had improved after the exposure to the videos in the intervention, (see 4.2.3).

8. Statistical analysis of test results (descriptive statistics, Shapiro-Wilk tests, parametric t-tests and nonparametric t-tests) were undertaken (see 4.3.1). DMDA was refined to include relevant and incorporate additional systems to further explore findings indicated from the statistical analysis. State Transition Machines were produced and results were analysed, incorporating information from the statistical analysis. This led to the data from the statistical analysis being presented separately in Chapter 5, and the DMDA presented in Chapter 6.
9. DMDA, test data and contextual information were triangulated, (see 4.3.3). Triangulation was included in Chapter 6 in 6.6.

This illustrates that at the method and analysis level, the implementation was concurrent. The quantitative analysis indicated which of the various multimodal systems in the DMDA were the most integral for the context, but the methods initiated at the same time indicated that initially they were separated before the analysis stage. Thus, the statistical analysis informed the DMDA in this manner. As mentioned in the questions above, the video choice informed the pre- and post-test designs, which further informed the initial system choices for the DMDA before the methods were initiated. The system choices were refined in the analysis stage. This illustrates the concurrent nesting/embedding of the methods.

The data, however, was triangulated with the theory after this one-phase approach, see 4.3.3. The next part of this chapter discusses data collection and data analysis and they are separated by the two different methods and analyses. The analysis section in 4.3 further has a sub-section on the triangulation of these analyses and how this was implemented (4.3.3). Since the rest of this chapter is organised by method and analysis, an overview of the research procedures discussed, and their motivations are provided in 4.1.2.1.1. This is separated by method and analysis, namely the two-group experimental design method and statistical analysis (4.1.2.1.1) and DMDA (4.1.2.1.2). Due to the embedded nature of the quantitative study, it is discussed first in the overview in 4.1.3, in the data collection section in 4.2.2 after sampling (4.2.1) is discussed, and in the data analysis discussion in 4.3.1. The DMDA is subsequently discussed in each section (in 4.1.3.2 for overview, 4.2.3 for data collection and 4.3.2). Furthermore, as mentioned earlier in this section, the quantitative study is presented first in the analysis chapters as well in Chapter 5, while the DMDA and triangulation of both analyses with theory is discussed in Chapter 6.

Section 4.1 has provided an overview for the research approach and methodology. This includes elaboration on the rationale for the mixed-method design, and a concurrent embedded mixed-method design model and triangulation. Along with the two methods and analyses, this section described the two group experimental design with statistical analysis and the DMDA, and how they were selected for data and data analysis to complement one another within a critical realist perspective. The research procedures were also outlined. Since the DMDA framework was outlined in Chapter 3, the next section discusses the choice of the videos used in this study (independent variable), as well as the treatment of learner exposure to the videos (independent variable) and the tools of measurement for the literacy scores (dependent variable).

4.2. Data collection for intervention and video choice

This section describes the sampling methods (4.2.1) of the research site, including sampling for experimental and control groups (4.2.1.1) and the choice of videos for both intervention, test design and DMDA (4.2.2.2). The pre- and post-test design choices and the entire process of their data collection from learners, with motivations for choices in research design are discussed (4.2.2). Thus, this entire section is to elaborate on the data collection from the methods outlined in 4.1.2.1 and relate this to the research design (4.1.2) and paradigms (4.1.1).

4.2.1 Sampling

While the priority of this study was the DMDA of these types of videos, this thesis complements this with a two-group experimental research design method from a single research site, rather than collecting data from multiple sites. This is to gain a nuanced, fine-grained understanding of one context in the small-scale intervention which draws on both the granularity of empirical data as well as the richness and depth of the complex qualitative data. When too many contexts are present, there are too many variables that would make convergence of evidence more challenging, which is a reason as to why a particular school and single research site was chosen, rather than a multi-site research project. The following sub-section, 4.1.3.1, thus describes the research site for the small-scale intervention and elaborates on the motivations behind the site choice and the sampling for the two-group design method.

4.2.1.1 Research site

The research site is a school situated in peri-urban area, which is a cluster of informal settlements (Buffalo City Metropolitan Municipality 2020b). Thus, to understand the context of this area, as well as sampling choices and motivations for the two-group experimental method, this sub-section is divided into further sub-sub-sections. Section 4.2.1.1 describes the peri-urban area, Newlands, located in Ward 26, as well as the cluster of informal settlements and the research site's situation in the lower region of Newlands. Section 4.2.1.1.2 describes the research site and its selection. And section 4.2.1.3 describes the sampling of participants from the site and ethical considerations.

4.2.1.1.1 Ward 26 and Newlands

The learners in this study are from a no-fee paying school in one of the informal settlements in the Newlands area. Newlands is in Ward 26 (see Figure 6), a subdivision of the Buffalo City Metropolitan Municipality (BCMM), Eastern Cape (Buffalo City Metropolitan Municipality 2020a).

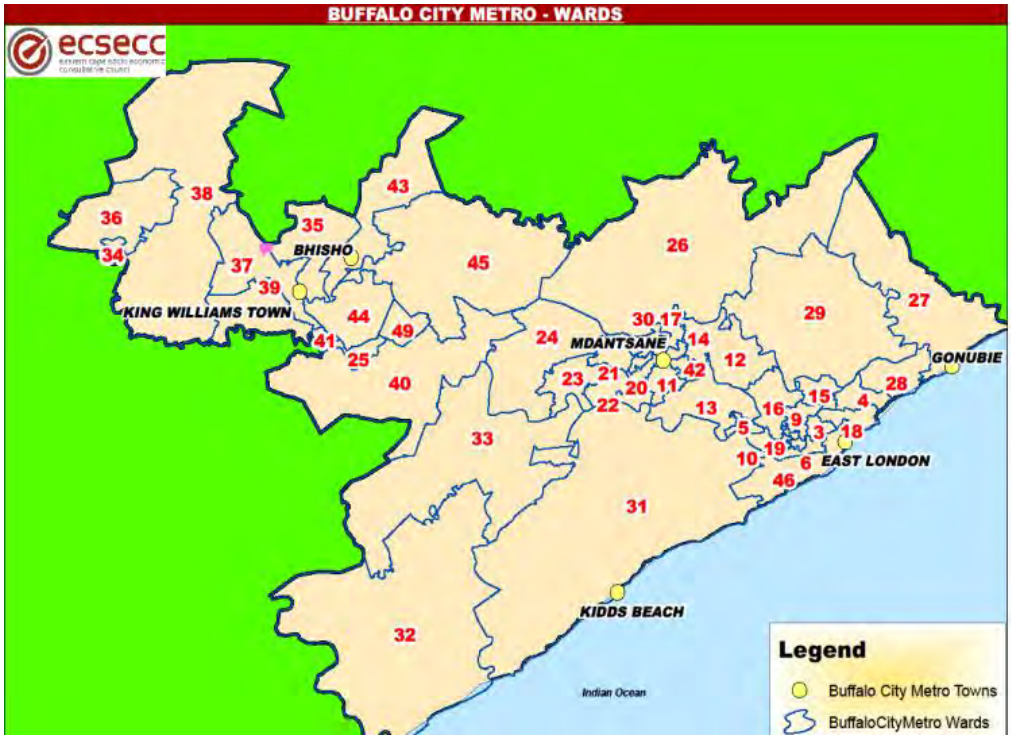


Figure 6 Ward map of Buffalo City Metropolitan Municipality (BCMM) from Buffalo City Metropolitan Municipality (2020a)

Newlands is a peri-urban area with a cluster of 14 informal settlements on the R102 branching out northwards from Mdantsane (Buffalo City Metropolitan Municipality 2020b). These are Eluxolweni, Cuba, Gwigi (Gwiqi), KwaMpundu, St Mary (or St Mary's), Nqonqweni, Zikwaba, Kwetyana, Mzonkeshe, Msobumvu, Ncalukeni, Ekhupumleni, Matanga, and Smiling Valley (Buffalo City Metropolitan Municipality 2020b). Peri-urban areas comprise 20% of the population in in the BCMM (Buffalo City Metropolitan Municipality 2021).

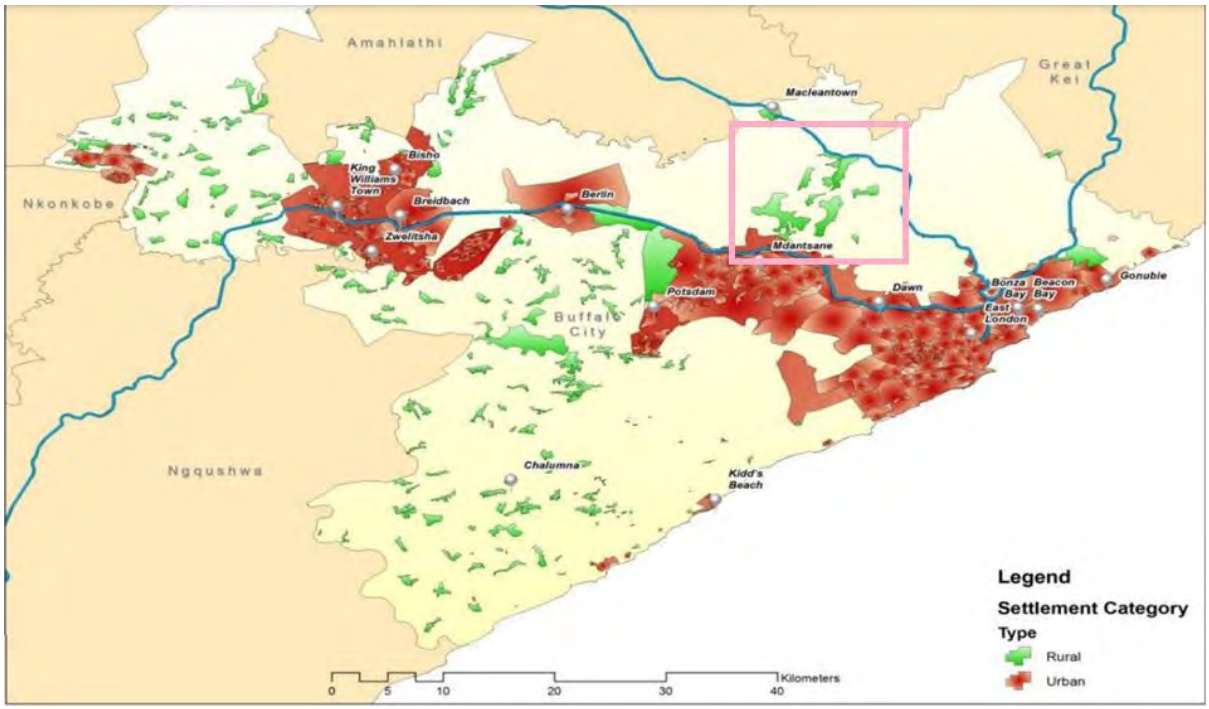


Figure 7 Settlement category within BCMM adapted from Buffalo City Metropolitan Municipality (2021)

Former informal settlements in Newlands area, such as Sandisiwe, perhaps still described by that name by residents, seems to have been integrated into neighbouring informal settlements. However, Sandisiwe was added to Figure 8, as there are statistics that describe the Newlands area from Sandisiwe to Maclean Town (Buffalo City Metropolitan Municipality 2020/2021). Maclean Town is at the uppermost border of Ward 26 (see Figure 7).

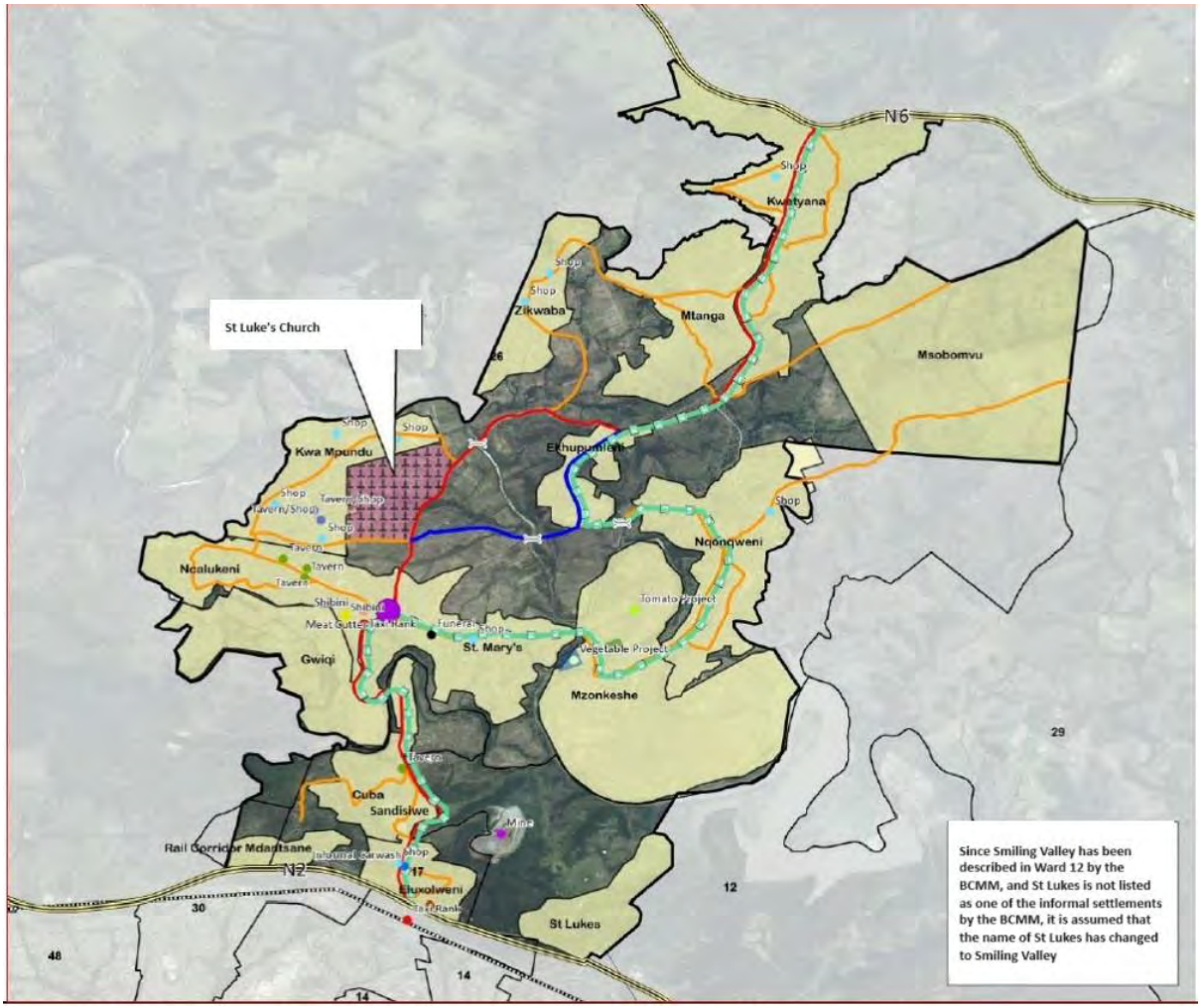


Figure 8 Adapted map of Newlands area from (Buffalo City Metropolitan Municipality 2020b)

The school in question was one of two original missionary schools set up in 1858, which is now public or run by the government. Its location caters for children from the informal settlements in the lower regions of Newlands towards Mdantsane. As one can observe from Figure 8, there is a significant geographical space between the informal settlements near the bottom Newlands boundary, close to Mdantsane, and the informal settlements at the topmost Newlands boundary, which gravitate further towards Maclean Town. The informal settlements in the lower region of Newlands are served by six primary schools, including the one in this study, and two high schools. As such, if children attend secondary schooling, they often have to leave the Newlands area and attend school in Mdantsane, the closest urban settlement. For the sake of anonymity, as one school is located often per informal settlement, the settlements are referred to as the entire Newlands area, unless specifying something specific to the lower region of the Newlands area. This school is relatively small in comparison to many counterparts

of the region, with around 20 children per grade and two classes per grade, as it was initially a missionary school with little geographical space to develop.

Finding accurate representative statistics for the regions and Newlands area have been incredibly difficult, due to the last stratified Census conducted in 2011 that was stratified by places, and not by municipality or province, not corresponding to the entire Newlands area (see Firth 2013; Statistics SA 2011). In this Census, areas were often erroneously demarcated and some areas within the lower regions of Newlands were simply marked with numbers that did not truly correspond with other settlement numbers. Furthermore, the Census was conducted exactly 10 years from the year in which the data was collected from the research site (2021) and a lot would have changed in the region, including the influence of COVID-19 starting from 2020. Consequently, this has resulted in reporting the BCMM statistics for the area, dubbed "The Macleantown, Sandisiwe Sub-metro Region" (Buffalo City Metropolitan Municipality 2020/2021:42), which seems to be from Sandisiwe in Figure 8 to Maclean Town (see Figure 7). It is assumed that, due to Eluxolweni's proximity to Mdantsane, it was excluded, although it could also have been included and just not explicitly mentioned. Nevertheless, one can see from Figure 9 that the population density between Eluxolweni and Ncalukeni, thus the lower region of Newlands, containing 801-1300 people per square kilometer in 2018 (Buffalo City Metropolitan Municipality 2021). This seems to be the maximum, as this decreases to 0-800 people per square kilometer the further one travels to the upper region of Newlands, where we see less of a population density, along with the rest of the ward (Buffalo City Metropolitan Municipality 2021).

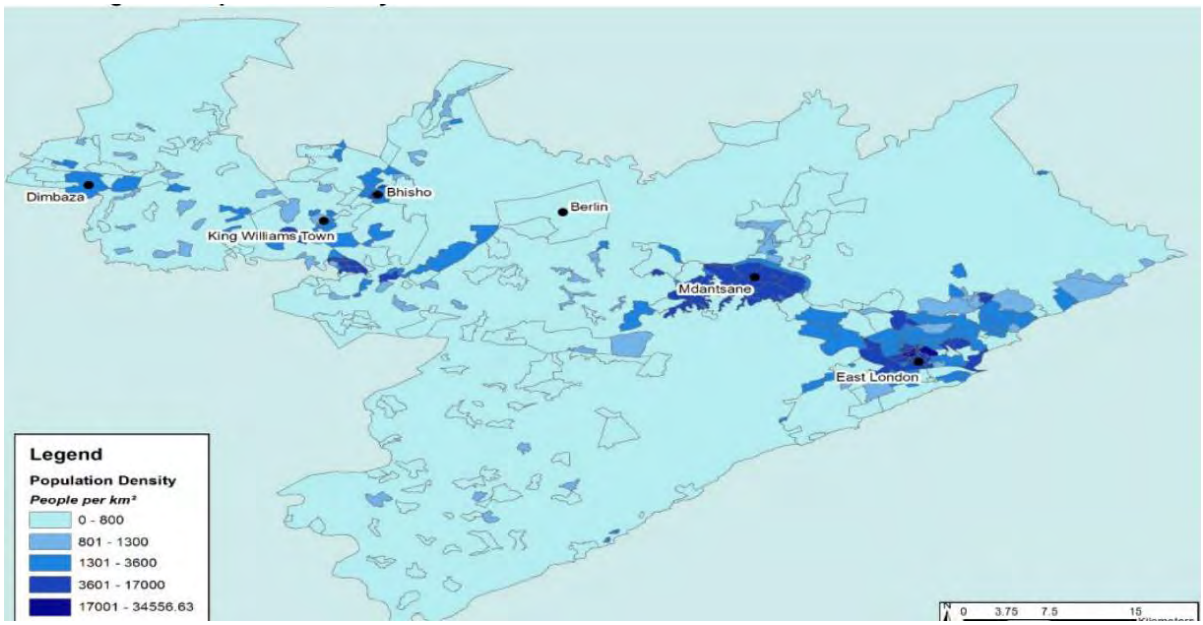


Figure 9 Population density of BCMM from Buffalo City Metropolitan Municipality (2021)

The Macleantown, Sandisiwe Sub-metro Region, which includes Newlands, had a recorded population of 65 100 people, which has increased dramatically since the government census in 2011 (Buffalo City Metropolitan Municipality 2020/2021). The majority demographics in these areas were reported to be isiXhosa L1 speakers in the 2011 Census (Statistics South Africa 2018). While this was not reported on in 2018 by the BCMM, the community survey undertaken in the Eastern Cape overall confirmed that isiXhosa was the predominant L1 for the province (Statistics South Africa 2018). This region has the highest percentage of people living in poverty in the BCMM based on statistics from 2018 (Buffalo City Metropolitan Municipality 2020/2021). Interestingly, this region had the lowest unemployment rates of other regions in the municipality, but one has to take into account that this was before COVID-19, as data was collected in 2018 (Buffalo City Metropolitan Municipality 2020/2021). This region also had the lowest annual total personal income than other regions, with very low economic growth (despite the presence of a quarry in Newlands in Figure 8) and the lowest labour force of the BCMM (Buffalo City Metropolitan Municipality 2020/2021). Most homes in this area are constructed by the government (Buffalo City Metropolitan Municipality 2020b).

According to the BCMM in 2018, functional literacy was at its lowest for this region in comparison to the rest of the municipality, however it was reported as high at just over 85.5% functionally literate readers in the population, with 14.5% of the population in the region reported as illiterate (Buffalo City Metropolitan Municipality 2020/2021). These statistics,

however, do not illustrate how they measure this, nor whether it was in English or the populations home language or even another of the many languages in South Africa. Most of the people living in this region seem to have a high school certificate (National Senior Certificate or Matric Certificate), or Grades 10-11, or Grades 7-9 (Buffalo City Metropolitan Municipality 2020/2021). Not many people in this region have a tertiary education degree or diploma (Buffalo City Metropolitan Municipality 2020/2021). While the municipality has mentioned many development plans for the region, my own observations of the research site and prior research in the region (see Figure 7) indicate its large-scale underdevelopment. A 2020 review of the goals of the last local spatial development framework for Newlands established in 2015 indicated previous goals for the development of Newlands (Buffalo City Municipality 2020b). These goals included full ranges of facilities, maintenance, an introduction of a formal cemetery and off-grid wastewater solutions and energy sources such as solar were not achieved (Buffalo City Municipality 2020b). This region had the lowest amount of flushing toilets, running water and electricity in the BCMM (Buffalo City Metropolitan Municipality 2020/2021). I now describe my own observations of the particular school and research site.

4.2.1.1.2 Research site description and selection

The school runs a feeding scheme, and many children are wearing uniforms donated from an East London partner school and there were issues with access to water observed. While 20 learners a class is an ideal and not an everyday reality for most low-income schools, the school is plagued by the same challenges of lack of resources as many other schools. There is only one laptop that teachers have at their disposal, children are not exposed to computers themselves and the school shares one projector. This is, however, better than most schools in the region. When video resources are projected to children, they are projected in the staff room of the school. Internet access is not present for children in the classroom. Classrooms are side-by-side, but not connected to each other, which requires children to go outside on concrete platforms to enter and exit classes. There are no classroom libraries. Classrooms are rather bare, with some posters on the wall but not many, and the whiteboard is the teacher's primary tool. There is no playground for the children except a stretch of lawn where they play football or run around amongst the cars parked there. Teachers take the role of counsellors, teachers and parents often for the children, as there has been research that has mentioned issues of Gender-Based Violence in Newlands (Jilingisi 2013). Furthermore, with working parents and

parents with little education, even if the functional literacy rate reported was high, children's primary access for reading and learning is at school.

I was introduced to this school from a non-government organisation (NGO) in East London that had collected EGRA results in partners schools in Newlands prior to this study. Teachers had been trained, as well as the research assistants, who had been involved in mentoring learners for literacy skills at this school. This had been conducted in former projects by assisting learners with reading books, but not with any technological resources or subtitled videos prior to this study. After COVID-19, all projects had ceased at the school, resulting in no co-occurring projects during the time of this study. The selection of L1 isiXhosa research assistants from this NGO is described in 4.2.2 in greater detail, but it was to ensure they were familiar to learners and resulted in less observational bias, as well as to ensure communication with learners were done in isiXhosa. It was further to ensure accuracies in the first phase of data collection, which I as the researcher then confirmed.

The school was chosen due to its location outside of the area in which I as researcher live, as areas around the university that I attend tend to be overly saturated in regard to literacy research. This would have compromised understanding this intervention's efficacy on learner literacy levels in comparison to other literacy interventions that may co-occur. As established from describing the context of Newlands, very little has been written and researched about this context. In addition, the context of this school represents many schools in peri-urban areas of the country, and a type of school that was predominant in the PIRLS (2006) results, which further promotes the investigation of this school. The smaller sample size, despite its limitations on statistical analyses and less representative of the overcrowding of classrooms in South Africa, was selected to ensure learners engaged with the video and removed a potential independent variable that could have resulted in ambiguity of results. Thus, the type of sampling conducted was purposively sampling, however the learners were given an opportunity to participate voluntarily and with parental permission, and thus, within the purposive sampling was also self-selected sampling. The next section describes participant sample size, the ethics and how this self-selected sampling occurred, as well as the arrangement of learners into experimental and control groups.

4.2.1.1.3 Participant sampling and ethics

As discussed in 4.1.2, the sample size consisted of 70 learners in Grade 2 and 3 that consented to participate in the research, in order for me to gain a representative sample of the grades for the study. Grade 2 was selected as I wished to observe literacy scores after a full school year (Grade 1) and prior research described in 1.1.1 and Chapter 2 indicates that Grade 1 learners can often be in pre-literacy phases (Schaefer and Kotze 2019). Grade 3 was also included due to the small sample size of Grade 2 participants in this school. Since the Grade 2 and 3s in this school had no prior exposure to SLS videos and no co-occurring literacy interventions or programmes were being implemented during the collection of this data, they were selected. This was to ensure no other programmes had any influence on test scores and this current intervention was the only one conducted with these learners.

I, as the researcher, was guided by research ethical principles throughout this thesis, as we were working with children, a vulnerable research group. First, ethical clearance from the university was requested and obtained from the university's ethics committee. Permission from the Eastern Cape Provincial Department of Education had to be obtained, as well as permission from the principal. These were provided in the form of letters from both university ethics committee and stakeholders.

Permission was also granted by learners' parents. Consent forms that were in English and isiXhosa were given to children in each of the four classes prior to the research, for parents to sign. This is based on the understanding that children present limitations when making decisions (Caldeira *et al.* 2021). However, sole permission from parents limits the view of children's capacity of making decisions (Caldeira *et al.* 2021). To ensure children agree with participating in research and the intervention within these limitations of decision-making capacity, informed assent was requested from learners.' One cannot expect learners still learning to read to read and understand an assent form, thus this was obtained and negotiated between the learners and the research assistants in spoken language during the sessions. If the learners needed to leave at any time during the session for any reason and experienced any risks such as fear of embarrassment, boredom, or tiredness, and consequently wished to opt out completely, they were given the opportunity to do so. This was explained to them as each task was described in their home language by research assistants. Data collected was kept confidential and only examined by myself and research supervisors. Data was also anonymised,

in terms of school as described in 4.2.1.1 and in terms of participants, which were coded from names to participant numbers once their data was collected.

In terms of separating participants into experimental and control group, all of the videos were shown per class rather than separating children to watch videos and leave other class members out of the activity, thus the experimental and groups were purposively sampled from each class per grade. However, there was self-sampling within this, as children's participation in the study was voluntary. All the children provided written consent and spoken assent, and this resulted in all of the children in Grades 2 and 3 at this school participating, resulting in a representative sample. The combined experimental (N=37) and control groups (N = 33) were 70 participants in total. By Grade, the two groups were Grade 2 experimental (N = 18) and control (N = 19) and Grade 3 experimental (N= 19) and control (N = 14). The next section of the data collection describes the selection of the videos used in this research.

4.2.1.2 Video choice

Two channels on *YouTube* contain SLS isiXhosa videos that are in the public domain: *Xhosa Kids* and *MPBXhosa*. While the channel *Xhosa Kids* contains isiXhosa songs, the videos come from English nursery rhymes and songs and were just dubbed and translated over to isiXhosa, with many spelling errors, for example, "Igqirha lendlela ngu qongqotwane" instead of "IGqirha lendlela nguqongqothwane" (the diviner of the road is the beetle) (Xhosa Kids 2019; Simelane-Kalumba 2014:43). This was the reason why videos from this channel could not be used in this study. The channel selected for this study, *MPBXhosa*, has translated stories from English to isiXhosa, but the isiXhosa translations are mostly linguistically accurate. The videos in this study were designed from a collaboration between Moving Picture Books and the Project for the Study of Alternative Education in South Africa (PRAESA), an NGO based in Cape Town. The reason *YouTube* was chosen as the platform is that they are readily available on the public domain and easily accessible, provided one has a data connection.

The videos in this study are animated fictional stories in isiXhosa which are also subtitled in isiXhosa. The primary selection criterion was that the video was to be under 5 minutes in duration. There is evidence that children may have the ability to maintain focus in their attention if classroom activities have a short duration (Godwin *et al*. 2016). Furthermore, studies have shown that participants can concentrate on videos for longer and comprehend

more information without intense cognitive load when videos have a shorter duration (Guo *et al*. 2014; Slemmons *et al*. 2018). Since sessions of video-watching would be repeated to ensure the length of exposure was not too short, I selected videos that were shorter, that could be repeated twice a session, thus ensuring shorter times of engagement, but exposure repeated at regular intervals during the intervention.

The second selection criterion was that the video has to have minimal presence of English in the visuals of the videos, or words that were left untranslated. Some of the eighteen videos on the channel were not eligible for use in the study as they contained too many English words such as *On Top of Spaghetti*, where written English words such as a menu and an entire oral song in English appears.

I then decided to exclude videos for specific reasons that were related to content that was video-specific. I excluded songs such as *Eency Weency Spider* (which also had no audio and was excluded on that basis) and *Eight Silly Monkeys*, in case the children had been exposed to them before and this resulted in another variable introduced that would cause ambiguity in results. This was the same for *I Know An Old Lady*, although this video further led me to exclude potentially triggering topics like death and monsters in *I See A Monster.* A child was presented as sad and lonely, if she was a grown up, with no parents to kiss her goodnight, in *When I Grow Up*, and I excluded the video in case it was potentially triggering for any children who may not be raised by their parents. I also avoided *Who Stole the Cookie* to avoid introducing exposure to behaviours of stealing. I excluded *Bible Stories* as I did not want to impose religious beliefs on the learners. This left me with the five videos that I selected, which appeared to be on relatively neutral topics and examined a wide variety of topics, vocabulary and trends.

I selected videos with a duration of no more than 5 minutes each from the *YouTube* channel *MPBXhosa* based on topics and vocabulary that are age and grade appropriate for a Grade 2 learner. The choice of five videos further mirrors Kothari *et al.* (2002)'s research design in that they selected five Hindi videos for their analysis. The further had to gauge whether vocabulary would be fit for a Foundation Phase level, and this was informed by the word lists and texts within the EGRAs designed for that level, as well as the CAPS documents for the Foundation Phase.

The procedure for video analysis in terms of DMDA and the statistical analysis of the pre- and post-tests are discussed in the following section.

4.2.2 Test design and dependent variable measurement

As previously described in 4.1.2.1.1, tests were designed and completed by learners in order to measure the literacy learners' skills as dependent variables to the intervention (independent variables) within the two-group experimental design method. These were conducted before and after the intervention, and were the same test, however, to specify exactly which scores I refer to, I separate them as pre-test and post-test, referring to the test conducted before the intervention and after the intervention.

The pre- and post-tests were designed in step 3 of research procedures (see 4.1.3) after sampling and video choice. They occurred before the DMDA in order for the analyst to not demonstrate any bias when designing the tests. Quantitative written tests were conducted with learners instead of qualitative methods, due to the two-group experimental method potentially eliciting a more suitable indicator of literacy influence for the purpose of this thesis and research questions than qualitative interviews. As illustrated in 4.1.3, they were conducted in steps 3 and 7 of research procedure. Pre-test data collection occurred in the beginning of March 2022, and post-test collection occurred at the end of July 2022. The tests used for pre-tests and post-tests comprise three components: (1) to gauge word recognition in relation to semiotic awareness skills and multimodal literacy skills, (2) Oral Reading Fluency of isolated words (lexemes, see 1.2.1.2) for word recognition and automaticity in decoding, and (3) the Oral Reading Fluency rate and automaticity of decoding words and syllables (orthographic words, see 1.2.1.2) within a Long Passage text. These components are labelled Section A and B of Appendix B.

Section A is the comprehension component. It consists of a Picture-Word Match test, which was given to Grade 2 and 3 learners to complete simultaneously and individually by research assistants. An example is illustrated in Figure 10.

Figure 10 Picture-Word Match test

Children can recognise mental representations of an object early in their development and will examine pictures when words refer to them and during activities involving stories (Schmithorst *et al.* 2007). Variants of Picture-Word Match tests have been used for Word Recognition and Fluency in Finnish, another agglutinative language (Lepola *et al.* 2005; Psyridou *et al.* 2020). Along with these skills, this test is further a test for multimodal literacy as multiple types of processing and comprehension are required for the various modes in question. These modes include the linguistic, particularly the linguistic within the visual mode as written language, and the visual image. These modes straddle the phonological, visual object and semantic domains (see Huettig and McQueen 2007 on the phonological, visual and semantic hypotheses). The visual aspect of processing and comprehension further relates to the reading of multimodal texts with the binding of visual features (identified as words or objects) or location indexing in the multimodal reading model outlined by Liao *et al*. (2021).

Thus, while this test technically examines word recognition, it does not do it solely at a level of decoding, but rather this test examines word recognition in relation to visual recognition, and semantic awareness of both the written word, the image, and a connection between them. Thus, this entire system of multimodal literacy that this test examines is semiotic awareness. As mentioned in 1.2.3, semiotic awareness is a relationship between semiotic resources

between two or more modes and the ability to connect meanings conveyed between multiple modes (Abrams 2016) This is a simple aspect of semiotic awareness. A more advanced awareness of semiotic systems would include understanding how meaning affordances of modes construct potentials for meaning creation (Lim and Toh 2020). To examine word recognition and word fluency in relation to a multimodal context and investigate semiotic awareness, a Picture-Word Match test was included in the battery of pre- and post-tests.

The Picture-Word Match test was adapted from the EGRAs and a Picture-Word Match test designed for Sesotho by Sefotho (2019). This test design was exploratory, as no similar tests exist for this isiXhosa. The Sesotho test in the study conducted by Sefotho (2019) consisted of 10 words with four pictures per word in order to measure semantic awareness. As discussed, written words and picture match tasks investigate and encompass more skills than just semantic awareness for the linguistic word. Thus, the Picture-Word Match test in this thesis for isiXhosa is focused on, rather than an understanding of the meaning of the word alone, the connection of a word meaning to the meaning of an image that is present in the videos. This activity has words from the video on the left-hand-side in a column, and a single corresponding image for each word taken from the videos on the right-hand side. Images and words do not correspond horizontally. The learner needed to draw a line from the word to the corresponding image or vice versa. The words selected for this test were taken from Word List 1 and tried to include a mixture of root verbs and nouns. The simplification of the word to bi-syllabic verb root and non-prefixed forms and their corresponding images from the five videos were to ensure that learners could perform without reaching floor effects. While Sefotho (2019) used simple words and phrases for her Sesotho tests, the majority of her participants still presented a low mean score of either 3 or 4. Each video was provided two words and two images, in order to represent words and images from all five videos in the test. Learners had 20 minutes to complete the test at their own pace, but with time constraints in place.

Section B is the decoding component and comprises of two separate lists of 50 words each, that were either bi-syllabic or tri-syllabic (see example in Figure 11). The learner had to read as many words as possible aloud in one minute correctly for each word list. The first list in this thesis is titled Word List 1 and the second list is titled Word List 2 for ease of discussion. Word List 1 has words from the *YouTube* videos (the videos screened in the intervention) and Word List 2 has words that are not in the videos but are within the original EGRA. They were based on the Early Grade Reading Assessment (EGRA), which is an international assessment for

reading and literacy measurement (Spaull 2020). The isiXhosa version is what has led to the isiXhosa average benchmark WCPM reading rates at 20 WCPM for Grade 2 and 35 WCPM for Grade 3 (Ardington *et al.* 2020). This is the reason these tests were chosen to measure learners' literacy skills in this study. These tests are based on the Familiar Word Reading and Oral Reading Fluency (titled Long Passage Reading, as these tests are both tests for the Oral Reading Fluency skill) sections of the Early Grade Reading Assessments (EGRAs) (Spaull 2020). A research assistant explained the tests within this section to a learner in isiXhosa.

mama	tata	atsho	cula	kuba
kati	thetha	senza	funa	cinci
bona	apha	hlala	jonga	atsho
ndiya	thanda	inja	hamba	kwethu
siza	kodwa	khaya	engwe	cinga
hayi	vala	intethe	tsala	ibali
wabona	izinto	imini	umhlobo	fihla
zuba	phezu	kakhulu	sila	phumla
nantso	kusasa	isele	fika	sithi
thula	cwaka	naxa	isikolo	funda

Figure 11 Word List 1 adapted from EGRAs

As the learner reads aloud, the research assistant made notes on how many words read accurately and took an audio recording of their reading for me to be able to confirm syllable and word count for each learner. The word and syllable count were noted because isiXhosa is an agglutinative language and, as mentioned in 2.3.2, the grain size in isiXhosa is at the level of the syllable (Probert 2019). Furthermore, the study by Kothari *et al*. (2002) in India, on which this study was based, also took note of the syllable and word count with a similar test, albeit the number of words was greater. I limited the word count to the original EGRA word count, but since the Long Passage text had 53 words, and not 60 words like the Familiar Word Reading, I used 50 words in the Familiar Word Reading section, to better compare the differences between shorter lexical unit reading rates and longer word reading rates. Furthermore, the benchmarks at 20 WCPM and 35 WCPM were less than the 50 WCPM, and

the assessment was mainly to see whether they were reading at rates appropriate for their grades, as well as whether they read Word List 1 faster than Word List 2.

The limitation of Word List 2 is that, while "sela" means drink and was not in the videos, the syllable cluster "sela" existed in the word "balisela" in the videos, which means to tell. There is one other word like this in Word List 2, in which "zisa" (bring) is within the word "phanyazisa," in which the "-is-" in "zisa" is a morpheme reflecting ditransitivity in the verb "phanyaza" (to shine). There were only two instances like this in Word List 2. However, since their lexical meaning as isolated words were not present in the videos, they were kept in Word List 2, while the rest of the words were specifically chosen to avoid syllable clusters found in the videos that resembled the isolated words in Word List 2.

Due to issues of lexical words and syllabic phonological reading in isiXhosa and smaller lexical words not naturally existing in isolation, I also conducted the Long Passage Reading test (see Figure 12).

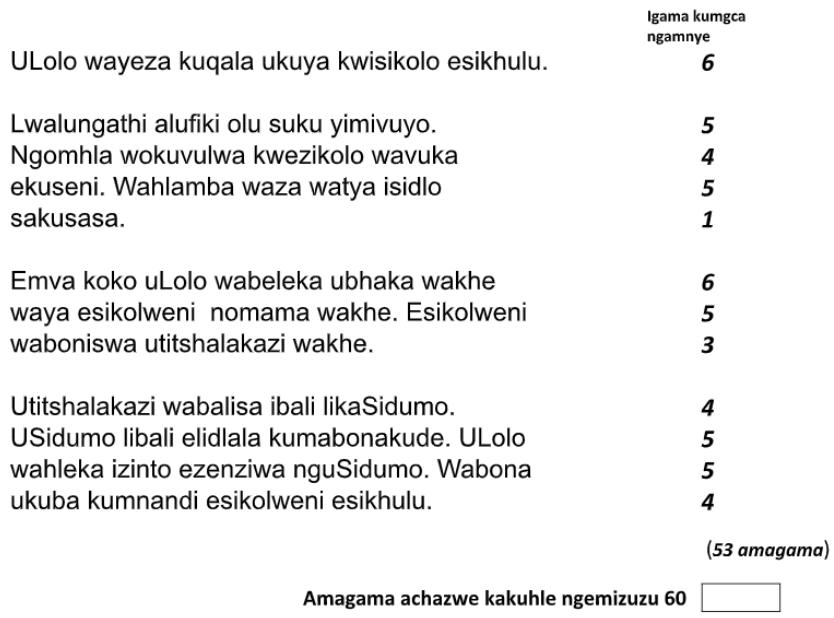


Figure 12 Longer Passage Reading from EGRAs

The learner had to read as many words as possible aloud in one minute correctly for the passage. As the learner reads aloud, the research assistant took an audio recording of their reading for

me to be able to confirm syllable and word count for each learner and made notes on word count. This test is from the EGRA and was left unchanged. However, the test has verb forms in conjoint (short) form and the language in the videos also presented disjoint (long) form (see 1.2.1.1). The presence of potentially shorter words in the test than in the videos, which may be less morphologically complex, was considered. Test scores would reflect a slightly faster reading rate for learners than their reading scores for the subtitles. The word and syllable count were noted because isiXhosa is an agglutinative language and, as mentioned in 2.3.2, the grain size in isiXhosa is at the level of the syllable (Probert 2019). This test does include a comprehension test, but since none of the learners could read the complete text in one minute, and due to the content not corresponding to the videos, this was excluded. A comprehension test in this study would have had too many variables to consider – the speech rate and visuals could have allowed learners to comprehend the videos and not shown whether the subtitles alone allowed for better comprehension. They could not read the entire text within a minute, thus comprehension tests were not undertaken. Further research in this area can examine the comprehension components of ORF tests in this context.

Oral Reading Fluency acts as a better indicator of reading comprehension than silent reading fluency (Fuchs *et al.* 2001, see 2.1.1). This is due to silent reading rates relying on the learner to self-report this information (Fuchs *et al.* 2001, see 2.1.1), which may not be objectively accurate (Probert 2016). Thus, despite subtitles usually being read silently, ORF tests to provide a more accurate estimate of reading rates are used in this thesis. It further seeks to understand whether learners phonologically decode accurately when reading. In addition, eye-tracking methods do not necessarily confirm whether decoding and comprehension in silent reading occurred, even if it tracks eye movement along the subtitle. Eye-tracking software was not seen as the best research tool for this investigation and purpose of this study, as this software, aside from the finances involved, is incredibly difficult to use in a peri-urban context in terms of mobility and electronic access. Thus, there is no objective method to accurately test for a silent reading rate.

Furthermore, section A, the Picture-Word Match test, could be conducted in experimental and control groups, while section B was conducted individually, which allowed for the simplest and most efficient way of collecting the data. The completion of section A was closely monitored by research assistants to minimise learners influencing one another in their answers,

as they were instructed to complete the tests on their own. The questions asked of learners at the beginning of this section were conducted orally in groups for learners to feel more comfortable to share opinions on videos in comparison to them being on their own in the presence of research assistants.

4.2.3 Video intervention

I took the roles of researcher and analyst, while two research assistants from Newlands, the area in which the learners are situated, were employed to collect the data from the learners. This was to limit the observer's paradox as much as possible because both research assistants coming from Newlands and isiXhosa L1 speakers would be more relatable to learners than I, who am an English L1 speaker and not a local resident of Newlands. A training session for research assistants was conducted and a meeting with the principal and teachers was held to explain the study prior to the intervention. The next section discusses the context of Newlands and the school in more detail.

Once the above pre-tests had been administered to learners, research assistants then screened videos in the intervention to the Grade 2 learners in the experimental group and control group (step 4 of research design), on a data projector from a laptop once a week in a 15-session intervention that spanned from the end of March to the end of July. Video screening regularity is based on the study conducted by Kothari and Bandyopadhyay (2014) in India, as while the study conducted by Kothari *et al*. (2002) conducted sessions three times a week for three months, their later longitudinal study illustrated that once a week viewing was sufficient. The control groups were not exposed to the videos during the intervention, rather the videos were exposed to the experimental groups. After the test was conducted with both groups post-intervention, videos were shown to the control groups to ensure that the control groups received the same benefits as the experimental groups. Research assistants screened the videos twice, no longer than five minutes, once a week. The intervention thus ran for a maximum of 10 minutes in one day a week for 15 sessions in total. Each of the five videos was screened in three sessions of the 15 sessions.

For each video, research assistants asked learners two questions in isiXhosa orally after the screening: “What was the story in the video?” and “How did the video make you feel?” For every session, a short discussion occurred. The answers were not recorded, as the purpose was

to encourage learner engagement with the videos and not to elicit data for the study. However general observations during the intervention are described. The Grade 2 experimental group were more active in reflecting on how they felt from the videos than the Grade 3 experimental group. They demonstrated happiness and excitement upon watching the videos and this was present throughout the intervention. In the beginning phases of the intervention, the Grade 2 experimental group were more willing to answer discussion questions than the Grade 3 experimental group. They also had more nuanced answers than the Grade 3 experimental group, noticing more complex details about the videos' content than the Grade 3 experimental group. This pattern was consistent, and by the end of intervention the learners illustrated more nuanced responses when discussing the videos' stories in comparison to the beginning of the intervention, as well as their personal preferences of the videos and reasons for their preferences.

After the last session of screening, the post-test phase was conducted in July 2022. After the post-test phase, the control groups were exposed to the videos for the same duration as the experimental group was during the intervention, to ensure that the children in those groups could still enjoy and watch the videos that the children in the experimental group had watched.

4.3 Data analysis

The types of analysis employed in this thesis were the t-test statistical analysis of pre- and post-intervention tests (4.3.1) and a DMDA of the videos (4.3.2).

4.3.1 Statistical analysis of test scores

Data from the experimental and control groups underwent further statistical analysis, which included the use of t-tests. T-tests compare the mean scores (Kim 2015) of two groups. The t-tests used in this statistical analysis were on one sample (the learners at the research site) split into experimental and control group by grade. The first groups were exposed to the intervention and the second groups were not exposed to the videos until after the post-test. Parametric and nonparametric are the two types of t-tests. For parametric t-tests to be used, the assumption of reasonable sample size should be followed i.e., greater than 30 learners (Browner *et al.* 2007). Furthermore, other assumptions required to use parametric t-tests include normality, equal

variance and independence (Kim 2015). A Shapiro-Wilk test was conducted to validate the condition of normality. If the assumption of normality is not met, nonparametric t-test methods have to be used, which include the Wilcoxon Signed-Rank test for paired samples and the Mann-Whitney U test for independent samples (Kim 2015). Paired sample t-tests (parametric or nonparametric) are used when the compared groups are dependent on the other. Independent sample t-tests (for both parametric and nonparametric data) are conducted when the compared groups are independent of one another.

All descriptives and t-tests, except for Wilcoxon Signed-Rank tests, were conducted using Jamovi. There was an error with Jamovi running birank serial correlation on the data for Wilcoxon Signed-Rank tests, as it was providing negative mean differences and effect sizes when they should have been positive. Furthermore, Jamovi does not allow one to examine how the figures were calculated and what of the three formulae for Wilcoxon Signed-Rank tests were used (see Tomczak and Tomczak 2014). This was confirmed when conducting the Wilcoxon Signed-Rank tests using Statistics Kingdom (2017).

The EGRA-based test and Picture-Word Match scores underwent the Shapiro-Wilk test to determine whether there was normal distribution. The overall data split between grades and experimental and control groups illustrated a nonnormal distribution through a Shapiro-Wilk test, recommended for small sample sizes (Mishra *et al*. 2019). However, initial descriptives of this data split by grade and group (discussed in 5.1) is included to calculate the average mean scores per grade, in order to compare this to the subtitle rate within the videos. The t-tests, however, were conducted on Grades 2 and 3 experimental groups combined and Grade 2 and 3 control groups combined, in order to achieve sample sizes greater than 30 (Browner *et al.* 2007). For the Wilcoxon Signed-Rank tests, sample sizes could be smaller and greater than 16 (as Dwivedi *et al.* 2016), thus Wilcoxon Signed-Rank tests were run separately by grade. However, their findings showed the same as combined group test scores, therefore, it was decided that it was not necessary to include both.

For the independent sample data, the data was non-normally distributed, thus Mann-Whitney U tests were conducted. For the paired sample data of the experimental group combined grades, Word List 1 (WCPM and SCPM) was normally distributed thus parametric t-tests were used. However, in this data, Word List 2 (WCPM and SCPM), Long Passage (WCPM and SCPM) and Picture-Word Match tests were non-normally distributed, thus a nonparametric t-test was

used (Wilcoxon Signed-Rank test). For the paired sample data of the control group combined grades, Word List 1 (WCPM and SCPM) and Word List 2 (WCPM and SCPM) was normally distributed, thus parametric t-tests were used. However, in this data, Long Passage and Picture-Word Match tests were non-normally distributed, thus nonparametric t-tests were used.

These tests compare the means or medians of two independent, nonnormal distributions (Fagerland and Sandvik 2009), and thus can illustrate whether mean literacy scores improved after video treatment or not.

A Mann-Whitney U test was conducted between experimental and control group scores to determine whether there was a significant difference between groups. Furthermore, Mann-Whitney U tests and Wilcoxon Signed-Rank tests were also run between the Word Lists 1 and 2[21] to ensure there was no significant difference in scores between tests before intervention in the first testing phase. Since a significant difference was established, the data were not transformed and the categorisation between grades has been maintained in this study. In a similar three-group study by Kothari *et al*. (2002), the *p* value was set at 0.05 for 46 children per group. Thus, the *p* value for significance has been set to 0.05.

This chapter examined the entire research design and methodology, including motivations for video, sample and experimental design selection, as well as choices made on pre- and post-test and intervention design. This section discussed how the videos were analysed with DMDA and *Multimodal Discourse Analysis* and how the data was analysed using a variety of tests for statistical analysis.

4.3.2 DMDA

The systems of DMDA discussed in 3.1.1 were created in the software *Multimodal Analysis Video*, which was used to annotate the five videos and produce State Transition Machines (STMs) for the research (see Figure 14). I labelled each of the five videos, which were *The Icky Sticky Frog* (Video 1), *They All Loved Sue* (Video 2), *The Moon Followed Me Home* (Video 3), *Jungle Tumble* (Video 4) and *I Can Go To School* (Video 5). This was the order they were shown to the learners. I annotated Video 1-5 sequentially. I further annotated each

[21] Word List with words from the video and the Word List with ones not in the video

video consistently, first for language from the ideational metafunctional systems, then the interpersonal metafunctional systems and finally the textual metafunction. I then analysed the visual mode in the same way, by examining sub-systems of the ideational, interpersonal and textual metafunctions sequentially. I had to analyse each mode individually to gain a better understanding of how they were similar or contrasted in meanings for Intersemiotic Texture. After this was completed, I analysed the Intersemiotic Texture for Comparative Relations and calculated the subtitle rate. Instances of other types of meaning relations were noted, not annotated, as the purpose of this thesis is to focus on the Comparative Relations and Similarity and Contrast of meanings. Figure 13 DMDA catalogs and represents the catalogs and systems used in the software *Multimodal Analysis Video*.

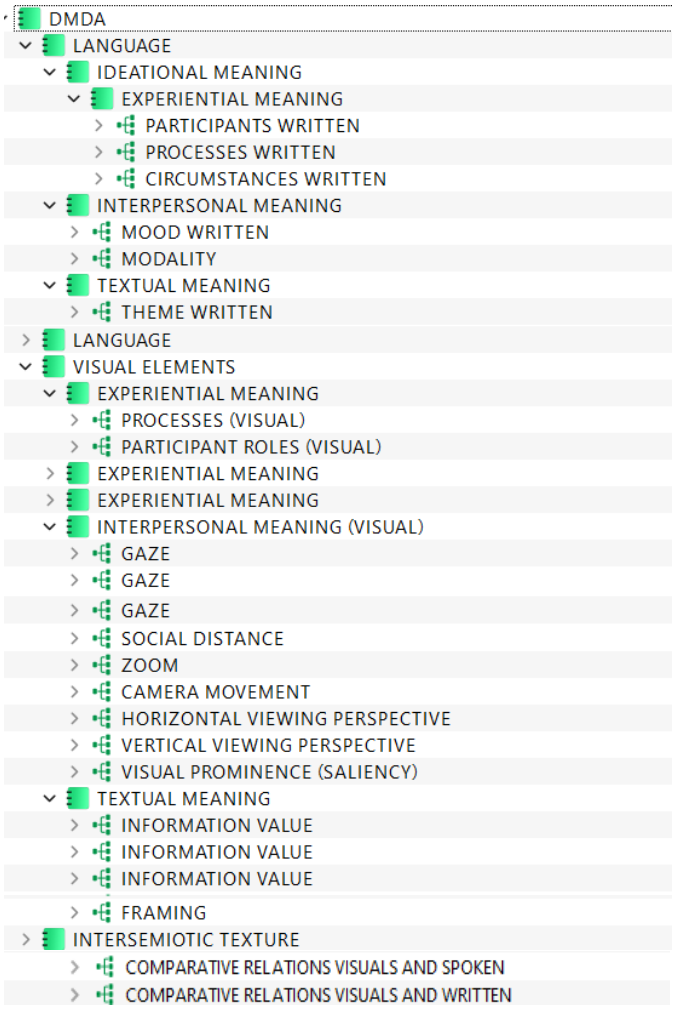


Figure 13 DMDA catalogs and systems

According to the Library in *Multimodal Analysis Video* that I used, which is represented in Figure 13, one can see there is a repetition of 3 EXPERIENTIAL MEANING catalogs. These

catalogs contain repeated systems of Processes (Visual) and Participant Roles (Visual). There is further repeated systems of Gaze (three of them) and INFORMATION VALUE. This is due to software errors, as when one attempts to code, for example the presence of a particular visual Participant in one line as overlapping co-occurrences, the software does not recognise those annotations. Thus, they had to be separated with distinctive labels for the researcher to distinguish them from each, see example in Figure 14.

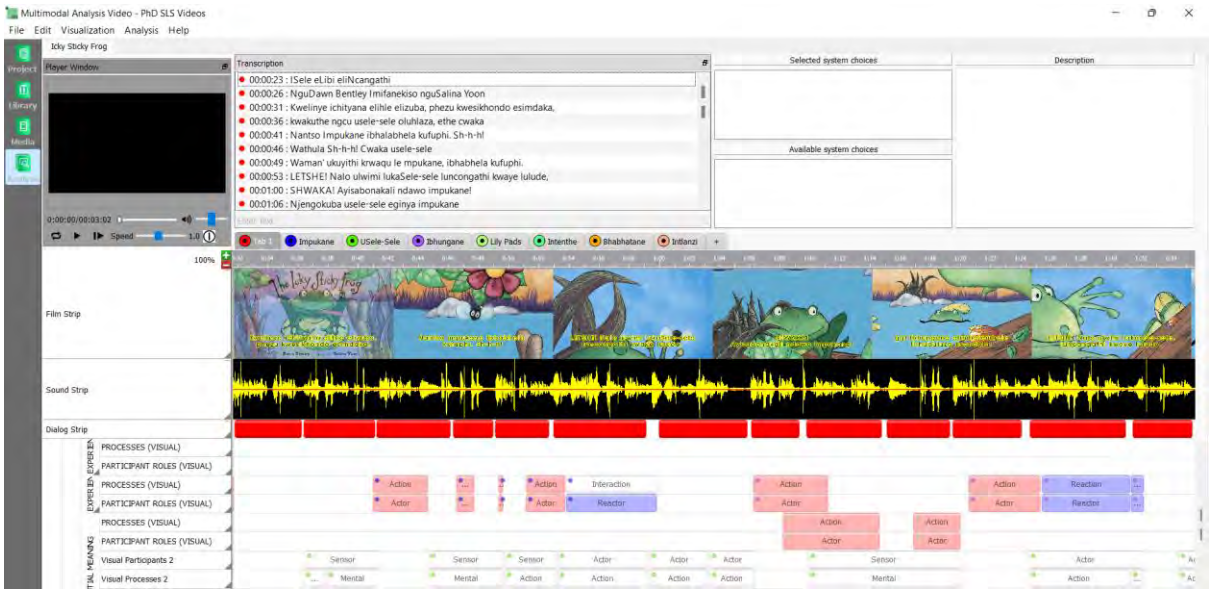


Figure 14 Example annotations in Multimodal Analysis Video

In order to ensure results were still accurate, certain Participants were coded in one of three systems within the catalogs, and others that potentially co-occurred were coded in another of the three systems. For example, in Figure 14, the lily pads as Visual Actors and their actions of moving across the water co-occur with the visual Actor and Material Processes of the beetle (*ibhungane)* and the frog (*USele-sele*). The first set of systems, the visual beetle and insects and their Processes that did not co-occur were coded, in the second set of systems, the visual lily pads and their Processes were coded and in the third set of systems, the visual frog and its Processes were coded. This was to ensure accuracy of the state transition diagrams while avoiding the error from the software.

System choices for language (4.3.2.1), visuals (4.3.2.2), Intersemiotic Texture (4.3.2.3) and subtitle rate (4.3.2.4) are discussed in more detail in the following sections, along with state transition machine diagrams (4.3.2.5).

4.3.2.1 Language

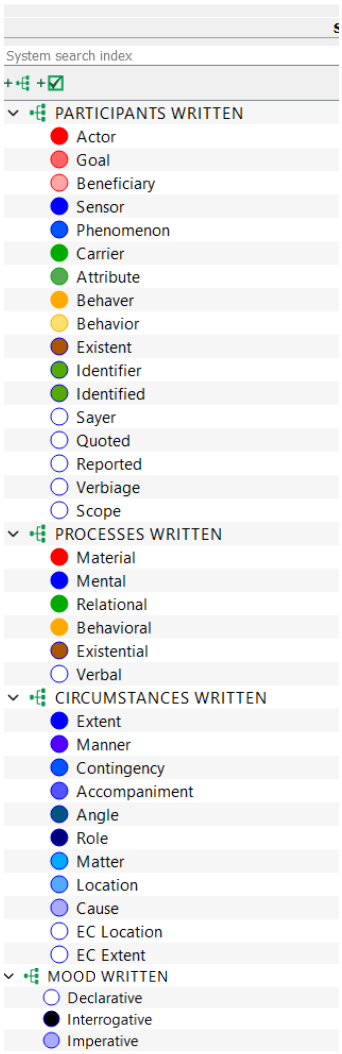


Figure 15 Systems and their system choices for language in all three metafunctions

This was annotated as language unfolded and represented the duration and frequency of system choices spoken language. For subtitles, they stayed on the screen for a specific amount of time. Thus, instances of language that occurred as it unfolded in the spoken language were coded together within the duration of each subtitle, as it is more realistic of how long every system remains on the screen. It also results in the percentages of systems in the STMS (described in 4.3.2.5) being higher for written language than spoken language (see Chapter 6). The details of these choices within the systems in Figure 15 for language (used for both spoken and written) are described in 3.1.1.

I annotated prefixes as Participants as they were present in the meaning in the verb complex, even though they were not present as individual noun Participants. Even though they are often agreement markers, they also have pronominal properties and consequently were annotated as if they were Participants that stood on their own with the Process, in which they were involved. For example, in the sentence in Video 2 "**U**Sue **wa**ye**zi**thanda **izi**lwanyana" (Sue loved animals), the "wa-" subject marker refers to Sue and the "-zi-" refers to the direct object, which is animals.

The complexity of embedded constructions, which often include a lot of adjectivised and nominalised Participants, Processes and Circumstances, reflect Embedded Circumstances of Location and often Circumstances of Extent. I chose to code components in the Embedded Circumstance separately, as well as the entire clause as EC Circumstance. This explains the additional choices of EC Circumstances and EC Extent under the system CIRCUMSTANCES WRITTEN in Figure 15. For example, in Video 3, "Siye sithi naxa sihambela kude okanye kufutshane nekhaya" (Even when we are walking far or close to home), is an embedded clause Circumstance of Location: Time with the use of the conjunction even when. Choices, which I made in coding were checked with my supervisors, and particular nuances in meanings and confirmation of translations were checked with two different L1 isiXhosa speakers and linguists.

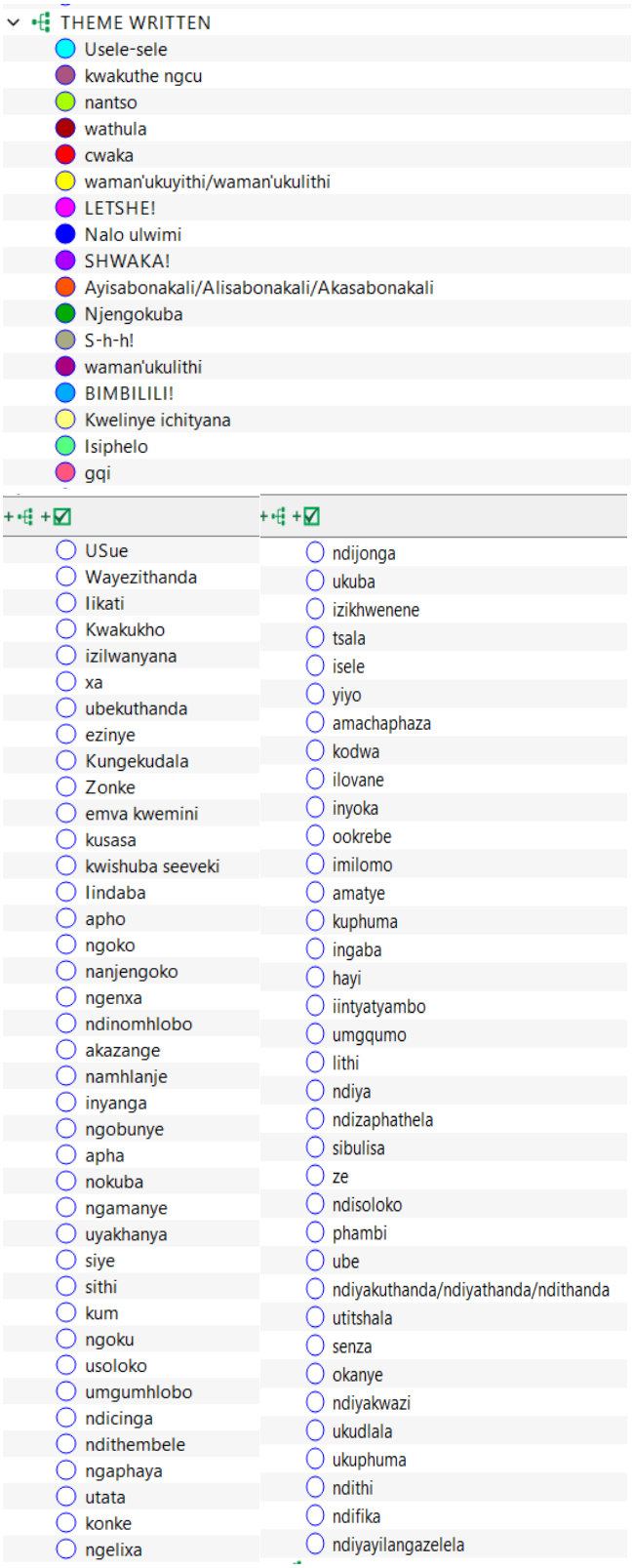


Figure 16 Linguistic Theme system choices for all 5 videos

To annotate the Themes, I had to write each Theme down manually and add it to the Theme system, then annotate them as they appeared in the videos (see Figure 16). Similar words and

phrases that just differed by subject were grouped together. The choice was made to highlight the entire word as Theme, and not necessarily a prefix representing a Participant on its own, rather the entire Process and details within the word itself.

4.3.2.2 Visuals

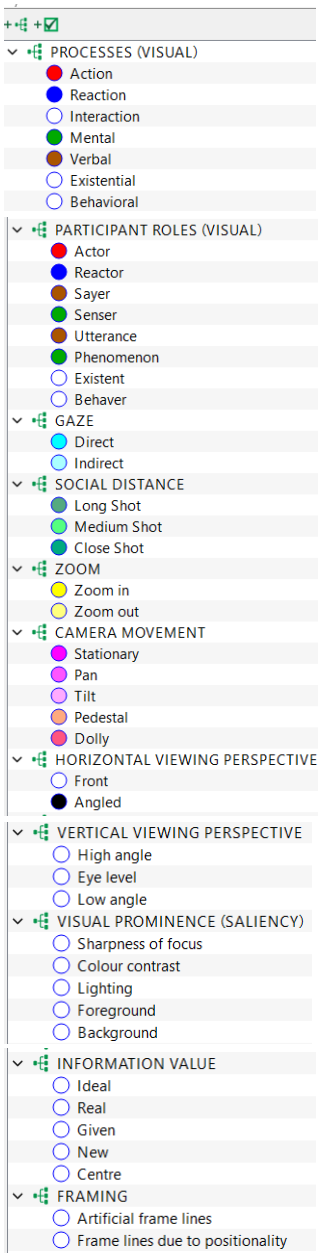


Figure 17 Systems and their system choices for visuals

Figure 17 illustrates choices within the visual systems described in 3.1.1. As mentioned in 4.3.2.1, duplicate systems, i.e., two sets of systems to describe concurrent choices, were

required for the visual, as there were multiple instances of co-occurring systems from different visual Participants. The unit of annotation for the visual mode was the shot.

4.3.2.3 Intersemiotic Texture

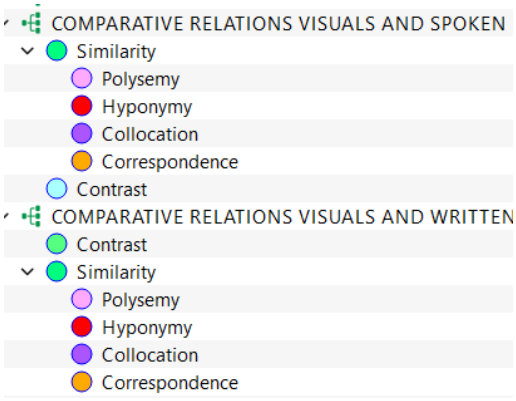


Figure 18 Systems and their system choices for Intersemiotic Texture

Figure 18 illustrates choices within Intersemiotic Texture and Comparative Relations described in 3.1.3. Annotations of spoken language were duplicated from earlier annotations for the first system COMPARATIVE RELATIONS VISUALS AND SPOKEN and annotations of written language were duplicated from earlier annotations for the second system COMPARATIVE RELATIONS VISUALS AND WRITTEN. They were then categorised for types of Similarity and/or Contrast. For example, Similarity appears in Video 1 for most of the video when the subtitle "LETSHE! Nalo ulwimi lukaSele-sele luncangathi kwaye lulude" (SHOOT! Out goes Sele-sele's sticky, long tongue) appears with the presence of Sele-sele and his tongue in the visuals. However, in Video 1, there is an instance of this subtitle portraying both Similarity and Contrast simultaneously, something which Liu and O'Halloran (2009) claim cannot occur. This is due to a shot transition occurring to a different visual while the subtitle is still on the screen, thus the tongue is no longer in the visual mode. Therefore, while Similarity and Contrast may be binary choices, in which only one can occur at a time, for the spoken language, they can co-occur in a written subtitle, particularly when examining smaller units of Parallel Structures for similar meanings.

4.3.2.4 Subtitle rate

The spoken words and subtitles are exactly the same, and since the focus was on the subtitle as written language that may influence literacy skills, the speech rate was not analysed, rather the subtitle rate. The subtitle rate is calculated according to De Linde and Kay (2014). Subtitle word per minute (WPM) rate was calculated from the number of words in a subtitle being divided by subtitle presentation rate, which was calculated in the *Multimodal Analysis Video* software. According to De Linde and Kay (2014:45), subtitle presentation rate is "a measure of how quickly a subtitle appears and disappears from screen" and is measured in words per minute. While isiXhosa is an agglutinative language and the words are considerably longer than English words, the language with which this formula was first used, the syllable per minute rate was also counted to ensure that adequate averages were calculated and accounted for these discrepancies in linguistic differences. Subtitle syllable rate per minute (SPM) was calculated by dividing the number of syllables in a subtitle by presentation time. The average duration of subtitles was downloaded from Excel and the annotated duration is used as the total subtitle presentation rate. If the presentation rate was two minutes one second, then the number of seconds had to be divided by 60 to convert the number into a decimal that could be used as the subtitle presentation rate. If words had hyphens, they were included as a single word. In terms of clause separations between subtitles, I did not count instances of these per video, rather I noted the videos where this was a frequent issue that could potentially negatively influence literacy development. For example, in Video 2, "iikati, izinja, amafudo, iintaka, izilwanyana zalo, naluphi na uhlobo zazingavumelekanga kweso sakhiwo sakhe" (cats, dogs, tortoises, birds, all kinds of animals were not allowed in her building) is separated across three subtitles.

As discussed in 2.1.1, frequency of occurrence and duration are factors that relate to the potential of repeated exposure of SLS videos to influence L1 isiXhosa literacy. Frequency of occurrence and duration are two primary attributes of all events in all modalities, however there are distinct differences between duration and frequency (Lantz *et al.* 2020). Further described in 4.1.2.1, duration is described in every method and subsequent analysis in different ways. The ways in which duration and frequency of occurrence manifest are attributes of the event of the intervention in the actual domain and infer causative mechanisms in the real domain that may influence literacy (4.1.1).

Video subtitle rate, aside from an affordance of subtitles as semiotic resources (see 1.2.3), are a measurement of subtitle duration (calculated to a minute as WPM and SPM). This

measurement of duration was compared to another measurement of duration (also calculated to a minute as WCPM and SCPM), which is learners' mean ORF pre- and post-test scores of Word Lists and Longer Passage reading. This was to relate DMDA findings to theories of multimodal literacy and learning regarding features of design (3.2) and, thus, elaborate on the potential influence of SLS videos on L1 isiXhosa literacy skills.

A further aspect of duration and frequency of occurrence examined in the DMDA is in relation to meaning system choices. The duration and frequency of meaning occurrences are represented and analysed in the DMDA by State Transition Machine Diagrams (STMs), with states representing meaning(s) duration in the video and transitions representing meaning(s) frequency of occurrence. This was also to relate DMDA findings to theories of multimodal literacy and learning regarding features of design (3.2) and, thus, elaborate on potential influence of SLS videos on L1 isiXhosa literacy skills. STMs, their interpretation and organisation are described in the next sub-section.

4.3.2.5 State transition machine diagrams

The software *Multimodal Analysis Video* was used to annotate the videos and produce state transition machine (STM) diagrams. According to O'Halloran and Fei (2014), the state transition visualisation illustrates with nodes and their transitions the overall patterns of systemic choices in the video and time allocated to choice combinations in percentages. They are composed of states (circles) and transitions (lines with arrows) as per the example in Figure 19.

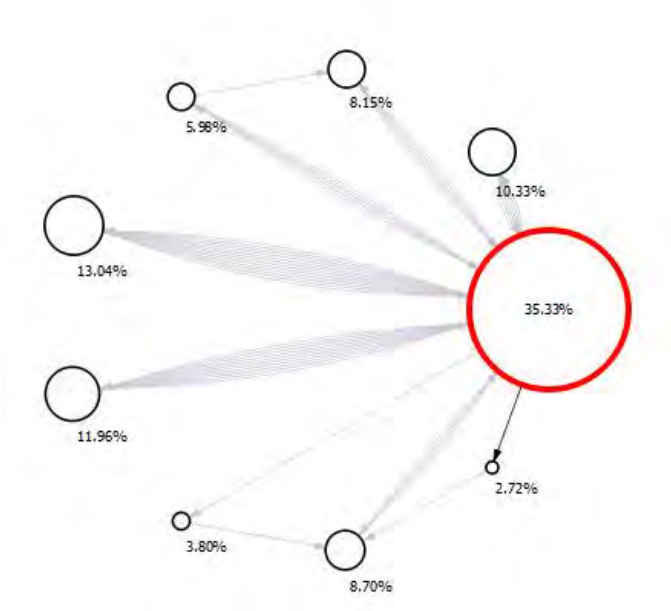


Figure 19 Sample STM

Each state represents either a system or the percentage of the video duration where no systems are present (such as the 35.33% state in Figure 19). Each state can also represent a co-occurrence of multiple systems, and the duration of co-occurrences of meaning systems in percentage form (see 2.1.1 for discussion of duration). For example, let us assume that Figure 19 depicts system choices relating to the Participant system in the written linguistic mode. If the largest state, for example illustrates a co-occurrence of Actor and Goal Participants, it would indicate these co-occurrence(s) appeared for a total duration of 13.04% of the video. They do not indicate how many distinct co-occurrences (frequency) there are, just the overall duration, which implies a frequency of occurrences at the level of their duration for this meaning.

All states circled in red are the states where there are no system choices that occur in the video at all for that specific system (e.g., in Figure 19, there are no other Participant system choices for this mode for 35.33% of the video). The transition lines illustrate a linear unfolding of states and their relationship with each other as the video unfolds, indicating broader patterns. Furthermore, the number of transition lines also indicate frequency of occurrences of system choices (see 2.1.1). For example, let us assume that Figure 19 indicates the system choice of Sensor Participant(s), which comprise 8.70% of the video's duration. The 8.15% state may indicate the total video duration which includes Phenomenon Participant(s). Sensor Participants appear slightly longer in the video overall than Phenomenon Participants. However, the amount of transition lines between the state (total duration of the video) where

there are no Participant choices (35.33%) indicate that there are more frequent individual occurrences of Phenomenon Participants than Sensor Participants. There is often a correlation between duration and frequency of occurrences (such as with the highest state in Figure 19 having the highest amount of frequency of occurrences). However, the presentation of both in STMs ensure that we do not conflate duration and frequency and can truly represent nuances between these attributes for different system choices within each mode. Furthermore, transition lines between states indicate less gaps between no system choices in that system, and while this is uncommon in spoken language, it is more common between visual modes. This is particularly integral in the analysis of subtitles, as it illustrates an imperceptible gap between one subtitle and another. This may often happen in a clause that has been split up, due to its length. This is discussed in detail for each video in Chapter 6.

The more likely a system choice is repeated (its frequency), and the longer system choices are presented in the video (its duration), the greater the potential for better retention (see 2.1.1). This contrasts with a system choice appearing far less often and longer in the video, consequently it could be less salient to the viewer and less influential in relation to L1 isiXhosa literacy skills (see 2.1.1).

4.3.2.5.1 Organisation of STMs, their impact on state and transition interpretation and limitations

As discussed in 4.3.2.1, the nature of spoken and written linguistic modes as distinct, particularly in understanding their co-occurrence in the videos for literacy (see discussions of this in 1.2.3, 3.1 and 3.2). The distinction between spoken and written linguistic modes result in separate modal analyses and separate STMs. They are examined in comparison to one another and separately presented as STMs in the linguistic mode. They are differentiated by metafunction: the ideational metafunction (first sub-heading of each video in Chapter 6), as well as interpersonal and textual metafunctions (second sub-heading of each video in Chapter 6). In order to not create STMs that were highly complex, and consequently unreadable, separate STMs for each system for these metafunctions in the linguistic mode (see 4.3.2.1) are presented.

In terms of spoken language and written language STMs, one thus has a Participant STM, a Process STM and a Circumstance STM in the ideational metafunction for each video (see 3.1.1.1 and 4.3.2.1). The state percentages in a particular STM relate total duration of a system

choice or their co-occurrences in comparison to other system choices of that system. For example, the 35.33% state in Figure 19 indicates the total duration of the video, in which there are no Participant system choices, but there could be Process and Circumstance system choices within that state.

Thus, a state representing no system choices does not equate to a percentage of no spoken language in the videos overall, as well as no other system choices within the spoken linguistic mode and other modes. This further applies to other system STMs in both linguistic modes, including any STMs that represent interpersonal and textual metafunction systems, such as Mood and Themes. Within these categories, due to a lack of variance in Mood, or low frequencies and duration in different types of Modality and Themes in many of the videos, these systems are often described, rather than their STMs included in the DMDA. This also may occur with discussions of visual systems, where there were, for example, a lack of Framing devices within the visual mode in some videos, but a presence of them in others.

For spoken linguistic STMs, as discussed earlier, the states represent the total duration of a system choice(s) in relation to other choices in that specific system (e.g., Participant System Choice). The transition lines indicate frequency of meaning occurrences (states) between choice(s) and their co-occurrences (states) within a specific system, e.g., phonological Participants in a separate STM, Processes in a separate STM and Circumstance in a separate STM. The state, in which no system choices are present, and transition lines from this state to other states indicate frequency between system choices in that STM. These are Participants in that STM (spoken nouns and affixes), Processes in that STM (verb complexes and predicates) and Circumstances in that STM (mostly spoken adverbs, adjectives, or nouns relating to description). For example, in a STM representing spoken Participants, the transition lines (frequent meaning occurrences) indicate nouns and affixes that represent them in the verb complex. Although these affixes comprise a minimum duration, they show occurrences of meaning representation when the Participant it represents may be elided.

Processes are annotated for the entire verb complex, or other predicative expression (such as “kukho,” meaning “there is” or ideophones, see 1.2.1.3). Thus, the STMs do not represent frequency of grammatical categories, but frequency of meaning occurrences in relation to meaning system choices. This is in order to analyse the features of design that may potentially influence L1 isiXhosa literacy skills.

For written language STMs, as discussed earlier, the states represent the total duration of system choice(s) in comparison to the video conveying none of that particular system choice (e.g., Participant System Choice). They are annotated according to each subtitle as its presented (each subtitle duration), and as a result, there can be more co-occurrences than in the spoken linguistic mode, in which language unfolds chronologically in smaller units of duration. This results in, for example, the same systems present for a shorter time span in the videos in the spoken language in comparison to the written subtitles, which contain the same language content. The duration of meanings within the subtitles are represented by states in STMs, which are annotated according to subtitle duration. However, the meanings of system choices may not necessarily be retained by a viewer for as long as the subtitle is on-screen, as the viewer is decoding subtitles in part as well as a whole, which is similar to how spoken language unfolds over time (see 2.1, 2.2.3.1 and 3.2). This relates to the potential complexity of meanings and forms from the written language and other modes, which is an aspect examined in Chapter 6.

The transition lines indicate frequency of meaning occurrences (states) of a system choice between subtitles in one STM. Thus, transition lines to and from a system choice for the written linguistic mode indicate the frequency of specific system choice(s) in relation to other system choices of that same system. Therefore, an occurrence (state) of no system choices does not equate to a percentage of no written language in the videos at all, or frequencies within one subtitle, within the STMs. Frequencies between language overall are better represented in the spoken language mode, which contains identical content to the subtitles. This way of annotating the subtitles indicates the frequency of co-occurrences within a single subtitle (by transition lines), and their entire duration on the screen (by states), which are ultimately longer than spoken linguistic system choices. It indicates the complexity of meanings afforded by the semiotic resource of subtitles, considering the duration and frequency of meanings at different linguistic modes. Linguistic modal affordances respond to the research questions 1, 2, and 4, and the responses to these questions based on analysis findings from Chapter 6 are discussed in 7.1, 7.2, and 7.4. The linguistic mode, in particular, and differences between spoken and written annotations indicate the presentation and processing of the subtitle and its complexity, but also accounts for reading horizontally, which is similar to how spoken language unfolds over time (see 2.1, 2.2.3.1 and 3.2). This relates to the potential complexity of meanings and forms from the written language and other modes, which is an aspect examined in Chapter 6.

To further mediate STM readability, separate STMs for the visual mode were created. The visual mode had already been analysed separately from the linguistic mode with its own separate system choices in the visual mode (see 4.3.2). These were also differentiated by metafunction: the ideational (third sub-heading of each video in Chapter 6), interpersonal (fourth sub-heading of each video in Chapter 6) and textual metafunctions (fifth sub-heading of each video in Chapter 6). For visual STMs, as discussed earlier, the states also represent the total duration of system choice(s) in comparison to the video conveying none of that particular system choice (e.g. Participant System Choice). However, there are naturally systems that occur in this mode that are different to the written and spoken linguistic modes, such as Gaze in the interpersonal metafunction. All system choices within visual systems are annotated according to each system choice as it's presented (each type of shot). As a result, there might be fewer transition lines (frequencies) between no system choices for this mode, and more transition lines between states (as Long shot changes to a Medium shot, for example). The duration of these tend to be far longer than durations in the linguistic modes. Zoom and Social Distance are often presented together, as Zoom is often a transition between Social Distance system choices.

STMs for Intersemiotic Texture Comparative Relations are represented as two distinct STMs per video (sixth sub-heading of each video in Chapter 6), according to how the meanings in these two linguistic modes each relate to the meanings in the visual mode. This is due to the differences in duration and frequency of occurrences in system choices between spoken and written linguistic modes. Thus, the annotations of spoken language by system choice in the ideational metafunction were annotated together for the different systems of Similarity, as well as Contrast. The spoken linguistic mode and visual mode relations are closer to a horizontal type of reading (see 3.2), as it examines how the spoken language and visual meanings correspond and diverge as the spoken language unfolds and how words would be read phonologically (see 2.1.2). Naturally, these annotations are meaning-focused, due to the focus on the motivated sign, comprised of both material form and meaning in modes (Jewitt *et al.* 2016). While phoneme-to-grapheme mapping when reading is from form to meaning (decoding), not meaning to form, the spoken linguistic annotations may indicate a horizontal reading closer aligned to phoneme-to-grapheme mapping. Transition lines indicate Similarity or Contrast at the level of spoken system choice, and their frequencies correspond the closest to word form. Processes in isiXhosa are usually one unit, usually the verb complex. Participants

are usually individual nouns and agreement affixes in the verb complex, which are of a low state duration, and Circumstances are between 1-4 words in length at most.

Similarity and Contrast (system choices) between the written linguistic mode and visuals are represented in separate STMs for each video. These STMs represent Similarity and Contrast duration (states) per subtitle and frequency of types of Similarity and Contrast between subtitles (transitions). In both spoken-to-visual and written-to-visual STMs, the states also represent the total spoken language and written language durations in comparison to no language at all in the videos. While written mode annotations are from meaning to form, not form to meaning, the written linguistic annotations may indicate better opportunities for integration between two modes, which could lead to advanced semiotic awareness affordances (see 1.2.3). Naturally, systems that occur longer and are frequent in one STM, such as the visual Participant of Sensor, and other systems such as the linguistic Participant of Sensor in another STM could complement one another, thus leading to higher Similarity rates. Similarity and Contrast between modes are further represented by their own STMs in Intersemiotic Texture in the sixth sub-heading of each video analysis in Chapter 6. Similarity could further assist literacy development, alternatively, meanings in Contrast could potentially cause a diversion. This corresponds to Mayer's (2009) Spatial and Temporal Contiguity Principles, as Similarity contributes to this Principle and Contrast detracts from this principle, as stated in 3.2.1. This makes STMs well-suited for the purpose of responding to the research questions of this thesis that concern a DMDA and examining the relationships of meanings between modes (research question 4).

Nevertheless, there are limitations to STMs. The dynamic nature of the STMs allow for logogenesis to be examined, however, due to technical limitations of the software, the STMs could not be presented in video form, rather as static figures in this thesis. The STMs in this thesis, with their states and their percentages, illustrate the frequency and duration of systems and the STMs illustrate this in a synoptic way. However, the STMs do not account for different meanings outside a particular system without resulting in the STM becoming too complex to decipher. Thus, the meanings within the semiotic resources had to be annotated according to separate system choices. The Comparative Relations STMs in Intersemiotic Texture, however, allow for a better representation of meanings across system choices and across modes. The next section discusses the triangulation of the findings from the DMDA and two-group experimental method.

4.3.3 Triangulation of the findings from both methods

To triangulate the findings from both methods, I examined the findings from the statistical analysis (mean scores and t-test results) and compared those with the WCPM rate of the subtitles to gauge whether learners could read the subtitles. This is described in the seventh sub-heading of each video. I examined the results from the Picture-Word Match test to understand whether children would have been able to navigate similar meanings from two different modes, thus illustrate an advanced aspect of semiotic awareness (see 4.1.2.1). From these results, I then examined the patterns of semiotic resources and the predominant ones for each system and mode in STM results. I related both the speed of the subtitles and the findings from the STM results and Picture-Word Match tests to Mayer's (2009) principles (see 3.2.1) and multimodal literacy theories (see 3.2). The triangulation of DMDA and test result findings to theories of multimodal literacy and previous literature were to explore features in design, as well as modal affordances (3.2). This describes and elaborates on the potential influence of SLS videos on L1 isiXhosa literacy skills, and I set up the hierarchy of principles for this context in 7.9.1. Furthermore, I reflected on this triangulated information to arrive at responses to the research questions based on the concurrent embedded mixed-method research design. The analysis from this triangulation is described in 6.6.

4.4 Chapter summary

This chapter described the concurrent embedded mixed-method design, as well as outlined the two methods: DMDA and two-group experimental design with statistical analysis and the way data from both methods was analysed and triangulated. This type of mixed-method design provides a holistic picture of potentials of the videos to assist and influence L1 literacy development in isiXhosa in the Foundation Phase. This is from the DMDA and an understanding of whether the videos assist or do not assist L1 literacy development in isiXhosa for a particular context of L1 literacy development in isiXhosa that has been reported to have low literacy scores. DMDA integration with quantitative methods such as t-test results do not often occur in this manner and for learning potentials. This research design further highlights a realisation of the critical realist paradigm. In the conceptualisation stage of this research, two inferences of potential causal mechanisms were identified. These two inferences of causal mechanism are (1) the ability for learners to follow subtitles adequately to influence literacy

development and (2) the nature of learners integrating meanings between visual and linguistic modes to assist in literacy development.

Furthermore, sampling methods examine participants from a context, i.e., Newlands area, that has a lack of published research, as well as materials such as the SLS isiXhosa animated story videos, which lack research. This methodology is thus useful for establishing an alternative methodology to previous ones outlined in the literature in Chapter 2 for areas of literacy, subtitle research and multimodality. It combines methods from different disciplines, to provide a holistic and comprehensive understanding of the context of this research.

Chapter 5 – Intervention results

This chapter of data results and analysis is a discussion of the intervention data. The data presented in 5.1. is divided into Grade 2 and 3 descriptive statistics of words correct per minute (5.1.1), Long Passage Reading (5.1.2), and Picture-Word Match scores (5.1.3). In 5.2, the t-test results pre-and post-test between experimental and control groups are discussed, firstly combining Grades 2 and 3 by experimental and control group, in order to ensure accurate results by sample size (5.2.1). T-test results are then separated by post-test Word List 1 and 2 in experimental group in 5.2.2, with average reading scores from pre- and post-test for each grade reported in 5.2.2.1. The statistics in this chapter are an explorative additional pilot study to investigate the analysis outlined in the DMDA findings.

5.1 Initial descriptive data

The Grade 2 and 3 learners in the school comprised a total of 70 learners. This is a full representative sample of the research site for these grades and the Grade 1 learners are often of a preliterate stage. Thus, it was deemed inappropriate to include them for the sake of a greater sample for concerns that they could not complete the test. The goal of this data is to provide an additional layer of context to the DMDA undertaken on the videos and whether the subtitles in the videos could potentially aid with literacy within 3 months or not. Furthermore, we use the average reading score of learners and compare that to the objective subtitle rate within the videos to deduce whether the subtitle rate in the videos would be at an adequate speed for learners to read. This is to answer research questions 5-9.

An explanation of abbreviations in this chapter are as follows: 2A represents the Grade 2 experimental group and 3A represents the Grade 3 experimental group. 2B represents the Grade 2 control group and 3B represents the Grade 3 control group. WCPM stands for Words Correct Per Minute (bi- or trisyllabic lexical units) and SCPM stands for Syllables Correct Per Minute. This refers to the Word Lists for Section B of the tests (see Appendix B), with Word List 1 including words only from the videos and Word List 2 including words not in the videos. The other component of that test, the general reading passage, is titled "Long Passage." Section A's Picture-Word Match test is titled Picture-Word Match in the tables.

In order to see the effects in a larger sample size, the Grade 2 experimental group and Grade 3 experimental group were combined. The Grade 2 control group and Grade 3 control group were also combined. This is illustrated in Tables 1 and 2. When grades are combined, they are represented as EXP for experimental group and CTRL for control group.

									Shapiro-Wilk	
	Group	N	Missing	Mean	Median	SD	Minimum	Maximum	W	p
Pre-test Word List 1 WCPM	EXP	37	0	19.19	19.00	15.27	0.00	46.0	0.910	0.006
	CTRL	33	0	21.82	19.00	16.28	1.00	50.0	0.917	0.015
Pre-test Word List 1 SCPM	EXP	37	0	40.92	39.00	33.97	0.00	103.0	0.898	0.003
	CTRL	33	0	47.39	38.00	36.66	2.00	110.0	0.910	0.010
Pre-test Word List 2 WCPM	EXP	37	0	20.16	20.00	16.09	0.00	49.0	0.902	0.003
	CTRL	33	0	23.27	26.00	16.73	0.00	50.0	0.926	0.026
Pre-test Word List 2 SCPM	EXP	37	0	43.22	42.00	35.52	0.00	108.0	0.903	0.004
	CTRL	33	0	51.76	52.00	39.00	0.00	122.0	0.930	0.035
Pre-Test Long Passage WCPM	EXP	37	0	17.59	20.00	12.99	0.00	39.0	0.906	0.004
	CTRL	33	0	18.64	21.00	15.51	0.00	45.0	0.889	0.003
Pre-Test Long Passage SCPM	EXP	37	0	58.35	67.00	43.68	0.00	139.0	0.909	0.005
	CTRL	33	0	60.58	67.00	50.91	0.00	157.0	0.911	0.010
Pre-Test Picture-Word Match	EXP	37	0	7.41	8.00	2.80	0.00	10.0	0.818	< .001
	CTRL	33	0	8.39	10.00	2.40	3.00	10.0	0.693	< .001

Table 1 Combined experimental and control group pre-test descriptives

A table of normality test in Table 3 (Shapiro-Wilk) illustrates that the data for all pre-tests was not normally distributed, as the p value was less than 0.05. The p value that was used in this study was adopted from the p value used in Kothari *et al*. (2002). The nonparametric distribution was not only due to the small sample size, but rather due to a pattern of bimodal distribution of the sample illustrated in this section for most of the data, which is visible in the Figures in 5.1. The only exception of bimodal distribution in the pre-tests for experimental and control groups (combined grades) is represented in Figure 20, which seems to indicate tri-modal distribution in SCPM for Long Passage and Word List 2 tests in the control group. However, the post-test results resume bimodal distribution. This indicates that many readers who could not read (at SCPM score 0) stayed reading at that score while good readers who could read at higher scores improved, ultimately leaving their illiterate or underperforming classmates behind.

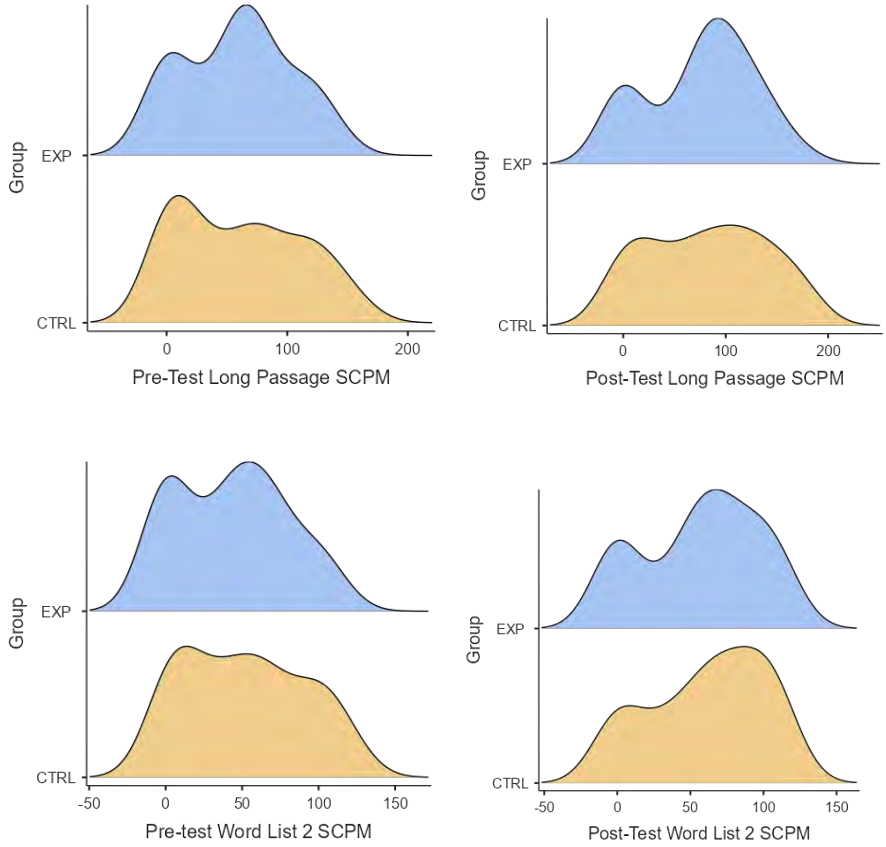


Figure 20 Long Passage SCPM pre-and post-test for experimental and control groups (combined grades) and Word List 2 SCPM pre-and post-test for experimental and control groups (combined grades)

The post-test data was also not normally distributed, which is observed in Table 2 and contained a pattern of bimodal distribution, which is illustrated by the Figures later in this section.

	Group	N	Missing	Mean	Median	SD	Minimum	Maximum	Shapiro-Wilk	
									W	p
Post-Test Word List 1 WCPM	EXP	37	0	26.24	30.0	17.57	0.00	50.0	0.878	< .001
	CTRL	33	0	30.36	34.0	18.22	0.00	50.0	0.864	< .001
Post-Test Word List 1 SCPM	EXP	37	0	58.27	65.0	39.99	0.00	114.0	0.891	0.002
	CTRL	33	0	67.58	73.0	41.36	0.00	114.0	0.876	0.001
Post-Test Word List 2 WCPM	EXP	37	0	25.59	30.0	17.01	0.00	50.0	0.888	0.001
	CTRL	33	0	28.18	32.0	17.31	0.00	59.0	0.925	0.026
Post-Test Word List 2 SCPM	EXP	37	0	57.08	63.0	39.01	0.00	112.0	0.896	0.002
	CTRL	33	0	61.76	68.0	38.69	0.00	112.0	0.907	0.008
Post-Test Long Passage WCPM	EXP	37	0	22.84	27.0	15.00	0.00	51.0	0.886	0.001
	CTRL	33	0	24.09	25.0	16.72	0.00	50.0	0.923	0.023
Post-Test Long Passage SCPM	EXP	37	0	75.78	84.0	51.07	0.00	174.0	0.909	0.005
	CTRL	33	0	81.70	78.0	57.81	0.00	179.0	0.936	0.053
Post-Test Picture-Word Match	EXP	37	0	8.16	10.0	3.05	0.00	10.0	0.663	< .001
	CTRL	33	0	8.09	10.0	3.06	0.00	10.0	0.670	< .001

Table 2 Combined experimental and control group post-test descriptives

Despite the small group sizes per grade, in order to understand the mean scores and differences between grades, the Shapiro-Wilk values for individual pre- and post-tests are presented in Table 3. This table illustrates that within grades, the data is non-normally distributed and contains a bimodal distribution outlined in the sub-sections of 5.1.

		Pre-Test Shapiro-Wilk		Post-Test Shapiro-Wilk	
	Group	W	P	W	P
Pre-Test Word List 1 WCPM	2A	0.929	0.183	0.890	0.039
	3A	0.877	0.019	0.845	0.06
Pre-Test Word List 1 SCPM	2A	0.913	0.097	0.912	0.095
	3A	0.876	0.018	0.851	0.07
Pre-Test Word List 2 WCPM	2A	0.882	0.028	0.894	0.045
	3A	0.853	0.08	0.870	0.015
Pre-Test Word List 2 SCPM	2A	0.904	0.067	0.922	0.143
	3A	0.861	0.010	0.870	0.015
Pre-Test Long Passage WCPM	2A	0.888	0.035	0.855	0.010
	3A	0.845	0.06	0.878	0.020
Pre-Test Long Passage SCPM	2A	0.889	0.037	0.885	0.031
	3A	0.864	0.011	0.896	0.041
Pre-Test Word Picture-Word Match	2A	0.815	0.03	0.780	<0.01
	3A	0.766	<0.01	0.537	<0.01

Table 3 Descriptives of Shapiro-Wilk test between Grade 2 and 3 experimental groups.

In order to ensure that both experimental and control groups for both grades were at the same starting point, Table 4 illustrates the Mann-Whitney U independent sample t-tests conducted.

		Statistic	P
Pre-test Word List 1 WCPM	Mann-Whitney U	548	0.465
Pre-test Word List 1 SCPM	Mann-Whitney U	547	0.458
Pre-test Word List 2 WCPM	Mann-Whitney U	534	0.370
Pre-test Word List 2 SCPM	Mann-Whitney U	531	0.349
Pre-Test Long Passage WCPM	Mann-Whitney U	573	0.658
Pre-Test Long Passage SCPM	Mann-Whitney U	585	0.768
Pre-Test Picture-Word Match	Mann-Whitney U	458	0.057

Table 4 Independent Samples Mann-Whitney t-test of all pre-tests between experimental and control groups for combined Grades 2 and 3.

As Table 4 illustrates, there is no significant difference between the experimental and control groups pre-tests for the individual Word Lists. This is visible for both WCPM and SCPM. For the pre-test Word List 1 WCPM between control and experimental Groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean WCPM score of EXP and CTRL U = 548, p = 0.465. For the pre-test Word List 1 SCPM between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean WCPM score of EXP and CTRL U = 547, p = 0.458. For the pre-test Word List 2 WCPM between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean WCPM score of EXP and CTRL, U = 543, p = 0.370. For the pre-test Word List 2 SCPM between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean SCPM score of EXP and CTRL, U = 531, p = 0.349. This was to ensure both experimental and control groups for both grades were at the same starting point and that there was no significant difference between groups' pre-test scores.

Table 4 further illustrates, there is no significant difference between the experimental and control groups pre-test for the individual Long Passage and Picture-Word Match tests. This is visible for both WCPM and SCPM. For the pre-test Long Passage WCPM between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean WCPM score of EXP and CTRL U = 573, p = 0.658.

For the pre-test Long Passage SCPM between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean WCPM score of EXP and CTRL U = 585, p = 0.768. For the pre-test Picture-Word Match between control and experimental groups (combined Grade 2 and 3 groups), the results indicated that there was no significant difference between the mean score of EXP and CTRL U = 458, p = 0.057. There was no significant difference between these pre-test scores between control and experimental Groups for combined Grade 2 and 3, which further indicates any differences between WCPM and SCPM reading rates between experimental and control groups were arbitrary and insignificant.

5.1.1 Grade 2 and 3 isolated words correct per minute scores

This section discusses the section B Word List results, initially with Grade 2 groups and then with Grade 3 groups. Table 5 illustrates the descriptive statistics for Grade 2A and 2B (experimental and control groups).

(2A is Experimental Group and 2B is Control Group)

			Pre-Test Word List 1					Post-Test Word List 1				
	Group	N	Mean	Med	SD	Max	Min	Mean	Med	SD	Max	Min
WCPM	2A	18	14.7	17.0	11.0	40.0	0.0	24.2	28.5	15.6	49.0	0.0
	2B	19	15.6	12.0	13.6	43.0	2.00	24.2	25.0	18.7	50.0	0.0
SCPM	2A	18	30.4	35.0	24.3	90.0	0.0	52.7	60.5	34.8	113.0	0.0
	2B	19	33.4	25.0	30.5	94.0	3.00	54.0	55.0	42.3	114.0	0.0
			Pre-Test Word List 2					Post-Test Word List 2				
	Group	N	Mean	Med	SD	Max	Min	Mean	Med	SD	Max	Min
WCPM	2A	18	15.6	18.5	11.8	34.0	0.0	24.3	28.0	15.5	47.0	0.0
	2B	19	16.5	13.0	14.8	48.0	0.0	22.5	26.0	17.0	44.0	0.0
SCPM	2A	18	32.8	37.5	25.9	80.0	0.0	53.7	60.5	35.9	111.0	0.0
	2B	19	36.2	26.0	34.2	122.0	0.0	50.5	52.0	39.7	105.0	0.0

Table 5 Descriptives for Grade 2 experimental and control group

In the Word List 1 pre-test, the WCPM mean scores are 14.7 (SD = 11.0) and 30.4 syllables (SD = 24.3) for the experimental group. For the control group, the WCPM mean scores are 15.6 (SD = 13.6) and 33.4 syllables (SD = 30.5) for the control group. The higher standard deviation rates are expected given the non-normal and often bimodal distribution. The learners scored slightly better on average on Word List 2 (words not in the videos) with a WCPM mean score of 15.6 (SD = 11.8) for the experimental group and 16.5 (SD = 14.8) for the control group. This was also reflected in the SCPM mean scores of 32.8 SCPM (SD = 25.9) for 2A and 36.2 SCPM (SD = 34.2). Nevertheless, these mean scores only differ by a single word or three syllables, which indicates they read the tests at similar rates in the pre-test.

There was a steady improvement by the post-test learners in 2A and 2B, who could read at a mean score of 24.2 WCPM (SD = 15.5) and 24.2 WCPM (SD = 17.0). This is 52.7 SCPM for 2A (SD = 35.9) and 54.5 SCPM for 2B (SD = 39.7). For Word List 2, the mean scores are 24.3 WCPM (SD = 15.5) for 2A and 22.5 WCPM (SD = 17.0) for 2B. Comparing mean scores, there was only a SCPM difference of two syllables between 2A and 2B mean scores. There was also no WCPM difference in mean scores between 2A and 2B. Between Word List 1 and 2, mean scores seemed similar in experimental group between lists but it seems the control group scored a mean score of two words less in a minute than in the pre-test. Whether this is statistically significant is discussed in 5.2.

While the average mean scores of the readers did improve during the intervention between pre- and post-test, one can see the minimum score for these tests was 0, illustrating learners unable to read a single word from the tests. The pre-and post-test density graphs in Figure 21 illustrate the large number of learners that could not read in Grade 2, which established the non-normal and bi-modal distribution data pattern. Differences or lack thereof between experimental and control group post-tests are discussed in 5.2.

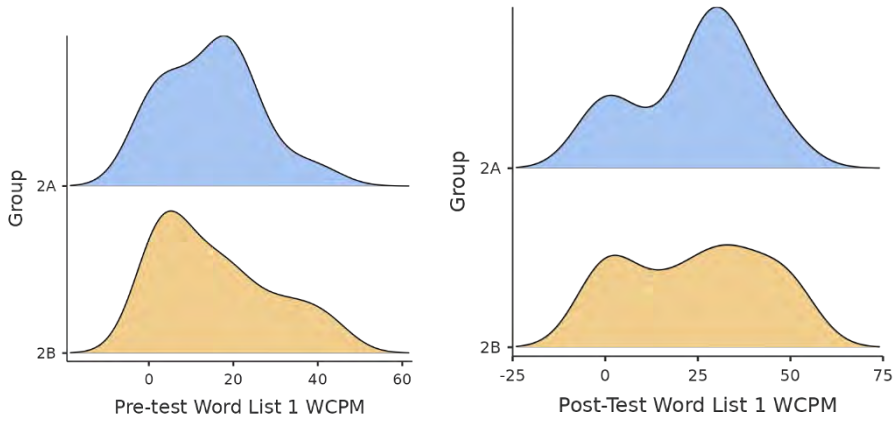


Figure 21 Pre- and post-test WCPM for Word List 1 (words in videos) between Grade 2 experimental group (2A) and control group (2B)

This bimodal distribution pattern was clustered together as learners were on similar levels in the pre-test for WCPM in 2A, with less good readers than illiterate readers in 2B. One can see from the post-test Word List 1 graphs that the peaks are slightly further apart. This illustrates, that while there was an overall improvement in scores and less illiterate readers based on post-test results, that many readers who could not read (at WCPM score 0) stayed reading at that score. Furthermore, good readers who could read at higher scores improved, ultimately leaving their illiterate or underperforming classmates behind. This pattern is also found in Word List 2 (see Figure 22).

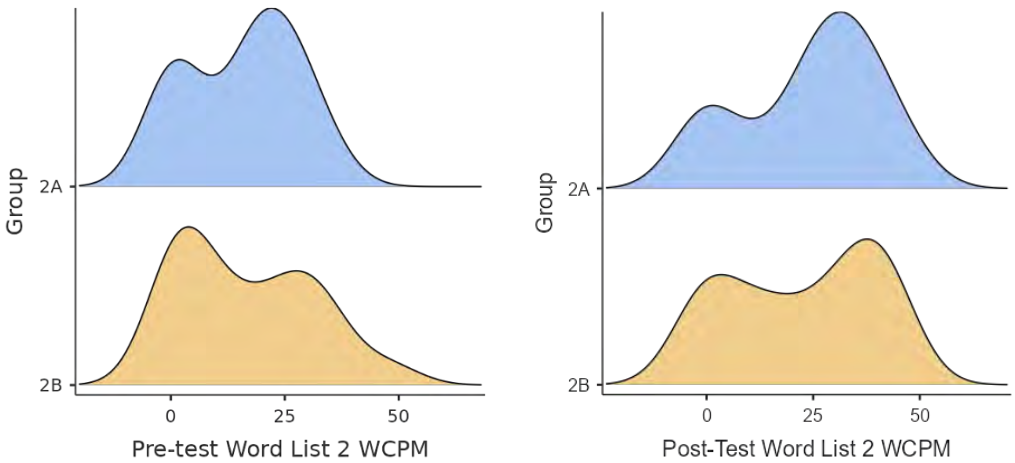


Figure 22 Pre- and post-test WCPM for Word List 2 (words not in videos) between Grade 2 experimental group (2A) and control group (2B)

As one can see from the syllable density graphs in Figure 23, they are similar to the WCPM rate (basically SCPM is double the WCPM which is common for the lists having bisyllabic words) and consequently SCPM rates for Word Lists are only discussed in relation to their longer passage counterparts in 5.1.2.

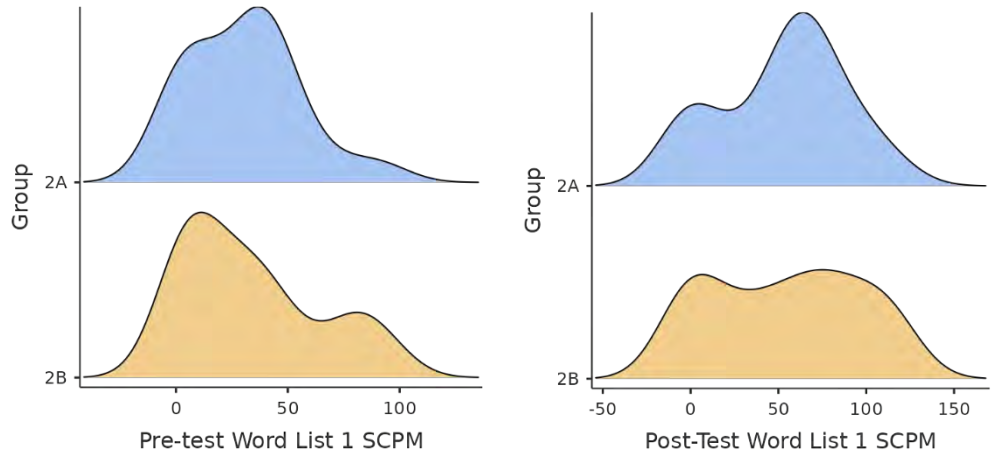


Figure 23 Pre- and post-test SCPM for Word List 1 (words in videos) between Grade 2 experimental group (2A) and control group (2B)

As mentioned in 5.1 and illustrated by the Mann-Whitney U test of pre-test scores in Table 4, while the control group mean scores seem slightly higher than experimental group, there is no significant difference between experimental group and control group pre-test scores. Table 6 illustrates the descriptive statistics for Grade 3A and 3B (Grade 3 experimental and control groups). The pre-test mean scores of Group 3A are 23.5 WCPM (SD = 17.6) and 50.8 SCPM (SD = 39.2) on Word List 1. For Word List 2, the pre-test mean scores of Group 3A are 24.5 WCPM (SD = 18.6) and 53.1 SCPM (SD = 41.0). The pre-test mean scores of Group 3B are 30.2 WCPM (SD = 16.2) and 66.4 SCPM (SD = 36.6) for Word List 1. For Word List 2, the pre-test mean scores of Group 3B are 32.4 WCPM (SD = 15.1) and 72.9 SCPM (SD = 35.9). Particularly in the Grade 3 data, the pre-test mean scores in Word List 2 seem to be higher than Word List 1 and paired sample t-tests and Mann-Whitney U tests of post-test results are conducted to examine whether this was significant in 5.2.

A similar pattern of control group performing better than experimental group is found in the post-test. For Word List 1 post-test, the Grade 3 experimental group mean scores are 28.2 WCPM (SD = 19.5) and 63.6 SCPM (SD = 44.6). For Word List 2 post-test, the experimental

mean scores are 26.8 WCPM (SD = 18.6) and 60.3 SCPM (SD = 42.5). The control group mean scores are 38.7 WCPM (SD = 14.3) and 86 SCPM (SD = 33.1) for Word List 1 post-test. For Word List 2 post-test, the control group mean scores are 35.9 WCPM (SD = 15.1) and 77.1 SCPM (SD = 32.6). A Mann-Whitney U tests of post-test results are conducted to examine whether this was significant in 5.2.

			Pre-Test Word List 1					Post-Test Word List 1				
	Group	N	Mean	Median	SD	Max	Min	Mean	Median	SD	Max	Min
WCPM	3A	19	23.5	24.0	17.6	46.0	0.0	28.2	33.0	19.5	50.0	0.0
	3B	14	30.2	32.0	16.2	50.0	1.0	38.7	44.0	14.3	50.0	0.0
SCPM	3A	19	50.8	42.0	39.2	103.0	0.0	63.6	72.0	44.6	114.0	0.0
	3B	14	66.4	70.0	36.6	110.0	2.0	86.0	97.0	33.1	114.0	0.0
			Pre-Test Word List 2					Post-Test Word List 2				
	Group	N	Mean	Median	SD	Max	Min	Mean	Median	SD	Max	Min
WCPM	3A	19	24.5	30.0	18.6	49.0	0.0	26.8	30.0	18.6	50.0	0.0
	3B	14	32.4	38.0	15.1	50.0	1.0	35.9	34.0	15.1	59.0	0.0
SCPM	3A	19	53.1	65.0	41.0	108.0	0.0	60.3	66.0	42.5	112.0	0.0
	3B	14	72,9	84.0	35.9	114.0	2.0	77.1	78.0	32.6	112.0	0.0

Table 6 Descriptives for Grade 2 experimental and control group

A similar bimodal distribution is visible with both Grade 3 and 2 learners which alludes to an overall improvement in reading scores from pre- to post-test. However, there are still many illiterate readers with a score of 0 and the peaks do not separate further as we saw within the 3 months for Grade 2 learners. This indicates good and bad readers are already defined as early as Grade 2 without serious intervention. One can also see the higher scores with more competent readers in the control group in Grade 3. Table 4 illustrates that, while the control

group scored higher in the pre-tests, results were arbitrary and insignificant, thus supporting equal and random initial scores for experimental and control group samples.

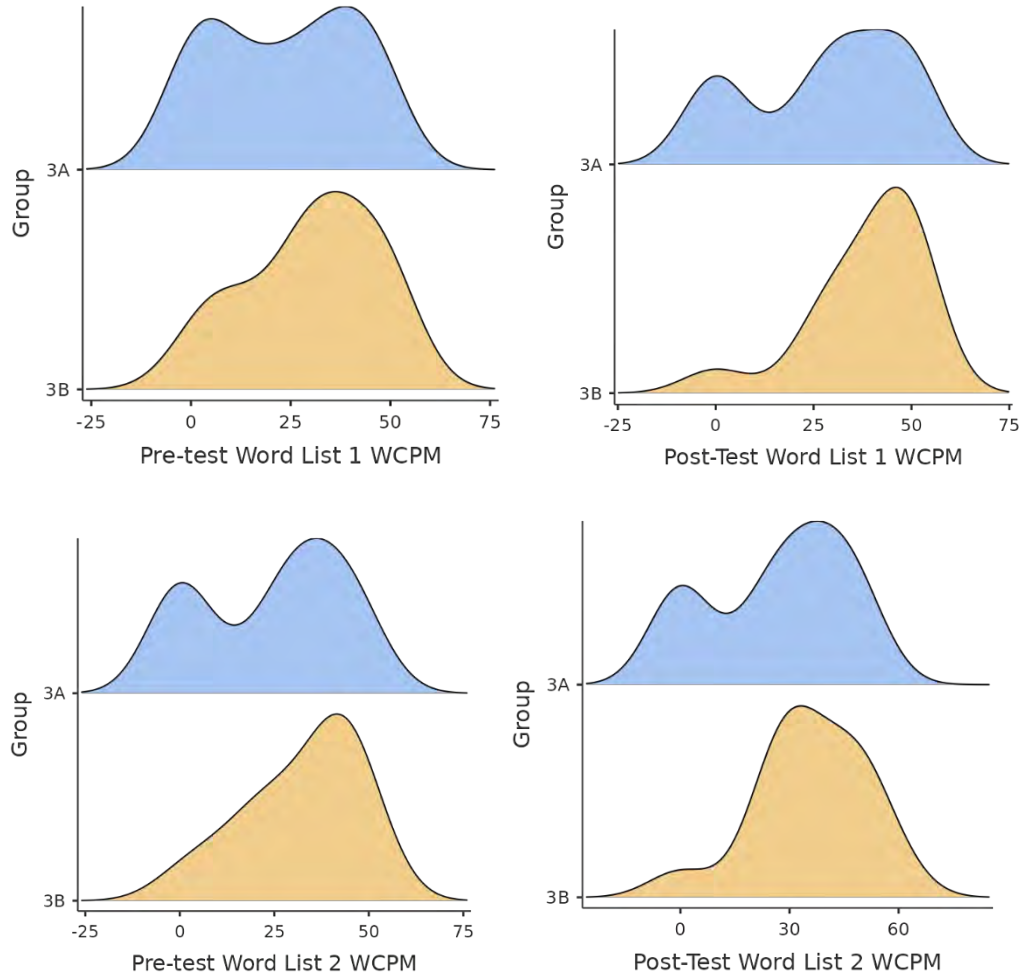


Figure 24 Pre- and post-test WCPM for Word List 1 (words in videos) and Word List 2 (words not in videos) between Grade 3 experimental group (3A) and control group (3B)

The next section examines the descriptive statistics for the Long Passage reading scores for Grade 2 and 3 learners.

5.1.2 Long Passage Reading

WCPM in this reading passage indicate the written word as a unit separated by spaces (see Appendix B), while WCPM in the Word Lists are separated in tables (see Appendix B). The mean scores pre-test for the experimental group in Grade 2 are 15.4 WCPM (SD = 9.77) and 50.6 SCPM (SD = 32.3). Their post-test mean scores are 21.8 WCPM (SD = 13.5) and 70.1 SCPM (SD = 44.8) in the Long Passage Reading test. The Grade 2 control group read at mean

scores of 13.2 WCPM (SD = 14.0) and 40.2 SCPM (SD = 40.4) in the pre-test and 18.1 WCPM (SD = 14.8) and 60.3 SCPM (SD = 49.9) in the post-test. These results are illustrated in Table 7. While the experimental group did slightly better in their reading than the control group, it is discussed in 5.2. whether this was a result of the videos. The mean scores for Long Passage reading are significantly less than the Word Lists. This could be attributed to readers processing connected text and the words are often longer in length than in the Word Lists. Mid-way through the year, the Grade 2 experimental group had reached the benchmark for the end of Grade 2 outlined by Ardington *et al*. (2020) of 20 WCPM, while the control group had reached slightly under the benchmark.

Descriptives for Grade 2 Experimental and Control Group

			Pre-Test Long Passage					Post-Test Long Passage				
	Group	N	Mean	Med	SD	Max	Min	Mean	Med	SD	Max	Min
WCPM	2A	18	15.4	19.5	9.77	34.0	0.0	21.8	26.5	13.5	41.0	0.0
	2B	19	13.2	6.0	14.0	36.0	0.0	18.1	21.0	14.8	41.0	0.0
SCPM	2A	18	50.6	63.5	32.3	116.0	0.0	70.1	82.0	44.8	141.0	0.0
	2B	19	40.2	21.0	40.4	111.0	0.0	60.3	67.0	49.9	138.0	0.0

Table 7 Descriptives for Grade 2 experimental and control group

Density graphs for these statistics are illustrated in Figure 25. As with the previous Word Lists, there is a bimodal distribution that appears for the longer reading passage similar for both words and syllables, where learners who could not read a word of the passage showed no significant improvement and remained at that level.

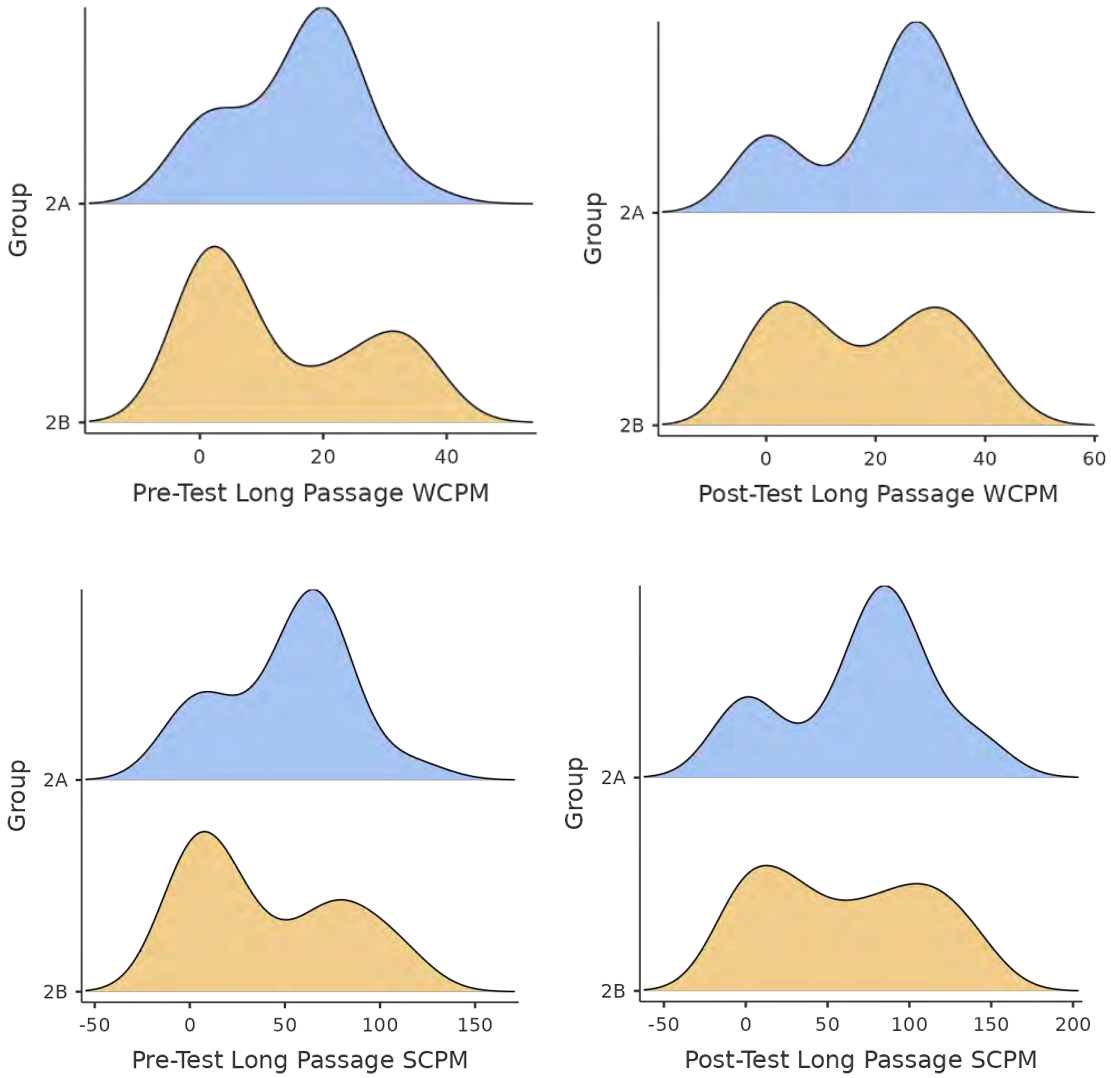


Figure 25 Pre- and post-test Long Passage WCPM and SCPM Grade 2

In Figure 26, the pre-test mean scores of the Grade 3 experimental group are 19.7 WCPM (SD = 15.4) and 65.7 SCPM (SD = 52.1). The post-test mean scores of the Grade 3 experimental group are 23.8 WCPM (SD = 16.6) and 81.2 SCPM (SD = 57.1). The pre-test mean scores of the Grade 3 control group are 26 WCPM (SD = 14.8) and 88.2 SCPM (SD = 51.8). The post-test mean scores of the Grade 3 control group are 32.2 WCPM (SD = 16.1) and 111 SCPM (SD = 56.6). These results are illustrated in Table 8. The Grade 3 learners in both groups did not reach the benchmark for the end of Grade 3 outlined by Ardington *et al*. (2020) of 35 WCPM.

Descriptives for Grade 3 Experimental and Control group

(3A is Experimental Group and 3B is Control Group)

			Pre-Test Long Passage					Post-Test Long Passage				
	Group	N	Mean	Med	SD	Max	Min	Mean	Med	SD	Max	Min
WCPM	3A	19	19.7	23.0	15.4	39.0	0.0	23.8	31.0	16.6	51.0	0.0
	3B	14	26.0	28.0	14.8	45.0	0.0	32.2	34.5	16.1	50.0	0.0
SCPM	3A	19	65.7	67.0	52.1	139.0	0.0	81.2	98.0	57.1	174.0	0.0
	3B	14	88.2	98.5	51.8	157.0	0.0	111.0	119.0	56.6	179.0	0.0

Table 8 Descriptives for Grade 3 experimental and control group

As with the Word Lists, similar bimodal distribution is visible in the Grade 3 learners and the Grade 2 learners (see Figure 26). The mean scores seem to improve dramatically in the control group, but it seems the control group in Grade 3 started out reading faster than the experimental group based on the mean scores. Table 4, however, illustrates that in pre-test, there were no significant differences between the experimental and control groups, when Grade 2 and 3 experimental groups are combined. This is also discussed in terms of individual grades and their experimental and control groups in 5.2. The next section examines the descriptive statistics for the Picture-Word Match test for Grade 2 and 3 learners.

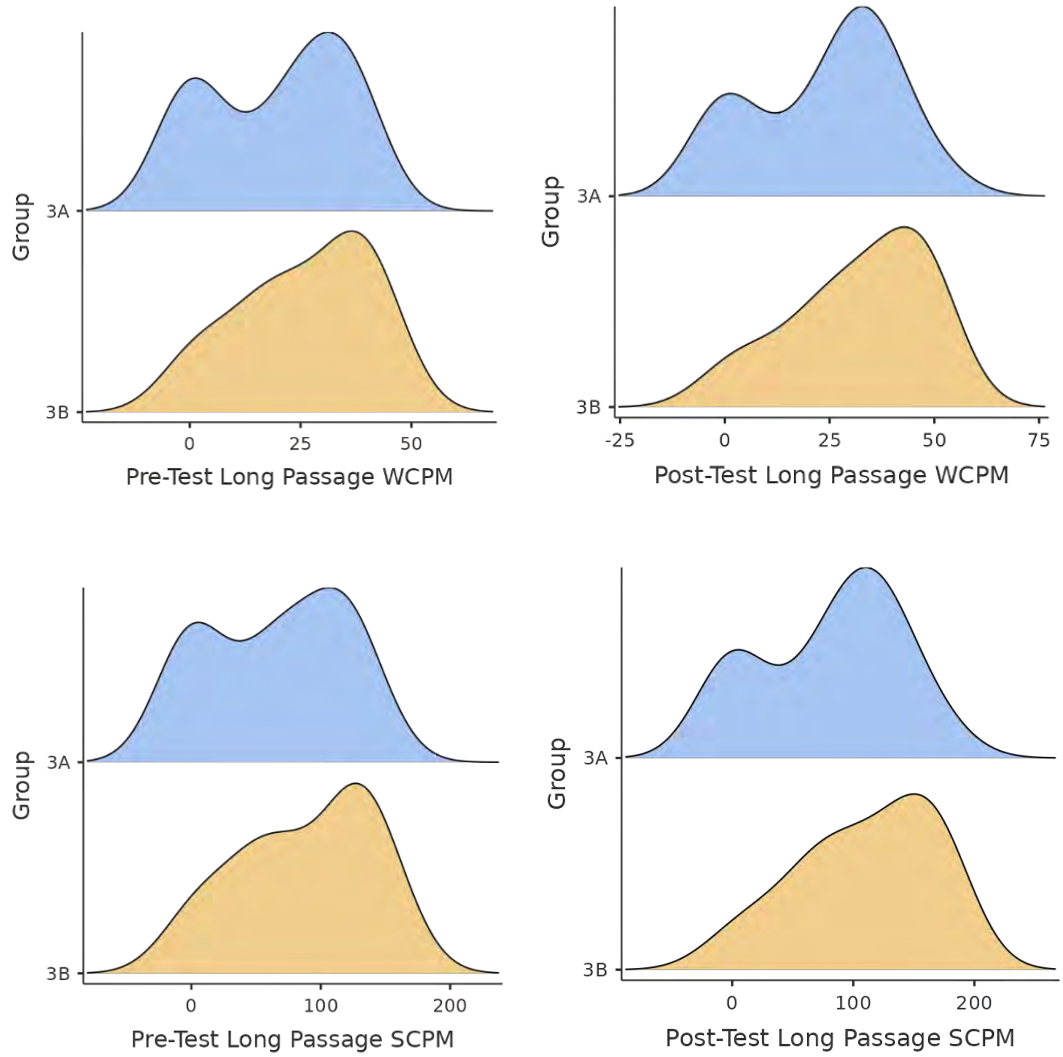


Figure 26 Pre- and post-test Long Passage WCPM and SCPM for Grade 3 experimental group (3A) and control group (3B)

5.1.3 Picture-Word Match scores

Out of a list of 10 words with corresponding images from the videos, the pre-test mean scores as depicted in Table 9 for Grade 2 learners are 7.11 correct matches (SD = 2.54) in the experimental group 7.42 correct matches in control group (SD = 2.65). Post-test mean scores are 7.89 correct matches (SD = 2.81) for the experimental group and 8.37 correct matches (SD = 1.89) for the control group. Mean scores were similar for both groups. Such a high score indicates learners could comprehend simple word meanings and match this to a picture with a corresponding meaning.

Descriptives for Grade 2 Experimental and Control Group

(2A is Experimental Group and 2B is Control Group)

			Pre-Test Picture-Word Match					Post-Test Picture-Word Match				
	Group	N	Mean	Median	SD	Max	Min	Mean	Median	SD	Max	Min
Score	2A	18	7.11	8.00	2.54	10.00	2.00	7.89	9.00	2.81	10.00	0.00
	2B	19	7.42	8.00	2.65	10.00	3.00	8.37	9.00	1.89	10.00	5.00

Table 9 Descriptives for Grade 2 experimental and control group

Judging from the scores themselves without t-test results, one can see there was no significant increase or difference in these scores for Grade 2 learners post-video, indicating the videos did not really influence their ability to better connect linguistic and visual meanings, as this was already a high score pre-video exposure. The density graphs in Figure 27 indicate bimodal distribution pre-test, but results improve post-test.

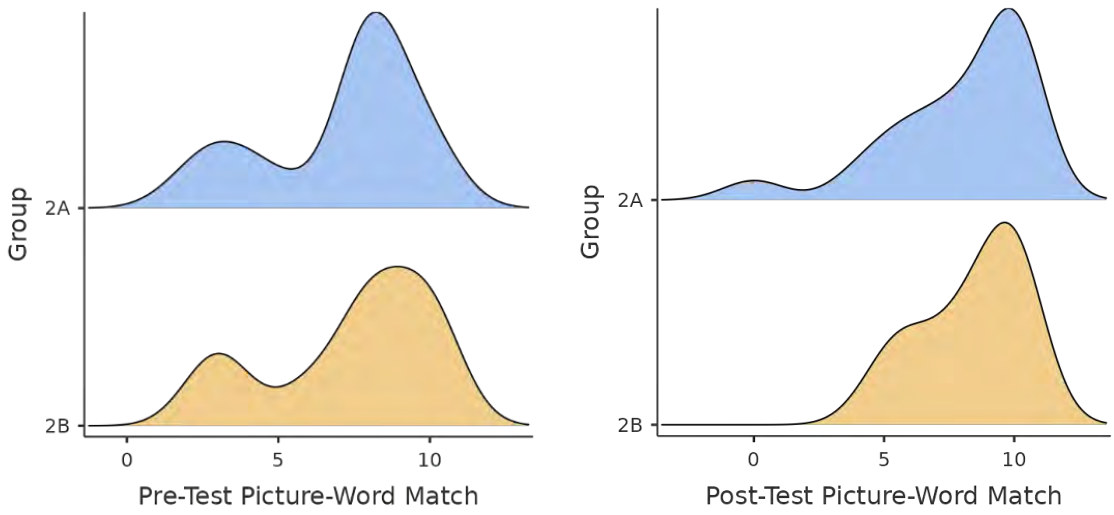


Figure 27 Pre- and post-test Picture-Word Match for Grade 2 experimental (2A) and control (2B) group

The Grade 3 learners also scored relatively high on average out of 10 (see Table 10), with pre-test mean scores of 7.68 (SD = 2.54) for experimental group and 9.71 for control group (SD = 1.07). For the post-test out of 10, the mean score Grade 3 experimental group is 8.42 (SD = 2.81) and 7.71 (SD = 4.21). The mean score was less for the Grade 3 control group post-video

viewing, and while this may indicate that the video was worse for this test, learners still scored relatively high and illustrated that they could connect word meaning to a corresponding visual prior to having seen the videos. Whether the decrease in score in the control group was significant is discussed in 5.2. More research could examine whether the speed of recognition here was affected instead of the results only, as this was not the scope of this thesis.

Descriptives for Grade 3 Experimental and Control Group

(3A is Experimental Group and 3B is Control Group)

			Pre-Test Picture-Word Match					Post-Test Picture-Word Match				
	Group	N	Mean	Median	SD	Max	Min	Mean	Median	SD	Max	Min
Score	3A	19	7.68	8.00	2.54	10.00	2.00	8.42	9.00	2.81	10.00	0.00
	3B	14	9.71	10.00	1.07	10.00	6.00	7.71	10.00	4.21	10.00	0.00

Table 10 Descriptives for Grade 3 experimental and control group

The density graphs in Figure 28 illustrate less errors and more ceiling effects for the test post-video viewing, however, there are more learners who could not complete the test in the control group post-video than before the video was shown.

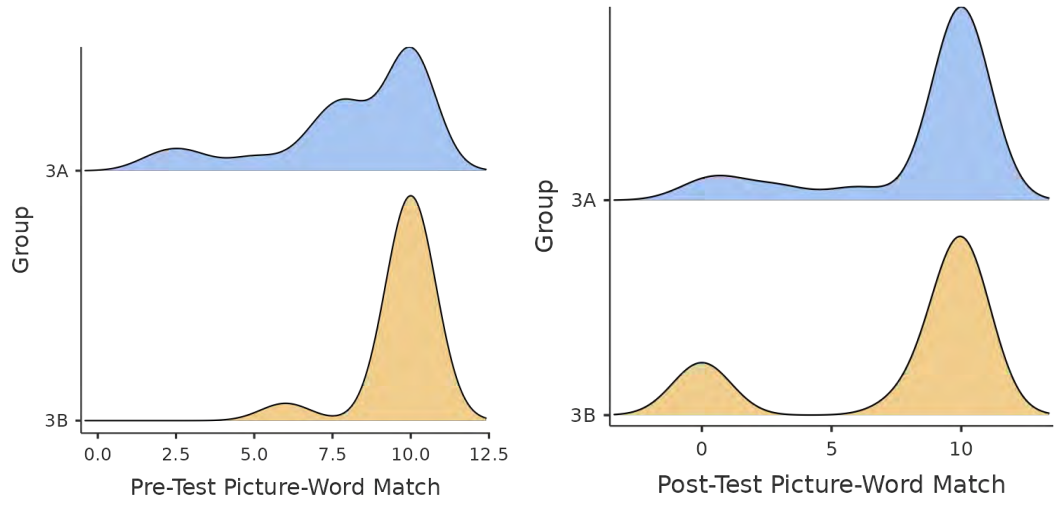


Figure 28 Pre- and Post-Test Picture-Word Match for Grade 3 Experimental (3A) and Control (3B) Group

The next section illustrates the paired-sample t-tests results that compare all pre- and post-test results.

5.2 T-test results and analysis

This section illustrates the paired sample t-tests conducted on the above descriptives and their results. As the sample size was small, the findings of the tests for Grade 2 and 3 are combined for both experimental and control groups to see if there was a statistically significant difference between pre- and post-test results. This is to ensure t-test requirements are met for accuracy in statistical analysis overall to determine whether there was an improvement in ORF reading scores and whether this was due to the intervention. Table 4 presented Mann-Whitney U test results for the bimodal distributed pre-test data and was presented in 5.1 to illustrate that despite occurrences in which the control groups had slightly higher initial mean scores than the experimental group mean scores, there was no significant difference between any mean scores at the start of the intervention.

5.2.1 describes results that focus on all pre- and post-test results for experimental and control groups which combined Grades 2 and 3, including t-tests and nonparametric Wilcoxon Signed-Rank tests**.** The paired sample t-tests are to examine if there was an improvement in literacy scores during the treatment period for both groups. The Mann Whitney U of post-test results is also presented in this section to determine whether any improvements were due to the intervention and video exposure. 5.2.2 describes non-normal paired sample t-tests (Wilcoxon Signed-Rank) for post-test Word List 1 and 2 results. Average reading scores from pre- and post-test Long Passage scores are calculated for each grade in 5.2.2.1. This is to ascertain if there was a significant difference in ORF scores between words from the videos and words that were not included in the videos. This is a further test to examine whether the intervention did or did not assist in learners ORF scores. This is to answer research questions 5-9.

5.2.1 Pre- and post-test results for experimental vs control groups for combined grades

This section first draws on Mann-Whitney U test results to ascertain whether there was a significant difference between experimental and control group (combined grades) post-intervention. Table 4 had already indicated there was no significant difference in pre-test results between experimental and control groups (combined grades). Independent sample t-tests of post-tests results illustrate that the intervention was not the reason for the significant difference in means between pre- and post-test results. This is illustrated in Table 11. Descriptives for these tests are in Table 2 in 5.1.

Independent Samples T-Test

		Statistic	P
Post-Test Word List 1 WCPM	Mann-Whitney U	511	0.239
Post-Test Word List 1 SCPM	Mann-Whitney U	514	0.253
Post-Test Word List 2 WCPM	Mann-Whitney U	557	0.531
Post-Test Word List 2 SCPM	Mann-Whitney U	569	0.624
Post-Test Long Passage WCPM	Mann-Whitney U	568	0.615
Post-Test Long Passage SCPM	Mann-Whitney U	573	0.662
Post-Test Picture-Word Match	Mann-Whitney U	582	0.707

Table 11 Post-Test Mann-Whitney U independent sample t-tests for experimental and control groups combining grades

Results from paired sample parametric and nonparametric t-tests are discussed in this section, in order to determine whether mean scores improved between pre- and post-test. For easier reading, paired sample results and discussion are divided into experimental group (5.2.1.1) and (5.2.1.2).

5.2.1.1 Experimental group (combined grades) paired sample t-tests

Based on the results for normality in Table 12, standard t-tests were conducted on results for pre-test Word List 1 for both WCPM ($p = 0.107$) and SCPM ($p = 0.334$).

			W	P
Pre-test Word List 1 WCPM	-	Post-Test Word List 1 WCPM	0.951	0.107
Pre-test Word List 1 SCPM	-	Post-Test Word List 1 SCPM	0.967	0.334
Pre-test Word List 2 WCPM	-	Post-Test Word List 2 WCPM	0.897	0.002
Pre-test Word List 2 SCPM	-	Post-Test Word List 2 SCPM	0.891	0.002
Pre-Test Long Passage WCPM	-	Post-Test Long Passage WCPM	0.896	0.002
Pre-Test Long Passage SCPM	-	Post-Test Long Passage SCPM	0.854	< .001
Pre-Test Picture-Word Match	-	Post-Test Picture-Word Match	0.909	0.005

Note. A low p-value suggests a violation of the assumption of normality

Table 12 Shapiro-Wilk normality test for experimental group (combined Grade 2 and 3) for paired sample t-tests

Shapiro-Wilk test results are presented in Table 12. Nonparametric Wilcoxon Signed-Rank tests were conducted on the results for pre-tests Word List 2 WCPM (p = 0.002) and SCPM (0.002), pre-test Long Passage WCPM (p = 0.002) and SCPM (p = <.001), and pre-test Picture-Word Match (p = 0.005). Descriptives for paired samples are presented in Table 13. The nonparametric tests results are presented in Table 14.

	N	Mean	Median	SD	SE
Post-Test Word List 1 WCPM	37	26.24	30.00	17.57	2.888
Pre-test Word List 1 WCPM	37	19.19	19.00	15.27	2.510
Post-Test Word List 1 SCPM	37	58.27	65.00	39.99	6.574
Pre-test Word List 1 SCPM	37	40.92	39.00	33.97	5.584
Post-Test Word List 2 WCPM	37	25.59	30.00	17.01	2.796
Pre-Test Word List 2 WCPM	37	20.16	20.00	16.09	2.645
Post-Test Word List 2 SCPM	37	57.08	63.00	39.01	6.414
Pre-test Word List 2 SCPM	37	43.22	42.00	35.52	5.839
Post-Test Long Passage WCPM	37	22.84	27.00	15.00	2.465
Pre-Test Long Passage WCPM	37	17.59	20.00	12.99	2.135
Post-Test Long Passage SCPM	37	75.78	84.00	51.07	8.395
Pre-Test Long Passage SCPM	37	58.35	67.00	43.68	7.181
Post-Test Picture-Word Match	37	8.16	10.00	3.05	0.502
Pre-Test Picture-Word Match	37	7.41	8.00	2.80	0.461

Table 13 Descriptives for experimental paired sample t-tests (combined grades)

The results from the pre-test (M = 19.19, SD = 17.57) and post-test (M = 26.24, SD = 15.27) for Word List 1 WCPM indicate that the learners isolated word reading score significantly improved, $t(36) = 5.98$, $p = <.001$. A significant improvement is also observed for the Word List 1 SCPM pre-test (M = 40.92, SD = 33.97) and post-test (M = 58.27, SD = 39.99), $t(36) = 6.31$, $p = <.001$. This is illustrated in Table 14. However, Table 11 has illustrated that this improvement was not due to the video exposure from the intervention.

Paired Samples T-Test

			Statistic	Df	P	Mean difference	SE difference		Effect Size
Post-Test Word List 1 WCPM	Pre-test Word List 1 WCPM	Student's t	5.98	36.0	< .001	7.05	1.18	Cohen's d	0.983
Post-Test Word List 1 SCPM	Pre-test Word List 1 SCPM	Student's t	6.31	36.0	< .001	17.35	2.75	Cohen's d	1.038

Table 14 T-test results for post-test Word List 1 WCPM and SCPM in experimental group (combined grades)

As illustrated in Table 15, there is a significant difference between the pre- and post-tests for Word List 2 WCPM (W = 45.5, z = 3.83, p = <.001), Word List 2 SCPM (W = 47.0, z = 3.93, p = <.001). There is also a significant difference between the pre- and post-tests for Long Passage WCPM (W = 29.0, z = 3.96, p = <.001), pre-test Long Passage SCPM (W = 34.5, z = 3.95, p = <.001) and Picture-Word Match (W = 97.5, z = 2.01, p = 0.045) in the experimental group. However, Table 11 has illustrated that this improvement was not due to the video exposure from the intervention.

Paired Samples T-Test

			Statistic	P	Mean difference	Z	Effect Size
Pre-Test Word List 2 WCPM	Post-test Word List 2 WCPM	Wilcoxon W	45.5 [a]	< .001	5.43	3.839	0.701
Pre-Test Word List 2 SCPM	Post-test Word List 2 SCPM	Wilcoxon W	47.0 [b]	< .001	13.86	3.932	0.706
Pre-Test Long Passage WCPM	Post-Test Long Passage WCPM	Wilcoxon W	29.0 [d]	< .001	5.24	3.957	0.748
Pre-Test Long Passage SCPM	Post-Test Long Passage SCPM	Wilcoxon W	34.5 [e]	< .001	17.43	3.947	0.733
Pre-Test Picture-Word Match	Post-Test Picture-Word Match	Wilcoxon W	97.5 [f]	0.045	0.76	2.008	0.394

[a] 7 pair(s) of values were tied

[b] 6 pair(s) of values were tied

[d] 9 pair(s) of values were tied

[e] 8 pair(s) of values were tied

[f] 11 pair(s) of values were tied

Table 15 Paired t-test experimental group combined Grades 2 and 3 (Wilcoxon Signed-Rank)

5.2.1.2 Control group (combined grades) paired sample t-tests

Based on the results for normality in Table 16, standard t-tests were conducted on pre- and post-test results for Word List 1 for both WCPM (p = 0.107) and SCPM (p = 0.204). They were also conducted for Word List 2 for both WCPM (p = 0.321) and SCPM (p = 0.333). Due to the violation of the normality assumption, Wilcoxon Signed-Rank tests were conducted on Long Passage WCPM (p = 0.008), SCPM (p = 0.005) and Picture-Word Match (p = <.001).

Normality Test (Shapiro-Wilk)

			W	P
Pre-Test Word List 1 WCPM	-	Post-test Word List 1 WCPM	0.948	0.117
Pre-Test Word List 1 SCPM	-	Post-test Word List 1 SCPM	0.956	0.204
Pre-Test Word List 2 WCPM	-	Post-test Word List 2 WCPM	0.963	0.321
Pre-Test Word List 2 SCPM	-	Post-test Word List 2 SCPM	0.964	0.333
Pre-Test Long Passage WCPM	-	Post-Test Long Passage WCPM	0.907	0.008
Pre-Test Long Passage SCPM	-	Post-Test Long Passage SCPM	0.899	0.005
Pre-Test Picture-Word Match	-	Post-Test Picture-Word Match	0.840	< .001

Note. A low p-value suggests a violation of the assumption of normality

Table 16 Shapiro-Wilk normality test for control group (combined Grade 2 and 3) for paired sample t-tests

Descriptives are presented in Table 17. T-test results are presented in Table 18. The nonparametric tests results are presented in Table 19 and Table 20.

Descriptives

	N	Mean	Median	SD	SE
Pre-test Word List 1 WCPM	33	21.82	19.0	16.28	2.833
Post-Test Word List 1 WCPM	33	30.36	34.0	18.22	3.172
Pre-test Word List 2 WCPM	33	23.27	26.0	16.73	2.913
Post-Test Word List 2 WCPM	33	28.18	32.0	17.31	3.013
Pre-test Word List 1 SCPM	33	47.39	38.0	36.66	6.381
Post-Test Word List 1 SCPM	33	67.58	73.0	41.36	7.200
Pre-test Word List 2 SCPM	33	51.76	52.0	39.00	6.790
Post-Test Word List 2 SCPM	33	61.76	68.0	38.69	6.735
Pre-Test Long Passage WCPM	33	18.64	21.0	15.51	2.701
Post-Test Long Passage WCPM	33	24.09	25.0	16.72	2.911
Pre-Test Long Passage SCPM	33	60.58	67.0	50.91	8.863
Post-Test Long Passage SCPM	33	81.70	78.0	57.81	10.064
Pre-Test Picture-Word Match	33	8.39	10.0	2.40	0.417
Post-Test Picture-Word Match	33	8.09	10.0	3.06	0.532

Table 17 Descriptives for control paired sample t-tests (combined grades)

There is also a significant difference between the control group pre-tests (M = 21.82, SD = 16.28) and post-tests (M = 30.36, SD = 18.22) (combined grades) for Word List 1 WCPM (illustrated in Table 18), $t(32) = 5.14$, $p = <.001$. The results from the pre-test (M = 47.39, SD = 36.66) and post-test (M = 67.58, SD = 41.36) for Word List 1 SCPM indicate that the learners isolated word reading score significantly improved, $t(32) = 5.48$, $p = <.001$. A significant improvement is also observed for the Word List 2 WCPM pre-test (M = 23.27, SD = 16.73) and post-test (M = 28.18, SD = 17.31), $t(32) = 3.71$, $p = <.001$. Furthermore, a significant improvement is also observed for the Word List 2 SCPM pre-test (M = 51.76, SD = 39.0) and post-test (M = 61.76, SD = 38.69), $t(32) = 3.11$, $p = 0.004$. However, Table 11 has illustrated that this improvement was not due to the video exposure from the intervention.

			Statistic	Df	P	Mean difference	SE difference		Effect Size
Post-Test Word List 1 WCPM	Pre-test Word List 1 WCPM	Student's t	5.14	32.0	< .001	8.55	1.66	Cohen's d	0.895
Post-Test Word List 1 SCPM	Pre-test Word List 1 SCPM	Student's t	5.48	32.0	< .001	20.18	3.68	Cohen's d	0.954
Post-Test Word List 2 WCPM	Pre-test Word List 2 WCPM	Student's t	3.71	32.0	< .001	4.91	1.32	Cohen's d	0.645
Post-Test Word List 2 SCPM	Pre-test Word List 2 SCPM	Student's t	3.11	32.0	0.004	10.00	3.22	Cohen's d	0.541

Table 18 T-test results for post-test Word List 1 and 2 WCPM and SCPM in control group (combined grades)

As illustrated in Table 19, there is a significant difference between the pre- and post-tests in the control group for Long Passage WCPM (W = 75.5, z = 3.063, p = 0.002), pre-test Long Passage SCPM (W = 57.5, z = 3.450, p = <.001). However, Table 11 has illustrated that this improvement was not due to the video exposure from the intervention.

Paired Samples T-Test

			Statistic	P	Mean difference	Z	Effect Size
Pre-Test Long Passage WCPM	Post-Test Long Passage WCPM	Wilcoxon W	75.5 [a]	0.002	5.45	3.063	0.569
Pre-Test Long Passage SCPM	Post-Test Long Passage SCPM	Wilcoxon W	57.5 [a]	< .001	21.12	3.450	0.641

[a] 4 pair(s) of values were tied

Table 19 Long Passage paired samples t-test Grade 2 and 3 control groups combined (Wilcoxon Signed-Rank)

Furthermore, as illustrated in Table 20 there was no significant difference between pre- and post-test Picture-Word Match results (W = 72.0, p = 0.849) in the control group.

Paired Samples T-Test

			Statistic	P
Pre-Test Picture-Word Match	Post-Test Picture-Word Match	Wilcoxon W	72.0 [a]	0.849

[a] 16 pair(s) of values were tied

Table 20 Picture-Word Match paired samples t-test Grade 2 and 3 control groups combined (Wilcoxon Signed-Rank)

5.2.2 Post-test comparison of Word Lists 1 and 2 for the groups combined grades

This section first summarises Mann-Whitney U test results to ascertain whether there was a significant difference between experimental and control group per Grade 2 and 3 post-interventions. Table 4 had already indicated there was no significant difference in pre-test results between experimental and control groups (combined grades). Table 11 further already indicated there was no significant difference in post-test results between experimental and control groups (combined grades). Therefore, there was no significant difference between experimental and control group post-test, indicating that the videos did not significantly influence learners' literacy development.

This following sub-sections examine a comparison of post-test Word List 1 (words in the video) WCPM and SCPM and Word List 2 (words not in the video) WCPM and SCPM for combined grades. This is to illustrate whether there was a significant difference between post-tests Word List 1 and 2 WCPM and SCPM scores in the experimental group within the treatment.

Based on the results for normality in Table 21, nonparametric t-tests were conducted on results for post-test Word List 1 and 2 for WCPM (p = 0.012) and SCPM (p = <.001).

Normality Test (Shapiro-Wilk)

			W	P
Post-Test Word List 1 WCPM	-	Post-Test Word List 2 WCPM	0.921	0.012
Post-Test Word List 1 SCPM	-	Post-Test Word List 2 SCPM	0.881	< .001

Note. A low p-value suggests a violation of the assumption of normality

Table 21 Shapiro-Wilk normality test for Grade 2 experimental group for paired sample t-tests

Descriptives are presented in Table 22. Nonparametric t-test results for post-test Word List 1 and 2 WCPM and SCPM are presented together in Table 23 for the experimental group.

Descriptives

	N	Mean	Median	SD	SE
Post-Test Word List 1 WCPM	37	26.2	30.0	17.6	2.89
Post-Test Word List 2 WCPM	37	25.6	30.0	17.0	2.80
Post-Test Word List 1 SCPM	37	58.3	65.0	40.0	6.57
Post-Test Word List 2 SCPM	37	57.1	63.0	39.0	6.41

Table 22 Descriptives of Grade 2 experimental group for paired sample t-tests

There is no significant difference in particular between Word List 1 (M = 26.2, SD = 17.6) and Word List 2 (M = 25.6, SD = 17.0) post-test scores for WCPM (see Table 23), W = 132.0, p = 0.410 and SCPM (see Table 23), W = 138.0, p = 0.340. Despite the average reading score improvement for these learners in 3 months, there is no significant improvement between Word List results for the experimental group (combined grades). This further indicates that, for both Grade 2 and 3 experimental groups, there is no significant improvement between Word List results. These results demonstrate that the videos as a simple passive addition were not helpful in improving reading rate.

Paired Samples T-Test

			Statistic	p	Mean difference	Z	Effect Size
Post-Test Word List 2 WCPM	Post-Test Word List 1 WCPM	Wilcoxon W	132 [a]	0.410	0.648	0.823	0.165
Post-Test Word List 2 SCPM	Post-Test Word List 1 SCPM	Wilcoxon W	138 [b]	0.340	1.189	0.955	0.187

[a] 12 pair(s) of values were tied

[b] 11 pair(s) of values were tied

Table 23 Paired samples nonparametric test (Wilcoxon Signed-Rank) of post-test between experimental group (combined grades)

5.2.2.1 Average reading scores from pre- and post-test results by grade

I calculate overall grade mean scores for Long Passage reading to triangulate with the subtitle rate in the videos to determine whether learners could read the subtitles at the beginning of the year and post-intervention. I have done this for both experimental and control group, as Table 11 illustrated no significant difference from the intervention on reading scores and to get a representative mean of each grade pre- and post-test. The mean scores pre-test for the experimental group in Grade 2 are 15.4 WCPM (SD = 9.77) and 50.6 SCPM (SD = 32.3). The Grade 2 control group read at mean scores of 13.2 WCPM (SD = 14.0) and 40.2 SCPM (SD = 40.4) in the pre-test. The mean pre-test for both experimental and control group added together and divided by 2 is 14.3 WCPM (45.40 SCPM) for Grade 2.

The post-test mean scores for the Grade 2 experimental group are 21.8 WCPM (SD = 13.5) and 70.1 SCPM (SD = 44.8) in the Long Passage reading. The Grade 2 control group read at 18.1 WCPM (SD = 14.8) and 60.3 SCPM (SD = 49.9) in the post-test. The mean post-test for Grade 2 overall is 19.95 WCPM (65.20 SCPM).

Since there was no significant difference from the videos, I calculate overall Grade 3 mean scores to triangulate with the subtitle rate to determine whether learners could read the subtitles at the beginning of the year and three months into the school year. The pre-test mean scores of the Grade 3 experimental group are 19.7 WCPM (SD = 15.4) and 65.7 SCPM (SD = 52.1). The pre-test mean scores of the Grade 3 control group are 26 WCPM (SD = 14.8) and 88.2 SCPM (SD = 51.8). The mean pre-test for both experimental and control group added together and divided by 2 is 22.85 WCPM (76.95 SCPM).

The post-test mean scores of the Grade 3 experimental group are 23.8 WCPM (SD = 16.6) and 81.2 SCPM (SD = 57.1). The post-test mean scores of the Grade 3 control group are 32.2 WCPM (SD = 16.1) and 111 SCPM (SD = 56.6). The mean post-test for Grade 3 overall is 28 WCPM (96.10 SCPM). For easy reference, this is presented in Table 24.

	Grade	Pre-Test EXP Mean	Pre-Test CTRL Mean	Total Pre-Test Mean	Post-Test EXP Mean	Post-Test CTRL Mean	Total Post-Test Mean
Long Passage WCPM	2	15.4	13.2	14.3	21.8	18.1	19.95
Long Passage SCPM	2	50.6	40.2	45.40	70.1	60.3	65.20
Long Passage WCPM	3	19.7	26	22.85	23.8	32.2	28.00
Pre-Test Long Passage SCPM	3	65.7	88.2	76.95	81.2	111.0	96.10

Table 24 Average total mean scores for Grades 2 and 3 pre- and post-test

5.3 Conclusion

This section discussed intervention results and highlighted that, while reading scores improved between pre-and post-tests for both grades and for both experimental and control group, the Mann-Whitney U tests confirmed this was not due to treatment. Furthermore, there was no significant difference found between pre-tests between groups and grades, illustrating they were all at a similar initial point when the intervention was conducted. There was also no significant difference found between post-tests in the treatment, and no significant difference found between post-test Word Lists 1 and 2. This could be due to prior exposure in the pre-test and external teaching factors that were unrelated to the intervention, and the intervention did not cause a decrease in WCPM or SCPM rates for any tests in the experimental group. Maximum WCPM rates, irrespective of word length, were no more than 50 WCPM for Grade 2 and 3, with average scores lower for Grade 2 learners than Grade 3. Mid-way through the year, the Grade 2 experimental group had reached the benchmark for the end of Grade 2 outlined by Ardington *et al*. (2020) of 20 WCPM, while the control group had reached slightly under the benchmark. The Grade 3 learners in both groups did not reach the benchmark for the end of Grade 3 outlined by Ardington *et al*. (2020) of 35 WCPM.

Due to the nature of the video texts as longer texts, the average mean scores of pre- and post-tests per grade in 5.2.2.1 of longer passage texts are used when examining subtitle rate for the videos in Chapter 6.

The high scores in the Picture-Word Match test indicate the learners' ability prior to video exposure to relate meaning correctly between visual and linguistic modes. In the Word List and

Long Passage oral reading tests, there was improvement in average reading scores in the three months of the intervention, however there was no significant difference found between any of the tests and Word Lists indicating that the videos may not have influenced reading scores or visual comprehension at all. Findings were difficult to determine with the small sample size, even though the size in this study represented the entire school. Further research would include more participants to confirm the results, as well as examine other contexts. Nevertheless, the reason the videos did not assist in reading could be due to the lack of prior subtitle exposure, the additional nature of the resource that is not integrated into the curriculum, and potential principles related to the video design and subtitle rate.

There could be a number of reasons why the videos were not effective:

1. As Dal (2010) indicated with adult FL learners, videos as passive tools do not generate enough opportunity for learning to occur. If the videos were integrated as part of the curriculum, it could potentially be more of an aid than an additional resource.
2. The language isiXhosa has no prior subtitle culture and the learners have no subtitle exposure (cf. Kuosmanen 2020 in 2.2.2). Subtitles differ from static text in their limited screen time and the lack of prior exposure may cause difficulty in reading subtitles and fully utilising them to assist in improving reading rate. This further supports the Pre-Training Principle stipulated by Mayer (2005) which states learners will find content easier if they have prior knowledge of it. This not only includes words and meanings the learners have not been exposed to before, but also the subtitles as a mode of learning.
3. This is further compounded with the problem of learners unable to follow the subtitles simply due to the subtitle speed being over their reading rate. The subtitle rate of these videos is examined in Chapter 6. to determine if this would have been a factor causing the videos to not assist in reading.
4. This research seems to support Mayer (2005)'s Redundancy and Modality Principle. These principles state that visual text, narration and visuals often cause less learning to occur when all three are present and that spoken language and visuals are better for learning. The presence of three modes to process that may cause extra cognitive load may be a reason why the videos were effective resources for aiding literacy for these learners.
5. Mayer's (2005) Extraneous Principle states that the more unnecessary information for the goal of instruction present in media, the less cognitive capacity a learner may have for important information. The Signaling Principle further indicates that when units of

meaning are highlighted, learning is improved. In Chapter 6, the DMDA shall highlight the time spent on emphasising specific meanings in the video and this can illustrate whether their design adheres to these principles, or whether the design of the videos could be improved in order for them to be helpful for learners.

6. This is also true of the Spatial and Temporal Contiguity Principles highlighted by Mayer (2005) that state if visuals and words are presented closer together or simultaneously, learning is improved. This is shown in Chapter 6. with the amount of Similarity in meaning between modes.

Whether the video design adheres to prime learning conditions outlined by Mayer (2009) is examined in Chapter 6. The next section describes the DMDA of the five videos used in this intervention.

Chapter 6 – DMDA and triangulation

This chapter examines the five isiXhosa videos that were used in the intervention using a Digital Multimodal Discourse Analysis (DMDA). This is in order to answer research questions 1-4. The diagrams used are STMs as mentioned in 4.3.2.5.

Each of the sections 6.1, 6.2, 6.3, 6.4 And 6.5 are the analyses of Video 1, 2, 3, 4, and 5 respectively. They all have the same layout with the first sub-heading describing STMs for the ideational metafunction for the spoken and written linguistic mode. This refers to linguistic Participants, Processes and Circumstances which explain the predominant characters, objects and actions in the scene. This is relevant to understand the predominant linguistic systems of meaning in the story and whether they co-occur and have a high frequency in the visual mode, consequently either adhering to Mayer's (2009) Spatial and Temporal Contiguity Principles and potentially assisting children in their reading. The second sub-headings discuss the interpersonal and textual metafunctions for this mode, which relates to Mood and Theme STMs specifically. This is discussed in order to understand whether the information is given or requested and to see what is further emphasised or not to assist in learning, as this could assist adherence to Mayer's (2009) Signaling Principle. The third sub-headings discuss the visual ideational metafunction, which discusses visual Participants and Processes, further highlighting predominant actions and characters and how the information in the video is presented. The fourth sub-headings discuss the visual interpersonal metafunction and the systems of Social Distance, Zoom, Gaze, Camera Movement with Horizontal and Vertical Perspective and Salience, which further illustrates particular Participant emphasis, particularly in relation to the viewer. This also could illustrate adherence to Mayer's (2009) Signaling Principle in video design. The fifth sub-headings examine the visual textual metafunction, specifically Information Value and Framing, to further illustrate visual organisation and emphasis and further could assist in adherence to Mayer's (2005) Signaling Principle. The sixth sub-heading discusses Intersemiotic Texture and how the visual, spoken and written linguistic modes create either similar or contrasting meanings and how this occurs. This could assist in adherence to Mayer's (2009) Coherence, Spatial and Temporal Contiguity Principles. The final sub-heading relates to the subtitle rate of each video.

Section 6.1 describes the first analysis of Video 1, *The Icky Sticky Frog* or *ISele eLibi eliNcangathi.* This video includes the story of a frog, "Usele-sele", named like the isiXhosa

word for frog (*isele*). For ease of reference, in-text I discuss "Usele-sele" by the translated name Sele-sele. Sele-sele eats first a fly, then a locust and then tries to eat a butterfly before being swallowed himself by a fish. The video ends with the butterfly flying towards the viewer.

Section 6.2 describes the analysis of Video 2, *They All Loved Sue.* The title, translated to *Bonke babemthanda USue*, has been said to be an incorrect translation by native speakers as the prefixes are not human, but animals. This video includes the story of Sue (*USue* in the title) who loves all kinds of animals but is not allowed to have any at home. She would meet many different animals, particularly dogs and cats, along the way to work and would start bringing snacks for them, thus befriending them. Soon they would follow her to work, making Sue very happy and she opened her own business to look after the animals.

Section 6.3 describes the analysis of Video 3, *The Moon Followed Me Home.* The title is translated to *iNyanga iNdilandele ukuya eKhaya*. This video is told by the narrator, a lion cub in the visuals, who describes his friend the moon (*inyanga* in the title) and how it is always with him. He also talks about his father who appears in the video and activities they do together, while the moon is with them.

Section 6.4 describes the analysis of Video 4, *Jungle Tumble.* The title is translated to *iiNtshukumo zaseNdle.* This video includes a wide variety of animals in the jungle: birds, monkeys, frogs, a hippo, leopards, a chameleon, insects, snakes, crocodiles, a giraffe, a mouse, a zebra, an elephant and a tortoise. The video reveals some hidden animals in the jungle, first a parrot behind a branch, then a frog behind a stone, then a leopard behind a bush, then a chameleon. Furthermore, the video reveals a snake behind a branch, then crocodiles behind bushes, a monkey and another leopard. There is no narrative plot unfolding chronologically in this video, as depicted events randomly occur.

Section 6.5. describes the analysis of Video 5, *I Can Go To School.* The title is translated to *Ndingaya esikolweni*. This video includes the story of the narrator, a little girl in the visuals, who describes her day at school from being dropped off by her father to talking to her teacher and the activities that she and the other children do throughout the day. These include drawing, reading to sleeping. The video ends with her being picked up by her mother.

The translation for the text of the videos is found in Appendix A.

6.1 Video 1 – The Icky Sticky Frog or ISele eLibi eliNcangathi

The analysis of this video is divided into spoken and written ideational metafunction STMs for language (6.1.1), and the interpersonal and textual metafunctions in these modes (6.1.2). Section 6.1.3 examines the ideational metafunction STMs in the visual mode and 6.1.4 examines the interpersonal metafunction STMs in the visual mode. Section 6.1.5 also examines the textual metafunction STMs in the visual mode and 6.1.6 examines Intersemiotic Texture. There is a discussion of subtitle rate in 6.1.7, which highlights the subtitle rate and how the rate may result in learners unable to read the subtitles. Lastly, 6.1.8 concludes the analysis of this video as having a high Similarity rate in Intersemiotic Texture and patterns that co-occur between modes that may be conducive to learning, according to Mayer's (2009) Principles of Spatial and Temporal Contiguity. Important spoken and/or visual Participants to note for this video include Sele-sele (the frog), the fly, the butterfly, locust, beetle and fish.

6.1.1 Spoken and written language ideational metafunction STMS

This section describes the STMs for the spoken and written Participants initially, before discussing spoken and written Process and Circumstance STMs.

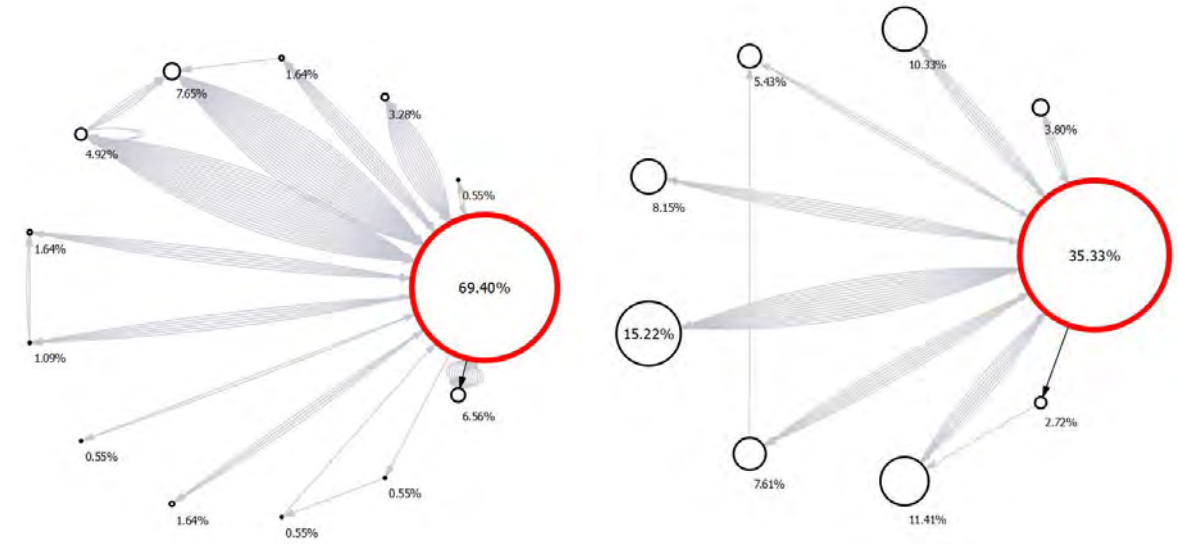


Figure 29(a) The Icky Sticky Frog spoken Participant STM (top left) and (b) written Participant STM (top right) with (c) screenshot subtitled "Usele-sele swallowed the beetle"

The predominant spoken Participant is Goal, which is 7.65% of the video (see Figure 29(a)), and these co-occur with Actor only for 1.64% of the video (the topmost state connected to the 7.65% state by a transition line). Thus, Goal Participants are the most predominant spoken Participants in terms of video duration (state) and frequency of occurrence (transition lines). In the written mode, 8.15% of the video is Phenomenon, however they occur in six subtitles of the video (transition lines) in Figure 29(b). This is in stative constructions, for example "Ayisabonakali" (the fly can no longer be seen). However, both Goal and Phenomenon

Participants are often the insects which Sele-sele eats, for example, the beetle in the sentence "Usele-sele wayiginya ibhungane" (Sele-sele swallowed the beetle) in Figure 29(c). Within Figure 29(b), Goal Participants are 3.80% of the video in the written mode, while the 15.22% represents the duration of written co-occurrence(s) of the Actor and Goal Participants. The predominant Actor is Sele-sele as it eats the insects (4.92% spoken language and 5.43% in the written language), as in the example sentence in Figure 29(c). While the 5.43% state in Figure 29(b) indicates Actor and is less than the other states in both duration and frequency, the full duration and frequency of the Actor Participant includes the 15.22% state and its transition lines. This illustrates the predominance of concrete actions in the video, which is typical of children's narratives (Guijarro and Sanz 2008). This higher duration and frequency of Goal Participants, particularly in the spoken language, is understandable considering the nature of the isiXhosa language. In isiXhosa, the noun can be omitted from a sentence and be present within the verb complex as a prefix, which often occurs with the agent noun in a clause. Furthermore, because nouns are often omitted, a higher percentage of Participant duration relates to direct objects. This indicates that Material Processes are predominant in the video. This does, however, cause the linguistic mode to have greater emphasis on the insects being eaten by Sele-sele in the video. This could assist with adherence to Mayer's (2009) Signaling Principle which states that highlighting essential words guide learners when processing knowledge and lead to more efficient learning. This is also important in Similarity and Mayer's (2009) Principles of Coherence, Spatial and Temporal Contiguity discussed later in 6.1.6.

For the spoken mode, 69.40% is uncategorisable for Participants and this lack of system choice is 35.33% in the written mode, as the subtitle duration is longer on screen than spoken units of language that unfold in smaller chunks over time. This causes two parallel sets of segmenting that, if they are presented simultaneously or closer to one another, learning is improved, which may assist the adherence of the Principles of Spatial and Temporal Contiguity (Mayer 2009). Furthermore, it may be optimal for learning if subtitles are presented at the pace of the learner, which adheres to Mayer's (2009) Segmenting Principle. This is discussed in 6.1.7. The narrator pauses often between specific words and clauses, which could assist emphasis and adherence to Mayer's (2009) Signaling Principle. It also allows the pace of the video to be easier for the learner as per Mayer's (2009) Segmenting Principle.

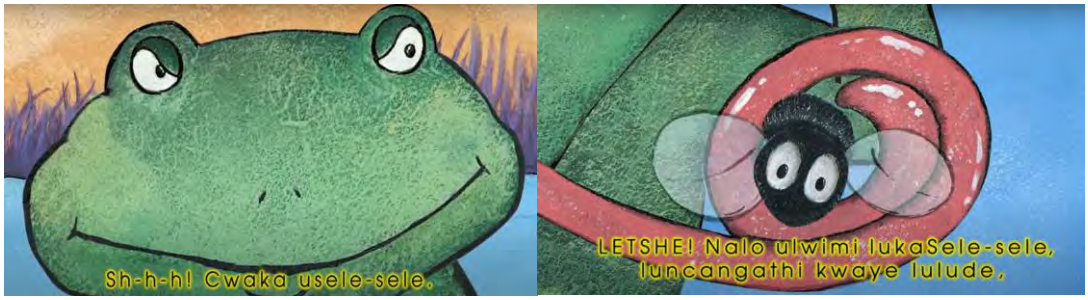


Figure 30 Screenshots from The Icky Sticky Frog, (a) subtitled (left): "Sh-h-h! Quiet Sele-sele," (b) subtitled (right): "LETSHE! This tongue of Sele-sele, sticky and long."

Another frequent Participant durationally is an Existent (6.56% spoken and 10.33% written). This occurs repeatedly with the phrase, for example Figure 30(b), "Nalo ulwimi lukasele-sele, luncangathi kwaye lulude" ([There is] that tongue of Sele-sele, sticky and long), where the demonstrative "nalo" (that/there is) acts as an Existential Process and the Existent is "ulwimi" (tongue). Existent Participants occur slightly less frequently (transition line) in the spoken mode than the Actor and Goal Participants. Between subtitles the 10.33% state of Existent Participant may not seem to occur frequently in the videos (few transition lines). However, in the written mode, 11.41% of the video has a combination of Actor and Existent Participants, where they co-occur as the second-most frequent system choice in the video after the co-occurrence of Actor and Goal. Figure 29(b) indicates that Participants for Material Processes occur more frequently (transition lines) and are more prevalent durationally (states) than Participants for Existential Processes.

An example of a co-occurrence of Existent and Actor Participants is in "gqi ibhungane elimbejembeje lirhubuluza lisondela" (a colourful beetle appears, crawling nearby). There is a 2.72% co-occurrence durationally in the written mode of Existent, Sayer and Verbiage, for example in Figure 31. According to Peña (2020:2), "deictic words are words that require contextual information and a point of reference in order to be interpreted correctly." Deixis can be categorised as time, person, and space (Levinson 2004). Time deixis consists of words that are specific to a date, time or reference point in which an event occurs (discussed with Circumstances of Location: Time later in this section). Person deixis consists of pronouns referring to specific Participants (these are simply categorised with Participants in this analysis). Existential Processes like "nalo" act as deixis of space which refers to a specific Participant in relation to its space and in comparison, to any other similar Participants around it. This kind of deixis requires contextual and spatial cues. Thus, the predominant Participants in the spoken and written modes are Actors, Goals and Existents. Material Processes and the

Participants may be predominant in children's narratives (Guijarro and Sanz 2008). Existents and Existential Processes conveyed by "nalo ulwimi" (there is/that tongue) allow for a deictic focus, which may assist adherence to Mayer's (2009) Signaling Principle for better literacy development.

Figure 31 Screenshot from The Icky Sticky Frog, subtitle "perched green Sele-sele, he kept quiet"

Other frequent Participants durationally are Behaver, which is the 1.64% state in the spoken language in Figure 29(a) with the transition line to the 1.09% Behaviour state. Behaver occurs marginally more frequent (transition lines) than Behaviour in the spoken mode, nevertheless, in the subtitles, they only appear together (7.61% of the video durationally, and third-most frequently). This is when the word "cwaka" (silent) appears on its own with Sele-sele. The nature of "cwaka" in isiXhosa as an adjective-like word with predicate properties has been discussed in previous research as an ideophone (see 1.2.1 and Andrason (2020)), e.g., in the phrase "Cwaka usele-sele" (see Figure 30(a)). The meanings of this ideophone allows us to assign it a Behavioural Process and Behaviour with Sele-sele as the Behaver. This word also appears in the phrase "ethe cwaka" (see Figure 31) which can be translated to "He remained silent" and in this case it has been categorised as Verbiage of the Verbal Process "ethe" and the Sayer is Sele-sele. Another ideophone construction within Material Processes includes "Waman' ukuyithi krwaqu le mpukane, ibhabhela kufuphi" (He kept glancing at this fly, flying nearby). In this example, "Waman'ukuyithi krwaqu," (He kept glancing at it [the fly]) which is another ideophone construction. The verb "-thi" in isiXhosa normally means "to say," however when combined with an ideophone like "cwaka" (silent) or "krwaqu" (glance), this causes the "-thi" to become semantically bleached and take on the meanings of the ideophone. There is very little research on how children read or learn to read ideophone constructions in

isiXhosa and further research is required to determine whether they are appropriate for a literacy resource in the Foundation Phase.

Figure 32(a)The Icky Sticky Frog Spoken Processes STM (left) and (b) Written Processes STM (right)

The Participants mentioned earlier are connected to specific predominant Processes durationally in the videos. The Actors and Goals exist with Material Processes such as "wayiginya" (he swallowed it) in Figure 29(c). Spoken Existential Processes (like "nalo") comprise 12.02% of the video durationally. The second-most predominant spoken Process is the Material Processes, which comprises 10.93% of the video durationally in Figure 32(a). However, there is a discrepancy between Process duration in the spoken and the written mode, as 24.04% of the video comprises of Material Processes in Figure 32(b) and written Existential Processes comprise 13.66% of the video. Nevertheless, there are slightly more frequency of occurrences (transition lines) for spoken Material Processes than spoken Existential Processes. This indicates that Material Processes occur more frequently in both linguistic modes than Existential Processes and that they are both predominant Processes in this video. As in the case of "nalo", words like these do not only act as Existential Processes that state that a tongue exists, but also have a deictic quality to them like a demonstrative, which would cause the sentence to be translated like "[There is this tongue] of Sele-sele, that is long and sticky)".

Written Behavioural Processes (e.g. “cwaka” in Figure 31(a)), are 7.65% of the video durationally in Figure 32(b). There are slightly more frequencies of occurrence (transition lines) between subtitles than Mental Processes (stative constructions) and the co-occurrence of Material and Existential Processes in both modes. Spoken Behavioural Processes comprise 1.64% of the video durationally in Figure 32(a). The spoken co-occurrence between Material and Existential Processes is in 0.55% of the video durationally and the topmost 8.20% written state in Figure 32(b). Mental Processes (stative constructions) only occur for 2.19% of the video durationally in the spoken mode and 8.20% (bottommost state) durationally in the written mode. Thus, written Mental Processes (stative constructions) and the co-occurrences of Material and Existential Processes occur with a higher duration than written Behavioural Processes (ideophones). Mental Processes (stative constructions) occur with a higher duration in the spoken mode than Behavioural Processes (ideophones) and Material-Existential co-occurrences. However, Behavioural Processes (ideophones) in both spoken and written modes occur more frequently (transition lines) than Mental Processes (stative constructions) or the co-occurrence of Material-Existential Processes. Verbal Processes also occur 2.73% of the video durationally in the written mode.

The predominant co-occurrences of multiple types of Processes are in a subtitle, as its duration is longer than spoken language. Nevertheless, co-occurrences, as well as discrepancies between duration and frequency of system choices may create complexity for the learner and may cause challenges for comprehension. This may increase learners’ cognitive load, which is not conducive for learning according to Mayer’s (2009) limited capacity assumption, which may influence adherence to Mayer’s (2009) Principles of Coherence, Redundancy, Spatial and Temporal Contiguity. However, due to the repetition of these co-occurrences in combination with representation in the visual mode, this might be alleviated more than in other videos. As discussed in 2.3.1.3, frequency of exposure may not be as integral for literacy development as Similarity and Contrast co-patterning between modes (Tseng and Djonov 2023). Thus, this thesis focuses on Comparative Relations and a Similarity of meanings between modes as a potentially more accurate indicator of literacy development than frequency or duration. Further research in this context is required.

The high frequency of Material and Existential Processes correlate with the frequent appearance of their Participants. This emphasises the story of the video, which is focused on the actions of Sele-sele in eating the insects and consequently being eaten himself. According

to Guijarro and Sanz (2008), this is typical of other genres of children's stories like picture books, illustrating that this is not unique to isiXhosa stories but that the story seems to adhere to broader genre narrative patterns.

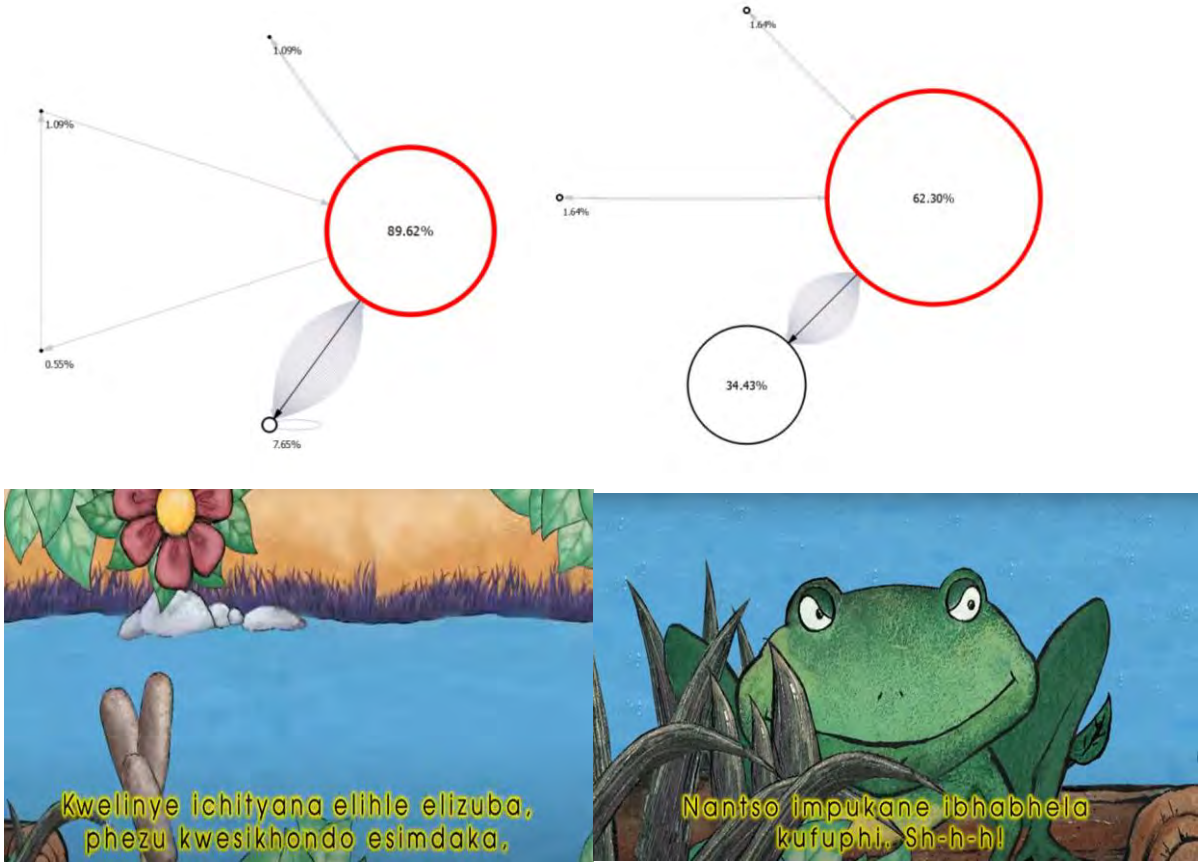


Figure 33 (a) The Icky Sticky Frog Spoken Circumstances STM (top left) and (b) Written Circumstances STM (top right) with screenshot (c) subtitled: "in some little blue lake, on a brown trunk," (left) and (d) "there is a fly flying nearby, Sh-h-h!" (right).

In Figure 33(a), the 7.65% state represents Location or Time, while this is represented by the 34.43% state in Figure 33(b) (see 3.1.1.1 on how Circumstance of Location is divided into Location: Place and Location: Time). As one can see, Location appears for a similar length in both modes in the video and is the most frequent Circumstance. Examples of Location Circumstances include "Kwelinye ichityana elihle elizuba" (in some little blue lake) in Figure 33(c) and "kufuphi" (nearby) in Figure 33(d). This is to emphasise both the setting (the lake) and the location of the insects and Sele-sele in the scenes. A predominance of Location Circumstances would be typical to try and gain a child's attention watching the video, as deictic words often have the role of directing someone's attention to a particular area or object in a space (Peña 2020). The duration of these predominant Circumstances in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In

Figure 33(b), the topmost 1.64% state is categorised as Manner, with the topmost 1.09% state in Figure 33(a). For example, in Figure 29(c) and Figure 34, “ngendlela efanayo” (in the same way as) in “Usele-sele walinginya ibhungane ngendlela efanayo naginya ngayo impukane” (Sele-sele swallowed the beetle in the same way he swallowed the fly).

In addition, there is a further categorisation of 1.09% of Circumstances of Manner that permeate across spoken embedded clause structures pictured in the leftmost state in Figure 33(a). The leftmost 1.64% state in the written mode is the co-occurrence of Circumstances of Manner within Embedded Clause Circumstances of Manner. For example, see Figure 34. In 0.55% of the video, there is a Manner Circumstance within a spoken Embedded Clause entirely categorised as Circumstance of Manner, for example in Figure 34. The complexity of embedded constructions, which often include a lot of adjectivised, and nominalised Processes might create challenges in decoding, due to the strain it may place on the cognitive load of the learner according to Mayer’s (2009) limited capacity assumption.

Figure 34 Screenshot from The Icky Sticky Frog, subtitlted "in the same way as Sele-sele swallowed the fly"

As mentioned in 3.1.1.1, meanings referring to knowledge of the world and ideas are conveyed by the ideational metafunction. There is a high frequency of Goal Participants, which is consistent with the language. In isiXhosa, the noun can be omitted from a sentence and be present within the verb complex as a prefix, which often occurs with the agent noun in a clause. Furthermore, because nouns are often omitted, a higher percentage of Participant duration relates to direct objects. Furthermore, because subjects are often omitted, a higher percentage of Participants are direct objects, and Material processes are predominant in the video. This does, however, cause the linguistic mode to have greater emphasis on the insects being eaten by Sele-sele in the video. As is described in 6.1.3, many linguistic Material Processes are echoed in the visual mode where Sele-sele swallows each individual insect which then adds a layer of similar meaning and emphasis to the language used. While Mayer and Moreno (2003)

warn against the redundancy of meaning repetition in three modes (spoken, written and visual imagery), repetition as a learning tool could assist learners and the emphasis on the Goal Participants in the visual and linguistic mode. This could not only adhere to Mayer's (2009) Signaling Principle, but also to Mayer's (2009) Principles of Spatial and Temporal Contiguity. Kothari *et al.* (2002)'s empirical research with the presence of same language subtitles suggests that one can learn in videos better with on-screen narration present. Particularly, if a learner is unfamiliar with the written word, having a spoken word to accompany the written word, along with the visuals, could assist in comprehending that word and larger phrases.

There are many Circumstances of Location. These seem to be a tool the producer is using to draw the viewer's attention to particular objects and maintain the viewer's attention in conjunction with the use of deictic words. The Similarity and Contrast between visuals and language is discussed more in 6.1.6 in relation to Intersemiotic Texture. The pattern of a high frequency of Goals and Actors, Material and Action Processes and Circumstances of Location as a multimodal coupling seems to be a syndrome within all five videos. This points to a pattern within children's literature and narrative stories that attempts to direct the children's attention to and create emphasis on various Participants and the Setting within the video. This emphasis may act as a type of signal which, according to Mayer' (2008) Signaling Principle, is conducive to learning.

6.1.2 Spoken and written language interpersonal and textual metafunction STMs

This section briefly discusses the predominant Mood within the linguistic interpersonal metafunction, as well as predominant Themes in the linguistic textual metafunction, for both spoken and written modes.

The Mood in this video was mostly Declarative (3.83% in the spoken mode and 67.21% in the written mode). The large discrepancy in percentages is due to the longer duration of subtitles on screen than the duration of these units in the spoken language. As there are no other Moods besides Declarative, a Mood STM is not pictured here, however this predominant Mood is typical for a children's narrative patterns offering information (Guijarro and Sanz 2008). While different Moods offer the benefit to learners of better engagement with the video, these Moods have different structures. Decoding and comprehending the information conveyed in the

different Moods may be challenging for isiXhosa L1 learners in the Foundation Phase, as there is limited research in this area. Too many variations in structure may create cognitive overload (see Mayer's (2009) limited capacity assumption).

Modality may, when it relates to Moods like Interrogative and Imperative, cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. If the Mood is Declarative, it may increase the load on cognitive capacity that violates the limited capacity assumption. Since there are only Declarative statements in *The Icky Sticky Frog*, this video has realisations of Modality that could increase learners' cognitive load, as they create complexity in the language. *The Icky Sticky Frog* has particularly complex realisations of Modality due to the *-s-* (negative of *-nga*) and *-akal-* (stative), as well as *-mana* (do continuously) affixes. For example, see Figure 41 in 6.1.5, "Waman' ukulithi krwaqu eli bhungane lirhubuluza kufuphi" (He kept glancing at the beetle crawling nearby). This specific modal construction realises usuality with high intensity (see 3.1.1.2) and with positive polarity, as mentioned in 6.3.1. This specific modal construction is repeated for each insect Sele-sele swallows, and thus Modality is repeated in each of these constructions. This is further repeated for the stative "Ayisabonakali" (It cannot be seen anywhere), with only the subject markers that vary according to the insect, to which they refer. This stative construction realises probability with high intensity and negative polarity. See (3) in 1.2.1.2 and (15) in 1.2.1.3. While the *-s-* affix triggers the realisation of Modality and not the stative *-akal-*, the complexity of this construction may limit learners' cognitive capacity to process other components of the video (see 3.2).

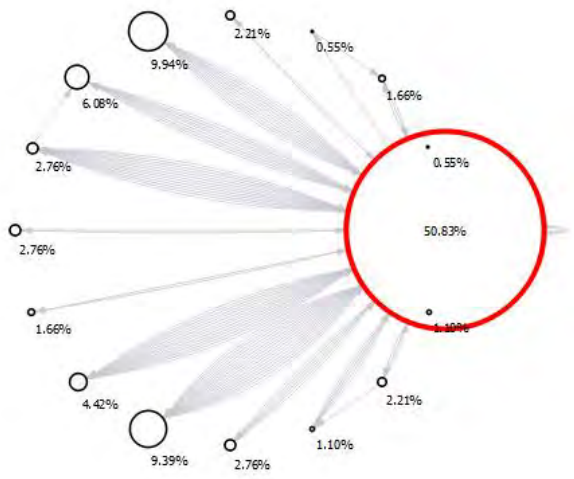


Figure 35 The Icky Sticky Frog spoken Theme STM

In Figure 35, the predominant Theme in the spoken language is “Usele-sele” (Sele-sele) (9.94%). Another predominant Theme is “nalo ulwimi” (there is that tongue, see Figure 30) which always follows the onomatopoeia “LETSHE!” (SHOOT!, in capitals as it occurs in the subtitles). This is present 9.39% of the time within the spoken mode. The latter state occurs more frequently (transition lines) than “Usele-sele” (Sele-sele). This is due to the lack of necessity in isiXhosa to always present a noun if it is already present in the verb complex as a prefix, thus the phrase “LETSHE! Nalo ulwimi” occurs more frequently than “Usele-sele” as Theme. Furthermore, the tongue belonging to Sele-sele indicates that the Theme, even though relating to the Existent Participant “ulwimi” (tongue) is directly related to the Participant “Usele-sele” (Sele-sele).

There is a 6.08% state that represents the Theme “waman’ukuyithi/ukulithi krwaqu” (He kept glancing). The verbs -yithi and -lithi with the ideophone “krwaqu” translate to “he kept glancing at it”, the different object markers relate to the different noun class prefixes of the insect words. The 4.42% state is a combination of the ideophone “SHWAKA!” (GONE!)[22] and “Ayisabonakali/Alisabonakali” (translated to “it can no longer be seen”). This structure, “Ayisabonakali/Alisabonakali,” is a stative construction repeated throughout the video (see 1.2.1.2). It further explains the high amount of Phenomenon Participants in the video, where the insect can no longer be seen (by anyone), but these types of constructions are challenging for Grade 3 children to understand, even if they were to decode them correctly (Rees 2016). Furthermore, in the visuals, these constructions not having any Participant visible creates Similarity, even when the Processes and Participants in these shots are not the same, as the overall meaning between the visual and linguistic modes is similar.

The 2.76% state is related to “S-h-h!” and the ideophone “cwaka” (silent). In terms of frequency of Theme occurrence, the ideophone “SHWAKA!” (GONE!) and the stative “Ayisabonakali/Alisabonakali” (it can no longer seen) occur the most frequently. The second-most frequent Theme occurrence is “S-h-h!” and the ideophone “cwaka” (silent) and the third-most frequent Theme occurrence is the other ideophone Theme “waman’ukuyithi/ukulithi krwaqu” (He kept glancing at it). Thus, aside from the ideophones “SHWAKA!” (GONE!) and onomatopoeia “Sh-h-h!” and “LETSHE,” the most frequently occurring Theme in relation to

[22] Andrason (2020)

Participant is Sele-sele, followed by the insects he eats, which are represented in the repeated stative constructions.

Thus, the predominant Themes relate to the predominant Actor Participant in spoken, written and visual modes, namely Sele-sele. The predominant Themes are usually Actors and Existents (e.g. “Usele-sele” (Sele-sele) and “nalo ulwimi” (there is that tongue)) within Action and Material Processes, and frequent repeated demonstratives and sound effects may assist in maintaining childrens’ attention. Nevertheless, there is a lack of simple consistency of regular Themes and this variation instead of consistency and simplicity could create further complexity in the plot and divert attention for children. Consequently, they may find it more difficult to comprehend meanings in the videos than if Themes were more frequent and consistent throughout the video. This video, however, is more consistent in its Themes than the other four videos. The Themes are discussed in relation to the visual textual metafunction components of Information Value and Framing in 6.1.4.

Another challenge and aspect to these videos in terms of the subtitles is the separation of a clause, often within linguistic groups or syntactic phrases. According to Lappalainen (2021), “words belonging together” (e.g. phrases in syntax and/or groups like nominal groups in SFL) should be in the same line of subtitle as per SLS conventions. In this video alone, there are multiple instances where, due to the length of words in isiXhosa, clauses are separated and while they often occur in the line underneath them (as there are often two lines of subtitle per subtitle), they can also occur in the following subtitle. This leads to their separation which could lead to confusion for viewers. For example, in Figure 45, “naginye ngayo impukane, emva koko wabona” is part of the bigger phrase “Usele-sele wayiginya intethe ebitsiba-tsibela kufuphi, ngendlela efanayo nale aginye ngayo ibhungane, naginye ngayo impukane, emva koko wabona, ibhabhathane elihle” (USele-sele swallowed the grasshopper jumping nearby in the same way he swallowed the beetle, and in the same way he swallowed the fly, after that he saw a beautiful butterfly). The beautiful butterfly, the Phenomenon Participant here, is only visible in the next subtitle with no marker for it in the verb to indicate it in the linguistic mode. Learners would have to assume by the presence of the visual Participant that the butterfly is the Phenomenon, but it does not reflect in the first segment of subtitle. Consequently, this could cause confusion due to the absence of the corresponding linguistic Participant.

6.1.3 Visual ideational metafunction STMs

This section examines the visual Participant and Process STMs and describes briefly how they relate to linguistic Participants and Processes. This and Intersemiotic Texture (and the relationship to the linguistic mode) is further described in 6.1.6.

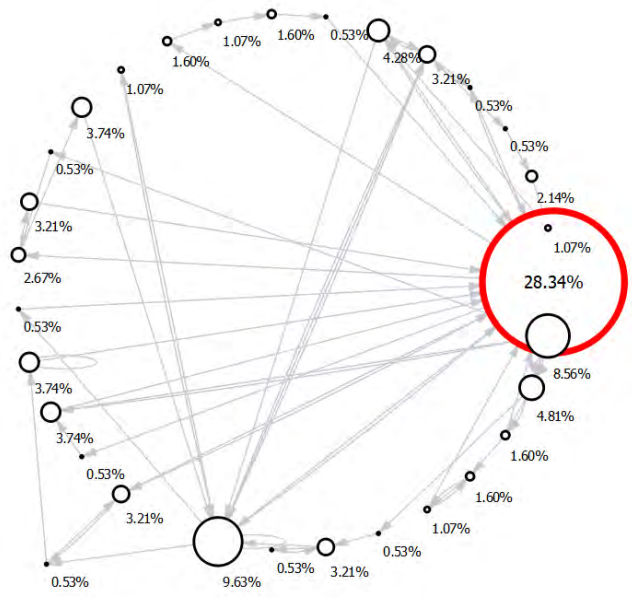


Figure 36 The Icky Sticky Frog visual Participants and Processes

The visual Participants and Processes are very similar to the Participants and Processes in the spoken and written modes, and, therefore, are presented together in one STM (see Figure 36). In Figure 36, 28.34% of the video is uncategorisable for visual Participants. The predominant visual Process is Action, with 9.63% relating to Sele-sele, who also appears as the predominant Actor and 8.56% of the video is a co-occurrence of Sensor and Actor for Sele-sele. Other predominant visual Action Processes and Participants occur in the 4.81% state of visual Action Processes, with Sele-sele as Actor and Sensor and fly as Actors. In 4.28% of the video, the butterfly is Actor. The 3.74% states from bottom to top consists of (1) Sele-sele and the lily pads as Actors and (2) Sele-sele as an Actor with the beetle as a Reactor and (3) Sele-sele and the locust as Actors. The 3.21% states from bottom right include only Sele-sele as Actor, the fly as Actor and Reactor, the beetle as Actor and Sele-sele, and the butterfly as Actors, and the locust as Actor and Sele-sele as Sensor. The left 2.67% state relates to the locust as Actor. There is also a 5.98% state relating to Mental Processes where Sele-sele thinks of all the insects he has consumed. This does not correlate to the Material Process in the language here and thus

is an instance of Contrast. It would have been preferred for learning had the language also contained Mental Processes, as this is rarely portrayed in visuals in videos that does not include Participants seeing objects or other Participants. The challenge of examining Mental Processes in the visuals is discussed in 6.6. Most of the Action Processes are related to Sele-sele swallowing the insects or any of the Participants that are moving. This correlates to the high frequency of Material Processes in the language, which assists in adherence to Mayer's (2009) Spatial and Temporal Contiguity Principles for learning with multimedia. This Similarity between Processes and Participants may help ease of reading. The Similarity between Processes and Participants between modes is discussed in more detail in 6.1.6.

6.1.4 Visual interpersonal metafunction STMs

The components within the interpersonal metafunction discussed in this section are Social Distance (Long, Medium and Close shots) and Zoom (In and Out). Thereafter Gaze (Direct or Indirect) is discussed, followed by the STMs of Camera Movement (Stationary, Pan, Tilt), Vertical Perspective (High angle, Eye Level, Low angle) and Horizontal Perspective (Front or Angled). Finally, Salience is discussed in relation to the Foreground, Background and Sharpness of Focus. This illustrates particular Participants that have a greater emphasis in the videos, and this may increase the focus on Participants, Processes and Circumstances in a manner that language alone cannot (Guijarro and Sanz 2008). This is due to the meanings establishing a relationship between the viewer and Participants and Processes within the story (Guijarro and Sanz 2008).

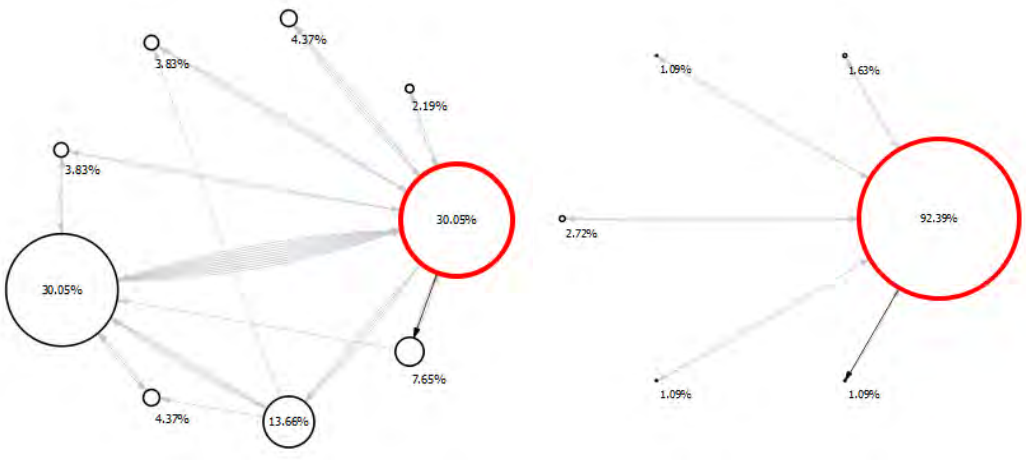


Figure 37(a) The Icky Sticky Frog Social Distance STM (left) and (b) Zoom STM (right)

Figure 37(a) illustrates that the majority of shots within Social Distance are focused on Sele-sele as Close (30.05%), Medium (7.65%) or Long (13.66%). Social Distance STMs do not account for trinary choices of shots, as some states indicate duration of shot transitions, in which there are co-occurrences of system choices. Thus, we only focus on the three system choice states, which are Close, Medium and Long. The duration of these predominant Shots and Zoom choices in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Long shots usually include another insect or the fish. The 4.37% states from the top state are Close shots of the butterfly and fly and 3.38% states from the top state are Close shots of the locust and the beetle. The Close shots here add a layer of intimacy and identification between the viewer and the characters (See Smith and Adendorff 2021; Guijarro and Sanz 2008). Figure 37(b) illustrates different emphasises created by Zoom in the video. Zoom In occurs 2.72% of the video durationally on Sele-sele, 1.63% of the video durationally on the butterfly and 1.09% durationally of the video comprises general Zoom In. Zoom Out occurs 1.09% of the video durationally from the butterfly. General Zoom Out occurs for 1.09% of the video durationally. This indicates that the majority focus is on Sele-sele. The next predominant focus is on the butterfly and the fly in comparison to the other insects.

Figure 38 The Icky Sticky Frog Gaze STM

In Figure 38, the 23.76% and 6.08% states belong to Sele-sele and he only ever has Indirect Gaze. The duration of these predominant Gaze choices in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). The other Participants all have Direct Gaze (the butterfly, 6.63% and the fly 4.42%), except the beetle (3.87%) and the locust (3.31%). Often, Sele-sele is present in the shot with other Participants, such as the fly (Indirect Gaze, 7.18% state) and the locust (Indirect Gaze, 4.97%). This Direct Gaze with the

audience is often made by the victims or prey of Sele-sele, which could lead the viewer to sympathise more with them, thus portraying Sele-sele as the antagonist or enemy character. While this seems to request information more than offer information and does not align with the Declarative Mood, if a Participant has Direct Gaze, the viewer is more likely to be engaged with them (Guijarro and Sanz 2008). This could assist in the adherence of Mayer's (2009) Signaling Principle of Participants with Direct Gaze that may be better for learning.

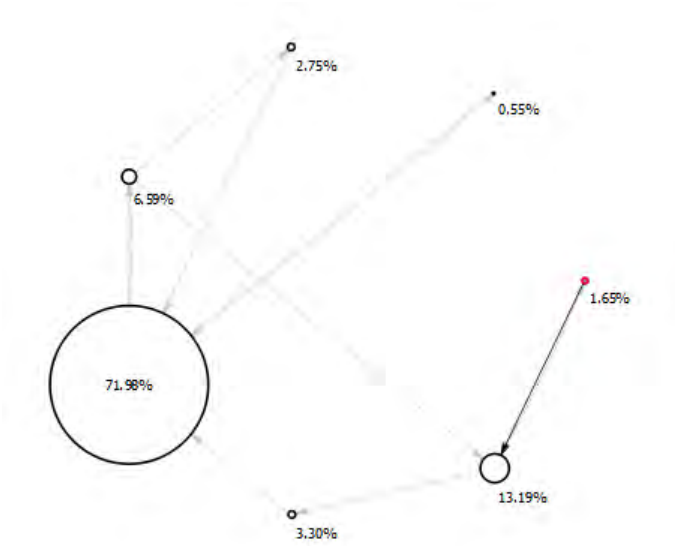


Figure 39 The Icky Sticky Frog Camera Movement, Horizontal and Vertical Perspective STM

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 39, the frequency of occurrences are not pertinent to the analysis of these visual systems. The duration of these predominant Shots and Zoom choices in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). There are Front, Eye Level and Stationary shots for 71% of the video durationally (see Figure 39 for percentages). This is similar to Guijarro and Sanz's (2008) study, which may indicate the producer trying to elicit the viewer to relate and identify with main characters, similar to the prevalence of Close shots mentioned earlier in this section. However, previous literature on children's stories often saw animal Participants as not being very relatable to children (Zohrabi *et al*. 2019). Nevertheless, perhaps attempts at engaging the child's attention is further elicited by Close shots and Front and Eye Level shots (Zohrabi *et al*. 2019). The 13% state refers to Front shots only, 6.59% refers to Eye-level and Front, 2.75% refers to Front, Eye-level and Pan (left to right) and 0.55% refers to Front, Eye-level and Tilt. Ultimately, it is clear that the majority of shots durationally are Front, Eye-level and Stationary, which allows the viewer to engage as

equals with Participants (Kress and van Leeuwen 2006). Viewer engagement is usually achieved more in these shots than if they were Angled, High and/or Low. Front and Eye Level shots may allow for the viewer to feel a part of the story and are similar to natural interactions (Kress and van Leeuwen 2006). Consistency of these system choices may cause less distraction for learners, as this type of movement in addition to shot transitions, animations, subtitle transitions and other types of visual animation and movement could lead to learners focusing their gaze on that instead of decoding and comprehending subtitle content.

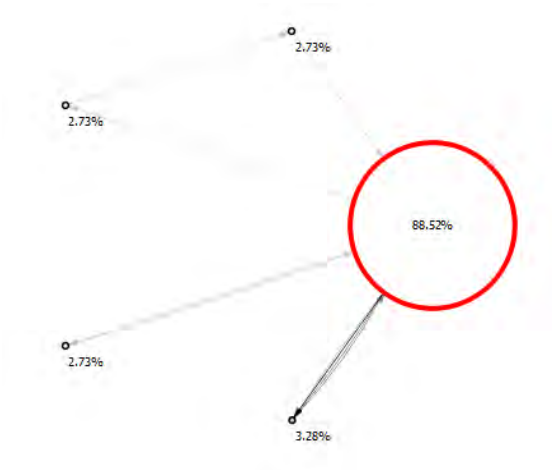


Figure 40 The Icky Sticky Frog Saliency STM

In terms of Salience, Figure 40 illustrates that there is not a high duration or frequency of particular salient effects in the videos. The duration of these predominant Processes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In 3.28% of the video durationally, Sele-sele is emphasised in the Foreground, and from the bottom to the top of Figure 40, the 2.73% states position Sele-sele in the Background, the locust in the foreground and there is a Sharpness of Focus on Sele-sele. While the duration of these states is low, again this illustrates the predominant focus on Sele-sele. This allows for adherence to Mayer's (2009) Signaling Principle, which states that highlighting particular words guide learners to process them, and this may improve learning. Thus, the presence of salient Participants that are integral to the story may help learners to direct and maintain attention on them to assist their learning.

6.1.5 Visual textual metafunction STMs

This section examines firstly Information Value, or in other words the positioning of visual Participants as Given or New information, as well as Centre of the frame. Information Value, as well as Framing devices further place emphasis on the visual Participants in the video, potentially assisting adherence to Mayer's (2009) Signaling Principle.

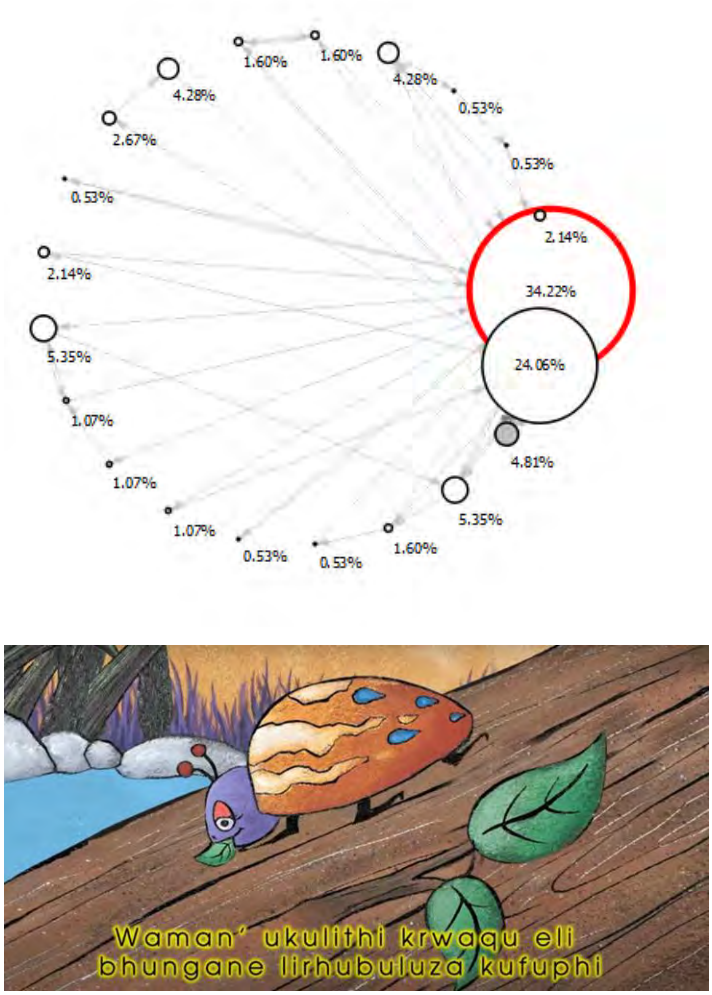


Figure 41(a) The Icky Sticky Frog Information Value STM (top) with (b) screenshot (bottom) subtitled "he kept glancing at the beetle crawling nearby"

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 41(a), the frequency of occurrences are not pertinent to the analysis of these visual systems. The above Figure 41(a) illustrates that the predominant Participants, excluding Sele-sele, are presented as New information. Sele-sele is presented mostly in the Centre (24.06%) or as Given information durationally. The only exceptions are the butterfly in the Centre (4.28%) and the beetle in the Centre (2.14%). There is also a 4.28% state where Sele-sele is Given and the locust is New information. The 5.35% states from the bottom to the top of Figure

41(a) include Sele-sele as Given information and a co-occurrence of Sele-sele as Given Information and the beetle as New information. The 4.81% state is where Sele-sele is Given information and the locust is New information. Only Sele-sele was a Theme and none of the other visual Participants were, and, therefore, the high duration of Sele-sele as Given or Centre correlates to the high duration and frequency of Sele-sele as Theme in the video.

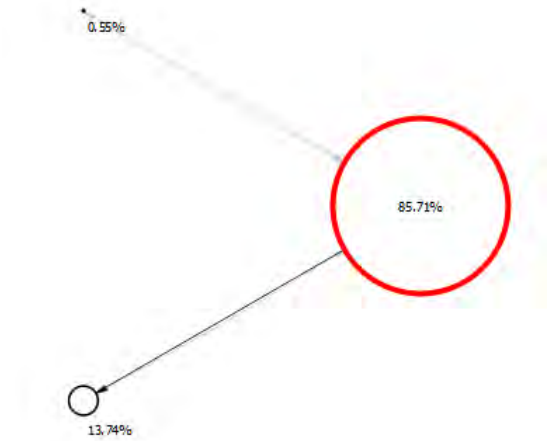


Figure 42(a) The Icky Sticky Frog Framing STM (top) and (b) screenshot from The Icky Sticky Frog (bottom), subtitled: "LETSHE! This tongue of USele-sele, sticky and long."

Figure 42(a) illustrates Framing devices in the video. The duration of these Framing devices in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). There is a 13.74% state which represents the Framing device of a log that separates the beetle from the rest of the scene (see Figure 41(b)). This is also established to a lesser extent with the locust in the leafy shrubbery in the Foreground. There is also a small 0.55% instance where Sele-sele is again emphasised as he enters the mouth of the fish and this framing device (see Figure 42(b)), although brief, further places emphasis on him. System

choices in the interpersonal and textual metafunctions for both modes create greater emphasis on Sele-sele and his actions. This further corresponds to the title of the video that indicates that it is focused on Sele-sele. The high duration of these system choices for Sele-Sele allows adherence to Mayer's (2009) Signaling Principle and learners may consequently draw their attention to a particular integral Participant to the story. The video is visually simpler for learners, potentially allowing them to focus on the language with less distractions from the visual mode. The simplicity and Similarity of meanings between modes may further assist in using the visual mode to comprehend linguistic mode content, see 6.1.6. This may assist in learning if the literacy development areas are focused on the narrative of the video and vocabulary around this.

6.1.6 Intersemiotic Texture STMs

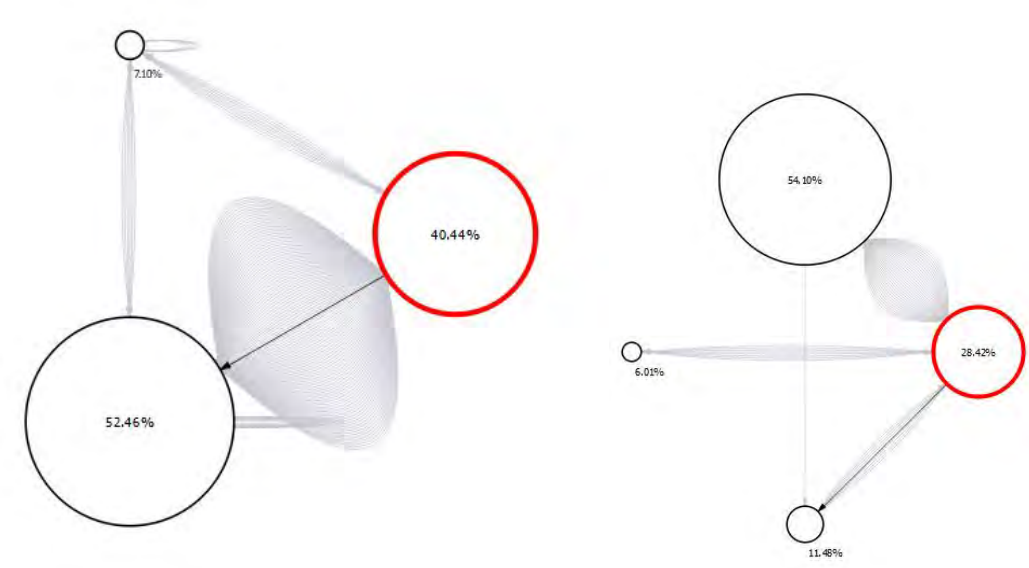


Figure 43(a) The Icky Sticky Frog Comparative Relations STMS with Spoken Language and Visuals (left) and (b) Written Language and Visuals (right)

The duration of these predominant system choices for Comparative Relations in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). This section discusses the Similarity and Contrast between spoken language and visuals (Figure 43(a)) and written language (subtitles) and visuals (Figure 43(b)). There is a 52.46% state in Figure 43(a), which shows a high amount of Similarity for Participants, Processes and Circumstances between spoken and visual modes. This is in comparison to the 7.10% state

representing Contrast between spoken language and visuals. This illustrates a high level of Similarity. This may assist in adherence to Mayer's (2009) Spatial and Temporal Contiguity Principles for ideal cognitive learning, which argues that the linguistic and visual meanings should correspond (have a high Similarity rate) and occur simultaneously or sequentially.

For Figure 43(b), there is a high 54.10% Similarity rate and only 6.01% of the video comprises Contrast. In 11.48% of the video, Similarity and Contrast appear simultaneously and this may cause potential confusion for learners. This is observed in the below screenshots. This is illustrated in Figure 44, where the subtitle remains on screen during a shot transition. The first shot shows the frog and his tongue being swallowed, but the tongue is no longer visible in the second shot while the subtitle is still present on the screen.

Figure 44 The Icky Sticky Frog Shot Transition for "LETSHE! This tongue of Usele-sele, sticky but long"

Nevertheless, the Similarity rate is still almost five times higher than this score, and the Similarity consisting of more than half the video duration may be conducive for literacy

development, as it further adheres to Mayer's (2009) Spatial and Temporal Contiguity Principles.

All Similarity relates to Correspondence in which the visual and linguistic items correspond to one another in meaning, such as a picture of a frog when Sele-sele is mentioned and the action of him swallowing the insects, as they are no longer visible.

Intersemiotic Temporal Relations map out a successive relationship between semiotic resources of different modes, such as manual instructions being portrayed linguistically, but with a further sequence of images to expand on a single linguistic step (Liu and O'Halloran 2009:383). This is illustrated in the chronological unfolding of the actions and subtitles in Video 1. When the subtitle comprises an onomatopoeic word or description of the frog's tongue and the disappearance of the insect, the visual mode depicts the frog either catching an insect with its tongue or swallowing it sequentially after the subtitle has appeared. In Tseng and Djonov (2023), children from ages 4-5 could understand such a chronological unfolding of events in a narrative. All the videos have a temporal successive relationship as described between the written language (subtitle) and the visuals. However, this thesis focuses on Comparative Relations as it is a more fine-grained system choice in Intersemiotic Relations that may allow for better understanding of the adherence to Mayer's (2009) Principles in the design of the videos. In Video 1, the repetitive and recurring frequency may not only be good for interpreting time in the video, but if this is used as a literacy resource, it could assist in other areas of vocabulary learning.

6.1.7 Subtitle rate

This section examines subtitle rate. The subtitle rate is calculated according to De Linde and Kay (2014). The average duration of subtitles were downloaded from Excel and the annotated duration is used as the total subtitle presentation rate. According to De Linde and Kay (2014:45), subtitle presentation rate is "a measure of how quickly a subtitle appears and disappears from screen" and is measured in words per minute. IsiXhosa is an agglutinative language and the words are considerably longer than English words, the language with which this formula was first used. Thus, the syllable per minute rate was also counted to ensure that adequate averages were calculated and accounted for these discrepancies in linguistic differences.

The total duration of subtitles (subtitle presentation rate) was 2 minutes 1 second and the total number of words (separated by spaces) is 154 words. Therefore, 1 divided by 60 = 0.02 + 2 minutes = 2.02. If words had hyphens, they were included as a single word. Thus, 154 words divided by 2.02 = 76.23 WPM. If the maximum reading threshold of 50 WCPM (Ardington *et al*. 2020) holds for Grade 3 learners, not only is this 20 words per minute above the benchmark and maximum reading score (Ardington *et al*. 2020). It is also more than three times the Grade 3 pre-test average reading score for a longer passage of static text (22.85 WCPM) and more than double the post-test average score (28 WCPM). It is also three times the average benchmark for Grade 2 (20 WCPM) and double the average benchmark for Grade 3 (35 WCPM) (Ardington *et al*. 2020). In terms of Grade 2, it is five times the average pre-test reading score (14.3 WCPM) and four times the average post-test score (19.95 WCPM). Thus, due to the fast WPM and SPM rates, this video may not adhere well to Mayer's (2009) Segmenting Principle.

Figure 45 Screenshot from The Icky Sticky Frog, subtitled "just as he swallowed the fly, after that he sees"

The total syllables per minute (SPM) were calculated with a similar method. The calculation is 494 syllables divided by 2.02 = 244.55 SPM. The maximum reading score for Grade 2 readers was 141 SCPM and Grade 3 readers was 179 SCPM. The mean scores, however, which reflect average reading rate is far lower at 45.4 SCPM pre-test and 65.20 SCPM post-test for Grade 2, and 76.95 SCPM pre-test and 96.10 SCPM post-test for Grade 3. The syllable rate in the videos is still not at a pace the learners would comfortably manage. This violates this thesis' expanded definition of Mayer's (2009) Segmenting Principle. While silent reading is faster than reading

out loud, the large difference between the subtitle presentation rate and Oral Reading Fluency scores indicates the likelihood that the average learner could not read the subtitles for this video quickly enough. Even if a few learners reached higher average reading scores, this would not account for the average reader in these grades. A limitation of this study was the inability to measure silent reading rate, as this is entirely dependent on the learner self-reporting this information, which is not necessarily accurate or objective (Probert 2016). Furthermore, eye-tracking methods do not necessarily confirm whether phonological decoding in silent reading took place, even if it tracks the movement of the eye along the subtitle. Thus, there is no objective method to accurately test a silent reading rate.

6.1.8 Conclusion and summary of Video 1 analysis

In Video 1, Sele-sele is the predominant Actor Participant. He is also the only Participant that is a Theme and is also emphasised with Zoom, Salience and Social Distance (Close shots that are often Centred), when not presented as Given information. Nevertheless, the majority of Participants are mainly Goals and insects, but this is also due to the nature of the isiXhosa language that can omit subject Participants. They are often presented as New information and the Direct Gaze of these Participants also allows the viewer to sympathise more with their characters being swallowed by Sele-sele. Mostly Action processes, repeated sequences, ideophones and onomatopoeia illustrate the producer's attempt to maintain the childrens' attention. This video's repetition of meanings allows for this video to convey the information in a manner, which may be conducive for learning and young children. This adheres to Mayer's (2009) Coherence and Signaling Principles. This may enhance learning.

However, as Guijarro and Sanz (2008) state, too much repetition and focus on the same components could become tedious for learners. Guijarro and Sanz (2008) state that it is adequate to use such language for children that are five years old, yet these children are seven to nine years of age, and this poses the question of whether children would find the content boring or repetitive and not learn from these videos. Due to the low scores of children's reading, this is suggested as another potential reason for the lack of improvement in reading scores from video exposure. As the chances of children being unable to follow the subtitles could also lead to boredom, but it could lead to an inability to learn anything from the videos. This would be due to their inability to read and comprehend any information from the linguistic mode. Complex grammar such as stative constructions and ideophones could have also caused

learners to have challenges in reading and comprehending the subtitles. Even though Video 1 (*The Icky Sticky Frog*) illustrates an ideal correspondence between modes for simple meanings to be conveyed and less Contrast. The presence of complex grammatical structures and a fast subtitle rate could have caused learners to struggle to follow the pace of the presented information. This could have potentially hindered learning from this resource and is discussed as the potential primary cause for no significant difference in literacy development from the videos in 6.6.

While the high Similarity rate in the video design may be conducive to learning, the high subtitle rate of 76.23 WPM is still too high for Grade 2 and 3 learners to adequately read it according to the benchmarks and pre- and post-intervention tests. Thus, while the video adheres to Mayer's (2009) Principles of Spatial and Temporal Contiguity, the subtitle rate being too fast for learners to potentially comprehend adheres to Mayer's (2009) Coherence and Redundancy Principle. The fast subtitle rate does not adhere, however, to Mayer's (2009) Segmenting Principle. Furthermore, the frequent amounts of Close and Medium shots in the video, as well as most shots being Front, Eye Level and Stationery could have assisted in adhering to Mayer's (2009) Personalization Principle. This is discussed further in relation to other videos in 6.6. The next section analyses *They All Loved Sue* or *Bonke babemthanda USue* (Video 2).

6.2 Video 2 - They All Loved Sue or Bonke babemthanda USue

As mentioned in the introduction to Chapter 6, *They All Loved Sue* includes the story of Sue who loves all kinds of animals but is not allowed to have any at home. She would meet many different animals, particularly dogs and cats, along the way to work and would start bringing snacks for them, thus befriending them. Soon they would follow her to work, making Sue very happy and she opened her own business to look after the animals. Important spoken and/or visual Participants to note for this video include Sue, and animals such as rabbits, tortoises, fish, dogs, cats, people, birds and squirrels. The analysis of this video begins with the discussion of spoken and written ideational metafunction STMs for language (6.2.1). I continue to discuss

the interpersonal and textual metafunctions in these modes (6.2.2). Section 6.2.3 examines the ideational metafunction STMs in the visual mode and 6.2.4 examines the interpersonal metafunction STMs in the visual mode. Section 6.2.5 examines the textual metafunction STMs in the visual mode and 6.2.8 examines Intersemiotic Texture. There is a discussion of subtitle rate in 6.2.7. Lastly, 6.2.8 concludes the analysis of this video.

6.2.1 Spoken and written language ideational metafunction STMs

This section focuses on the spoken and written language ideational metafunction STMs, starting with Participants, then Processes and finally Circumstances.

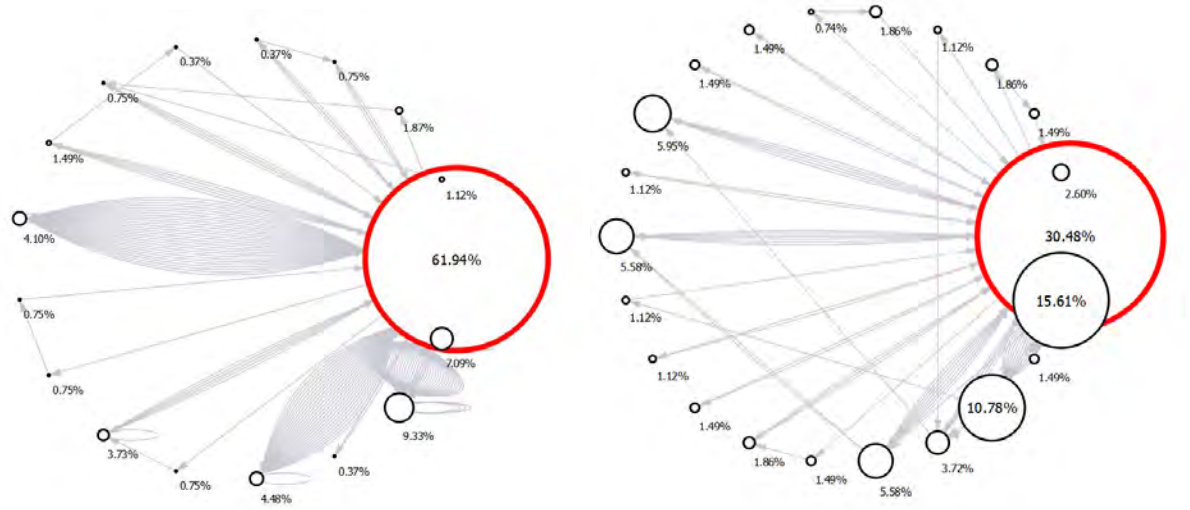


Figure 46(a) They All Loved Sue Spoken Participant STM (left) and 31(b) Written Participant STM (right)

The duration of these predominant Participants in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Only the spoken Goal (4.48%) and spoken Actor (4.10%) states share the same frequency of occurrence, which is consistent with both Participants relating to one Process (Material). In Figure 46(a), the 9.33% state represents Phenomenon and is the most frequent Participant in the spoken mode. As in 6.1, the high states for these types of Participants are consistent with the language. In isiXhosa, the noun can be omitted from a sentence and be present within the verb complex as a prefix, which often occurs with the agent noun in a clause. Additionally, because nouns are often

omitted, a higher percentage of Participant duration relates to direct objects. Furthermore, because subjects are often omitted, a higher percentage of Participants are direct objects, and Material processes are predominant in the video. However, unlike the predominant Participants and Processes of the genre typically being Material Processes, this video has a higher frequency of Mental Participants and Processes in both the spoken and written linguistic modes. Even though these types of Processes and Participants are predominant across both linguistic modes, there is a lack of different types of Mental Participants and Processes depicted in the visual mode. This excludes other Participants seeing objects and Participants in the video, which may cause a lack of Similarity, and thus a potential lack of adherence to Mayer's (2009) Principles of Spatial and Temporal Contiguity. This is discussed further in 6.2.4.

Figure 47 Screenshots from They All Loved Sue, subtitled top-left (a) "cats, dogs, tortoises," top-right (b) "birds, all kinds of animals" and bottom (c) "were not allowed in her building"

The predominant direct objects cause the linguistic mode to have greater emphasis on the animals. For example, in Figure 47, "iikati, izinja, amafudo, iintaka, izilwanyana zalo, naluphi na uhlobo zazingavumelekanga kweso sakhiwo sakhe" (cats, dogs, tortoises, birds, all kinds of

animals were not allowed in her building), all the animals are listed as Phenomenon Participants. This is also a stative construction, which discussed in Video 1 in 6.1, may cause challenges in comprehension for learners, as it may increase their cognitive load, according to Mayer's (2009) limited capacity assumption.

Figure 48 Screenshots from They All Loved Sue, (a) top-left to (b) right subtitled, "Soon Sue started to bring, dog snacks to work" Subtitles (c) bottom-left to (d) bottom-right "There were big, small, unusual/strange/rare, beautiful and ugly dogs"

The most prevalent states in Figure 46(b) are Sensor and Phenomenon (5.58% each with an 15.61% co-occurrence), then Actor (11.57%). For spoken language, Sensors comprise 7.09% of the video. This means that the subtitles that contain these Mental Processes and Participants are on the screen for longer than the Material Processes are. The Sensor Participants refer mostly to Sue, who loves the animals, although sometimes the animals are the Sensors that love Sue. The 4.48% state reflects the spoken Participant Goal (this is represented by 3.72% for the written mode in Figure 46(b)). There is 4.10% of the video which comprises spoken Actor Participants, for example in Figure 48(a-b), "USue waqalisa ukuphatha izimuncumuncu zezinja emsebenzini" (Sue started to bring dog snacks to work). There is 1.49% of the video which comprises the Spoken Attribute Participant, for example, "USue wayenabathengi abaninzi" (Sue had many customers). The 3.73% spoken Participant is Existent and is preceded by the word "kwakukho" (there was/were), for example in Figure 48(c-d), "Kwakukho izinja ezinkulu, ezincinane, ezingaqhelekanga, ezintle kunye nezinja ezimbi" (There were big, small,

unusual, beautiful and ugly dogs). The 5.95% written state is Actor and Goal Participants, the 2.60% state is a co-occurrence of Identifier, Identified, Actor and Goal, which appears in that order in the sentence ibi**yinto eqhelekileyo** ukuba **a**chithe **ixesha elininz**i kunye nazo (**it** was **a usual thing** for **her** to spend **a lot of time** with them). While Action Processes are still common in this video, there are also many Sensor and Phenomenon Participants (thus Mental Processes) which align with the story referring to Sue loving animals and the animals loving her back.

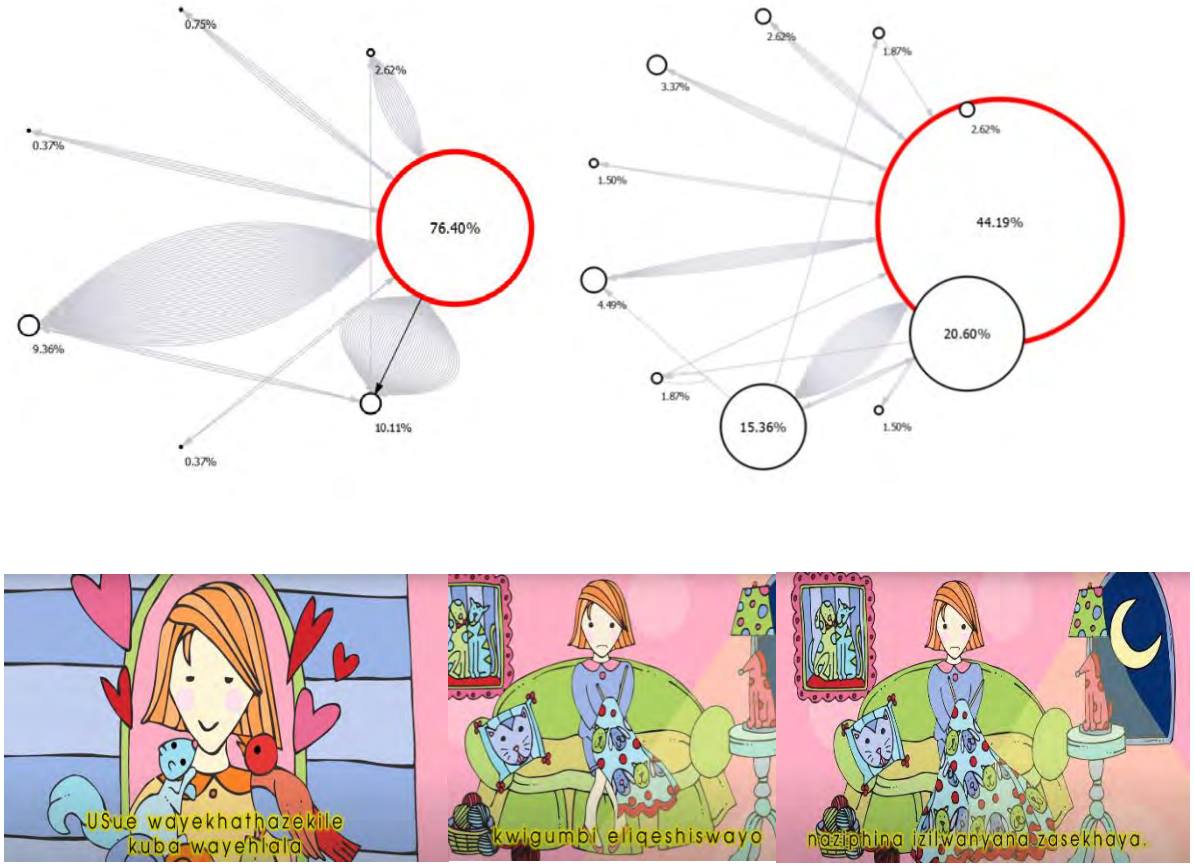


Figure 49(a) They All Loved Sue Spoken Processes (left) and (b) Written Processes (right) with (c) screenshot subtitled "Sue was sad/hurt because she lived in a rented room where no pets were allowed."

The duration of these predominant Processes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Material Processes are incorporated, as illustrated in Figure 49(a) and Figure 49(b), with the Material Process states comprising 9.36% for spoken and 15.36% for written. Spoken Mental Processes appear 10.11% of the time (20.60% of the time in the written mode). In the video, 2.62% comprises spoken Relational Processes (2.62% written) and lastly Existential, Verbal and Behavioural Processes occur for 0.37% of the video for the spoken mode. Verbal is 1.87% of the video and Existential

is 1.50% of the video for the written mode. The Material Processes remaining on the screen longer than the Mental Processes in the video might cause distraction and retention challenges for learners, as the most predominant spoken and written Participants are Mental. This would not adhere as well to Mayer's (2009) Principles of Spatial and Temporal Contiguity as Video 1. However, this would have to be tested further and specifically to confirm.

The topmost 3.37% state is a combination of Relational and Mental Processes, for example, "ubenamawaka amaninzi ezilwanyana ezimthandayo" (she had thousands of animals that loved her). Furthermore, in 3.37% of the video durationally, Material and Mental Processes occur simultaneously. For example, in the subtitle in Figure 49(c). The rightmost 1.87% state from right to left in Figure 49(b) are Material and Verbal Processes, for example "Ngenxa yoko uSue, wavula elakhe ishishini, elithi "IShishini leeNkonzo zokuGcinwa kweZilwanyana likaSue." (Because of that, Sue opened her own business called "Sue's Animal Keeping Services.). The other 1.87% state is only Verbal Processes, for example, "olufundeka ngolu hlobo" (that reads). There is a 1.50% co-occurrence of Behavioural and Mental Process in the written mode, e.g. Figure 49(a). This correlates with the frequency of Sensor, Phenomenon, Actor and Goal Participants which describe emotions and actions in the story of Sue and typical of children's narrative structures (Guijarro and Sanz 2008). However, one can see how long the subtitles are for this story and sentences have to be broken up along multiple shots (see Figure 47 for an example). This combined with multiple Processes in the subtitles might create a complexity that could cause challenges for learners in decoding and comprehending sentences with multiple concepts. The presence of multiple Processes in the written mode in Video 2 occurs less than Video 1. This means in this aspect of video design, Video 2 might be better for learning, as Video 2 has less complexity in meanings conveyed through Processes than Video 1. It may increase the cognitive load of learners and this is not conducive for learning according to Mayer's (2009) limited capacity assumption.

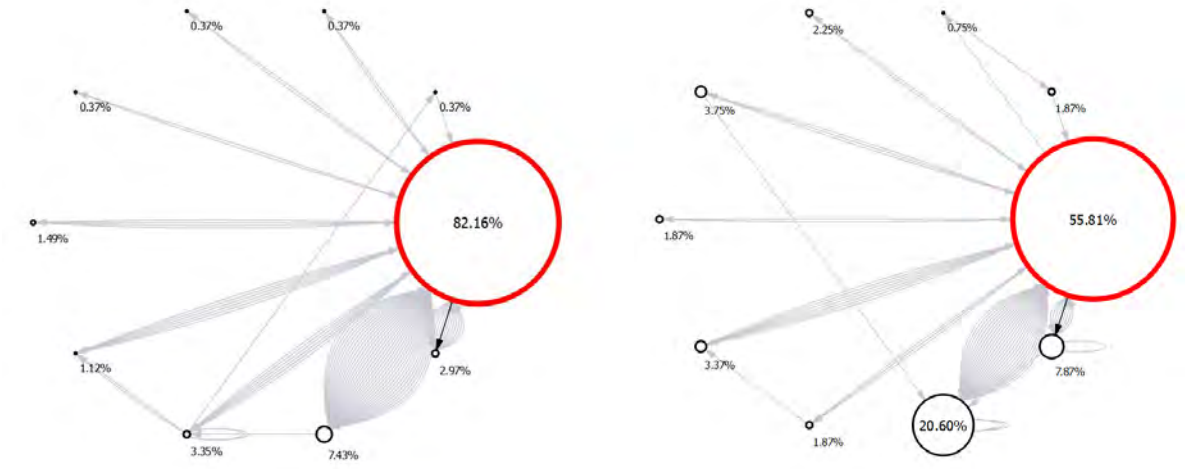


Figure 50(a) They All Loved Sue Spoken Circumstances STM (left) and (b) Written Circumstances STM (right)

The duration of these predominant Circumstances in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In terms of Figure 50(a), 7.43% of the video durationally comprises Circumstances of Location (20.60% in the written mode), which include either Location: Time, such as “kungekudala” (soon) or Location: Place, such as “emsebenzi” (at work). There is 2.97% of the video durationally which is comprised of spoken Circumstances of Manner (7.87% in the written mode), such as the word “kakhulu” (very). In the video, 1.49% comprises spoken Circumstances of Extent (1.87% in the written mode), such as “kwisithuba seeveki” (after a few weeks). The topmost 1.87% state in Figure 50(b) is Cause, e.g. “Ngenxa yoko” (because of that) and the 0.75% is Accompaniment, e.g. “kunye nazo” (with them). There is also a 2.25% in Figure 50(b) which represents Circumstances of Matter, e.g. “ngoSue” (about Sue) in the sentence “Iindaba zavakala kuzo zonke izinja nakuzo zonke iikati ngoSue” (The news reached all the dogs and all the cats about Sue).

In addition, there is a further categorisation of 3.36% of Circumstances of Location that permeate across spoken embedded clause structures (1.87% in the written mode). In 1.12% of the video, there is a Location Circumstance within a spoken Embedded Clause entirely categorised as Circumstance of Location (3.37% in the written mode), for example in Figure 50(a). The 0.37% state that connects via a transition line in Figure 50(a), illustrates a Circumstance of Manner within an Embedded Clause Circumstance of Location (3.75% in the

written mode), e.g., “USue ubengoyena mntu ugcina izilwanyana zasekhaya kakuhle kuso sonke eso sixeko (Sue was the most outstanding pet keeper in that whole city). The complexity of embedded constructions, which often include a lot of adjectivised and nominalised Processes might create challenges in decoding, due to the strain it may place on the cognitive load of the learner according to Mayer’s (2009) limited capacity assumption.

There is 3.76% of the video durationally which comprises combinations of written Circumstances of Manner and Location. As seen in Video 1, this pattern of a high number of Location Circumstances is typical for a video that is for children (Guijarro and Sanz 2008) and can be used to try and maintain a child’s attention, as discussed in 6.1.1 with Video 1. Circumstances of Location: Time relate to Time deictic words (Levinson 2004; Peña 2020).

This video has a predominance of Material Processes, yet a predominance of Sensor and Phenomenon Participants, which is different from the other videos analysed in this thesis. This may align with the story referring to Sue loving animals, but the discrepancy may not adhere to Mayer’s (2009) limited capacity assumption, which may influence adherence to Mayer’s (2009) Principles of Coherence, Redundancy, Spatial and Temporal Contiguity. While there is adherence with the Material Processes in the linguistic mode to the visual mode (see 6.2.3 and 6.2.6), the lack of visual representation of the Mental Processes and their Participants discussed in 6.2.3 means this adherence is not as strong as Video 1. The lack of Similarity for these Processes could result in distraction or lack of focus on them by learners.

As discussed in 2.3.1.3, frequency of exposure may not be as integral for literacy development as Similarity and Contrast co-patterning between modes (Tseng and Djonov 2023). Thus, this thesis focuses on Comparative Relations in Intersemiotic Texture (see 6.2.6 for this video) and a Similarity of meanings between modes as a potentially more accurate indicator of literacy development than frequency or duration. Further research in this context is required.

The high frequency pattern of Locations of Circumstance is further evident in this video and illustrates a broader syndrome in its coupling with Material Processes that seems to permeate the genre. This is a potential attempt by producers to maintain viewers’ attention to the videos and to specific Processes in the video. The maintenance and focus of viewer’s attention relates to Circumstances of Location: Time being Time deictic words (Peña 2020).

6.2.2 Spoken and written interpersonal and textual metafunction STMs

Approximately 50% of the video's duration includes spoken and written Declaratives, with no other mood, and, therefore, no STM is pictured for this system. This is typical for story videos that offer information, although more Imperative or Interrogative Moods could cause children to engage more with the videos. However, different Mood structures may pose a challenge for isiXhosa L1 learners to decode information according to Mayer's (2009) limited capacity assumption. There is a lack of research on Foundation Phase readers and their ability to distinguish between Mood structures.

Modality may, when it relates to Moods like Interrogative and Imperative, cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. If the Mood is Declarative, it may increase a load on cognitive capacity that violates the limited capacity assumption. Since there are only Declarative statements in *They All Loved Sue*, this video has realisations of Modality that could increase learners' cognitive load, as they create complexity in the language. This is realised in a variety of ways. For example, with the verb *-vumela* (to allow) in the subtitle Iikati, izinja, amafudo, iintaka, izilwanyana zalo, naluphi na uhlobo zazingavumelekanga kweso sakhiwo sakhe (Cats, dogs, tortoises, birds, all kinds of animals were not allowed in her building), see (11) in 1.2.1.2. This is also observed in the subtitle "USue wayekhathazekile kuba wayehlala kwigumbi eliqeshiswayo apho kwakungavumelekanga kubekho naziphi na izilwanyana zasekhaya" (Sue was sad/hurt because she lived in a rented room where pets were not allowed), see Figure 49(a-c). This verb realises a low intensity of modal obligation, and all these instances are demonstrating negative polarity. From the structure of these sentences alone, one can see their length and complexity of tenses, additional clauses and modality, which could lead to challenges for learners to not only comprehend these meanings, but also decode their written forms.

Another realisation of modality in this video is in the adjective *-qhele*. For example, in the sentence "Kwakukho izinja ezinkulu, ezincinane, ezingaqhelekanga, ezintle kunye nezinja ezimbi" (There were big, small, unusual, beautiful, and ugly dogs.), see Figure 48 This has negative polarity. Another instance in the video is in the sentence "Nanjengoko uSue ubethandwa kakhulu zizo zonke izilwanyana zasekhaya esixekweni, ibiyinto eqhelekileyo

ukuba achithe ixesha elininzi kunye nazo" (Since Sue was loved by all the pets in the city, it was a usual thing for her to spend a lot of time with them), see (8) in 1.2.1.2. This sentence contains modal realisations indicating positive polarity triggered by "eqhelekileyo" (usual) and "ukuba achithe ixesha elininzi" (to spend a lot of time). Both are examples of usuality modal realisations with median intensity. There are also instances of modality after the conjunction *ukuba*, which means "to be" (see 1.2.1.2). This was in the above example and another example includes, "izinja neekati zazifuna ukuba lapho akhoyo uSue" (Wherever Sue went, dogs and cats wanted to be where Sue was.). The modal realisation "zazifuna" seems to indicate a modal realisation of inclination from the dogs and cats, that are wanting to be with Sue wherever She goes. This modal realisation has low intensity and positive polarity. The example "Emva kwemini uSue uyewaqonda ukuba neekati ziyafuna ukukhathalelwa nazo" (In the afternoon Sue felt that the cats also needed to be cared for) further seems to contain two realisations of modality, see (4) in 1.2.1.2. "Ukuba neekati ziyafuna ukukhathalelwa nazo" is directly translated to "the cats wanted to be cared for," which implies a modal realisation of inclination with low intensity and positive polarity. However, this creates an obligation for Sue, who feels ("wayeqonda") the need to take care of them. This modal realisation has high intensity with positive polarity. Further research is needed to illustrate the finer nuances of how modality is realised in isiXhosa clauses such as these, and how they are perceived by learners. However, this was the beyond the scope of this thesis. The complex presence of modality in isiXhosa however in these instances may create complexity for learners in their comprehension in their early reading and may result in cognitive overload.

The *xa*-clauses further create complexity as they introduce conditional clauses (see 1.2.1.2). For example, "Zonke izinja zazisuke zize kuye zibaleka ukuza kumbulisa xa zimbona" (All the dogs would just run to greet her when they saw her). Another example includes the sentence "Kusasa zonke izinja zabamelwane zazimlandela uSue xa esiya emsebenzini." (In the morning all the neighbours' dogs would follow Sue to work). These are modal realisation of probability with median intensity and positive polarity. This is also seen in the clause: "Xa esiya emsebenzini, uSue wayesoloko ebonakalisa ububele kuzo zonke izilwanyana azibonayo" (On her way to work, Sue always showed kindness to all the animals she would see). "Wayesoloko" (Sue always) in this sentence realises usuality with high intensity and positive polarity and this is the only instance of the Circumstance of Location: Time -*soloko* (always). This clause is even more complex with two modal realisations of usuality and probability at different intensities and this could increase learners' cognitive load. The complexity of these

construction may limit learners' cognitive capacity to process other components of the video (see 3.2).

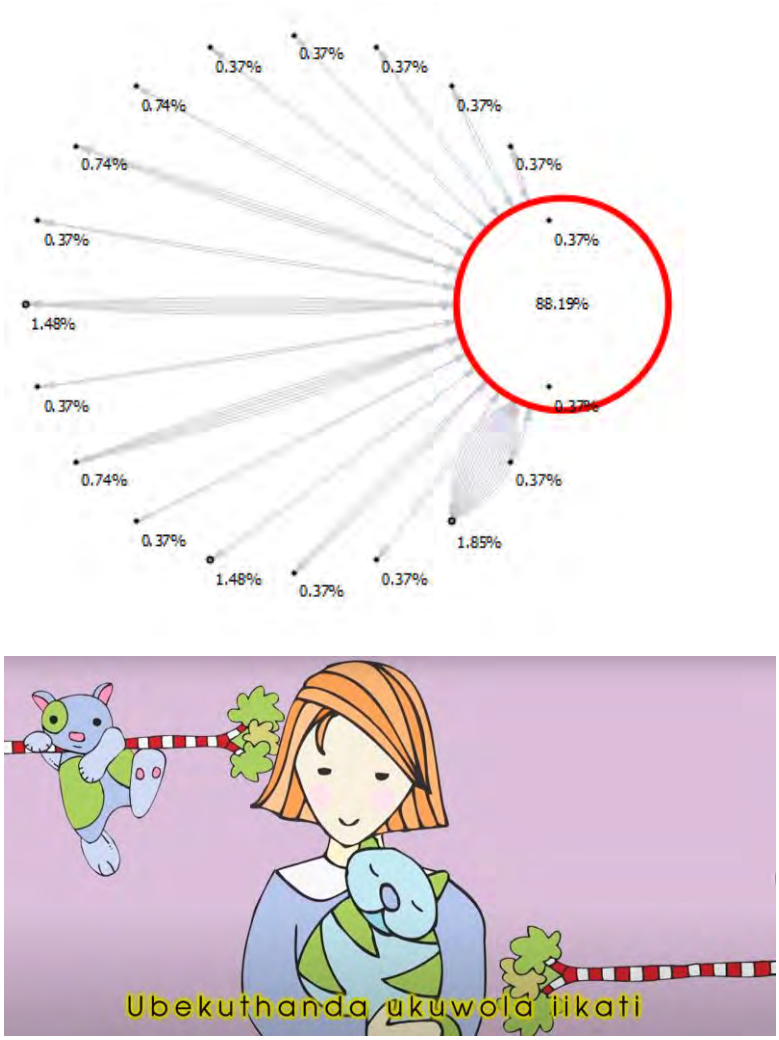


Figure 51(a) They All Loved Sue Theme STM with (b) screenshot subtitled "She loved to cuddle the cats"

The duration of these predominant Themes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Notable spoken Themes depicted in Figure 51(a) include Sue (1.85%), animals (leftmost 1.48% state), "kungekudala" (soon) for 1.48% of the video, see Figure 48(a). There is also "ubekuthanda" (she liked) for 0.74% of the video, see Figure 51(b). This corresponds to Sue and animals as a common Participant and further reflects the emphasis on Mental Processes and Location Circumstances in the video due to the repetition of these components. The lack of consistent Themes overall and the large number of them is similar to Video 1 and does not assist in emphasis and

reinforcement of particular Participants, even with Sue and the animals being at a higher frequency than other Participants. While there are a few Themes mentioned more than others like Sue, which assist in emphasis, it does not adhere to the Signaling Principle as strongly as illustrated in Video 1 (see 6.1.2).

Another challenge and aspect to these videos in terms of the subtitles is the separation of a clause, often within linguistic groups or syntactic phrases. According to Lappalainen (2021), "words belonging together" (e.g. phrases in syntax and/or groups like nominal groups in SFL) should be in the same line of subtitle as per SLS conventions. In this video alone, there are at least 19 instances where, due to the length of words in isiXhosa, noun phrases and similar phrases are separated. This is more than the instances in Video 1. While they often occur in the line underneath them (as there are often two lines of subtitle per subtitle), they can also occur in the following subtitle and be broken midway which can cause confusion. See Figure 49(c) for an example.

6.2.3 Visual ideational metafunction STMs

Figure 52 They All Loved Sue visual Participants and Processes

The duration of these predominant visual Processes and Participants in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). This section discusses the visual ideational metafunction, including visual Participants and

Processes. All the visual Processes in the video are Action Processes, however all the Participants are looking while they act and this does mean they qualify for Mental Processes and as Sensors, unless their eyes are shut. The challenge of examining Mental Processes in the visuals is discussed in 6.6. In 34.59% of the video, Sue and dogs are Actors. The dogs are often depicted as walking or Sue as bending to give them attention, their tails wagging and heads moving. This correlates with the high durational percentage of spoken and written Actors (see Figure 52). Other people Participants in the video except Sue are Actors for 7.52% of the video durationally, often walking, and animals are Existent for 6.39% of the video durationally. In 5.64% of the video durationally, Sue is Actor and the cats are Reactors. This is 5.26% for Sue (Actor) and dogs (Reactors). In 4.51% of the video durationally, Sue and the birds and squirrels are Actors simultaneously. In this video, Sue and the other Participants are constantly seeing while they perform actions, and while this counts and Mental Processes and causes them to be both Actors and Sensors, there is less illustration of other Mental Processes in visuals such as thinking and feeling. Thus, while the visual Action Processes correlate to the high duration and frequency of Material Processes in the linguistic modes, there is a lack of nuances in Mental Processes in the visuals to correlate to the Mental Processes in the linguistic mode. This causes Contrast to occur in the video. This could potentially cause less learning to occur than the potential in Video 1 where there could be a higher Similarity rate between Processes, which could better adhere to Mayer's (2009) Principles of Spatial and Temporal Contiguity.

6.2.4 Visual interpersonal metafunction STMs

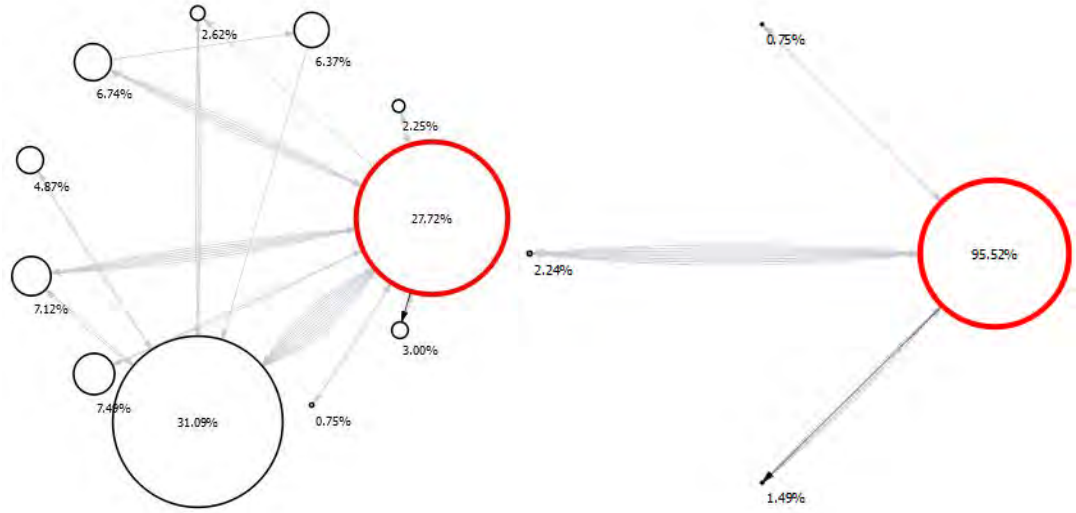


Figure 53(a) Social Distance STM (left) and (b) Zoom STM (right)

The duration of these predominant system choices for Social Distance and Zoom in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Social Distance STMs do not account for trinary choices of shots, as some states indicate duration of shot transitions, in which there are co-occurrences of system choices. Thus, we only focus on the three system choice states, which are Close, Medium and Long. This section discusses the visual interpersonal metafunction, namely Social Distance and Zoom, Gaze, Camera Movement and Horizontal and Vertical Perspective. Sue is viewed in a Long shot in 31.09% of the video. In 7.49% of the video durationally, there are general Medium shots of Sue and the animals. There is 7.12% of the video durationally which has Sue in Close shots, e.g. Figure 48(a-b) and 3% in Medium shots, such as in Figure 49(c). Of the Medium shots in the video, 6.7% are general and 4.87% are Close shots of dogs, such as in Figure 48(c-d). 1.49% of the video durationally has Sue categorised as Zoom In, and 2.24% of as Zoom out. The 0.75% state in Figure 53(b) reflects the cats categorised as Zoom out. Most of the video durationally includes Sue with dogs in Long shots to demonstrate the relationship between these two Participants, and Medium and Close shots are used to further emphasise these Participants. This seems to be typical of multimodal presented story narratives for children, with the Long shots potentially showing viewers that they do not belong within this world, but an imagined one (Guijarro and Sanz 2008). Sue herself as a visual Participant is depicted with white skin, which would be different for amaXhosa viewers, and already there is a layer of

distance between the character and the viewer from the visuals alone, in addition to the Long shots. This may be detrimental to learning, as there is less of an emphasis on specific Participants and Processes in terms of Close and Medium shots. This is less emphasised in comparison with Video 1 (see 6.1.4) which highlights the main character Sele-sele durationally and Close shots in conjunction with a high predominance of Long shots. Thus, in this system, this video adheres less to the Signaling Principle in this particular case than Video 1 and further may adhere less to this thesis' expanded definition of Mayer's (2009) Personalization Principle than Video 1.

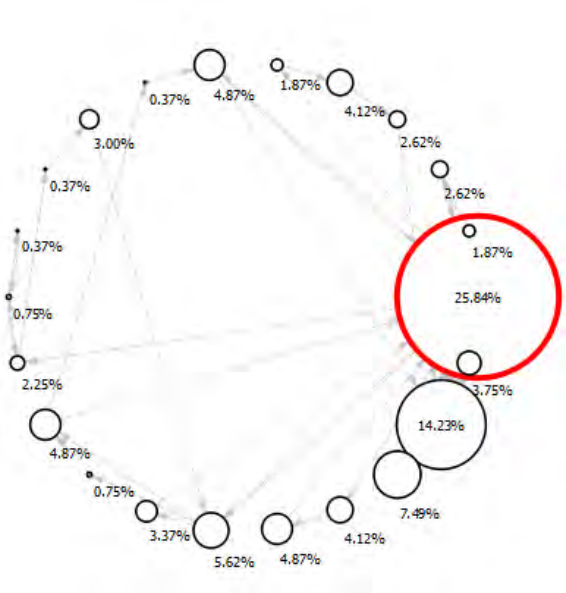


Figure 54 They All Loved Sue Gaze

Sue gazes Directly for 14.23% of the video and Indirectly for 5.62% of the video. In 7.49% of the video, there are Indirect Gazes from animals. The 4.87% states include (from bottom right to left states in Figure 54) Indirect Gaze from dogs, for example in Figure 48(c-d), and a combination of Direct Gaze from animals and Indirect from Sue. The 3.75% state comprises Direct Gaze from Sue and Indirect Gazes from birds and squirrels, for example in Figure 49(c). The 3.37% state represents cats with Direct and Indirect Gazes and Direct Gaze from Sue. The 3.00% state includes Indirect Gaze from the cats and Sue. There is a good combination in this video of Direct engagement with the audience and simply providing information from many of the Participants in the video with Direct Gaze. The most instances of Direct Gaze are from Sue, who is the main character in the story, despite the Long shots. This is different to Sele-sele, who had predominantly Indirect Gazes (see 6.1.4) and this is a case where this video may potentially adhere better to the Signaling Principle than Video 1, although it seems only in this system in comparison to both Social Distance, Theme, and Salience that were all used for Video

1. When more systems like this are in alignment in terms of their focus, this would further add another layer of Similarity in meaning systems that could assist in the adherence of Mayer's (2009) Spatial and Temporal Contiguity Principles. Thus, reducing cognitive load and keeping children focused on specific concepts and actions in the video.

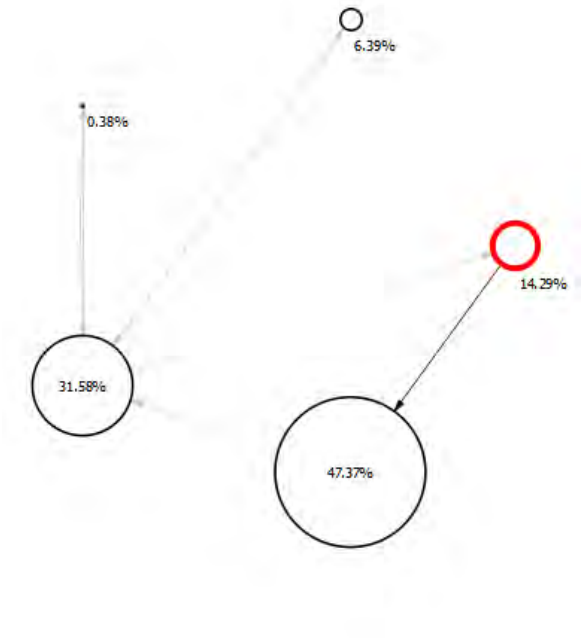


Figure 55 They All Loved Sue Horizontal and Vertical Perspective

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 55, the frequency of occurrences is not pertinent to the analysis of these visual systems. There is 47.37% of the video which includes Stationary, Front and Eye Level shots. Of the shots in Figure 55, 31.58% are Front and Eye Level and 6.39% are Pan, Front and Eye Level. The 0.38% state in Figure 55 represents Front, Eye level, and Tilt (towards Sue). Most of the video shots are uniform to illustrate Sue and the animals, and when there is Camera Movement, it moves to focus on the animals or Sue. Ultimately, it is clear that the majority of shots are Front, Eye-level and Stationary, which allows the viewer to engage as equals with Participants (Kress and van Leeuwen 2006). Furthermore, by allowing viewers to be more equal with Participants, this could have assisted in adhering to Mayer's (2009) Personalization Principle and foster learning.

Salience techniques such as Foreground, Background, Focus and Colour Contrast do not occur in this video and thus there is no STM for them. This is again another area where this video lacks in comparison to Video 1, which could have assisted with a better adherence to Mayer's

(2009) Signaling Principle. While Long shots and lack of Salience create more distance, there are many Medium and Close shots to emphasise visual Participants, combined with Zoom and Gaze which act as ways to maintain the child's attention to the videos.

6.2.5 Visual textual metafunction STMs

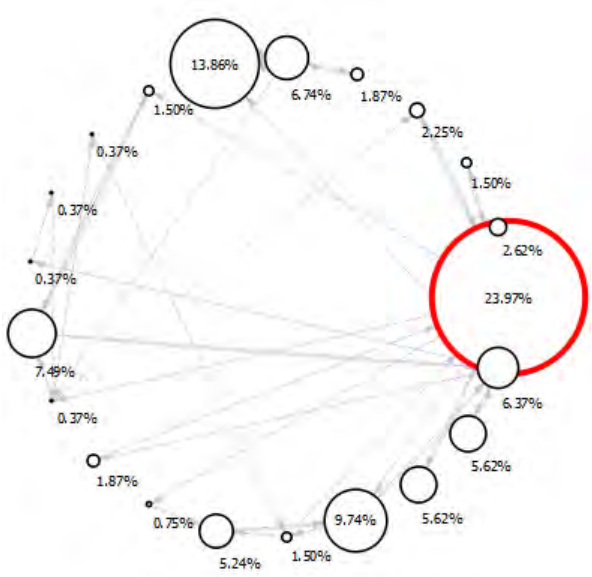


Figure 56 They All Loved Sue Information Value and Framing

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 56, the frequency of occurrences is not pertinent to the analysis of these visual systems. This section discusses visual textual metafunction STMs including Information Value and Framing. In 13.86% of the video durationally, dogs, cats and Sue are Centre. There is 9.74% of the video durationally which shows Sue and dogs at Centre and 7.49% of the video durationally that shows Sue and cats Centre. There is 6.74% of the video durationally which shows dogs and cats Centre and 6.37% comprises Sue at Centre. The 5.62% states from bottom right in Figure 56 shows Sue, the birds and squirrels Centre, as well as dogs as New information with cats as Given information. There is 5.27% of the video durationally which shows dogs Centre and Sue as Given Information. In most of the video, dogs and cats in particular are given more emphasis than other animals and Sue is also given emphasis with Centre placement. In other instances, in the video, Sue is presented as Given Information and animals as New information which correlates with the positionality of Sue as common Actor and animals as common Goal Participants. Both Sue and the animals are mostly in Centre position, and Sue is

Given information and the animals are New information at a similar rate in the video to Sele-sele and his insects. The animals being Centre allows their focus in the visual mode to be more coherent and consistent with their predominance as Participants, even when they are not present as regularly in the video as Themes. This may compensate in adhering to Mayer's (2009) Signaling Principle for visuals. While there is a Similarity between the linguistic and visual mode, a further addition of their predominance as Themes, such as in Video 1, would have been better in adherence to this Principle. It would create a more coherent visual design that signals, highlights and emphasises on specific Participants and Processes for the story and for learning purposes.

The duration of system choices for Framing devices in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In Framing, there are artificial frame lines with Sue at a window (10.86% of video), often with animals, such as birds and squirrels. This further place an emphasis on Sue and her love of animals, further contributing to a high Similarity rate, which is discussed in the next section on Intersemiotic Texture (6.2.6). As occurs with Video 1, this assists in adhering to Mayer's (2009) Signaling Principle**.**

6.2.6 Intersemiotic Texture STMs

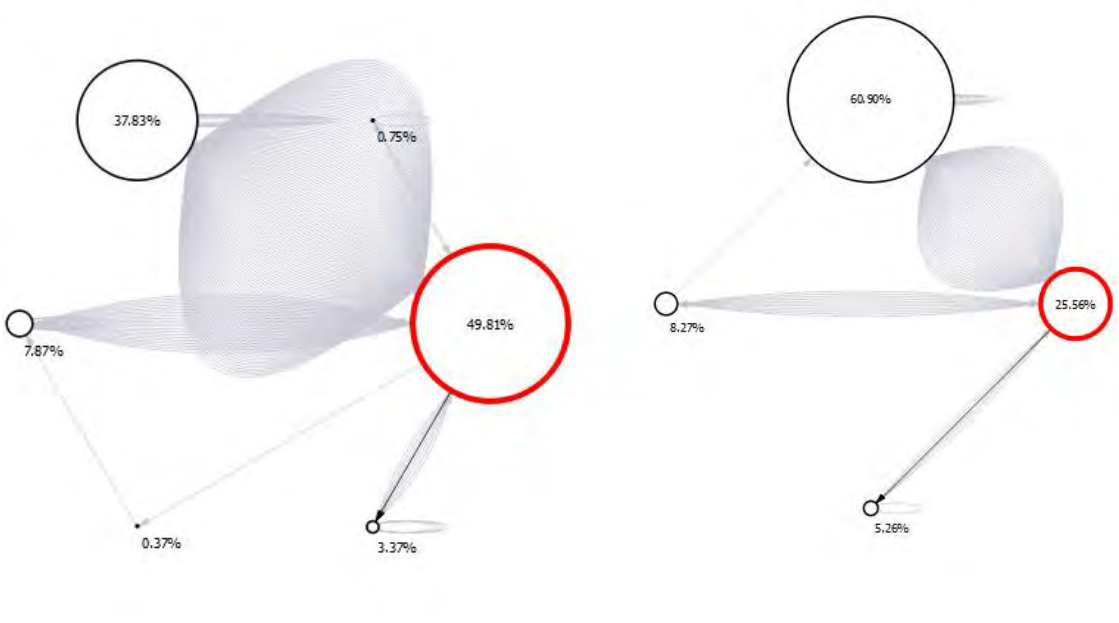


Figure 57(a) They All Loved Sue Comparative Relations STMS with Spoken Language and Visuals (left) and (b) Written Language and Visuals (right) with (c) screenshot subtitled "Sue loved animals."

The duration of these system choices for Comparative Relations in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). There is 37.83% of the video durationally which illustrates Similarity with Correspondence in Figure 57(a) (60.90% in written and visual Comparative Relations in Figure 57(b)). An example of this is seen in Figure 49(c) with "[USue] wayehlala kwigumbi eliqeshiswayo apho kwakungavumelekanga kubekho naziphi na izilwanyana zasekhaya" (Sue lived in a rented apartment where no pets were allowed) with visuals of Sue in her apartment alone without pets. There is 3.37% of the video durationally which illustrates Similarity with Hyponymy (5.26% in Figure 57(b)), for example in Figure 57(c), "USue wayezithanda izilwanyana" (Sue loved

animals) is accompanied by an image that illustrates only birds and squirrels for the category animals. The 0.75% state includes both Hyponymy and Correspondence appearing simultaneously “For example, in “uSue wayesoloko ebonakalisa ububele kuzo zonke izilwanyana azibonayo” (Sue always showed kindness to all the animals she would see), the visuals show her petting the animals but the animals are only represented visually by dogs. The 7.87% and 0.37% states on the left represent Contrast in Figure 57(a), 8.27% in total. They are only separate states due to a bug in the software. Despite the less adequate uses of the interpersonal and textual metafunctions as illustrated in Video 1, the high Similarity rate in comparison to Contrast in this video, like with Video 1, indicates that this video has a good pedagogic design as it could adhere to Mayer’s (2009) Principles of Spatial and Temporal Contiguity. Nevertheless, the lack of Mental Processes in the visual mode for the different types of Mental Processes in the linguistic mode may cause distraction and less attention on these Processes for children. This may not only influence the adherence to the Spatial and Temporal Contiguity Principles, but also to the Coherence Principle by including extraneous information. If Mental Processes are to be focused on by learners, then better visuals could be added to encourage their focus of these Processes. If they are not important for learning goals, then Processes only found in the language and not in the visual should ideally be kept to a minimum to avoid potential distraction for learners. The next section discusses subtitle rate (6.2.7).

6.2.7 Subtitle rate

The total duration of subtitles (subtitle presentation rate) was 3 minutes 5 seconds and the total number of words (separated by spaces) is 254 words. Therefore, 5 divided by 60 = 0.08 + 3 minutes = 3.08. If words had hyphens, they were included as a single word. The 254 words divided by 3.08 = 82.46 WPM. This rate is approximately six times the average pre-test reading rate (14.3 WCPM) and is more than four times the average post-test reading rate (19.95 WCPM) for Grade 2. The subtitle rate is also more than four times the average pre-test reading rate (22.85 WCPM) and more than three times the post-test average reading score (28 WCPM) for Grade 3. If the maximum reading threshold of 50WCPM is true, this is not only 30 words above the threshold and maximum reading score for Grade 3 learners. However, it is more than 3 times the best Grade 2 average reading score for a longer passage of static text (21.8 WCPM) and the average benchmark for Grade 2 (20 WCPM), and more than double the average benchmark for Grade 3 learners (35 WCPM).

The total syllables were calculated with a similar method: 853 syllables divided by 3.08 = 279.94 SPM. The maximum reading score for Grade 2 readers was 141 SCPM and Grade 3 readers was 179 SCPM. The mean scores, however, which reflect majority reading rate is far lower at 45.4 SCPM pre-test and 65.20 SCPM post-test for Grade 2, and 76.95 SCPM pre-test and 96.10 SCPM post-test for Grade 3. The syllable rate in the videos is still not at a pace the learners would comfortably manage. This violates this thesis' expanded definition of Mayer's (2009) Segmenting Principle. While silent reading is faster than reading out loud, the large difference between the subtitle presentation rate and Oral Reading Fluency scores indicates the likelihood that the learners could not read the subtitles for this video. A limitation of this study was the inability to measure silent reading rate, as this is entirely dependent on the learner self-reporting this information, which is not necessarily accurate (Probert 2016). Furthermore, eye-tracking does not necessarily state whether decoding took place, even if it tracks the movement of the eye along the subtitle, thus there are limitations in understanding an objective silent reading rate.

6.2.8 Conclusion and summary of Video 2 analysis

Action Processes are common in this video, as with Video 1, and there is also a large number of Mental Processes with Sensor and Phenomenon Participants which align with the story referring to Sue loving animals. However, while the visual Action Processes correlate with the high frequency of Material Processes in the linguistic modes, there is a lack of Mental Processes in the visuals to correlate with the different types of Mental Processes in the linguistic mode. Furthermore, the most predominant Participants relate to Mental Processes in the linguistic mode and not to Material Processes. This can cause Contrast to occur in the video. Nevertheless, the high Similarity rate in the video (Correspondence and Hyponymy) does indicate the video is well-designed for learning, adhering to Mayer's (2009) Principles of Spatial and Temporal Contiguity. Furthermore, most of the video includes Sue with animals in Long shots and Centre to demonstrate the relationship between these two Participants, and Medium and Close shots are used to further emphasise these Participants. Direct and Indirect Gaze are common for all Participants, which allow the viewer potentially to engage more with the video. However, the other systems of interpersonal and textual metafunction are less utilised for this purpose than in Video 1, which has more instances of adherence to Mayer's (2009) Signaling Principle.

However, the subtitle rate is extremely high at 83.28 WPM which, according to intervention test results and average benchmarks for Grade 2 and 3, indicates these subtitles could not be read by the learners, and is even faster than Video 1. A triangulation of reading and subtitle scores with DoE benchmarks is discussed in 6.6, as the videos adhere to Mayer's (2009) Coherence and Redundancy Principles and does not adhere to Mayer's (2009) Segmenting Principle. Thus far, Video 1 has more coherence in terms of design that is more adequate to principles of learning than Video 2 and Video 1 would be potentially more effective than Video 2, had the subtitles been at a slower rate. Video 2 may include human Participants, which according to Zohrabi *et al.* (2019), is more conducive to learning for children as they relate to them more. The human Participants like Sue are removed from the amaXhosa culture, combined with Long shots and less intimacy between viewer and Participants in terms of the interpersonal metafunction. This illustrates more potential for this video to be less effective than a video like Video 1*,* which has animal Participants, but utilises the interpersonal and textual metafunctions more frequently and with more repetition to emphasise these Participants and Processes.

6.3 Video 3 - The Moon Followed Me Home or iNyanga indilandele ukuya eKhaya

INyanga indilandele ukuya eKhaya, or *The Moon Followed Me Home* is told by the first-person narrator, who is a lion cub in the visuals. He describes his friend the moon and how it is always with him. He also talks about his father who appears in the video and activities they do together, while the moon accompanies them. The analysis of this video is divided into a discussion of the spoken and written ideational metafunction STMs for language (6.3.1), and then the interpersonal and textual metafunctions in this mode (6.3.2). Section 6.3.3 examines the ideational metafunction STMs in the visual mode and 6.3.4 examines the interpersonal metafunction STMs in the visual mode. Section 6.3.5 examines the textual metafunction STMs in the Visual Mode and 6.3.6 examines Intersemiotic Texture. There is a discussion of subtitle rate in 6.3.7. Lastly, 6.3.8 concludes the analysis of this video. Important spoken and/or visual Participants to note for this video include the moon, the father, any reference of the ndi- prefix, "I," which refers to the lion cub, the mother, insects and animals.

6.3.1 Spoken and written language ideational metafunction STMS

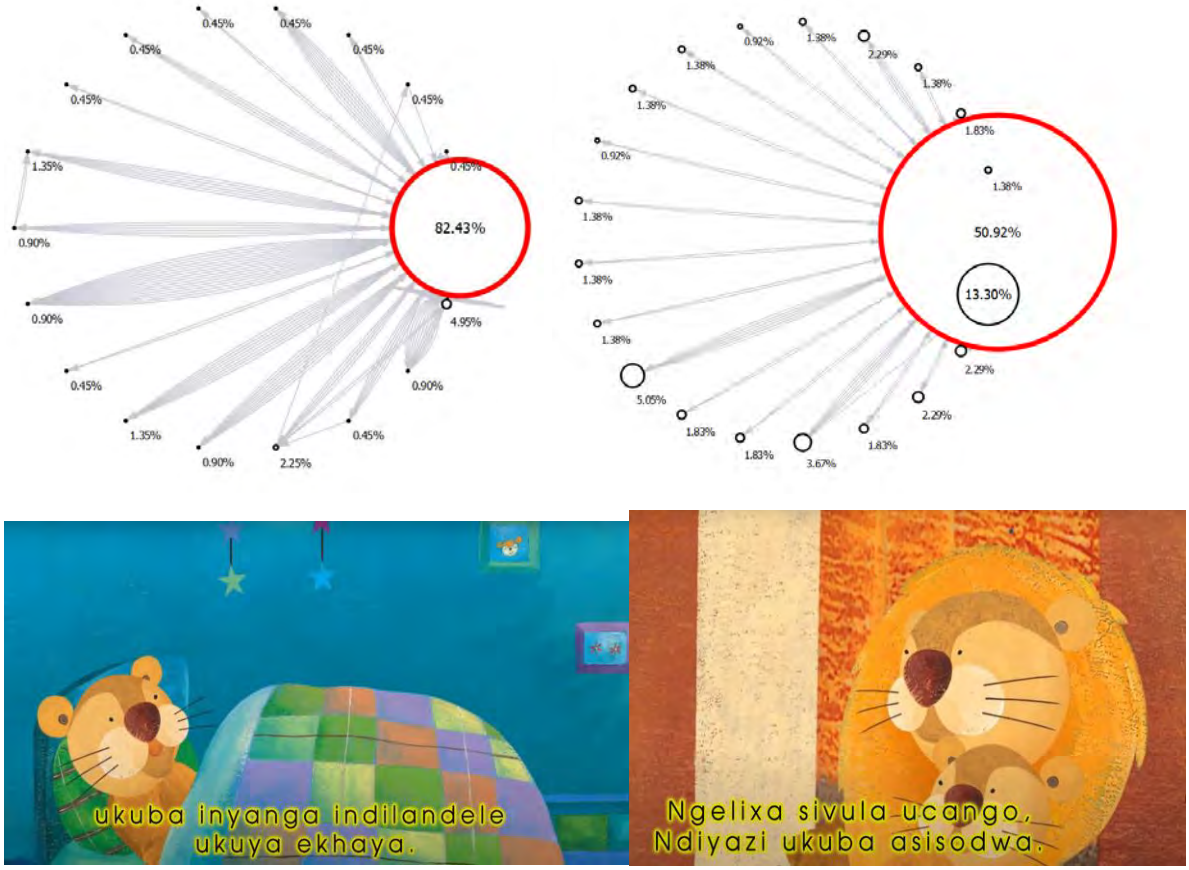


Figure 58(a) The Moon Followed Me Home Spoken Participant STM and (b) Written Participant STM with screenshots subtitled (c) "that the moon followed me home" and (d) "When we open the door, I know that we are not alone"

The duration of the written predominant Participants in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). There is not always a correlation between duration and frequency in the spoken linguistic mode and this discrepancy is discussed in this section.

In Figure 58(a), 4.95% of Participants durationally are spoken Actors, e.g. “iNyanga” (the moon) in “iNyanga indilandele ukuya eKhaya” (the moon followed me home). This is represented by the 13.30% in Figure 58(b) and mostly refers to the moon or father or a reference to the narrator, the lion cub. This is also the most frequently occurring spoken and written Participant.

Figure 59 Screenshots from The Moon Followed Me Home subtitled (a) "I have one friend of mine who comes out at night" and (b) "he has never lost me, he looks very beautiful."

The written and spoken STMs differ slightly in frequency of Participant types. The 5.05% in Figure 58(b) refers to the Actor and Goal appearing simultaneously, such as in the sentence “sivula ucango” (we open the door), with the lions being represented by the si- (we) prefix in “sivula” (we open) and “ucango” (door) is the Goal being opened. This is different to Video 1 and Video 2, as ordinarily Goal is most predominant over the co-occurrence of Actor and Goal Participants, due to the noun-omitting nature of isiXhosa. This might illustrate more than the others that this is a translated video from English, but it also highlights more complex clauses. See the discussion on the complex stative verb form found in Figure 59(b) later in this section. This may cause challenges to learners as it creates more complex structures, although having Actor and Goal present might be clearer than if the Actor was omitted and learners have to remember this information or understand it from the context. This should be tested with learners to better understand potential challenges.

In Figure 58(b), written Identifier and Identified Participants occur in 3.67% of the video durationally, for example, “umgumhlobo othuleyo” (he is a silent friend). While 2.29% durationally of the video (state in Figure 58(b)) is comprised of Carrier and Attribute Participants with Actor Participants, for example in Figure 59(a) “Ndinomhlobo wam oyedwa ophuma ebusuku kuphela” (I have one friend of mine who comes out at night). Another 2.29% state is found below the Carrier and Attribute state in Figure 58(b). This 2.29% state comprises Actor, Goal and Phenomenon Participants, for example in Figure 59(b) “Akazange alahlekane nam yaye ukhangeleka emhle kakhulu” (he has never lost me, he looks very beautiful). In English, "looks" in this context would be a Relational Process. In isiXhosa, you have the verb -khangela (categorised as a Mental Process; it can take a direct object). The -ek suffix reverses the Participant, in a stative verb form, so that the subject concord “u-” refers to the Phenomenon, not the Sensor. The subject concord “u-” here refers to “umhlobo” (friend) in

“Ndinomhlobo” (I have a friend), the preceding clause, which refers to the moon (iNyanga, normally with the subject concord “i-”). For readers to know who “u-” refers to, they must have read, understood and remembered the previous subtitle and understood the previously spoken clause. If this is at a rate that is too fast for learners to comprehend, they could be confused, as these subtitles should ideally, if they were not too long, be presented together for no loss of meaning to occur. This is discussed more in terms of placing “words that belong together” (Lappalainen 2021) in 6.3.7. Furthermore, this is present alongside “Akazange alahlekane nam” (he has never lost me), where “umhlobo” (friend) is represented in the negative subject concords “aka-” and “a-”, which makes this clause even more complex, as it has “friend” as both Actor and Phenomenon in the same clause. The other 2.29% state at the top of Figure 58(b) includes Sensor and Phenomenon. For example, “Ngaphaya kwefestile ndibona umama wam” (Through the window I see my mother).

Figure 60 Screenshots from The Moon Followed Me Home subtitled (a) my father tells me stories on the way and (b) my moon says hello to me, it does not forget

In Figure 58(a), however, 1.35% (the highest state after Actors at 4.95% and Attribute at 2.25%) belongs to Sayer Participants, thus occurring slightly less than in the written mode in

Figure 58(b). The top 0.45% is a co-occurrence of Attribute and Sayer, e.g., in the phrase "Konke kuzolile kuthe cwaka" (Everything is calm and silent) and the bottom 0.45% connecting to states with transition lines is Carrier in the spoken mode. An example Sayer Participant is seen in Figure 60(a).

Spoken Attribute may be high in duration (2.25%) but it has a lower frequency than the 0.90% states from bottom right to left, which are spoken Goal, Sensor and Phenomenon Participants. Thus, while spoken Attribute Participants are on the screen longer, they occur less frequently than Goal, Sensor and Phenomenon Participants, with Goal and Phenomenon Participants more frequent than Sensor Participants. Another Participant with a high frequency is Goal at 0.90% (lower duration) of the video (right bottommost state), such as "ucango" in "sivula ucango". Furthermore, the 0.90% (low duration) states from bottom right to left in Figure 58(a) comprises instances of Phenomenon only, Sensor only (low duration) and Identifier Participant only with Identified comprising 1.35% of the video (the topmost state). An example of an Identified Participant in the video is inkulu" (big) in "Ngamanye amaxesha inyanga yam inkulu" (Sometimes my moon is big). An example of a Phenomenon only is the "u-" subject concord (referring to "umhlobo," friend) in "ukhangaleka" (he looks) in Figure 59(a)**.** Actors and Goals are more frequent and co-occur in the subtitles and spoken language. While Actor and Goal Participants occur for a longer duration than other written Participants, the spoken Participants that are longer in duration are Sayer Participants of Verbal Processes and Attribute Participants of Relational Process. Thus, this presents less of an alignment between the spoken and written modes in terms of predominant Participants and Processes than the previous videos discussed, due to discrepancies in the frequency and duration in the spoken mode. This may influence adherence to Mayer's (2009) limited capacity assumption and Principles of Coherence, Redundancy, Spatial and Temporal Contiguity. As discussed in 2.3.1.3, frequency of exposure may not be as integral for literacy development as Similarity and Contrast co-patterning between modes (Tseng and Djonov 2023). Thus, this thesis focuses on Comparative Relations in Intersemiotic Texture (see 6.3.6) and a Similarity of meanings between modes as a potentially more accurate indicator of literacy development than frequency or duration. Further research in this context is required.

This may be an instance of potential cognitive overload for learners and an argument for the adherence to Mayer's (2009) Coherence and Redundancy Principle, which may be a potential distraction and disturbance for learners when learning with this video. It could further adhere

less to Mayer's (2009) Spatial and Temporal Contiguity Principles in this manner as the visual Participants with the second-highest duration are a co-occurrence of Actor and Existential Participants. This is even though the visual Actors and their Processes are the most frequent and have the longest duration. The applicative structure in Figure 60(a), combined with a co-occurrence of Participants and Processes, may cause challenges to learners as it creates more complex structures for learners to decode.

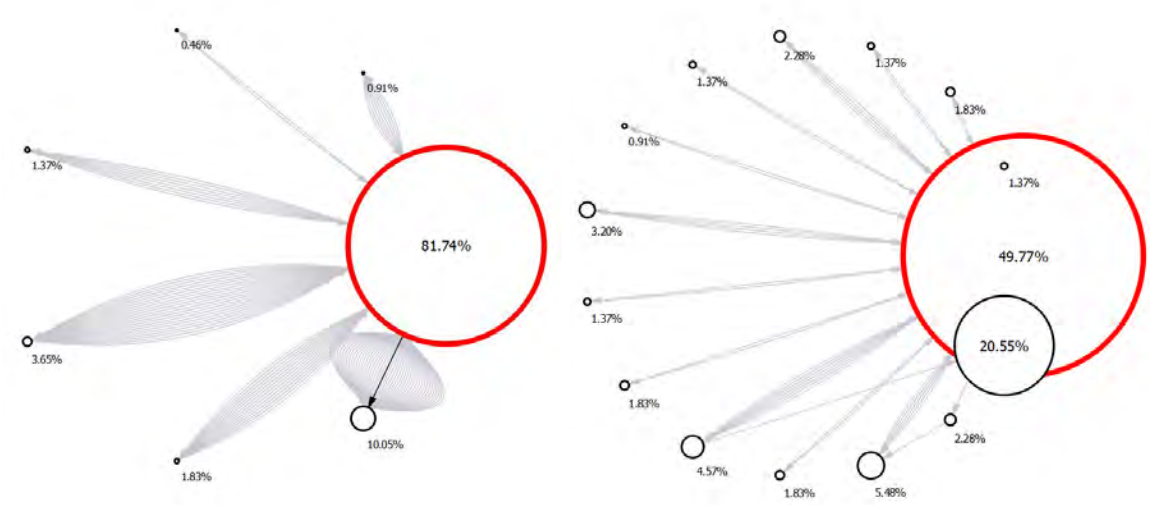


Figure 61(a) The Moon Followed Me Home Spoken Process (left) and (b) Written Processes (right)

The duration of these predominant Processes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In a similar manner and as highlighted by the STMs in Figure 61, the spoken and written Material Processes (see Figure 58(c) and (d) for examples) are most frequent (10.05% state in Figure 61(a) and 20.55% state in the Figure 61(b)). However, Mental Processes have the second-highest duration in the spoken mode at 3.65% (see Figure 61(a)), which is different to Participants of Verbal Processes, which have a longer duration than Sensor or Phenomenon Participants. In Figure 60(b), the 3.20% states illustrate written Verbal Processes. Relational Processes have the second-highest duration in the written mode (4.57%, see Figure 61(b)), which aligns with the Participants in the written mode). Video 2 only had a discrepancy between second-most frequent Participants of the spoken and written modes. The Material Processes are the predominant Processes in both spoken and written modes. Video 3 has a discrepancy in duration between Participants and Processes of spoken and written modes and a discrepancy in duration and frequency between spoken Participants. These discrepancies in repetition between

second-highest durational and second-most frequent Processes of the spoken and written modes could cause learners to not learn as effectively than if they were repeated at similar frequencies for a similar duration in both modes. This video design is less consistent in its system choices than Video 2, which makes Video 2 potentially better for learning in this regard in comparison with Video 3. This is potentially even more challenging for learners to learn these concepts as the Relational and Mental Processes are not often presented in the visual mode (see 6.3.3).

The 1.83% state is Relational and the 1.37% state is Verbal Processes in Figure 61(a), again presenting a discrepancy between duration of the spoken and written modes. There is another state of 5.48% in Figure 61(b) that represents co-occurring Material and Mental Processes, for an example see Figure 59(b). An example of Material and Mental Processes, which also co-occur in the written mode, is found in Figure 60(b), with "Inyanga yam ithi molo kum; ayilibali" (My moon says hello to me, he does not forget). This co-occurrence comprises the 5.83% state in Figure 61(b)). The 2.28% states from right to left in Figure 61(b) are (1) co-occurring Material and Relational, and (2) Mental only. An example of co-occurrence (1) is found in "Akazange alahlekane nam yaye ukhangeleka emhle kakhulu" (he has never lost me, he looks very beautiful). For an example of this (2) co-occurrence, see Figure 60(a). An example of a Mental Process is, "Ngaphaya kwefestile ndibona umama wam" (Through the window I see my mother). The presence of many co-occurrences in the written mode might cause distractions to learners as they are more complex than simple clauses with a single Process or Process type. It might create more of a cognitive load for learners, which is not conducive for learning according to Mayer's (2009) limited capacity assumption. Video 3 has roughly the same number of co-occurrences to Video 1, which is more than Video 2, however, Video 3 contains these co-occurrences for far longer than Video 1. Thus, Video 3 may be more challenging for learners to comprehend and decode than Videos 1 and 2.

Furthermore, Mental Processes are usually challenging to present in the visual mode and this often causes a Contrast between what is presented linguistically and visually. The addition of Mental Processes and ensuring learners comprehend these require the addition of different types of Mental Processes in the visual mode (see 6.3.3). Further tests for learners would need to be undertaken to confirm this. In comparison, for Figure 61(a), the 0.91% state is Behavioural Processes and the 0.46% state is Existential Processes, illustrating that these are of a lengthier time in the video and frequency than the other Processes that occur, but are presented for longer in the written mode. This further illustrates that while Material Processes

for both spoken and written mode are the most frequent and there is a similar pattern of Processes between modes, there is an imbalance in the reinforcement of these Processes in both modes. This is as the duration in the written mode is more on Relational Process than other Processes. Relational Processes are not as frequent in the spoken mode. There is thus a discrepancy between Processes and Participants of the spoken and written modes and this could lead to challenges in literacy development.

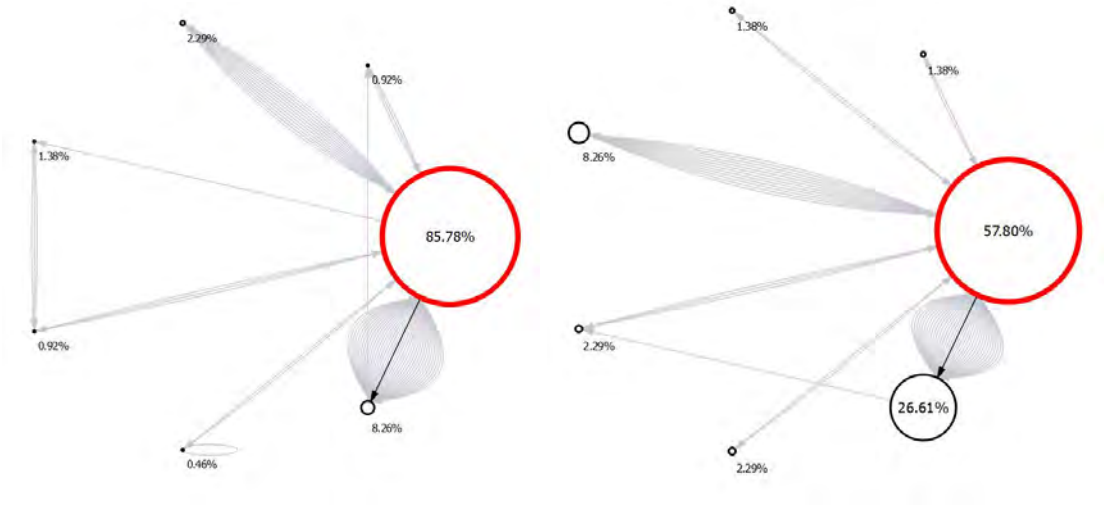


Figure 62(a) The Moon Followed Me Home Spoken Circumstances (on left) and (b) Written Circumstances (on right)

The duration of these predominant Circumstances in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Figure 62 illustrates Circumstances, which follow the same pattern of frequency in this video in comparison with the other videos. There is 26.61% of the video which comprises written Circumstances of Location, usually of Time and Place, for example, “ngobunye ubusuku” (on one night) in “ngobunye ubusuku uliceba, into encinci” (On one night he is a small piece) in Figure 63(c) and “ekhaya” ([to] home) in “inyanga indilandele ukuya ekhaya” (the moon followed me home). These Circumstances of Location are present in spoken language for 8.26% of the video durationally. In addition, there is a further categorisation of 1.38% of Circumstances of Location that permeate across spoken embedded clause structures. In 0.92% of the video, there is a Location Circumstance within a spoken Embedded Clause entirely categorised as Circumstance of Location (2.29% in the written mode), for example in “Apha ebengezelayo esibhaka-bhakeni” (where he shines in the sky). The complexity of embedded constructions, which often include a lot of adjectivised and nominalised Processes might create challenges in decoding, due to the strain it may place on the cognitive load of the learner according to Mayer’s (2009) limited capacity assumption. In the written mode, 1.38% of the video

comprises Circumstances are Extent in the spoken mode (8.26% in the written mode), such as “kude okanye kufutshane nekhaya” (far or close to home) in “Siye sithi naxa sihambela kude okanye kufutshane nekhaya” (Even when we are walking far or close to home). The other 0.92% state is Accompaniment in the spoken mode (in the written mode it is 1.38% and occurring with Circumstance of Location for 1.38% of the video). In the written Mode, 2.29% of the video durationally comprises Circumstances of Manner, for example, “kakhulu” (very) in “akazange alahlekane name yaye ukhangeleka emhle kakhulu” (he has never lost me, and he looks very beautiful). This comprises 0.46% in the spoken mode (see Figure 62(a). The frequent number of Circumstances of Location continues to follow the typical broader syndrome within children’s narratives of a high frequency of Material Processes and corresponding Participants and Circumstances of Location (Guijarro and Sanz 2008).

The high frequency of Circumstances of Location emphasises the focus on the moon and going home in the story, as well as the time of day when the moon appears, and this pattern is similar in Video 1 and 2. This would be typical as an attempt to gain a child’s attention watching the video and could assist in learning, again primarily due to the presence of words deictic and spatial in nature (Levinson 2004; Peña 2020).

This video might be more challenging for learners than Video 1 and Video 2 in terms of the ideational metafunction. Despite the high duration and frequency of Material Processes in both linguistic modes, which may be clearer for learners to comprehend, the discrepancy of frequency of the other Processes between these modes could cause confusion for learners. If this video was used to learn words related to Mental, Relational and other Processes, more visual cues may be needed (see 6.3.3) and clauses should not have as many co-occurrences within them. Both co-occurrences and the reciprocal verb form in Figure 59(b) and applicative verb forms (see e.g., Figure 60(a)) cause clauses to be more complex, potentially causing difficulties for learners to decode and comprehend. The challenges of discrepancy in frequency and duration may result in less adherence to Mayer’s (2009) limited capacity assumption, as well as Principles of Coherence, Redundancy, Spatial and Temporal Contiguity, when comparing this video with other videos. In terms of complexity, there is less adherence to the Coherence Principle. The high frequency of Material and Mental Processes, as well as Circumstances of Location: Place and Circumstances of Location: Time are typical of the narrative patterns in the children’s story genre (Guijarro and Sanz 2008). This would need further research with learners to confirm specific benefits and challenges.

Figure 63 Screenshots of The Moon Followed Me Home (a) sometimes the moon is big, (b) as it is tonight. (c) on one night, he is a small piece

6.3.2 Spoken and written language interpersonal and textual metafunction STMS

The only Mood in the video is Declarative at 52.75%. Since it is the only Mood, an STM is not illustrated here. Like Video 1 and 2, this video offers information more than it requests information, and this type of interaction may cause less engagement from the viewer than if more Moods were present. However, the processing of information in different Mood structures for isiXhosa L1 learners requires more research, as it may be challenging for learners according to Mayer's (2009) limited capacity assumption.

Modality may, when it relates to Moods like Interrogative and Imperative, cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. If the Mood is Declarative, it may increase a load on cognitive capacity that violates the limited capacity assumption. Since there are only Declarative statements in, *The Moon Followed Me Home*, this video has realisations of Modality that could increase learners' cognitive load, as they create complexity in the

language. For example, in future tense sentence that use -*za ku*- and -*yaku*-. An instance of this in the video include, “Ngoku sele siza kufika ekhaya, utata undibalisela amabali apha endleleni.” (Now as we are about to get home, my father tells me stories on the way) in Figure 60(a). This is a modal realisation of probability with median intensity and positive intensity. Another instance of this is included in the sentence “Kum uyakusoloko eqaqamba” (To me, he will always shine bright). The verb -*soloko* in “Uyakusoloko” (he will always) also indicates a modal realisation of usuality with high intensity and positive polarity.

The Circumstance of Location: Time -*soloko* also indicates usuality with high intensity and positive polarity in the example “Usoloko endenza ndihleke, ndincume, ngalo lonke ixesha sincokola” (He always makes me laugh and smile every time we chat). There are a total of three instances of the Circumstance of Location: Time -*soloko* in this video.

There is also an instance of -*soloko* (always) in a *xa*-clause in “Siye sithi naxa sihambela kude okanye kufutshane nekhaya, umhlobo wam unyanga usoloko endifumana nokuba sele sibhadula phi na” (Even when we are walking far or close to home, my friend, moon, always finds me wherever we wander). This is thus a modal realisation of probability with median intensity and usuality with high intensity, both with positive polarity. Another instance is realised between a clause and a *xa*-clause as “Ngelixa sivula ucango, ndiyazi ukuba asisodwa” (When we open the door, I know we are not alone.) in Figure 58(c-d). This is also a modal realisation of probability with median intensity and positive polarity.

Another complex modal realisation is in the clause complex “Sithi xa sifika ngasekhaya, akukho mfuneko yokuba ndijonge ukuze ndazi ukuba umhlobo wam unyanga uhamba ngqo emveni kwethu” (When we get close to home, there is no need for me to look behind us in order for me to know if my friend, moon, is directly behind us). Both *xa*, *ze* (do/did), *yokuba* and *ukuba* are present in this sentence and it results in multiple clauses. The first *xa* clause and its modal relations between clauses realise probability median intensity with positive polarity. The second *yokuba*-clause which is initiated by “akukho” (meaning there is not, translation starting from there is no need for) seems to realise obligation with low intensity and negative polarity. There is no requirement or need for the lion cub to look behind him, to see if the moon is following him. The phrase “ukuze (did) ndazi (knew) ukuba (that)” (in the translation, it is written “in order for me to know if”) indicates the next clause, which indicates a probability with a higher intensity, due to the presence of the modal realisation of obligation in front of it.

It is in a clause complex such as this that truly indicates how complex the meanings are in the videos, which could be incredibly difficult for early readers to grasp, as well as cognitively follow and retain. Another realisation of modality with *ukuba* includes "Nokuba ufihlakele kuloo mafu asibekeleyo, asiphosani" (Even when he is hidden under the clouds, we do not miss each other). The form has changed to "nokuba" (even when) and it realises a probability with median intensity and positive polarity.

Two other instances of modality in this video include the verb *-zange* (never) and the Circumstance of Location: Time *ngamanye amaxesha* (sometimes). For example, in "Akazange alahlekane nam yaye ukhangeleka emhle kakhulu" (He (the friend) has never lost me, and he looks very beautiful), see (10) in 1.2.1.2 and Figure 59(b). This is an instance of usuality with high intensity and negative polarity. In "Ngamanye amaxesha inyanga yam inkulu, njengokuba injalo ngobu busuku" (Sometimes my moon is big, as it is tonight), see Figure 63(a-b), "ngamanye amxesha" (sometimes) realises usuality with low intensity and positive polarity. This is the only instance of *ngamanye* amaxesha in this video and thus, there are more modal realisations of usuality with high intensity than with low intensity.

Figure 64 The Moon Followed Me Home Theme STM

As for the Themes in the video, represented in Figure 64, they all have the same balanced duration in the spoken mode, and in relation to frequency, each Theme occurs once, except

“utata” (father) which occurs twice. This illustrates that there is no particular emphasis on speech duration and frequency for any Theme in particular and that they are all balanced. This, however, causes Themes to lack emphasis and could also hinder focus on specific Participants for learning. This also does not assist in Similarity of meanings between modes, as this may be more conducive for learning than meaning that Contrast, according to Mayer’s (2009) Principles of Spatial and Temporal Contiguity. Examples of Themes include “utata” (father) and “inyanga” (moon), as well as Circumstances, such as Circumstances of Location: Time like “Ngamanye amaxesha” (sometimes) in Figure 63(a). However, none of these Themes appear frequently enough in the spoken mode to highlight any particular patterns. Thus, Themes are not a particularly useful system of analysis in this video. This is closer to Video 2 than Video 1 and another instance where this could have been improved, as this may improve learning with this thesis’s expanded definition of Mayer’s (2009) Signaling Principle. It could also lead to less focus on specific Participants and Processes by learners.

Another challenge and aspect to these videos in terms of the subtitles is the separation of a clause, often within linguistic groups or syntactic phrases. According to Lappalainen (2021), “words belonging together” (e.g. phrases in syntax and/or groups like nominal groups in SFL) should be in the same line of subtitle as per SLS conventions. In this video alone, there are multiple instances where, due to the length of words in isiXhosa, noun phrases and similar phrases are separated and while they often occur in the line underneath them (as there are often two lines of subtitle per subtitle as the maximum can be). They can also occur in the following subtitle and be broken midway which can cause confusion. For example, the subtitle “Ndijonga phezulu esibhakabhakeni ndibone…” (I look up at the sky and see…) is separated from “ukuba inyanga indilandele ekhaya” (that the moon followed me home). While there are less instances of this phenomenon in Video 3 in comparison with Video 2, which has the most occurrences, the presence of clause separations could still create confusion for learners.

6.3.3 Visual ideational metafunction STMs

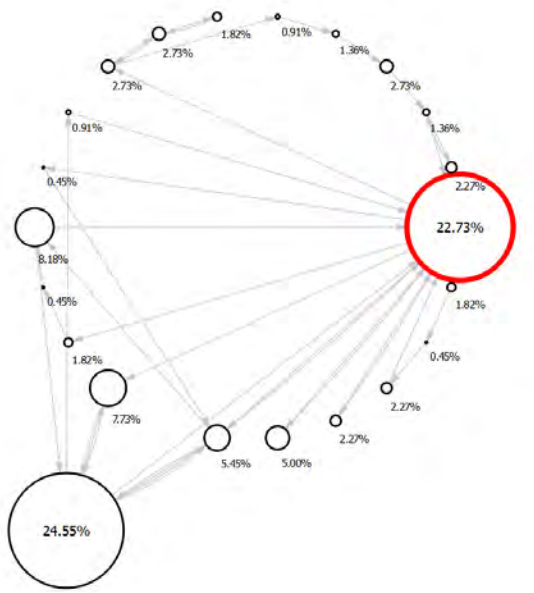


Figure 65 The Moon Followed Me Home visual Participants and Processes

The visual Processes in this video are Action mostly, with the 24.55% state in Figure 65 representing the lion cub, father and moon as Actors[23], all as a co-occurrence. Another co-occurrence, 8.18%, is represented by the moon as Existent, with the father and lion cub as Actors, which indicates a co-occurrence of Action-Existential Processes. Co-occurrences of Participants might be less distracting and more typical within the visual mode than the linguistic mode, provided the information in the linguistic mode is similar in meaning to the information mode. This aligns with Mayer's (2009) Principles of Spatial and Temporal Contiguity. In the video, 7.73% is represented by the moon as Actor. There is 5.45% of the video which includes only the lion cub as Actor and the father as Actor. In 5% of the video the moon is moving in the sky or only existing, and, thus, realises both Actor and Existent. The 2.73% states from right to left include the lion cub as Actor and the moon as Actor, the lion cub as Actor and the lion parents as Actors and the mother as Actor. The challenge of examining Mental Processes in the visuals is discussed in 6.6. This corresponds to the majority of spoken and written Actors of Material Processes in the video, increasing the Similarity rate discussed in 6.3.6. In terms of the visual Participants and Processes, there is less of a Similarity in meaning between the lions

[23] When visuals Participants are looking at something, this is coded as a Material Process, as we cannot truly indicate from the visuals the nuance between the action of looking at something and the perception of seeing something.

and the linguistic Participants of "utata wam" (my father) and the first-person use of I as the lion cub than with the other videos discussed. This adds further complexity to associating meanings between Participants that could make learning more difficult. If lions were used in the language here, it would have provided an opportunity for animal-related vocabulary development in terms of literacy through these videos. While family-related vocabulary is used in Video 3, it might have been better to illustrate this with a visual human family, preferably one closer in resemblance to the target culture, than the lion family (Zohrabi *et al*. 2019). The further into an imaginary space one enters in order to engage the child, the higher the risk that this thesis' expanded definition of Mayer's (2009) Personalization Principle may not be adhered to, as the world in the video becomes less relatable to the child.

6.3.4 Visual interpersonal metafunction STMs

The system of Gaze is discussed in this section, then Social Distance and Zoom, followed by Camera Movement and Horizontal and Vertical Perspectives. No Salient effects were found in this video and thus no STM on Salience is discussed. This could have assisted in a closer adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle. This may foster learning.

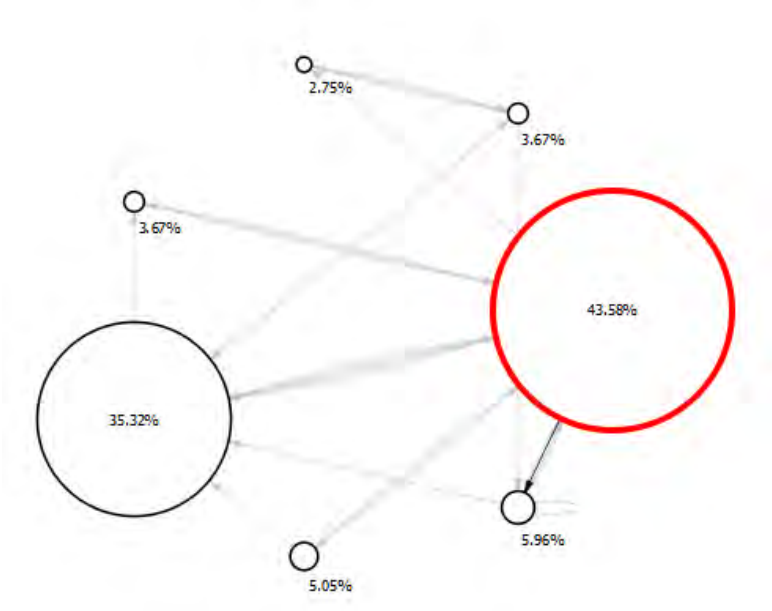


Figure 66 The Moon Followed Me Home Gaze STM

The duration of system Choices for Gaze in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In terms of Gaze depicted in Figure 66,

all instances of Gaze are Indirect and there is no engagement with the viewer, which for a children's video, could cause less attention in comparison to videos that contain instances of Direct Gaze. The 35.32% represents where father and lion cub have Indirect Gazes simultaneously, while the 5.96% refers to only the lion cub with an Indirect Gaze. The 5.05% state includes the father with Indirect Gaze. The 3.67% state from left to right includes animals and insects with Indirect Gazes, and the combination of father, lion cub and mother with Indirect Gazes. The 2.75% state includes the mother with Indirect Gaze. This further illustrates the more frequent presence of the father and lion cub than any other Participants (excluding the moon, which cannot be included in the Gaze system discussion, as a moon does not have eyes and cannot gaze). The lack of direct engagement with viewers also highlights that the world of the video is different from the world of the viewers in Video 2 and Video 1. The presence of Indirect Gaze does, however, correspond better to the sole use of the Declarative Mood in the language to offer information. There is, however, a context with the lions and the rondavel hut in this video that is more African in nature than in comparison to the other videos. The objects such as the hut and lions create items more contextually appropriate for a South African learning context than the other videos.

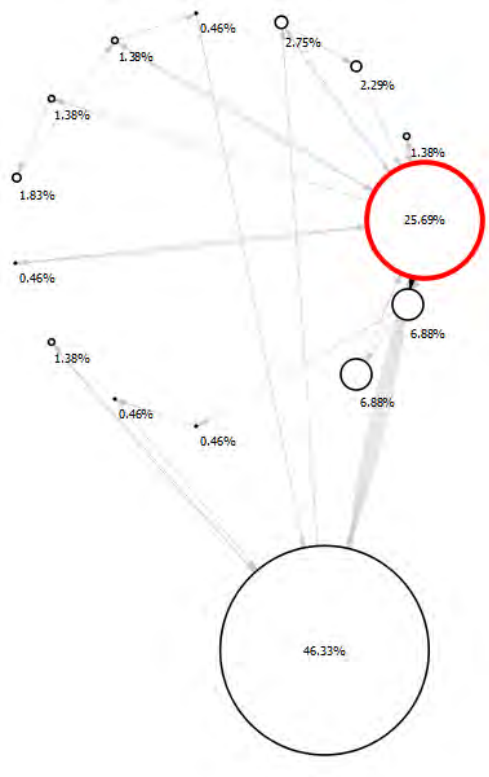


Figure 67 The Moon Followed Me Home Social Distance and Zoom STM

To illustrate how Zoom occurs in the context of the Social Distance Visual system, they are portrayed in the same STM for Videos 3-5. Social Distance STMs do not account for trinary choices of shots, as some states indicate duration of shot transitions, in which there are co-occurrences of system choices. Thus, we only focus on the three system choice states, which are Close, Medium and Long and their co-occurrence with system choices in Zoom. The duration of these system choices for Social Distance and Zoom in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In terms of Social Distance, most of the video is captured in Long shot indicating the illustration of the moon following the lion cub and his father home. Long shot also corresponds with the offer of information conveyed in the linguistic Declarative Mood and presence of Indirect Gaze in the video (see Guijarro and Sanz 2008). In Figure 67, the 46.33% state represents Long shots. When there are Medium shots, they are focused on the lion cub (6.88%), moon (6.88%) and the father (2.29%). When Zoom In occurs, it is around the lion cub (2.75%). This emphasises the Actions over particular Participants, but when the video is focused on Participants, this highlights the story's focus on the moon and the lion cub. However, more Medium and Close shots could have placed more emphasis on particular Participants and Processes and could have resulted in better engagement with the video by learners (Guijarro and Sanz 2008). This is more similar to Video 2 than Video 1, and results in potentially less adherence to the Signaling Principle. This may improve learning. However, the use of Long shots focuses on the entire phrase instead of specific words. This knowledge might assist constructing learning goals for the video and potentially not using it for specific vocabulary literacy development at the word level.

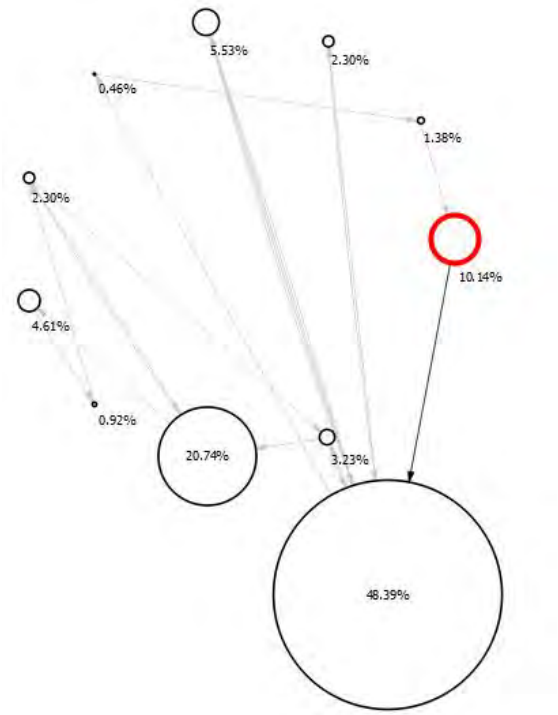


Figure 68 The Moon Followed Me Home Horizontal and Vertical Perspective STM

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 68, the frequency of occurrences is not pertinent to the analysis of these visual systems. Most of the video comprises Stationary shots that are Front and Eye Level (48.39% state depicted in Figure 68). All shots are either Eye-level or High, and thus 20.74% of the video durationally is Stationary, Front and High angle, usually depicting father and lion cub. There is also 4.61% of the video durationally that comprises High, Front and Dolly shots, where the camera follows the father and lion cub. The 2.30% state comprise Dolly, Front and Eye Level shots. The High angle shots clearly give a view from the moon's perspective, which assists in focusing the moon as a predominant Participant. This contributes to its emphasis and the Similarity between visuals and language for potential learning, according to Mayer's (2009) Principles of Spatial and Temporal Contiguity. A Pan with Front and Eye Level shots comprises 5.53% of the video durationally. In the video durationally, 3.23% further includes Stationary and Front shots. There is usually a high frequency of Long shots in children's narrative structures, as well as Front and Eye Level viewing to allow for children to be immersed in the world of the story (Guijarro and Sanz 2008). While the presence of other types of shots, such as High angle, may create a power distance (providing the viewer with more power over the scene) than the other videos, this could have been a choice due to the main Participant being the moon. The moon is higher in the sky and looks down on the father lion and his son. Therefore, it may not cause a lack of immersion and less adherence to this thesis'

expanded definition of Mayer's (2009) Personalization Principle. It may compensate for the lack of Medium and Close shots in the video durationally by narrowing the distance between the viewer and Participants. Further tests with learners are required to examine the effects of different systems in the interpersonal and textual metafunctions on learner engagement for isiXhosa literacy development.

6.3.5 Visual textual metafunction STMs

Figure 69 The Moon Followed Me Home Information Value

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 69, the frequency of occurrences is not pertinent to the analysis of these visual systems. As observed in Figure 69, most of the video presents the father and lion cub as Given information and moon as New information (23.64%). Even though Figure 64 illustrates that there are no specific dominant Themes, this STM of Information Value emphasises the high frequency of the moon, father and lion cub as Actors in the video in the visual, spoken and written Modes. It also mirrors predominant Subject Participants and sentence structure in the linguistic Mode. The convergence of meaning emphasis on these Participants and their actions from the linguistic and visual modes contribute to the high Similarity in this video, which is ideal for learning according to Mayer's (2009) Spatial and Temporal Contiguity Principles.

The moon is placed in Centre position for 8.18% of the video durationally, which also highlight the strong emphasis of the story on this Participant and may assist in the adherence to this thesis's expanded definition of Mayer's (2009) Signaling Principle. The 5.49% state shows the father and lion cub as Given information on their own and 4.09% shows the father and lion cub as New information. This is in contrast with 7.73% of the video durationally comprising a co-occurrence of father and lion cub as New information and moon as Given information.

Furthermore, in 2.75% of the video durationally, there are artificial frame lines of a window frame around the moon, which further emphasises this Participant, assisting in adherence to this thesis' expanded definition of the Signaling Principle. This also occurs in the other two videos, although it is less prominent durationally in this video in comparison with Video 1 and 2. Due to the lack of states (97.25% of the video durationally does not comprise any Framing devices), an STM for Framing was not included here. More Framing devices could have been used, like in Video 1 and Video 2, to allow for an emphasis of visual components, to potentially adhere better to this thesis' definition of Mayer's (2009) Signaling Principle. It may have also enhanced the meaning relations between the visuals and linguistic components in the story, potentially improving learning in these areas according to Mayer's (2009) Principles of Spatial and Temporal Contiguity.

6.3.6 Intersemiotic Texture STMs

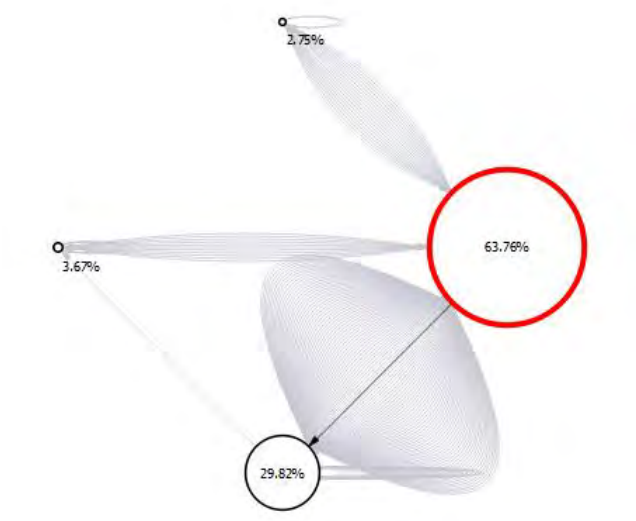


Figure 70 The Moon Followed Me Home Intersemiotic Texture Spoken language and visuals STM

The duration of these system choices for Comparative Relations in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In terms of spoken and visual modes, as represented by Figure 70, there are more instances of Similarity than Contrast. Only 2.75% of the video reflects Contrast. The 29.82% state reflects Similarity of Correspondence, such as visuals that show the moon following the lion cub and corresponding language. In Figure 70, 3.67% corresponds to Similarity of Collocation, such as when the word "utata" (father) is mentioned and the video illustrates an anthropomorphic father lion in the visuals, which is not quite the same as the language. However, there is a similar association of meaning attached between these two Participants. This is similar to the presence of the lion cub in the visual mode in Video 3 and the corresponding language is "I." The visual relationship between the lions are connected to the linguistic words and the semantic relationship of father and son, thus the image of these two lions have more than one meaning in this video, based on the language used. The image of the lions on their own would have been limited to a visual representation of lions and potentially more limited in conveyed meanings. Intersemiotic Similarity of Collocation may not be as effective for learning in comparison to Correspondence, as the relationship of similar meanings is less direct. It may require further processing and according to Mayer's (2009) assumption of limited capacity, less processing per channel is preferred to reduce the cognitive load on viewers.

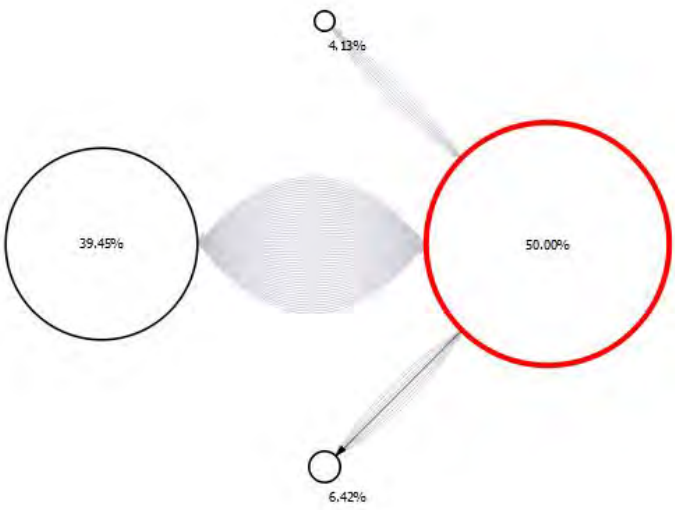


Figure 71 The Moon Followed Me Home Intersemiotic Texture Written language and visuals

The ratio of Similarity and Contrast between the spoken linguistic mode and the visual mode proportionate to the number of words in the video is higher and arguably better in Video 3 than in Video 1 and 2 (see 6.6). A comparative ratio of Similarity and Contrast is witnessed in the relationship between the written and visual modes in Figure 71, with 39.49% reflecting Similarity of Correspondence durationally, 6.42% reflecting Similarity of Collocation durationally and 4.13% reflecting Contrast durationally.

Intersemiotic Additive Relations convey different meanings, but the meanings relate to one another, i.e., when semiotic resources of one mode add information to another (Liu and O'Halloran 2009:379). For example, in the subtitle "Ndinomhlobo wam oyedwa (omnye) ophuma ebusuku kuphela" (I have one friend of mine who comes out only at night). The image reveals the `friend's identity (i.e. the moon). This suggests that the Intersemiotic Texture may facilitate learning, as this may adhere to Mayer's (2009) Spatial and Temporal Contiguity Principles and this thesis' expanded definition of Mayer's (2009) Signaling Principle.

6.3.7 Subtitle rate

The total duration of subtitles (subtitle presentation rate) is 1 minute 41 seconds. Therefore, 41 seconds divided by 60 seconds in a minute = 0.68 + 1 minute = 1.68 minutes. The total number of words (separated by spaces) is 173 words. If words had hyphens, they were included as a single word. 173 words divided by 1.68 = 102.98 WPM. If the maximum reading threshold of 50 WCPM) is true for Grade 3 learners. This is not only 52 words above the threshold and maximum reading score, but it is almost seven times the best Grade 2 average reading score for a longer passage static text in the beginning of the year (14.30 WCPM). Additionally, it is five times their ORF rate three months later (19.95 WCPM) and the average benchmark for Grade 2 (20 WCPM). The WPM subtitle rate in the video is more than double the average benchmark for Grade 3 (35 WCPM), three times the post-test ORF mean for Grade 3 learners (28 WPM) and four times their initial ORF rate (22.85 WPM).

The total syllables per minute (SPM) were calculated with a similar method. Therefore, 544 syllables divided by 1.68 = 323.81 SPM. The reading score for Grade 2 readers was 141 SCPM and Grade 3 readers was 179 SCPM. The mean scores, however, which reflect majority reading rate is far lower at 45.4 SCPM pre-test and 65.20 SCPM post-test for Grade 2, and 76.95 SCPM pre-test and 96.10 SCPM post-test for Grade 3. The syllable rate in the videos is still not at a

pace the learners would comfortably manage. While silent reading is faster than reading out loud, the large difference between the subtitle presentation rate and Oral Reading Fluency scores indicates the likelihood that the learners could not read the subtitles for this video. This violates this thesis' expanded definition of Mayer's (2009) Segmenting Principle. A limitation of this study was the inability to measure silent reading rate, as this is entirely dependent on the learner self-reporting this information, which is not necessarily accurate (Probert 2016). Furthermore, eye-tracking does not necessarily state whether decoding took place, even if it tracks the movement of the eye along the subtitle, thus there are limitations in understanding an objective silent reading rate.

6.3.8 Conclusion and summary of Video 3 analysis

There are mostly Material Processes in the spoken and written language in this video. While the written language focuses longer and more frequently on Relational Processes, there is a slight imbalance between linguistic modes as Mental Processes occur in the spoken language more frequently. The Relational Attribute Participant, while occurring less frequent than Mental Participants, occurs longer in duration than Mental Participants and Goal for spoken language. This discrepancy does not occur in the written language, which has Relational Participants for a longer duration and more frequently between subtitles than Mental Participants. This indicates discrepancies between Participants in relation to frequency and duration between modes. While these discrepancies and the presence of complex structures does not influence the Similarity Rate overall, it is less effective in comparison to Video 1 and Video 2 and may cause potential cognitive overload. It may also cause distractions to learners according to Mayer's (2009) limited capacity assumption.

Nevertheless, the high frequency of Material Processes correlates with the frequency of these Processes in the visual Mode, and the large number of Circumstances of Location seems to continue the typical pattern for this genre of video catered towards children. As per other videos, the presence of Front angles further attempts to maintain children's attention, and the presence of High angles is discussed as potentially assisting adherence to this thesis's expanded definition of Mayer's (2009) Personalization Principle and not distracting learners. Further tests in this area are required to confirm this.

The lion cub, the father and the moon as Participants are emphasised with Framing devices, High angles and Medium shots and the lions' statuses as Actors in the visual and linguistic modes are emphasised by their presence as Given information. There is less adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle than in the other videos and this could be improved by the addition of Framing devices. It may also be improved by the addition of more consistent Themes in the linguistic mode**.** There is a high amount of Similarity of Correspondence in the video, as well as Collocation, and thus this video adheres to Mayer's (2009) Principles of Spatial and Temporal Contiguity. The high Similarity rate establishes this video as a good video for learning. Similarity of Collocation may be similar but it could potentially cause learning difficulties and focus on different vocabulary than what the video offers.

However, the average subtitle rate for this video is the highest of the videos at 102.69 WCPM (323.81 SCPM) and far too high for the average Grade 2 and 3 learners according to benchmarks and intervention measurements. Thus, while the design is ideal for the video, the subtitles make the learning conditions unattainable. This adheres to Mayer's (2009) Redundancy Principle and violates Mayer's (2009) Segmenting Principle which is elaborated on in detail in the next section (6.6).

6.4 Video 4 - Jungle Tumble or IiNtshukumo zaseNdle

Jungle Tumble or *iiNtshukumo zaseNdle* reveals some hidden animals in the jungle, first a parrot behind a branch, then a frog behind a stone, then a leopard behind a bush, and a chameleon. Furthermore, the video reveals a snake behind a branch, then crocodiles behind bushes, a monkey and another leopard. There is no narrative plot unfolding chronologically in this video, as depicted events randomly occur independent of each other. Important spoken and/or visual Participants to note for this video include parrots, also discussed under birds, monkeys, frogs, a hippo, leopards, a chameleon, insects, snakes, crocodiles, a giraffe, a mouse, a zebra, an elephant and a tortoise. The analysis of this video is divided into spoken and written ideational metafunction STMs for language (6.4.1), followed by the interpersonal and textual metafunctions in these modes (6.4.2). The ideational metafunction STMs in the visual mode are examined in 6.4.3 and 6.4.4 examines the interpersonal metafunction STMs in the visual mode. The textual metafunction STMs in the visual mode are examined in 6.4.5 and 6.4.6 examines Intersemiotic Texture. There is a discussion of subtitle rate in 6.4.7. Lastly, section 6.4.8 concludes the analysis of this video.

6.4.1 Spoken and written language ideational metafunction STMs

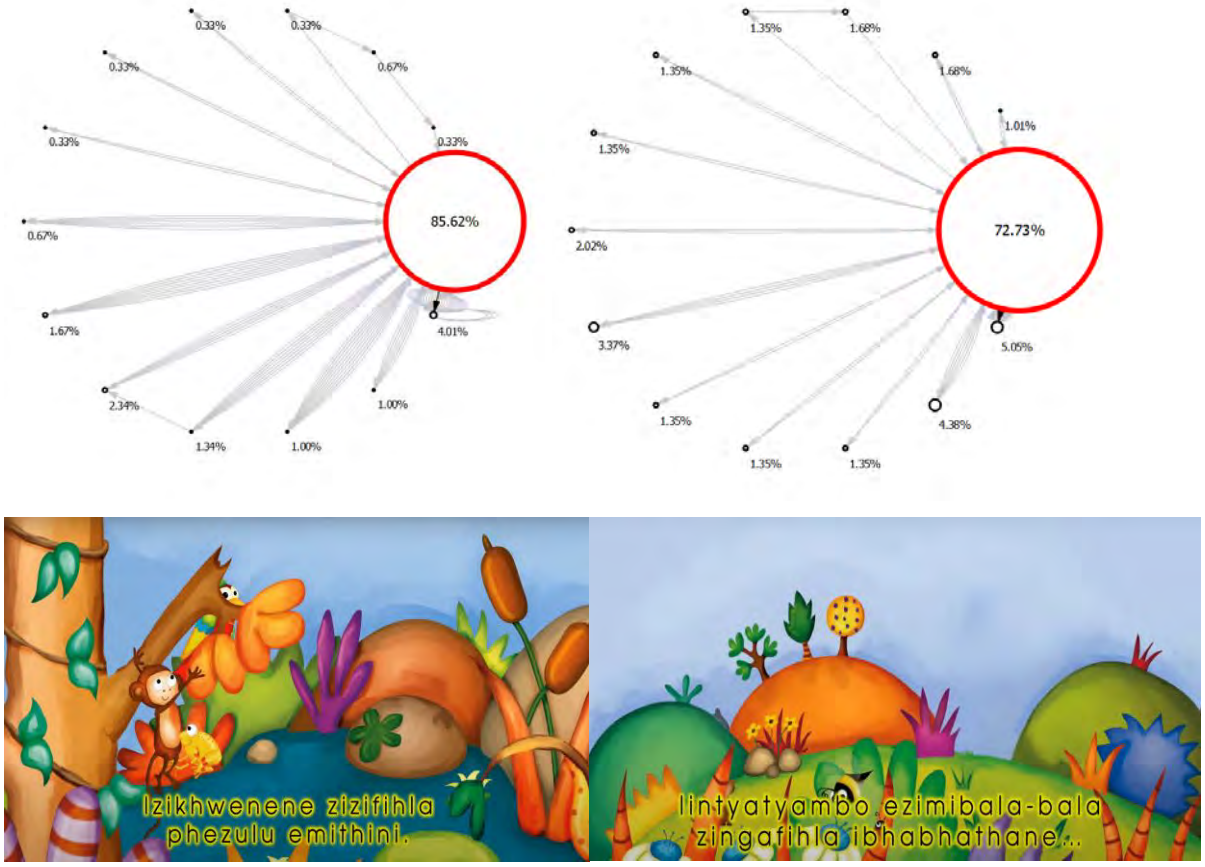


Figure 72(a) Jungle Tumble Spoken Participant STM and (b) Written Participant STM with screenshots subtitlted (c) "parrots hide themselves up in the tree" and (d) "multi-colured flowers can hide the butterfly for a long time"

The duration of the predominant written Participants in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). The frequency and duration do not, however, correspond in the spoken linguistic mode. In Figure 72(a) and (b), Actors are the predominant Participant durationally and by frequency in both modes (4.01% in Figure 72(a) and 5.05% in Figure 72 (b)), for example in Figure 72(c), the parrots in "izikwenene zizifihla phezulu emthini" (parrots hide themselves up in the tree). In Figure 72(b), 4.38% of the video durationally includes a co-occurrence of Actor and Goal. For example, in Figure 72(d) "iintyatambo" (multi-coloured flowers) is the Actor and "ibhabhatane" (butterfly) is the Goal in the clause "iintyatyambo ezimibala-bala zingafihla ibhabhatane ixesha elide" (Multi-coloured flowers can hide the butterfly for a long time). Goal is represented in Figure 72(a) as 1% (rightmost state) and 1.01% in Figure 72(b). While Goal Participants has the same duration as Sensor Participants in the spoken language (1.00%), Sensor Participants occur more frequently than Goal Participants in the spoken language. This discrepancy may cause

confusion for learners according to the adherence of Mayer's (2009) limited capacity assumption, as well as Principles of Coherence, Redundancy, Spatial and Temporal Contiguity, when comparing this video with other videos. Like Video 3, considering isiXhosa is a noun-omitting language, the higher duration Actors over Goals may originate from the English translation of this video. It may emphasise the Goals of the Material Process over the Actors and the predominance of Participants for the Material Process is typical of children's narrative structures (Guijarro and Sanz 2008).

Other Participants in the video occur as frequently as one another without any secondary predominant Participant. Existents in the video are represented by the 1.67% state in Figure 72(a), and the 1.68% state in Figure 72(b), e.g. "Kodwa ungabona…ukuba ikhona!" (But you could see…that it is there! [the leopard]). Existents are the third-most frequent Participant in the spoken mode after Actor and Sensor Participants. The 1.34% spoken state with the most transitions represents Carrier Participants and the 2.34% spoken state is Attribute. While the spoken Attribute Participant has a longer duration than the spoken Carrier Participant, it occurs less frequently than the spoken Carrier Participant. The duration of the Attribute is longer than the Carrier in the clause in Figure 73(c), as the Attribute "umbala womnyama" (any colour of the rainbow) contains more words than the Carrier "ilovane" (a chameleon). In the Figure 73(a), the other 1% state represents Sensor, for example, in "Lithi, "Ndiyakuthanda ukubukela iiNtshukumo zaseNdle!"" (it says, "I like watching Jungle Tumble!"). Co-occurrences in the written mode include Carrier and Attribute (3.37%), Actor and Existent (2.02%), and Goal and Sayer and Sensor (1.68%).

Figure 73 Screenshots from Jungle Tumble subtitled (a) The frog is green like a leaf, (b) "but you could see that it is there", (c) "the chameleon can look like any colour of the rainbow", (d) "it says, "I like watching Jungle Tumble""

This illustrates that like the previous three videos, this video has a high duration and frequency of Material Processes and Participants. This is further correlates to visual Participants in 6.4.3 and Intersemiotic Texture in 6.4.6. The correspondence between frequency and duration of Participants in both modes is similar in the written language, with Sensor Participants occurring more frequently than Goal Participants in the spoken language. However, Sensor Participants do not occur as frequently or for as long as duration as Goal Participants in the written language. Further discrepancies between Existent, Carrier and Attribute Participants between spoken and written modes and their co-occurrences may cause confusion for learners. This may be due to less of a potential adherence to Mayer's (2009) limited capacity assumption, as well as Principles of Coherence, Redundancy, Spatial and Temporal Contiguity, when comparing this video with other videos.

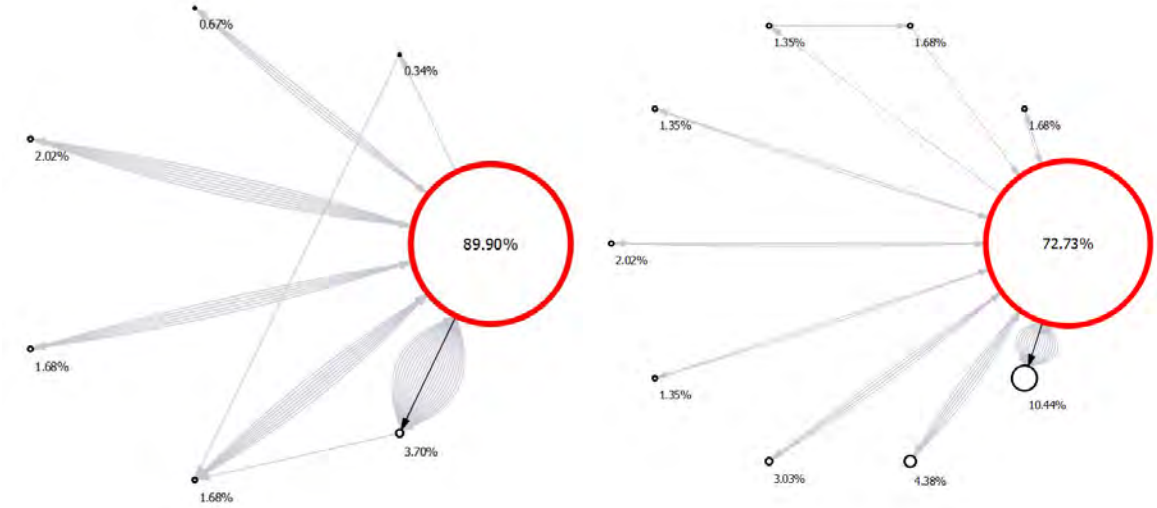


Figure 74(a) Jungle Tungle Spoken Process STM (left) and (b) Written Process STM (right)

The duration of these predominant Processes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Similar patterns are observed in the Processes for both modes as depicted in Figure 74, except that Relational and Mental have a higher rate durationally in spoken Processes than Behavioural does. The video shows 3.70% spoken Material Processes (10.44% written Material Processes), 2.02% spoken Existential Processes (1.68% Written) and 1.68% Relational Processes (1.35% written). In the video, there are also 1.68% Mental Processes and 0.67% Behavioural Processes (which is higher in the written mode at 1.35%). Furthermore, 0.34% of the video includes spoken Verbal Processes. The frequency of Processes in the spoken mode mirror the frequency of spoken and written Participants, as well as visual Processes in 6.4.3. However, the discrepancy between the frequency of Behavioural Processes in the spoken mode and the written mode, as well as the discrepancy between the frequency of these Processes and their Participants in both modes. This may increase learner's cognitive load and cause difficulties in decoding and processing the information, according to Mayer's (2009) limited capacity assumption. As discussed in 2.3.1.3, frequency of exposure may not be as integral for literacy development as Similarity and Contrast co-patterning between modes (Tseng and Djonov 2023). Thus, this thesis focuses on Comparative Relations in Intersemiotic Texture (see 6.4.6 for this video). It also focuses on Similarity of meanings between modes as a potentially more accurate indicator of literacy development than frequency or duration. Further research in this context is required.

Co-occurrences in the written mode include (1) Material and Existential for 2.02% of the video and the below 1.35% state of the video (see Figure 74(b)), (2) Material and Mental for 3.03%

of the video, for example in “Kuphuma…Ufudwazana eziva kamnandi,” (little tortoise comes out feeling nice/good) and (3) Material, Verbal and Mental for 1.68% of the video (see example in Figure 73(d)). It is not unusual based on the language for Existential Processes and Material Processes to co-occur (see Figure 73(b)). However, co-occurrences may create less cognitive capacity for learners, which may result in difficulties in learning according to Mayer’s (2009) limited capacity assumption.

There are multiple instances of reciprocal verb constructions in this video, for example see the *-en-* suffix in Figure 72(a). There is little research to understand how early learners comprehend reciprocal constructs in literacy, thus, the complexity of these constructions for this level is not well-known. This requires more research to determine if this construction is appropriate for these grades and level of literacy.

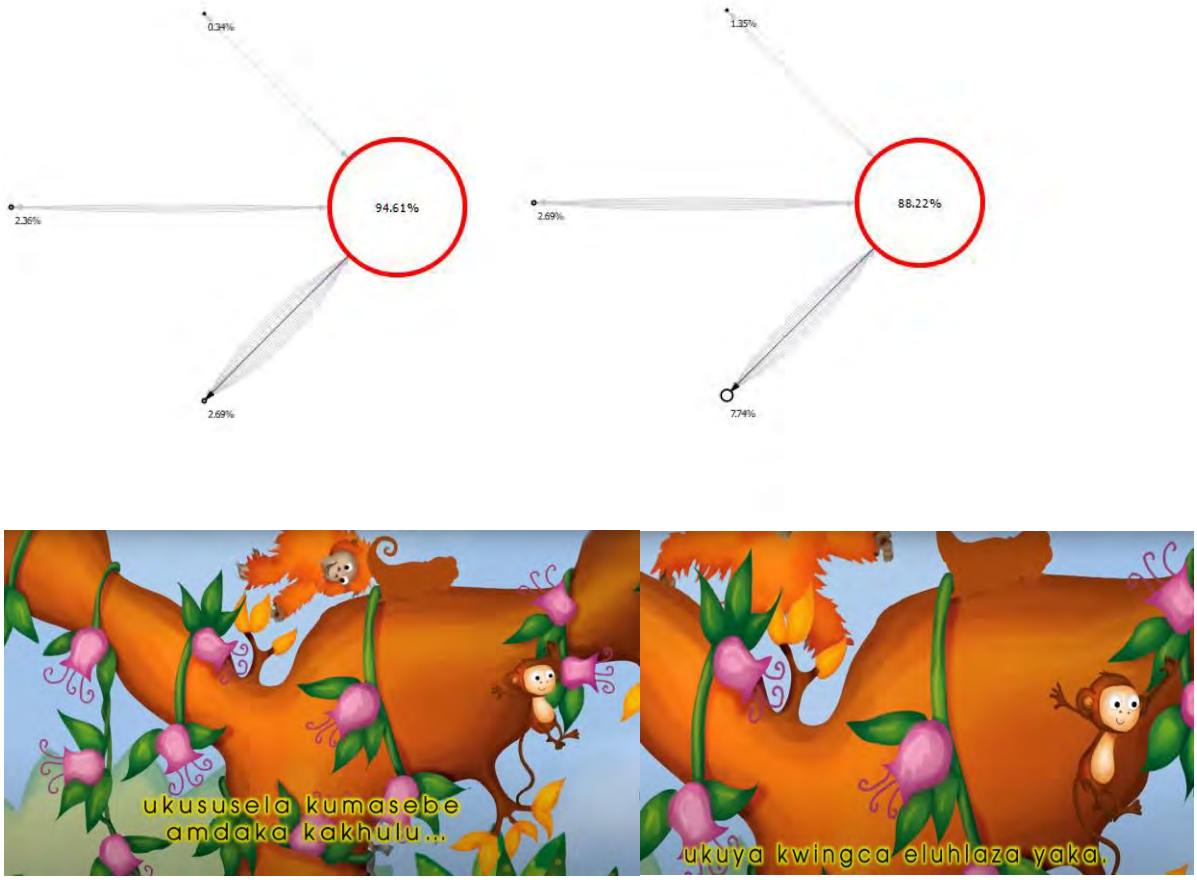


Figure 75(a) Jungle Tumble Spoken Circumstances STM (left), (b) Written Circumstances STM (right), with screenshots subtitled (c) "ranging from very dark brown branches" and (d) "to very green grass"

The duration of these predominant Circumstances in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). As with previous videos, there is a high frequency of Circumstances of Location: Time and Circumstances of Location: Place (2.69% in Figure 75(a) and 7.74% in Figure 75(b)), for example, “phezulu emthini” (up in a tree) in “izikwenene zizifihla phezulu emthini” (parrots hide themselves up in the tree). Circumstances of Extent represent 2.36% of the spoken video (2.69% in written mode), for example, “ukususela kumasebe amdaka kakhulu…ukuya kwingca eluhlaza yaka” (ranging from very dark brown branches to very green grass) in Figure 75(c-d). The 0.34% state represents spoken Circumstances of Accompaniment (1.35% in written Mode), for example, “imilomo yabo ivala “ngoKram!”” (their mouths close with a Kram sound), in Figure 76. The emphasis of the setting is typical of children’s narrative structures (Guijarro and Sanz 2008). Furthermore, frequent Circumstances of Location: Time may assist in maintaining the attention of a child as they are Time deictic words (Levinson 2004; Peña 2020).

Figure 76 Screenshot from Jungle Tumble subtitled "their mouths close with a "Kram!""

6.4.2 Spoken and written interpersonal and textual metafunction STMs

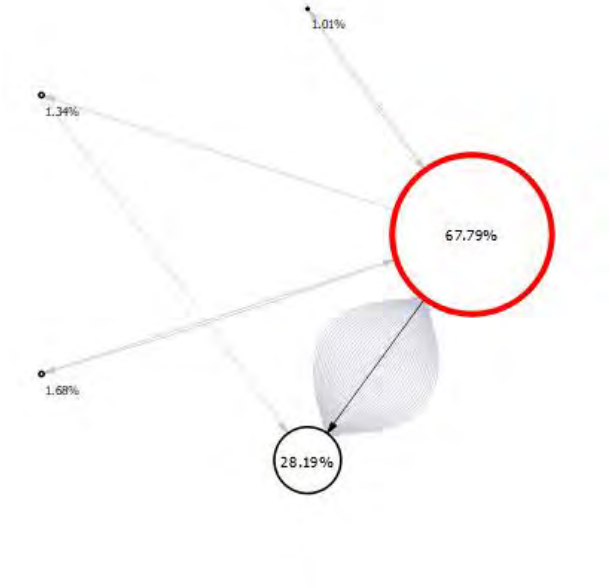


Figure 77 Jungle Tumble Mood STM

While all three Moods are present in this video, indicating a high level of viewer interaction and engagement, the most predominant Mood is still Declarative (28.19% of the video, see Figure 77). Declarative and Imperative clauses are present together for 1.68% of the video. Interrogatives are 1.34% of the video and Imperatives are 1.01% of the video. This conforms to the narrative pattern typical of children's stories, which offers information (Guijarro and Sanz 2008). The addition of these Moods might assist in the adherence of this thesis's expanded definition of Mayer's (2009) Signaling Principle. This may improve literacy development. However, there is a lack of research on whether learners can fully comprehend meanings conveyed in different Mood clause structures. There is, however, a low frequency of Interrogative and Imperative Mood structures, therefore, it may not greatly hamper a learner's cognitive load.

Modality may, when it relates to Moods like Interrogative and Imperative, cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. For example, in the Imperative Mood, there is a clause "Tsala isebe lomthi…Uza kubona!" (Pull a tree branch… You will

see!). The future tense markers *-za-* and *-ku-* realise probability with median intensity and positive polarity. This is the only instance where Modality directly addresses the viewer of the video, and thus, overall, there is obligation with high intensity and positive polarity, in which the viewer is required to pull the tree branch in order to see the parrot. The sentence in the Interrogative Mood "Ingaba yinyoka emdaka le ihleka emthini?" (Is that a brown snake laughing in the tree?) seems to realise obligation with low intensity and positive polarity, as it is a yes/no question. As mentioned above, the presence of the Interrogative and Imperative may lead to the learner being more engaged in the video. The Modality with both high and low intensity between the speaker and the viewer creates a conflict of power relations. In the Interrogative, the learner is less obligated and thus, has more power and is made to feel a part of the scene (see Zohrabi *et al.* 2019). In the Imperative Mood, the learner is more obligated without as much choice, and the speaker has more power when presenting information. This causes a dimension of conflict of adherence to the Personalization Principle, as clause in the Interrogative Mood with low obligation would be more informal to use than the Imperative. The clause in the Imperative Mood results in a larger social distance between the learner and the video than the clause in the Interrogative Mood which narrows the social distance between speaker and learner. Thus, according to Mayer's (2009) Personalization Principle (see 3.2.1), the Interrogative Mood and low obligation may be better for learning, and for literacy development. This is beyond the scope of this thesis to investigate, however, and more research is required.

If the Mood is Declarative, it may increase a load on cognitive capacity that violates Mayer's (2009) limited capacity assumption, as they create complexity in the language. This is due to the prefix *-nga-* (see 1.2.1.2). For example, in "Amachaphaza engwe ayayinceda ukuba ingabonakali" (The spots of a leopard help it to not be seen) see (7) in 1.2.1.2, there is a modal realisation of probability with high intensity and negative polarity. There is also an instance of probability with high intensity and positive polarity in the clause, "Kodwa ungabona…ukuba ikhona!" (But you could see… that it is there!) in Figure 73(b). While it occurs in the video more than once, due to it being reflected in a single prefix, its duration is not very long, however, thus the STM is not included here. There is a lack of research on when and how isiXhosa L1 learners decode and comprehend Modality in verbs in the Foundation Phase. While it may be easy to decode, as it is represented by a prefix and one syllable, learners may not necessarily comprehend the change in meaning if it was removed (Rees 2016). A complex semantic construction including Modality may be challenging for learners to read if it demands

a greater cognitive capacity of the learner (see Mayer's (2009) limited capacity assumption in 3.2.1.). Furthermore, Modality is challenging to mirror in the visual mode to assist learners in their comprehension. There is a lack of research on when and how isiXhosa L1 learners decode and comprehend Modality in the Foundation Phase.

Figure 78 Jungle Tumble Theme STM

The Spoken Themes that are most prevalent include "iintyatyambo" (flowers, see Figure 72(d)) at 1.00% of the video and "umngqumo" (the roar) at 0.67% of the video, e.g. "Umgqumo wehlosi uvakala okweendudumo" (the roar of a leopard sounds like thunder). This is more due to the length of time they are uttered than their frequency. While the flowers may be visible, the roar is not, and this does not really engage the viewer's attention as well as other videos would to specific words and concepts in the visual mode. This video thus adheres less to this thesis' definition of Mayer's (2009) Signaling Principle than Videos 1 and 2 but is slightly better than Video 3, as it has some predominant Themes. The rest of the Themes all appear once for a similar duration in the video, similar to Videos 2 and 3. More predominant Themes could have assisted in additional emphasis on particular Participants and Processes in the video and may have assisted children in learning these words and potentially adhering more to this thesis's definition of Mayer's (2009) Signaling Principle.

Another challenge and aspect to these videos in terms of the subtitles is the separation of a clause, often within linguistic groups or syntactic phrases. According to Lappalainen (2021),

"words belonging together" (e.g. phrases in syntax and/or groups like nominal groups in SFL) should be in the same line of subtitle as per SLS conventions. In this video alone, there are at least three instances where, due to the length of words in isiXhosa, noun phrases and similar phrases are separated. Even though there are instances in this video, this occurs less in Video 4 in comparison with Videos 1, 2 and 3. While they often occur in line 2 (as there are often two lines of subtitle per subtitle), they can also occur in the following subtitle and be broken midway which can cause confusion. For example, "ukususela kumasebe amdaka kakhulu… ukuya kwingca eluhlaza yaka." (ranging from very dark brown branches… to very green grass) is separated into two subtitles by the ellipsis.

6.4.3 Visual ideational metafunction STMs

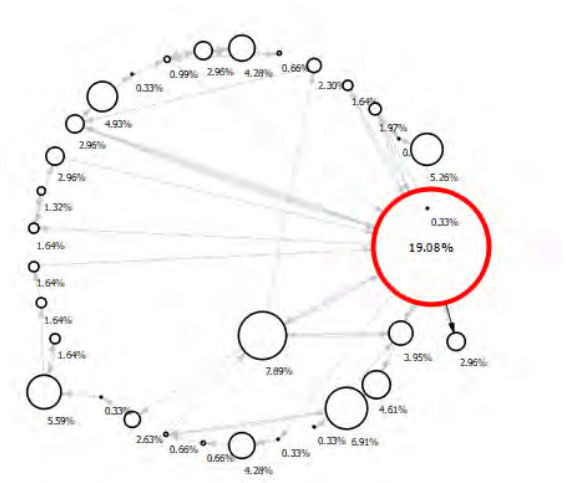


Figure 79(a) Jungle Tumble visual Participants and Process STM, with screenshot subtitled "No. It is a monkey's tail."

All the visual Processes in the video are Action Processes, however all the Participants are looking while they act and this does mean they qualify for Mental Processes and as Sensors,

unless their eyes are shut. The challenge of examining Mental Processes in the visuals is discussed in 6.6. Figure 79(a) illustrates that the predominant Actors are monkeys at 7.89%. There is a co-occurrence for 6.91% of the video of three Actor Participants in the same shots: monkeys, birds and frogs. The 5.59% state comprises insects as Actors. The top-most states in Figure 79(a) are also co-occurrences. For 5.26% of the video, leopards, insects and the elephant co-occur as Actors. For 4.93% of the video insects and a giraffe co-occur as Actors. For 4.61% of the video, monkeys and frogs co-occur as Actors and for 4.20% of the video, monkeys, the hippo and leopards co-occur as Actors. This correlates again with the high number of Material Processes in the spoken and written language, yet the spoken and written language do not have monkeys, insects, the elephant or the giraffe as Actors. The only mention of monkeys in spoken and written language is a reference to a tail in "Ngumsila we…nkawu" (it is a monkey's tail) in Figure 79(b). This video may not adhere to Mayer's (2009) Coherence Principle as these animals in the visual mode are extraneous information to the language and may potentially distract learners from the linguistic content and its visual Participants. This also causes this video to not adhere as well to Mayer's (2009) Principles of Spatial and Temporal Contiguity than the other videos that do not have as many extra visual Participants.

The ellipsis in the clause Figure 79(b) indicates tonal emphasis on the "we-" in "wenkawu" (of the monkey) and a long pause. This does occur in other subtitles in this video. It is an unconventional way to show this in isiXhosa, however, as this is usually written as a single word without any marks on tone and a child would have to be pre-trained (see Mayer's (2009) Pre-Training Principle in 3.2.1) before viewing the video to understand what the ellipsis represents. It could also cause spelling errors later if children start thinking they should spell words this way. While there is a correlation in the frequency between Actor Participants and Action Processes, as well as Action Processes and Participants and Material Processes between modes, the visual Action and linguistic Material Processes do not necessarily correspond with the same Actors. This causes less correspondence than if Participants mirrored one another between modes and may increase the cognitive load of learners according to Mayer's (2009) limited capacity assumption.

6.4.4 Visual interpersonal metafunctions STMs

Figure 80 Jungle Tumble Gaze STM

The duration of these system choices for Gaze in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Figure 80 illustrates Gaze, which is mostly Indirect for all Participants except the tortoise and birds. This aligns with the predominant Declarative Mood that offers information and is typical of the genre (Guijarro and Sanz 2008), however, it may cause learners to engage less with the video than if there was more Direct Gaze by Participants. This may relate to this thesis's expanded definition of Mayer's (2009) Personalization Principle. The 11.96% state comprises monkeys with Indirect Gaze and the topmost 7.31% state include a co-occurrence of birds with Direct Gaze and frogs with Indirect Gaze. The other 7.31% state comprises leopards with Indirect Gaze. The 5.65% state comprises insects with Indirect Gaze and the 5.32% state comprises frogs with Indirect Gaze. Other co-occurrences include (1) Indirect Gazes with insects and the giraffe (4.98%), (2) Indirect Gaze with the rat and Indirect/Direct Gaze with the tortoise (4.65% rightmost state), (3) Indirect Gaze with birds and hippo (4.32%) and (4) Indirect Gaze with the rat presented with the tortoise and its Direct Gaze (3.99%). While co-occurrences of Indirect Gaze simply emphasise the video offering information in the Declarative Mood, co-occurrences between Indirect and Direct Gaze, even if there is less Direct Gaze in the video overall, might cause the learner to focus less on Participants with Indirect Gaze in the shot than Participants with Direct Gaze. This requires further research to confirm.

There is also a 4.65% state of the crocodiles with Indirect Gaze. Birds and the tortoise are Participants that appear in the video and are first hidden away, but when they are revealed, their Gaze is directed at the viewer. This may provide them with more emphasis due to being revealed in the shot, as it may increase viewer engagement with them. In contrast, many Participants that are not suddenly revealed to the viewer gaze elsewhere, e.g. a frog has Indirect Gaze as it jumps around while the bird is being revealed from behind a branch and gazes directly at the viewer. Direct Gaze may assist in guiding the attention of viewers to particular Participants, which may assist in adhering to this thesis's expanded definition of Mayer's (2009) Signaling Principle. However, there are many Participants, such as the tortoise, that are not present in the linguistic mode. Thus, if Direct Gaze could have assisted in guiding a learner's attention, the lack of consistency of this throughout the video due to the discrepancy between linguistic and visual Participants may confuse learners. This may result in greater cognitive loads (see Mayer's (2009) limited capacity assumption described in 3.2.1).

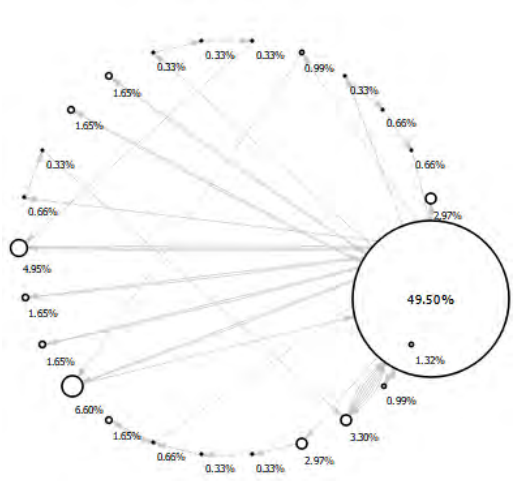


Figure 81 Jungle Tumble Social Distance and Zoom STM

Most of the video duration includes Medium shots (see Figure 81), mainly focused on monkeys (6.60%) and birds (4.95%), which further emphasises these Participants, even though their Gaze was not Direct in the video. Both these Participants are also animals that appear after being hidden. This further emphasises them and may cause higher engagement with them from viewers (Guijarro and Sanz 2008), which may assist in adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle. This may improve learning. In the video, 3.30% represents Long shots and Zoom Out, and occurs the most frequently, which portray the

jungle and multiple animals simultaneously, typical of the genre of children's narratives (Guijarro and Sanz 2008). A coupling of 2.97% includes a Long shot and Zoom In on birds, which further emphasises them as predominant Participants, however, the duration is less than the monkeys and birds. The topmost 2.97% state further shows leopards in Medium shot, which is also an animal that is revealed after hiding and thus this places greater emphasis on this Participant. All of these Participants are found in the linguistic mode, thus, shots and Zoom In or Zoom Out can be used to guide the learner's attention on these particular Participants. This assists in the adherence of Mayer's (2009) Principles of Spatial and Temporal Contiguity, and in the adherence of this thesis's definition of Mayer's (2009) Signaling Principle. By potentially adhering to these principles, this may foster literacy development.

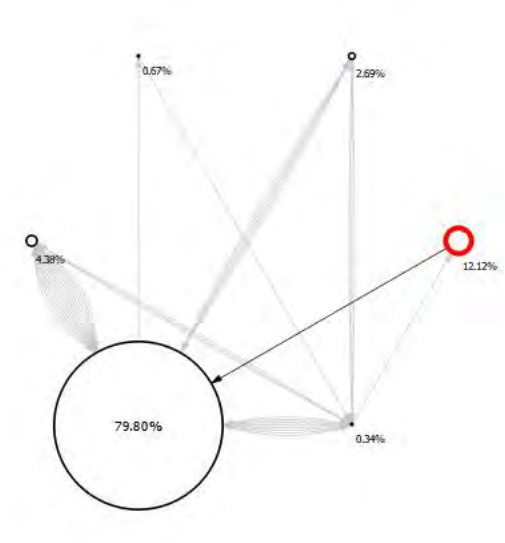


Figure 82 Jungle Tumble Horizontal and Vertical Perspective and Camera Movement

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 82, the frequency of occurrences is not pertinent to the analysis of these visual systems. Most of the video includes Front, Eye Level and Stationary shots (79.80%), with Eye Level and Stationary shots including Tilt, which tilts up to the trees with the monkeys, birds and snakes in them (2.48%). In Figure 82, the 0.34% state illustrates Front and Eye Level shots. Front, Eye Level and Stationary shots are typical of the genre and fosters learners' engagement as equal participants in the story (Guijarro and Sanz 2008). It may assist in the adherence of this thesis's expanded definition of Mayer's (2009) Personalization Principle. Shot Pans left to right of all the animals (2.69%) and shot pans on the frogs (2.69%) are also in the video. This further illustrates the emphasis on monkeys and birds in particular, but also additional emphasis

on snakes and frogs as Participants that appear after being hidden. These are also in the linguistic mode and this assists in the adherence of Mayer's (2009) Principles of Spatial and Temporal Contiguity, and this thesis' expanded definition of Mayer's (2009) Signaling Principle. Furthermore, this would relate to Information Value if there was an emphasis on Given and New information, however, this video does not really utilise this structure in the visuals as well as it could. This could have assisted in the closer adherence of the Signaling Principle.

6.4.5 Visual textual metafunction STMs

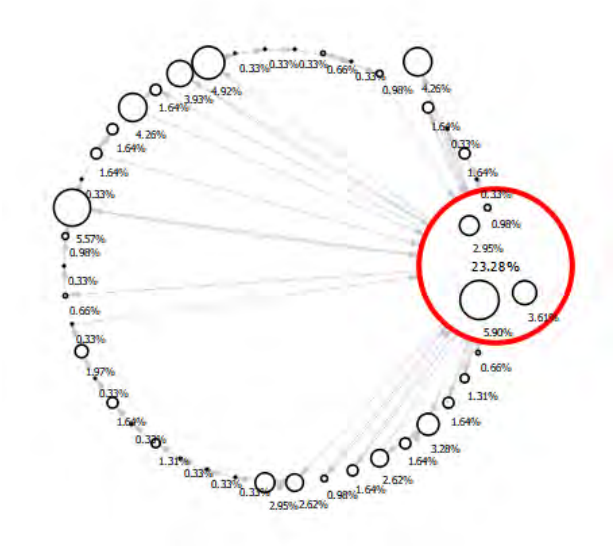


Figure 83 Jungle Tumble Information Value STM

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 83, the frequency of occurrences is not pertinent to the analysis of these visual systems. Most of the video durationally comprises monkeys as Given information (5.90% state) and insects as Centre (5.57%). This additional focus on these Participants is relevant as they are not the most predominant Participants in the language, where insects are not linguistic Participants. In Figure 83, the first 4.92% state comprises tortoise as Centre and the second comprises rat as Centre. This is further relevant as while tortoise is a linguistic Participant, the rat is only present in the visuals. Other co-occurrences include insects and the giraffe as Given information (3.93%), leopards as Given information, birds as Centre (3.61%) and birds and monkeys as Given information (3.28%). This aligns with the presentation of these Participants in the language, except for the presentation of the giraffe as Given information when it does not occur

in the linguistic mode. This causes a lack of adherence to Mayer's (2009) Coherence Principle and less of an adherence to this thesis's expanded definition of Mayer's (2009) Signaling Principle than the video could have had if there were less extra visual Participants. It may have also had more of an adherence to this thesis's expanded definition of Mayer's (2009) Signaling Principle if it used another system in the interpersonal metafunction, such Social Distance and Zoom, to guide learners' attention to particular visual Participants.

Thus, while many of the linguistic Participants are placed Centre in the visuals, many Participants are placed Centre or presented as Given information when they are not particularly relevant Participants to the linguistic storyline. They could potentially count as extraneous information which violates Mayer's (2009) Principle of Coherence. This may impede learning.

6.4.6 Intersemiotic Texture STMs

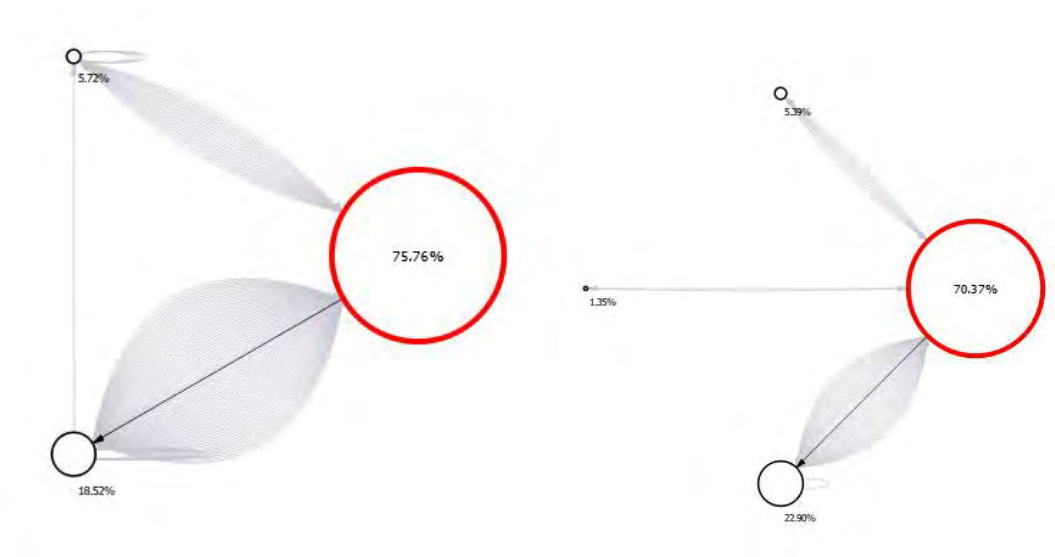


Figure 84(a) Jungle Tumble Intersemiotic Texture between spoken language and visuals STM (left), Intersemiotic Texture between written language and visuals (right)

The duration of these system choices for Comparative Relations in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Figure 84(a) illustrates that 18.52% of the video durationally represents Similarity of Correspondence between spoken language and visuals (22.90% between written language and visuals in Figure 84(b)). The 5.72% state in Figure 84(a) is Contrast between spoken language and visuals (5.39% between written language and visuals in Figure 84(b)). There is a further 1.35% co-occurrence in the Figure 84(b) where Similarity of Correspondence and Contrast occur

simultaneously. This is found in a sequence, where the first shot appears with the language “Isele liluhlaza okwegqabi” (The frog is as green as a leaf), seen in Figure 73(a). The leads the viewer to expect that a frog was hidden under a leaf in the next clause “Yiyo lento lizifihla…ngaphantsi!” (Therefore, it hides itself…below!). However, the frog is behind a stone and this is not explicitly said. This can cause slight confusion, however, occurrences such as these in the video are not frequent in comparison to the large amount of Similarity between modes in the video. This video, like the other videos, may have a conducive video design for learning, as the high frequency of Similarity may adhere to Mayer’s (2009) Principles of Spatial and Temporal Contiguity.

Figure 85 Screenshot from Jungle Tumble subtitled "Pull a tree branch...You will see!"

Intersemiotic Contingency within Intersemiotic Consequential Relations determines a possibility, with the aftereffect not a guarantee (Liu and O’Halloran 2009:382). An example of this is in Figure 85, where the words “Pull a tree branch…You will see!” do not specify that a bird is hidden behind this particular branch, although it was implied in the previous clause in Figure 72(c). From the previous information, it is implied that parrots hide in tree branches, and after this subtitle, a monkey pulls at the tree branch and a parrot is revealed and flies in the sky. In terms of potentially fostering learning, it could make learning more interactive for learners, which could maintain their attention and assist in the adherence to this thesis’ expanded definition of Mayer’s (2009) Signaling Principle.

6.4.7 Subtitle rate

The total duration of subtitles (subtitle presentation rate) was 1 minute 26 seconds. The 26 divided by 60 = 0.43 + 1 minute = 1.43. The total number of words (separated by spaces) is 90

words. If words had hyphens, they were included as a single word. The 90 words divided by 1.43 = 62.93 WPM. If the maximum reading threshold of 50 WCPM holds for Grade 3 learners, this is not only 10 words above the threshold and maximum reading score, but it is almost four times the best Grade 2 pre-test average reading score for a longer passage of static text (14.3 WCPM). Additionally, it is more than three times the Grade 2 average post-test reading score (19.95 WCPM). It is also almost three times the average benchmark for Grade 2 (20 WCPM). This subtitle rate is more than double the average pre-test reading score (22.85 WCPM) and the average post-test reading score (28 WCPM) for Grade 3. It is almost double the average benchmark for Grade 3 (35 WCPM).

The total syllable rate was calculated with a similar method. The 346 syllables divided by 1.43 = 241.95 syllables a minute. The maximum reading score for Grade 2 readers was 141 SCPM and Grade 3 readers was 179 SCPM. The mean scores, however, which reflect majority reading rate is far lower at 45.4 SCPM pre-test and 65.20 SCPM post-test for Grade 2, and 76.95 SCPM pre-test and 96.10 SCPM post-test for Grade 3. The syllable rate in the videos is still not at a pace the learners would comfortably manage. This violates this thesis' expanded definition of Mayer's (2009) Segmenting Principle. While silent reading is usually faster than reading out loud, the large amount above an Oral Reading Fluency score indicates the likelihood that the learners could not read the subtitles for this video. A limitation of this study was the inability to measure silent reading rate, as this is entirely dependent on the learner self-reporting this information, which is not necessarily accurate (Probert 2016). Furthermore, eye-tracking does not necessarily state whether decoding took place, even if it tracks the movement of the eye along the subtitle, so it also has limitations in understanding a silent reading rate.

6.4.8 Conclusion and summary of Video 4

While there is a high frequency of linguistic Material Processes and visual Action Processes in the video, not all of the visual Participants are present in the linguistic mode. This additional visual information may distract viewers and causes less of an adherence to Mayer's (2009) Principles of Coherence, Spatial and Temporal Contiguity**.** While predominant Processes in both modes are Material, Existential and Relational, there are discrepancies between duration and frequency of Mental and Relational Participants in the spoken mode. This may decrease a learner's cognitive capacity and may cause this video to be less conducive for learning

according to Mayer's (2009) limited capacity assumption. However, this video does not have as large a discrepancy between frequency and duration as Video 3.

Nevertheless, a high Similarity rate and a lack of Contrast may be conducive for learning in terms of Mayer's (2009) Principles of Spatial and Temporal Contiguity. However, whether the high Similarity and lack of Contrast can compensate for the distraction posed by the visual Participants is doubtful. This may be due to the potential lack of adherence to this thesis's expanded definition of Mayer's (2009) Signaling Principle. While linguistic Participants are emphasised visually through Medium shots, there is a lack of consistency when focusing on visual Participants using other systems of the visual interpersonal and textual metafunctions, such as Information Value and Gaze. The presence of reciprocal verbs may cause further reading challenges depending on if learners truly understand this structure and what it means at their grade and level of literacy. This may decrease a learner's cognitive capacity and may cause this video to be less conducive for learning according to Mayer's (2009) limited capacity assumption.

The addition of different Moods and the presence of viewer-directed Modality in this video in comparison to the sole use of the Declarative Mood in the other videos may increase viewer engagement for this video. In this manner, this video may adhere more to this thesis's expanded definition of Mayer's (2009) Personalization Principle. The pattern of Circumstances of Location: Time and Circumstances of Location: Place further emphasise the activities in the video and allow for viewer engagement in terms of Time deictic words (Levinson 2004, Peña 2020). The syndrome of Material Processes and Circumstances of Location are typical in children's narratives (Guijarro and Sanz 2008). While the high Similarity rate may be a good indicator of learning video design, there are more instances of a lack of adherence to Mayer's (2009) Principles than adherence and the high subtitle rate of 71.42 WPM is still too high for a Grade 2 and 3 learners to adequately read. Thus, in this video and all the videos in this thesis, this adheres to Mayer's (2009) Redundancy Principle and violates Mayer's (2009) Segmenting Principle which is elaborated on in detail in the next section (6.6).

6.5 Video 5 - I Can Go To School or Ndingaya esikolweni

Ndingaya esikolweni or *I Can Go To School* includes the story of the narrator, a little girl in the visuals. She describes her day at school from being dropped off by her father to talking to the teacher and the activities she and the other children do throughout the day from drawing, reading to sleeping and finally being picked up by her mother. Important spoken and/or visual Participants to note for this video include the girl, her father, her mother, her teacher and her children. It is relevant to this analysis that the word for teacher in isiXhosa can be either *utitshala* (male teacher) and *utitshalakazi* (female teacher). The video, despite only having a female teacher in the visual mode, uses both forms of the word, which could potentially result in confusion for viewers/learners. This might cause confusion for learners and adheres less to Mayer's (2009) Principles of Spatial and Temporal Contiguity. This is discussed further in 6.6.

The analysis of this video is divided into spoken and written ideational metafunction STMs for language (6.5.1), continuing to discuss the interpersonal and textual metafunctions in these modes (6.5.2). The ideational metafunction STMs in the visual mode are examined in 6.5.3 and 6.5.4 examines the interpersonal metafunction STMs in the visual mode. The textual metafunction STMs in the visual mode are examined in 6.5.5 and 6.5.6 examines Intersemiotic Texture. There is a discussion of subtitle rate in 6.5.7. Lastly, 6.5.8 concludes the analysis of this video.

6.5.1 Spoken and written language ideational metafunction STMs

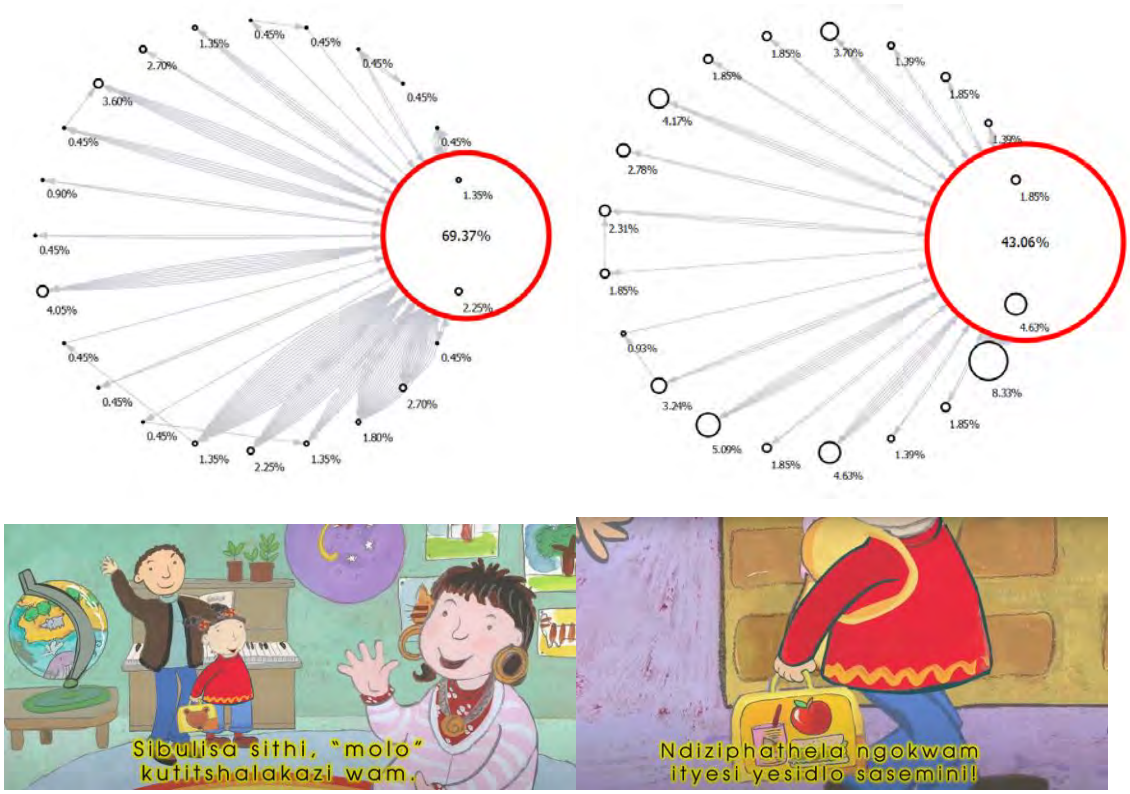


Figure 86 (a) I Can Go To School Spoken Participant STM and (b) Written Participant STM with screenshots subtitled (c) "We greet and say "hello" to my teacher" and (d) "I carry my own lunch box for the day"

The spoken Participants with the longest durations are Verbiage (4.05% state in Figure 86(a)) and Attribute (3.60% state). An example of Verbiage is "ngemisebenzi yemini" (activities of the day) in "Utitshala wam usixelela ngemisebenzi yemini" (My teacher tells us about the activities of the day) and an example of an Attribute is "nexesha" (time) in "Emva kwemini siba nexesha lokusebenza ngokuzolileyo" (In the afternoon we have time to work calmly). However, Verbiage is only present in 1.85% in the written mode, although it does co-occur with Sayer for 3.24% of the video. In the written mode, the most frequent Participants are (1) a co-occurrence of Actor and Goal (8.33%), and (2) a Sensor and Phenomenon state (5.09%) and (3). There is further 3.70% of the video durationally comprised of Behaver in the written mode, e.g. "si-" (we) in "sihlala" (we sit). This discrepancy of frequent Participants between linguistic and written modes may cause confusion for the learners. In addition, duration and frequency do not correlate in Figure 86(a) for spoken language, as the most frequent spoken Participants are Sayer (1.80%, Sensor (1.35), Goal (2.70%) and Actor (2.25). These discrepancies highlight that the video may not adhere as well as other videos in terms of

Mayer's (2009) limited capacity assumption, as well as Principles of Coherence, Redundancy, Spatial and Temporal Contiguity. The predominant Participants in the written mode of this video are typical of the genre (Guijarro and Sanz 2008).

Figure 87 Screenshot of I Can Go To School, subtitled "I like to play tag or ball with my friends"

The 2.70% states in Figure 86(a) from bottom to top include (1) Goal Participants and (2) Existent Participants. The 2.25% states in Figure 86(a) from bottom to top include (1) Actor Participants, which is the topmost 4.63% state in written mode, and has the same frequency as the 5.05% state for written Sensor and Phenomenon Participants. The next 2.25% state in Figure 86(a) includes (2) Beneficiary Participants. An example of a Goal Participant is "ityesi" (lunch box) in "Ndiziphathela ngokwam ityesi yesidlo sasemini" (I carry my own lunch box). An example of an Actor Participant is "utata wam" (my father) in Figure 88(a) "kusasa utata wam undisa eklasini yam" (In the morning my father takes me to my class). An example of an Existent Participant is in Figure 88(c) "Kukho oojingi, imityibilizi nendawo ethe gabalala yokudlala" (There are swings, slides and an open playing field). Existent Participants occur 2.78% in the written mode. The 2.27% state includes Beneficiary, e.g. in Figure 88(d) "umama wam" in "ndifika ndibonise umama wam omnye umsebenzi endiwenzileyo" (I always get to take my mother and show her some of the work I did). The 1.80% state includes Sayer (3.24% in written mode, along with Sayer and 2.31% along with Attribute), e.g. in Figure 88(b) "'Ube nemini emnandi' atsho utata (have a good day, father says). In the video, the 1.35% state comprises Sensor. These reflect differently to Figure 86(b), where one of the most prevalent states in the video is the co-occurrence of Sensor (4.63%), along with Sensor and Scope (4.17%), for example in Figure 87, "ndithanda ukudlala icekwa okanye ibhola nabahlobo bam" (I like to play tag or soccer with my friends). This discrepancy of frequent Participants between linguistic and written modes may cause confusion for the learners. This highlights that the

video may not adhere to adherence of Mayer's (2009) limited capacity assumption, as well as Principles of Coherence, Redundancy, Spatial and Temporal Contiguity, when comparing this video with other videos.

Figure 88 Screenshots from I Can Go To School subtitled (a) "In the morning my father takes me to my class," (b) ""Have a nice day" said father," (c) There are swings, slides and an open playing field "" and (d) I take and show my mom the work I did

Material Processes occur for the most time in spoken and the written mode, but Verbiage is the second-most frequent Participant in the written language, but not in the spoken language, which has longer occurrences of Sensors and Actors. The predominance of Goal Participants in the spoken language is typical for the subject-omitting nature of isiXhosa, but Actors and their co-occurrence with Goals being more frequent in the written language could be a result of the videos being translated from English. While this discrepancy may cause confusion, they still relate to the same Process so it would be less confusing in comparison with discrepancies between Participants of different Processes. The majority Actor and Sensor and Sayer Participant comprise the girl, or the girl co-occurring with her father, or mother. The children and the teacher occur as Goal or Beneficiary Participants more frequently in the linguistic mode than the visual mode. This is expected for a story that focuses on the girl and her parents as central characters and moving, thinking and speaking are activities she and other Participants do throughout the narrative. This is typical of the children's narrative genre (Guijarro and Sanz 2008).

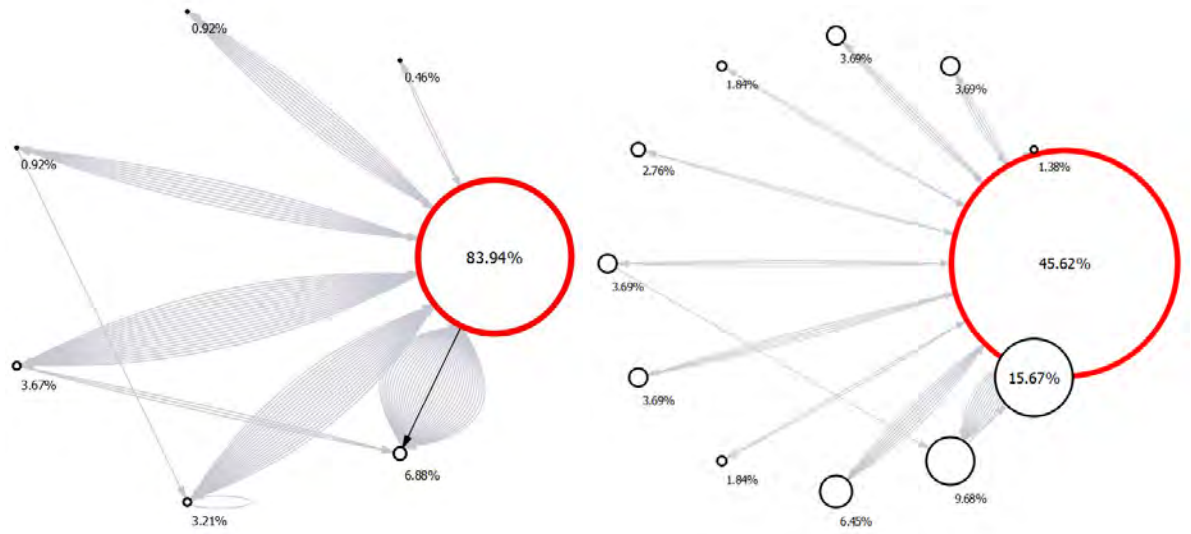


Figure 89(a) I Can Go To School spoken Processes (left) and written Processes (right)

The duration of these predominant Processes in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). In Figure 87(a), the Processes for both modes occur in a similar pattern to each other in terms of duration, with 6.88% of the video categorised as Material Processes (15.67% in the written mode in Figure 87(b)). Another 3.67% as Mental Processes (rightmost 3.69% state in Figure 87(b)) and 3.21% as Verbal Processes (9.68% in the written mode in Figure 87(b)). There is also 2.76% of the video with written Existential Processes. The 0.92% states from bottom clockwise include (1) Relational and (2) Behavioural Processes (the other topmost 3.69% state in written mode). The 0.46% state includes Existential Processes. Co-occurrences in the written mode include the 6.45% state of Material and Mental Processes, the leftmost 3.69% states of (1) Material and Relational and (2) Mental and Behavioural Processes. This high frequency of co-occurrences may cause challenges for learning as it may influence learners' cognitive load according to Mayer's (2009) limited capacity assumption. This corresponds to the high number of Actors, Goals, Verbiage and Sensors, and illustrates Material, Verbal and Mental Processes as the most predominant categories of Processes in the video. Verbal Processes are the second-most frequent Process in the written mode, but not as frequent in the spoken mode. This is in contradiction to Verbiage being the second-most frequent spoken Participant, but less frequent in the written mode. This discrepancy of frequent Participants and Processes between linguistic and written modes may cause confusion for the learners. This highlights that the video may not adhere to Mayer's (2009) Principles of Spatial and Temporal Contiguity as illustrated by its high Similarity rate.

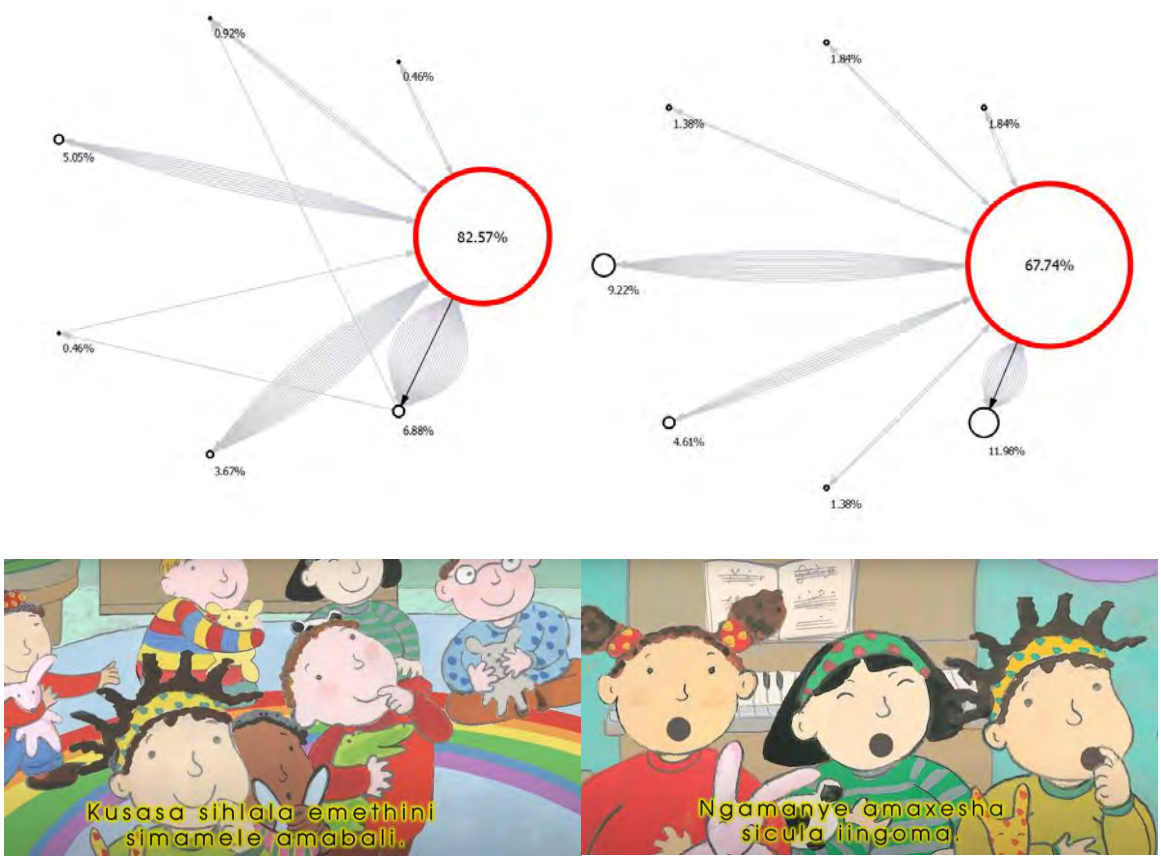


Figure 90 (a) I Can Go To School Spoken Circumstances STM (left) and (b) written Circumstances (right), with screenshots subtitlted (c) "In the morning we play on the mat, we listen to stories," and (d) “sometimes we sing songs”

As with other videos in this chapter, the predominant is Circumstance of Location: Place and Location: Time, for example in Figure 90(c) “kusasa sihlala emthini simamele amabali” (in the morning we sit on the mat, we listen to stories) at 6.88% in Figure 90(a) and 11.98% in Figure 90(b). This is followed by Matter (5.05% in Figure 90(a) and 4.61% in Figure 90(b)), e.g. “ngemini yam, nangabahlobo bam nezinto ezonwabisayo endizenzileyo” (about my day, my friends and the things that make me happy/that are enjoyable that I did.) in “Ndiyakuthanda ukumbalisela ngemini yam, nangabahlobo bam nezinto ezonwabisayo endizenzileyo” (I like telling her about my day, my friends and the things that are enjoyable that I did). Another predominant Circumstance is Circumstance of Extent (3.67% in Figure 90(a) and 9.22% Figure 90(b)), for example “ngamanye amaxesha sicula iingoma” (sometimes we sing songs) in Figure 90(d). There are also Circumstances of Accompaniment (0.92% in Figure 90(a), 1.38% Figure 90(b)), for example, “Ndiyakwazi ukuzoba emgangathweni ngetshokhwe” (I know how to draw on the ground with chalk) in Figure 91(a). Circumstance of Manner is also present in this video (0.46% and 1.83% in written), for example “Emva kwemini siba nexesha lokusebenza

ngokuzolileyo" (In the afternoon we have time to work calmly) in Figure 91(b). There is also an Embedded Clause Circumstance of Location for 0.46% of the spoken mode, e.g. "Phambi kokuba ahambe" (before he leaves). The complexity of embedded constructions, which often include many adjectivised and nominalised Processes might create challenges in decoding, due to the strain it may place on the cognitive load of the learner according to Mayer's (2009) limited capacity assumption.

Nevertheless, discrepancies between duration and frequency of system choices may create complexity for the learner and may cause challenges for comprehension. This may increase leaners' cognitive load, which is not conducive for learning according to Mayer's (2009) limited capacity assumption, which may influence Mayer's (2009) Principles of Coherence, Redundancy, Spatial and Temporal Contiguity. As discussed in 2.3.1.3, frequency of exposure may not be as integral for literacy development as Similarity and Contrast co-patterning between modes (Tseng and Djonov 2023). Thus, this thesis focuses on Comparative Relations in Intersemiotic Texture (see 6.5.6 for this video) and a Similarity of meanings between modes as a potentially more accurate indicator of literacy development than frequency or duration.

The presence of many Circumstances of Location and Extent further allows emphasis on the setting, time and extent of actions, which assist in maintaining the child's attention to the story and video (Levinson 2004; Peña 2020). The syndrome of Material Processes and their Participants and Circumstances of Location are typical for the genre of children's narratives (Guijarro and Sanz 2008).

Figure 91 Screenshots from I Can Go To School subtitled (a) "I know how to draw on the floor with chalk" and (b) "In the afternoon we have time to work calmly"

6.5.2 Spoken and written language interpersonal and textual metafunction STMs

The only Mood is Declarative (61.93% of the video) and since there is only this state, no STM is illustrated here. This again illustrates the genre of a story that is offering information to the viewer, but additional Moods, such as Interrogatives and Imperatives, may have assisted in viewer engagement (Guijarro and Sanz 2008). Modality may, when it relates to Moods like Interrogative and Imperative, cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. If the Mood is Declarative, it may increase a load on cognitive capacity that violates the limited capacity assumption. Modalised clauses not only create complex layers of meaning, they also tend to manifest in more complex forms, which may result in lengthier and more phonologically complex words. For example, the sentence "Ndiya esikolweni" (I go to school) in isiXhosa has two words amounting to a total of seven syllables. However, if this were changed to the Imperative Mood it would change to *Yiya esikolweni!* (Go to school!). While this sentence also has two words and seven syllables, the consonant cluster nd- representing the two phonemes [ɲd] is now no longer present, thus the Imperative clause is less complex to decode phonologically than the Declarative. "Ndingaya esikolweni" (I can go to school) is the title of the video. Here the modal infix *-nga-* (can) makes the clause two words and eight syllables long. There are also three complex digraphs (*nd-, ng-* and *lw-*). Thus, the Modality makes this clause more phonologically and semantically complex. The video has realisations of Modality that could increase learners' cognitive load, as they create complexity in the language. Interactions of Imperatives and Interrogatives may increase the complexity of the video, but they draw readers into the story, eliciting readers' attention and engagement, thus potentially resulting in better attention and retention of information by a younger audience than Declarative expressions on their own (Guijarro and Sanz 2008).

This video, however, has fewer realisations of Modality, which are less complex, as they are only present in the Circumstances of Location: Time *-soloko* (always) and *ngamanye amaxesha* (sometimes). The Circumstance of Location: Time realised in *-soloko* is found in the clauses "Ndisoloko ndonwabile kukubabona" (I am always happy to see them), "Ndisoloko ndivuya ukumbona" (I am always happy to see her) and "Usoloko ezingca ngam" (She is always proud of me). These three instances of *-soloko* are all modal realisations of usuality with high intensity and positive polarity. The Circumstance of Location: Time realised in *ngamanye amaxesha* is

found in the clauses “Ngamanye amaxesha sicula iingoma, okanye sabelane ngamabali obomi bethu” (Sometimes we sing songs, or we share stories about our lives) in Figure 90(d). It is also found in the clauses “Okanye ngamanye amaxesha ndithanda ukunyova ibhayisekile” (Or sometimes I like to ride the bicycle), “Ngamanye amaxesha ndihlala esitulweni esitofo-tofo ndifune incwadi (Sometimes I sit on the soft chair and read a book). The last instance is in the sentence “Ngamanye amaxesha ndizoba imifanekiso yosapho lwam” (Sometimes I draw pictures of my family), see (12) in 1.2.1.2. These four instances of *ngamanye amaxesha* realise usuality with low intensity and positive polarity. Thus, there are slightly more instances of low intensity than high intensity in relation to Circumstance of Location: Time. More research is required to investigate if isiXhosa learners could distinguish Circumstances of Location: Time and comprehension of time (such as in Tseng and Djonov 2023).

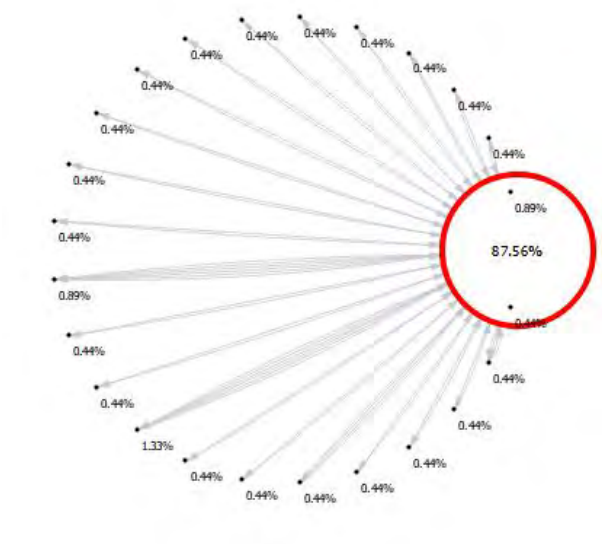


Figure 92 I Can Go To School Theme STM

The Themes with the most duration in spoken language include “ngamanye amaxesha” (sometimes), which is the leftmost 1.89% state in Figure 92 and “ndiyayilangazelala” (I am looking forward to it), which is the other 1.89% state. Another salient Theme is “ndithanda” (I like) or another form of the complex verb constructions with an object prefix (1.33% of the video). The most frequent Themes are “ndithanda” (the 1.33% state) and the modal Circumstance of Location: Time “ngamanye amaxesha” (sometimes). The rest of the Themes appear at simultaneous durations with one another and thus, these Themes are not as salient as the above. These Themes are either Circumstances of Location: Time or a Mental Process

which contributes to the high frequency of this Participant and Circumstance type in these videos. This is a way of emphasising Mental Processes, which are not always present in a video's visual mode, and this could assist in drawing learners' attention to them. This also could assist in the adherence of this thesis' expanded definition of Mayer's (2009) Signaling Principle. However, the Verbal Processes being the second-most frequent Process in the written mode, and not Mental Processes, as well as Verbiage being the second-most frequent Participant in the spoken mode and not Mental Processes may cause confusion for learners.

Another challenge and aspect to these videos in terms of the subtitles is the separation of a clause, often within linguistic groups or syntactic phrases. According to Lappalainen (2021), "words belonging together" (e.g. phrases in syntax and/or groups like nominal groups in SFL) should be in the same line of subtitle as per SLS conventions. In this video alone, there are at least four instances where, due to the length of words in isiXhosa, noun phrases and similar phrases are separated. In terms of the frequency of these instances, it is less than Videos 1, 2 and 3, but more than Video 4. While they often occur in the line underneath them (as there are often two lines of subtitle per subtitle), they can also occur in the following subtitle and be broken midway, which can cause confusion. For example, in "Ngamanye amaxesha ndihlala esitulweni esitofo-tofo ndifunde incwandi" (Sometimes I sit on the soft chair and read a book), the separation in subtitle occurs between esitofo-tofo (soft) and "ndifunde" (I read). Incwandi is usually spelt "incwadi" (book) in isiXhosa, and is thus spelt incorrectly, which may cause a transfer of spelling errors for the learner.

6.5.3 Visual ideational metafunction STMs

Figure 93 I Can Go To School visual Participants their Processes STM

All the visual Processes in this video are Action Processes with 12.22% of the video including the girl as Actor (see Figure 93). The children as Actors are depicted in 14.48% of the video and 8.60% comprises the co-occurrence of the father as Actor and the girl as Actor and Sensor. The co-occurrence of the girl and the other children as Actors and Sensors are represented by two states, the 6.79% state and the 4.13% state in Figure 93, thus this total co-occurrence is 9.63% of the video. The 5.43% state comprises the girl as Behaver and the 3.67% state comprises the mother and the girl as Actors. This correlates with the high frequency of Actors in the video in both modes, which may be ideal for learning these Processes and Participants, as this could assist in adherence to Mayer's (2009) Principles of Spatial and Temporal Contiguity Principles, possibly improving learning.

6.5.4 Visual interpersonal metafunction STMs

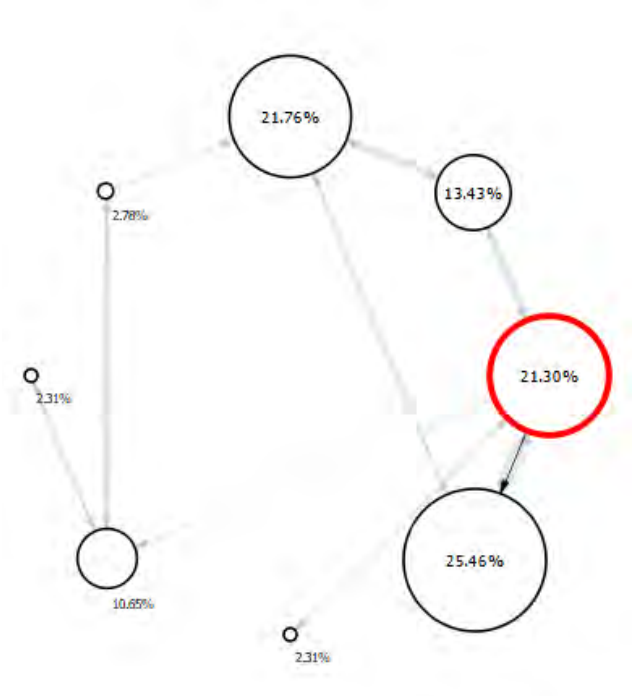


Figure 94 I Can Go To School Gaze STM

The duration of these system choices for Gaze in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Most of the Gaze in the video is Indirect, with the only exception being the girl (in the 2.31% state at the bottom of Figure 94). Indirect Gazes include the 25.46% state for the girl, the 21.76% state illustrates Indirect Gazes for the girl and the children together. The 13.43% state shows Indirect Gaze for the children, the 10.65% state shows Indirect Gaze for the girl and the father and the 2.78% state shows the girl, the children and the father. This shows less engagement with the viewer than other videos and highlights that the video mainly provides information, rather than requesting any information. Direct Gaze could have assisted in adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle in order to draw learners' attention to specific Participants in the video. The predominance of Indirect Gaze does align, however, to the Declarative Mood in the linguistic mode, which improves Similarity of meaning structure between modes. This would potentially assist learning by reducing learners cognitive load. This is in contrast to Video 4, which includes a clause in the Interrogative Mood (see 6.4.2) and more instances of Direct Gaze. Direct Gaze is also more visible in Videos 1 and 2.

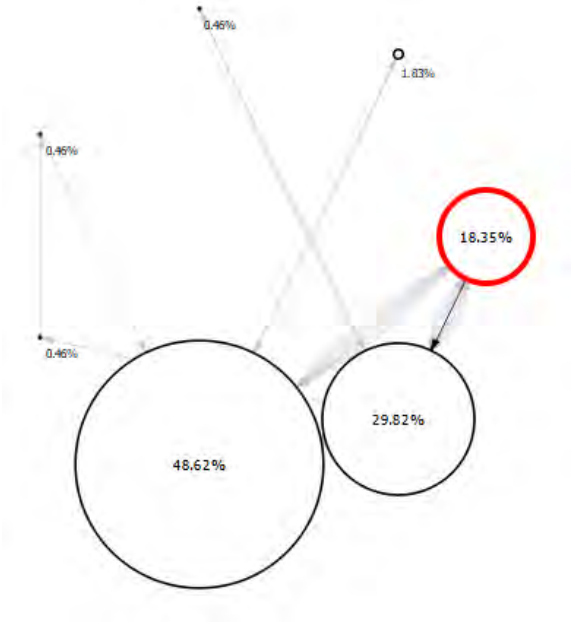


Figure 95 I Can Go To School Social Distance and Zoom

The duration of these system choices for Social Distance and Zoom in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Most of the video includes Medium shots at 48.62%, see Figure 95, and these shots include the girl or the children engaging in various activities. The 29.82% state in Figure 95 comprises Long shots. Long shots are typical of the genre and assist in maintaining the children's engagement in the story (Guijarro and Sanz 2008). The 1.83% state represents a Medium shot and Zoom In on the girl, while the 0.46% states include Zoom Out on the girl for Long shot and for Medium shot. The presence of Zoom effects further place emphasis on the girl, and the Medium shots on her and the other children emphasise their Action Processes. This could enhance learning as it may assist with adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle and Personalization Principle.

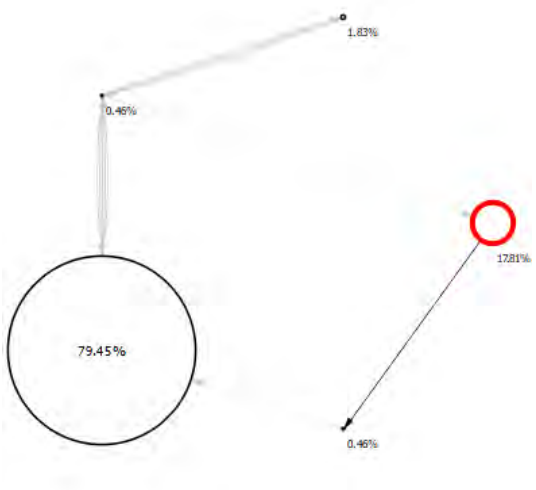


Figure 96 I Can Go To School Horizontal and Vertical Perspective and Camera Movement

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 96, the frequency of occurrences is not pertinent to the analysis of these visual systems. Most of the video includes Stationary, Front and Eye Level shots (79.45%). This is typical of the genre and allows for viewers to be placed in an equal relationship with the Participants in the video (Guijarro and Sanz 2008). This could assist in adherence to this thesis' expanded definition of Mayer's (2009) Personalization Principle. Any other shots include a Pan of the camera onto the children, with Front and Eye level angles (1.83%). The two 0.46% states include (1) Stationary and Front shots, and (2) Front and Eye Level shots. Aside from the Pan of the camera, which is short in duration, Camera movement is not used much in this video to create emphasis and there are no Salience features in this video to draw emphasis on any specific Participants or Processes. Thus, within these systems, this video has less adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle than other videos, such as Video 1, which may cause this video to not foster learning as well as Video 1 in this respect.

6.5.5 Visual textual metafunction STMs

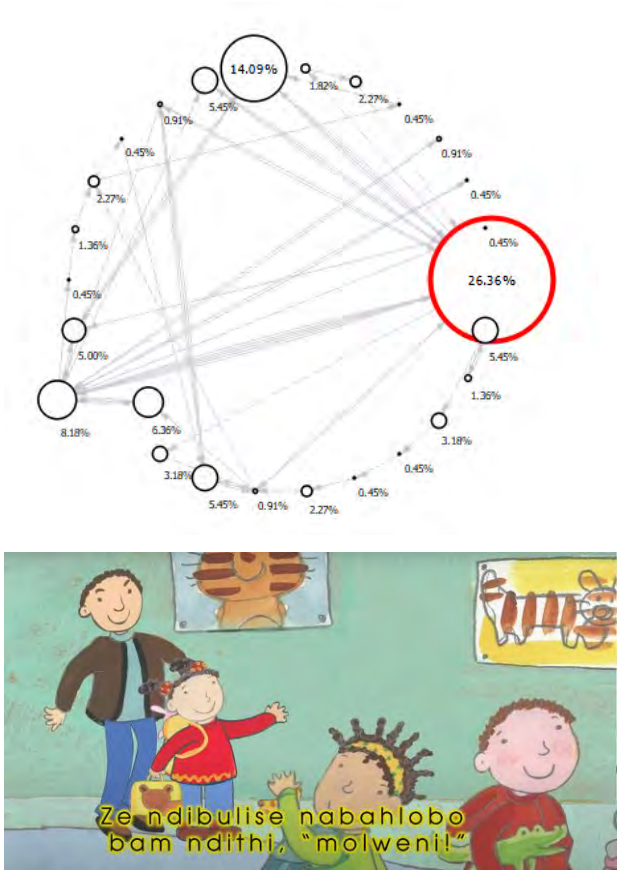


Figure 97(a) I Can Go To School Information Value STM with (b) screenshot subtitled "We greet my friends and say "hello everyone""

Due to the transition of shots represented in a sequential way by the STM transition lines in Figure 97, the frequency of occurrences is not pertinent to the analysis of these visual systems. Most of the video includes the girl as Centre (14.09%) or the children as Centre (8.18%). The 6.36% state comprises the children as Centre, but the girl as Given information. The three 5.45% states from the bottom and clockwise include (1) the girl as Given information, (2) the children as Centre, and (3) the co-occurrence of the children as Given Information and the girl as New information. The 5.00% state includes the children and the girl occurring together in the Centre and the 3.18% states from right to left include the co-occurrence of (1) the girl as Centre and the father as Given information, and (2) the girl and the father as Centre. Thus, when the girl or the children are not in Centre of the frame, they are mostly Given information which often correlates to the structure of the linguistic sentences presented with the visuals, where the girl is an Actor Participant. An example of this is in the sentence “Ze ndibulise nabahlobo bam ndithi, “molweni!”” (And then I also greet my friends and say, “hello all!”).

The predominance of Participants placed Centre further assist in the adherence of the video to this thesis' expanded definition of Mayer's (2009) Signaling Principle, and the similarity between structures assists in adherence to Mayer's (2009) Principles of Spatial and Temporal Contiguity.

There are artificial frame lines for 7.34% of the video durationally, namely a circle, around the girl and often the father along with her in Figure 88(a). Due to this being the only instance of Framing, a Framing STM is not illustrated here. This further emphasises the girl and her relevance to the story, which assists in the adherence with this thesis' expanded definition of Mayer's (2009) Signaling Principle.

6.5.6 Intersemiotic Texture STMs

Figure 98(a) I Can Go To School Intersemiotic Texture STM for spoken language and visuals and (b) written language and visuals

The duration of these system choices for Comparative Relations in the video (states) correlates to their predominance in terms of frequency of occurrence (transition lines). Figure 98 depicts relationship between the spoken and visual modes on the left and that between the written and visual modes on the right. There is 44.95% Similarity of Correspondence between the spoken language and visuals in the video with only 3.21% Contrast. There is a 56.42% Similarity of Correspondence between the written language and visuals in the video with only 3.67% Contrast. The ratios of Similarity and Contrast in proportion the linguistic content of the video cause Video 5 to have the highest Similarity rate and lowest amount of Contrast in comparison

with the other videos. This video adheres well to Mayer's (2005) Principles of Spatial and Temporal Contiguity and is a good example of a design conducive to learning. However, this does not mean Video 5 is the best video for learning of the five videos overall, as one has to take the other systems into account, for example the lack of Salience in the visual mode, the subtitle rate, the presence of a spelling error and the instances of clause separation in the subtitles.

Furthermore, there are instances of Intersemiotic Additive Relations in the visual mode, similar to the additional insects in the visual mode present in Videos 3, Video 4, 5. These do not add to the rate of Similarity in videos, like the animals and birds and squirrels in Video 2, but rather are additional and independent of the language. These are in the form of toys, often in the form of a stuffed rabbit in this video. This is not a Contrast, but due to the stuffed rabbits being animated, this could also pose a distraction to learners who might be focusing on processing this additional visual information that does not correspond to the linguistic information presented. If these additional visual elements were static and not animated, or animated during moments without subtitles, it could allow less focus on them and their addition may not pose as much of a distraction. The animated lily pads in Video 1, the insects animated in Videos 3 and 4, the additional visual animal Participants in Video 4 and the animated stuffed rabbit in Video 5 might shift the focus of the viewer onto their movement and take their eyes off the subtitles. In this way, the videos may adhere less well to Mayer's (2009) Coherence Principle. Further research would be needed, and this could be examined with eye-tracking equipment.

6.5.7 Subtitle rate

The total duration of subtitles (subtitle presentation rate) was 1 minute 59 seconds. The 59 divided by 60 = 0.98 + 1 = 1.98. The total number of words (separated by spaces) is 182 words. If words had hyphens, they were included as a single word. The 182 words divided by 1.98 = 91.91 WPM. If the maximum reading threshold of 50 WCPM holds for Grade 3 learners, this is not only 41 words above the threshold and maximum reading score. However, it is more than six times the best Grade 2 average pre-test reading score for a longer passage of static text (14.3 WCPM) and more than four times the Grade 2 average post-test reading score (19.95 WCPM). This is almost four times the average pre-test reading score (22.85 WCPM) and more than three times the average post-test reading score (28 WCPM) for Grade 3. This is also four times the

average benchmark for Grade 2 (20 WCPM) and more than double the average benchmark for Grade 3 (35 WCPM).

The total syllables were calculated with a similar method. Therefore, 633 syllables divided by 1.98 = 319.70 syllables a minute. The maximum reading score for Grade 2 readers was 141 SCPM and Grade 3 readers was 179 SCPM. The mean scores, however, which reflect the reading rate of the majority of learners, is far lower at 45.4 SCPM pre-test and 65.20 SCPM post-test for Grade 2, and 76.95 SCPM pre-test and 96.10 SCPM post-test for Grade 3. If the silent reading rate is faster and perhaps double the ORF rate, the average Grade 3 learner could not have followed the subtitle rate. This violates this thesis' expanded definition of Mayer's (2009) Segmenting Principle. A limitation of this study was the inability to measure silent reading rate, as this is entirely independent on the learner self-reporting this information, which is not necessarily accurate (Probert 2016). Furthermore, eye-tracking does not necessarily state whether decoding took place, even if it tracks the movement of the eye along the subtitle, thus, there are also limitations in understanding a silent reading rate.

6.5.8 Conclusion of summary Video 5

While the majority of Processes in both linguistic modes are Material, and the visual Processes are predominantly Action Processes, there is a discrepancy of second-most frequent Processes and Participants between spoken and written linguistic modes. This is predominantly between Verbal Processes and their Participants, and Mental Processes and their Participants. Furthermore, excluding the Material Process in the linguistic mode, the other linguistic Processes and Participants do not have corresponding visual Processes and Participants. Thus, there is a discrepancy between the meanings conveyed between modes. The majority Actor, Sensor and Sayer Participants in the video is the girl, or the girl and her father, or mother. The children and teacher occur as Goal or Beneficiary Participants more frequently. This is expected for the genre of a children's narrative and this is due to the presence of the syndrome of Material Processes and Participants, Action Participants and Participants, and Circumstances of Location in the video. Further emphasis is attracted with Medium shots, Information Value Centre and Themes, which may assist in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle, but there was a lack of Salience in the video to assist in the adherence of this Principle. Social Distance and Camera Movement assist in adherence to this thesis' expanded definition of Mayer's (2009) Personalization Principle, but the

predominance of Indirect Gaze does not. Instead, the predominance of Indirect Gaze and the Declarative Mood align with the offering of information and thus, cause a similarity of meaning structures between modes.

Nevertheless, like all the other videos, the subtitle rate was simply too fast for Grade 2 and 3 learner reading benchmarks and intervention test results, indicating that they are potentially unreadable for this age group. This adheres to Mayer's (2009) Redundancy Principle and violates Mayer's (2009) Segmenting Principle which is elaborated on in detail in the next section (6.6).

6.6 Triangulated analysis and chapter conclusion

The main research aim of this thesis is to examine SLS isiXhosa videos as a potential resource for L1 isiXhosa Foundation Phase literacy development in the context of an ongoing literacy crisis in South Africa. As established in Chapters 1 and 2, not much is known about the benefits and challenges of using this resource in this context, thus this thesis has taken a mixed-method approach (outlined in Chapters 3 and 4) to provide a holistic understanding of solutions. This thesis triangulates the intervention data results (ORF test results and the Picture-Word Match test to gauge learners' reading rate and broad semiotic awareness skills) and DMDA results of the five videos with the context by using theories of multimodal literacy and Mayer's (2009) principles as a lens. The results from Chapters 5 and 6 are thus discussed in relation to one another (triangulated) and consider them in relation to the research context, which stem from discussions in Chapters 1, 2, and 3.

The high results of the Picture-Word Match tests before the study indicates that learners had semiotic awareness skills to connect isolated written words with the pictures, as per Liao *et al.* (2021) description of a multimodal-integrated language framework for multimodal reading. This indicates they had developed Piaget's semiotic (symbolic) function to an extent (see 1.2.3). Furthermore, there was a high Similarity between visuals and language in the videos (see 6.6.6). However, while the Word Reading and ORF scores in the intervention significantly improved within the three months, the videos did not make a significant difference to learners' literacy development. Considering the word and syllable speeds calculated for each video, which are discussed in 6.6.4, all the videos had WPM and SPM speeds that were faster than the Grade 2 benchmark (20WCPM) and (35 WCPM). According to Kruger *et al.* (2022), fast subtitle speeds can influence word reading horizontally and vertically between the subtitles and visuals. Fast subtitles create difficulties in pairing meanings from semiotic resources between modes. They can cause a lack of comprehension in multimodal texts, and it appears, in this context, the subtitles were redundant information. Due to their fast speed, and their complexity in disjoint (long) form, where the EGRA-based tests were in conjoint (short) form, one can deduce that learners, despite the duration and frequency of similar meanings between modes, could not connect these meanings sufficiently due to the subtitle speed (Kruger *et al.* 2022). Further complexities from the language in the video include non-standardised conventions of written isiXhosa, lack of noun omission, complex verb forms such as passive, stative,

applicative, causative and reciprocal, frequency and duration (intramodal and intermodal linguistic discrepancies), system choice co-occurrences in the subtitles, ideophone constructions, clause separation in subtitles and modality.

The DMDA and Mayer's principles help to pinpoint where there are benefits and challenges for literacy development in relation to these findings. General benefits and challenges for literacy development are first discussed from the videos (6.6.1). Potential benefits are in relation to vocabulary. Potential challenges due to linguistic complexities in system choices and grammatical constructions are then described. They are described in relation to disjoint form, non-standardised conventions of written isiXhosa, lack of noun omission, complex verb forms such as passive, stative, applicative, causative and reciprocal, frequency and duration (intramodal and intermodal linguistic discrepancies), system choice co-occurrences in the subtitles, ideophone constructions, clause separation in subtitles and modality. This is to describe potential benefits and challenges, which could be remediated by Mayer's (2009) Pre-Training Principle, also discussed in this section. In the next section, 6.6.2, the patterns of syndromes, described by Guijarro and Sanz (2008), are discussed in the videos, particularly in relation to Circumstances and Modality, in order to highlight how these videos conform to typical genre norms. Whether these are suitable in our context, is described when discussing how the videos may adhere or not adhere to this thesis' expanded definition of Mayer's (2009) Personalization Principle in 6.6.7. Further research on this principle in particular is required to examine how it would fit into a hierarchy of principles for optimal SLS video design for literacy development.

The hierarchy of design principles for this context presented in 7.9.1 includes Mayer's (2009) Principles of Coherence (6.6.3), Redundancy (6.6.4), Segmenting (6.6.4), Signaling (6.6.5), Spatial and Temporal Contiguity (6.6.6). Mayer's (2009) Principle of Personalization is also expanded and discussed in relation to future research (6.6.7).

6.6.1 General benefits and challenges for literacy development with the videos

Each video had different potentials for various types of vocabulary learning. In Video 1, *The Icky Sticky Frog*, the main words relate to insects, the frog and various types of Action processes in the linguistic mode, e.g., ibhabela (it flies). In Video 2 and 4, there are many

animals, e.g., in Video 2, *They all Loved in Sue*, the nominal group "Iikati, izinja, amafudo, iintaka, izilwanyana zalo" (Cats, dogs, tortoises, birds, all kinds of animals) with corresponding visuals of the animals. This allows for learning of this category of vocabulary and word recognition. In Video 4, spoken and/or visual participants in this video include a variety of animals. These all relate to broader potentials for literacy development and reinforcing literacy around animal vocabulary. With *The Moon Followed Me Home*, Video 3, Similarity of Collocation appears but it could potentially cause learning difficulties and focus on a different type of vocabulary than what the video offers. Instead of a focus on animal vocabulary (e.g., the lions presented in the visual mode), it has a focus on family vocabulary (e.g., the use of "utata wam," my father, when the image of a lion is present). These are similar visually-contextualised meanings but contrastive enough to cause potential learning issues. Video 5, *I Can Go To School*, is the only video that does not deal with animal Participants. Consequently, if the focus in teaching and learning with these videos were to be reinforcing animal vocabulary, it may detract attention from that learning goal in comparison to the other 4 videos. Meanwhile, for reinforcing the types of vocabulary that are on the topic of school, Video 5 would be more suitable than any of the other videos. The potentials for various types of literacy improvement and reinforcement for each video may thus reinforce learning from different semantic domains in terms of literacy development in the classroom.

However, while these videos have the potential to assist in lexical vocabulary skills, due to the fast subtitle speed (and subtitles' linguistic complexity), automaticity in Word Reading tests was not significantly different due to SLS video exposure. There was no significant difference between post-test Word List 1 and 2 scores, indicating words present in the video were not read faster than words that were not present in the video. This could be due to these words, for example verb roots, being situated in a verb complex that a learner may not necessarily read lexically, but due to the different potential prefixes and suffixes, may have to read phonologically. This is made more challenging in the videos, which present isiXhosa in disjoint form, when the Foundation Phase (in EGRAs and CAPS documents) suggest that they should focus on reading words in conjoint form. The CAPS documents indicate that Foundation Phase learners should be able to read words in disjoint form, however, as discussed earlier in this thesis (see 1.2.2), the CAPS documents are markedly flawed in estimating what components of L1 isiXhosa literacy is developed in the Foundation Phase. Disjoint form, as illustrated in 1.2.1.1, results in verbs possessing a lot more affixes, including object markers, modal markers and verb extensions. This results in longer words in the disjoint form, and in the videos, than

in conjoint form longer passage texts. Thus, word lengths in the videos are longer than the words in the ORF Long Passage test, illustrating that learners would have had less words and syllable per minute reading scores for disjoint words than they scored in the conjoint ORF tests used in this thesis. This illustrates that, word lengths were often longer in the videos, particularly for verbs, and more linguistically complex. Word lengths from the disjoint form also cause instances of clause separations between subtitles, which would increase learners' cognitive load as they have to hold a lot of stored information from the previous subtitle(s) in the working memory (see 3.2) in order to make sense of the entire clause.

The Word Reading and ORF test scores and the lack of significance from the intervention confirms that isiXhosa is read phonologically at the grain size of the syllable in early grade readers (Probert 2019, see 2.1.2). Thus, fast subtitle speeds would cause difficulties in locating particular linguistic features (Kruger *et al.* 2022), as well as integrating meanings comprehended and decoded from other modes. This may be the reason that the words in the videos were not read faster or with more automaticity than words that were not in the videos. Small words like "kufuphi" (nearby) and small syllable nouns such as animals may be picked out in a short space of time, and this requires further research with a simpler video.

Picture-Word Match test scores implies they can connect modal meanings from written language mode and images. Of the readers that got zero in the ORF and Word Reading tests, half of the poor readers (50%) could still manage this test. It suggests that potentially having multimodal text semiotic awareness skills or intermodal semiotic awareness skills (vertical reading, see *3.2.*) could assist in reading the words enough to relate them to particular images. They may not have been able to read all of the words, but they processed the meanings they could from the visuals and made estimations that were still correct. This illustrates multimodal semiotic awareness skills may assist in literacy development, provided a multimodal text does not make the subtitles redundant or lead to a comprehension of a multimodal text without comprehending linguistic words. Children require vertical reading skills, which clearly some children had, at the word level at least. However, fast subtitles combined with complex and long words don't give them enough time, even repeated over 3 months, to improve written vocabulary knowledge and automaticity in ORF, even at the word recognition level. Perhaps more exposure, or guided exposure, could alleviate the lack of significant difference that the videos made on these skills and remediate the speed from the subtitles.

Shifts away from standardised conventions of written isiXhosa in the subtitles may cause learners to learn spelling and writing conventions for the language incorrectly. Spelling errors are not frequent, and one only occurs in Video 5, while the ellipsis occurs in Video 3, 4 and 5. For example, the spelling error of “incwadi” (book) as “incwandi” in Video 5. The occurrence of ellipsis (…) to indicate tonal emphasis in Video 4, and pauses in Videos 3, 4 and 5. In Video 3, this occurs in the clause complex Ndijonga phezulu esibhakabhakeni ndibone… Ukuba inyanga indilandele ukuya ekhaya (I look up at the sky and see…that the moon followed me home). In Video 5, Senza nezinye izinto ezonwabisayo ezifana… nokuthetha ngemo yezulu, nokuba loluphi usuku esikulo evekini (We do other fun things… such as talking about the weather and in which day of the week we are). Here "ezifana" can be translated "such as", so the "such as" should be before the ellipsis in the English translation. It has a higher frequency in Video 4 than in any other video, as it only happens once in Videos 3 and 5, but 9 times in Video 4. An example of this in Video 4 is in the clause Tsala isebe lomthi…Uza kubona! (Pull a tree branch…you will see!). This is most likely due to the English original stories having ellipses in the same positions in the videos as the isiXhosa translations and this was not removed in the isiXhosa translations. Thus, ellipses were used in a nonstandard way in the English and this was carried through into isiXhosa. While Video 5 has the highest Similarity rate and lowest amount of Contrast in comparison to the other videos, which is discussed in 6.6.6, the presence of incorrect spelling and writing conventions and its potential to cause errors in learning. This could mean that Video 5 is not the best video for literacy development. Furthermore, the incorrect prefixes were used in the title of *Bonke Babemthanda USue*. These prefixes refer to people, whereas for animals the title should have been *Zonke Zazimthanda USue*. Since this only occurred in the title of the video, it is not as disruptive as the spelling errors and ellipsis in the body of the text, but it does mean that Video 1 was the only video that had no instances of nonstandard isiXhosa writing conventions.

In order for subtitles to be used to develop literacy in this context, the subtitles should be at a pace the learners can follow, clear and accurate, while the visual and spoken linguistic modes are to assist learners in comprehending and developing literacy skills for unfamiliar phonological units, words or clauses. In the case of spelling errors, ellipses and incorrect prefixes, may lead to divergence between the mental representation of the word (phonological sound) and its form (alternative phonological), which influences feature binding (Liao *et al.* 2021, see 3.2). At a horizontal word level reading, learners may want to stop and try and make

sense of words they do not expect, to correctly decode it, but the subtitles being too fast in all the videos may result in learners miscomprehending (see Kruger *et al.* 2022 and discussions in 3.2). They could further take incorrect spellings and writing conventions as a standard, which could hamper their spelling and writing skills, or even reading skills. If words that are misspelt or infrequent uses of ellipses are concepts that they are unfamiliar or have not yet been exposed to, this may lead to learners' mental representation of the words forming inaccurately (see 1.2.3). Further research in this area is required.

Overall, none of the videos are suitable for learning due to their fast subtitle rate discussed in 6.6.4. However, if subtitle rate were to be excluded as a factor and the videos examined solely on the other aspects of video design discussed in this thesis, Video 5 would be a more adequate video for literacy development in terms of high Similarity rate. This is discussed further in 6.6.6. Due to ellipses and the incorrect spelling of "incwandi," Video 5 may not be the most preferred for literacy development. However, this is only one aspect of linguistic complexity. Further areas of linguistic complexity are discussed within the videos to examine which of the videos would be the most ideal for learning overall. Further aspects of linguistic complexity discussed include lack of noun omission, complex verb forms, discrepancies in frequency and duration, ideophone constructions, clause separation in subtitles and modality.

Video 2 has more Mental Processes and Participants by frequency and duration than the other videos, which have Material Processes as their dominant Processes and Participants. Mental Processes are not a common Process, depicted, however in the visual Mode and this is discussed further in 6.6.6. In terms of Participants within Material Processes, Videos 3 and 4 have a lower duration of single-state Goals than Actor Participants, which is not typical for a noun-omitting language that may often drop the subject. Actors and Goals do co-occur in these videos, particularly in the written language, however, this indicates that they are presented as prefixes within the verb complex and not as individual nouns for most of the videos. Thus, Videos 1, 2 and 5 place more emphasis on Goals than on the Actor in terms of duration and frequency. Video 2 further places emphasis on the Phenomenon in terms of frequency and duration within Mental Processes than on Sensors and Video 5 on Verbiage and Attribute rather than Carrier. Video 1 places further emphasis on Goal Participants in its utilisation of Theme and other systems within the interpersonal and textual metafunctions than the other videos, see 6.6.5. Videos 3 and 4 mostly place emphasis on the Actors by the duration and frequency of the language. This indicates translators adhering to English, which is not a subject-omitting

language. A higher frequency of Goals in comparison to Actor Participants in the other videos, even though the difference is not that great between them, may be more suitable for isiXhosa L1 learners as this structure follows the typical subject-omitting pattern of the language. Videos 1, 2 and 5 may then consequently be more natural for learners.

Nevertheless, stative verb forms such as the Mental Process in Video 1, e.g. “Akasabonakali” ([Usele-sele] cannot be seen anywhere) and in Video 3, e.g. “ukhangeleka emhle kakhulu” (he looks very beautiful), an attributive Relational Process (Halliday and Mattiessen 2004), are difficult grammatical constructions in any language. They are present in Videos 1, 2, 3 and 4 (see Appendix A). They not only emphasise the state, but the stative verb form further emphasises the object of the verb and not the agent of the verb in these constructions. They are different from passive constructions, another complex verb form, found in Video 2, for example in “uSue ubethandwa kakhulu zizo zonke izilwanyana zasekhaya esixekweni” (Sue was loved by all the pets in the city). Causative verb forms, further complex verb forms, are found in Videos 2, 3 and 5. For example, in Video 5 “kusasa utata wam undisa eklasini yam” (In the morning my father takes me to my class). Reciprocal verb forms in Video 4, “Izikhwenene zizifihla phezulu emithini” (Parrots hide themselves high up in the trees) are also difficult grammatical constructions in a language and applicative verb forms are found in Videos 2, 3, 4 and 5. For example, “utata undibalisela amabali apha endleleni” (my father tells me stories on the way) in Video 3. There is little research on when learners can decode and comprehend the passive, stative, causative, applicative and reciprocal verb forms in isiXhosa. Rees (2016) indicates that while the prefixes that represent these verb forms in the verb complex could potentially be decoded by Grade 3, they are not necessarily comprehended. Further research would be needed to determine the difficulty for learners to decode complex verb forms in their L1 for these levels. The complexity of language in these videos may decrease learners’ cognitive capacity, which is not conducive for learning according to Mayer’s (2009) limited capacity assumption. Since this research did not investigate which of these complex verb forms are more challenging than other verb forms for L1 literacy development in this context, all instances for each verb form were not counted. However, since all the videos contain one or more of these complex verb forms, all of the videos are determined as containing this aspect of linguistic complexity.

Another challenge in the linguistic mode that may result in a decrease in learners' cognitive capacity is the discrepancy of the frequency and duration between Processes and Participants in the spoken or written mode. Discrepancies between frequency and duration between Participants, Processes and Circumstances in a single linguistic mode, such as in the spoken mode, we shall label as intramodal linguistic discrepancies. While these are certain to occur, (see 4.3.2.5) due to the nature of language and their unique characteristics in duration and frequency, intramodal linguistic discrepancies relate to a system choice in one mode that are present for longer in the video. However they are not necessarily the most frequent and vice versa. For example, a spoken Actor Participants may be present for longer in the video, however it may not be the most frequent spoken Participant and vice versa. This particularly is highlighted in the spoken mode. This may naturally be a part of language, and Tseng and Djonov (2023) discuss that frequency and duration of exposure are not as significant for comprehension of time than co-patterning of temporal meaning relations. However, it is unknown whether this is true for this particular context, i.e., SLS videos for L1 isiXhosa literacy development in South Africa. Frequency and exposure short-term to the videos in the intervention described in 4.2.3 and the results in 5.2.1 indicate that short-term exposure and the frequency within the videos resulted in no significant difference for word recognition, ORF and specific word-level semiotic awareness skills. Such conditions of video-watching duration and frequency were not enough to surpass difficulties in following subtitle rate, even when Participant, Process and Circumstance meanings also appeared at different durations and frequencies in the videos. Nevertheless, if the subtitles could be followed by learners, further research is required to determine whether the frequency and duration are not as integral for literacy development in this context. Such as, for learning of time in the study Tseng and Djonov (2023), as co-patterning temporal relations, and thus Intersemiotic Texture and Comparative Relations in this context.

In terms of intramodal linguistic discrepancies, the only video in which duration and frequency correspond within the spoken mode and within the linguistic mode for Participants, Processes and Circumstances is Video 2. Videos with less intramodal linguistic discrepancies include Video 1 and 4. In Video 1, there is only a discrepancy in duration and frequency between spoken Processes, as Existential Processes appear longer than Material Processes durationally. However, Existential Processes occur less frequently than Material Processes, which aligns with Goals as the most frequent and highest Participant durationally. There are no discrepancies between frequency and duration in spoken Participants and in the written mode. In Video 3, 4

and 5 there is only a discrepancy in duration and frequency between Participants. In Video 4, Actors appear longer and is the most frequent Participant, but second-most spoken Participant durationally is Attribute and second-most frequent Participants are Sensor and Existent. Third-most spoken Participant durationally is Existent and third-most frequent spoken Participant is Attribute. There are no discrepancies between frequency and duration in spoken Processes and in the written mode for both videos. Since the most predominant Processes in the video are Material, Existential, Relational and Mental, this this discrepancy may not significantly influence learners' cognitive capacity. More research in this area is required.

In Video 3, Actors appear longer and most frequent, but second-most spoken Participant durationally is Attribute and second-most frequent Participant is Goal. Relational Processes are only the third-most frequent and third-highest Process in this video durationally; thus, the lengthy duration of an Attribute might result in confusion or particular salience. As it co-occurs with Relational Processes, this discrepancy may not significantly influence learners' cognitive capacity. However, this is not the only discrepancy. The third-most spoken Participant durationally is Sayer and the third-most frequent spoken Participant is Sensor. The second-most spoken Process and second-highest Process durationally is Mental. In Video 5, While Material Processes are most frequent in both spoken and written modes, Participants of Verbal, Relational and Mental Processes are more predominant in this video. There is an intramodal linguistic discrepancy between Participants in this video, with Verbiage as the highest Participant durationally and Sayer as the most frequent Participant. Attribute is the second-highest Participant durationally and Sensor is the second-most frequent Participant. Goal and Existent appear at the same duration and are third-highest Participant durationally, however, Goal appears third-most frequently, followed by Actor Participants. Thus, the frequency and duration of Participants and Processes in Videos 3 and 5 result in discrepancies which may cause cognitive overload for learners. If these intramodal linguistic discrepancies result in a greater cognitive load for learners, then Video 2 may be the best of the five videos in relation to this aspect of Mayer's (2009) limited capacity assumption. Videos with less intramodal linguistic modal discrepancies include Video 1 and 4, while Video 3 and 5 have greater discrepancies. More research in this area is required.

The only video with an intramodal linguistic discrepancy in Circumstances is Videos 5, which may further result in a greater cognitive load on learners. Thus, resulting in Video 5 being the worst of the videos in this aspect, with intramodal linguistic discrepancies in spoken

Participants, Processes and Circumstances. While Circumstances of Location: Place and Circumstances of Location: Time were most frequent in all of the videos, including Video 5, Videos 1, 2, had linguistic Circumstance system choices in each mode that corresponded by duration and frequency. In Video 5, the second-highest Circumstance durationally is Matter, but this Circumstance is the third-most frequent. The third-highest Circumstance durationally is Extent, but this Circumstance is the second-most frequent. This may result in Video 5 being the worst for a learners' cognitive load with relation to frequency and duration.

Discrepancies in frequency and duration of Participants, Processes and Circumstances between spoken and written linguistic mode, I shall label as intermodal linguistic discrepancies. These are certain to occur, as discussed in 4.3.2.5, due to the nature of language differences in spoken and written linguistic modes and they occur in all five videos. Spoken language unfolds in shorter phonological units over time, while more of the clause is visible in the written mode at one particular time than in the spoken mode. An examination of these discrepancies potentially highlights challenges in the written mode, particularly due to more co-occurrences and complexity for learners to decode. Furthermore, if there are more discrepancies in the spoken language (an intramodal linguistic discrepancy) relating to the content of the language, it may further cause intermodal linguistic discrepancies, which may result in more difficulties for learners to integrate meaning differences between linguistic modes. Furthermore, if they do not correspond in relation to the visual mode, it may result in an influence on Intersemiotic Texture. This may negatively influence literacy development with SLS videos, as it causes a video to potentially not adhere to Mayer's (2009) Principles of Spatial and Temporal Contiguity. Thus, these types of discrepancies, if they decrease learners' cognitive capacity, may not be conducive for learning according to Mayer's (2009) limited capacity assumption.

While all of the videos have instances of intermodal linguistic discrepancies, Video 5 has intermodal linguistic discrepancies in terms of frequency and duration between Participants, Processes and Circumstances. This may result in this video requiring more cognitive capacity from learners. In Video 5, spoken Participants in order of duration are Verbiage, Attribute, Goal and Existent, while written Participants in order of duration are co-occurrence of Actor and Goal, co-occurrence of Sensor and Phenomenon, Actor Participants only and Sensor Participants only. While Material Processes in Video 5 are the highest in duration and frequency between linguistic modes, the second-most spoken Mental Process does not correspond with second-most written Verbal Process durationally. Furthermore, third-most

spoken Verbal Process does not correspond to third-most spoken written co-occurrences of Material and Mental Processes. While Circumstances of Location: Place and Circumstances of Location: Time were most frequent in all of the videos, including Video 5, Video 5 contained discrepancies with the rest of its Circumstances. The second-highest Circumstance durationally is Matter in the spoken mode, but Extent in the written mode. The third-highest Circumstance durationally is Extent in the spoken mode, but Matter in the written mode. The fourth-highest Circumstance durationally is Accompaniment in the spoken mode, but Manner in the written mode and this is reversed for the fifth-highest Circumstances in both modes. Thus, intramodal and intermodal discrepancies in Video 5 may cause this video to be less conducive for learning. Further research is required to determine if these discrepancies in frequency and duration would influence learning more significantly than Similarity rate in Comparative Relations or if subtitles were presented at a slower speed.

Video 1 spoken Processes frequency aligns with written Processes duration and frequency, it's only in duration that there is a discrepancy, mainly between Existential and Material Processes as second-highest and third-highest Processes durationally. While Video 2 may have no intramodal linguistic discrepancy between duration and frequency, but there is an intermodal linguistic discrepancy between spoken and written modes. While Participants in Video 2 ordered according to predominance in the spoken mode are Phenomenon, Goal and Actor, in the written mode these are a Co-occurrence of Sensor and Phenomenon, Actor and Sensor. While Phenomenon is present more in this video in both modes, durationally, there is a discrepancy between the rest of the Participants between modes. This indicates that Video 1 has less of an intermodal discrepancy that may not necessarily result in a challenge for learners' cognitive capacity. Further research in this area is required.

Video 3 and 4, while not having as many intermodal linguistic discrepancies as Video 5, contain more than Videos 1 and 2. In Videos 3 and 4, Participants align more with frequency than duration, however there are still discrepancies between spoken duration, spoken frequency and written duration and frequency. In Video 3, Participants are ordered according to predominance in the spoken mode durationally are Actor, Attribute and Sayer, but in terms of frequency they are Actor, Goal and Sensor. In the written mode these are Actor, a co-occurrence of Actor and Goal, a co-occurrence of Identifier and Identified. In Video 4, Participants are ordered according to predominance in the spoken mode durationally are Actor, Attribute and Existent, but in terms of frequency they are Actor, Sensor and Existent (tied) and

Carrier. In the written mode these are Actor, a co-occurrence of Actor and Goal, a co-occurrence of Carrier and Attribute. Furthermore, there are further intramodal linguistic discrepancies between modes in terms of Circumstances in Video 3 and Processes in Video 4, both with no intramodal discrepancies. While Circumstances of Location: Place and Circumstances of Location: Time are predominant in both modes, Manner is the second predominant Circumstance in the spoken mode, but Extent is the second predominant Circumstance in the written mode. In Video 4, Processes are ordered according to predominance in the spoken mode durationally and by frequency are Material, Existential and Relational and Mental (tied). However, in the written mode, these are Material, Relational and a co-occurrence of Material and Mental. Thus, while Material is the predominant Process in both modes, there are discrepancies in the rest of the Processes for this video. This could also create difficulties for literacy development and particularly for Mental and other Processes that are not often presented concurrently in the visual Mode, such as Behavioural and Verbal Processes. Only Videos 1, 3 and 5 had Mental Processes in the visual mode that do not include Participants seeing objects or other Participants, even ones not included in the shot. If discrepancies hamper learning, then more research is needed to find out the amount of discrepancy between frequency and duration that would cause this. Nevertheless, a suggestion for video design to better adhere to Mayer's (2009) Spatial and Temporal Contiguity Principles would be to maintain similar Processes and Participants across modes to ensure there is less of a cognitive overload or chance of confusion for learners.

In addition, the co-occurrences of Processes and Participants in all of the videos may further increase the cognitive load for learners, thus hampering learning according to Mayer's (2009) limited capacity assumption. Co-occurrences of Processes and Participants, however, happen in language every day and whether this does decrease cognitive capacity to the point of influencing learning in a negative manner requires further research. Co-occurrences are discussed in 6.1-6.5, as well as earlier in this sub-section. Co-occurrences of different systems simultaneously, referred to as couplings and syndromes (see 3.1) are discussed in 6.6.2.

Another potential challenge for learners' cognitive capacity, which would not be conducive for learning according to Mayer's (2009) limited capacity assumption is the presence of isiXhosa ideophones in Videos 1 and 3. This raises questions on how this influences learning. Constant repetition of complex ideophone Processes (see 1.2.1.3), e.g., "ethe cwaka" (he remained silent) and "wa'man ukuyithi krwaqu le mpukane" (he kept glancing at the fly) in Video 1 pose

two questions for further research. Firstly, at what level do L1 isiXhosa children read and understand these predicates? As alone the verb -thi means to do or say, e.g., “Inyanga yam ithi molo kum” (My moon says hello to me) in Video 3. However, when combined with an ideophone, “-thi” becomes semantically bleached and takes the properties of the ideophone next to it. These are constructions are complex in meaning, but perhaps for L1 readers it may not be difficult in literacy development. This would require a battery of tests to understand at what level and age reading this construction is appropriate for a learner. Furthermore, tests would be required to ascertain whether complex ideophone Processes, can be identified and distinguished from the sole use of a verb like -thi.

Secondly, what ramifications does an ideophone Process have for SFL classification? Depending on the meaning of the ideophone, -thi is either a Material Process or a Verbal Process, depending on whether it means ‘do’ or ‘say.’ In a complex ideophone predicate, -thi is semantically bleached, carrying agreement for the predicate, but the actual semantic properties of the Process are on the word to its right and it is this word that assists in the classification of the Process. This is similar to Material Processes with Scopes in English, such as in the English sentence “Suzy took a bath.” In the example, ethe cwaka (he remained silent), I have classified it as Behavioural Process, with “cwaka” (silent) and the prefix e- representing Sele-sele as Behaver. In the example, “wa’man ukuyithi krwaqu le mpukane” (he kept glancing at the fly), the presence of the object concord -yi- which represents “impukane” (the fly) does not allow one to classify this Process as Behavioural. Instead, there is an emphasis by the modal prefix “-man” (keep doing) and “krwaqu” (glance) on the duration of the look, rather than on the action or what is seen, as this ideophone construction presents a modifier as well as a predicate. Consequently, this Process was classified as a Material Process with “le mpukane” as Goal and not as a Mental Process. Both ideophones further appear as common Themes in this video, which further emphasises them, and they are repeated often in the video. Further research on complex ideophone Processes that project off -thi, in terms of SFL classifications, how children learn to read them and at what level they are appropriate, is required.

A further instance that is specific to the written linguistic mode that may result in a decrease in learners’ cognitive capacity is the instance of clause separation in the subtitles, where a clause is separated between two subtitles. This is found in all five videos, although it is most prevalent in Video 2. Often, clause complexes are separated between subtitles due to the length of words and sentences, but this may not increase a learners’ cognitive load if the referents in the

following subtitles are clear. However, if a single clause is separated between two subtitles and components are separated, referents may not be clear and this may cause a loss of cohesion for the text and not adhere to SLS subtitle conventions in other agglutinative languages (see Lappalainen 2014). This may hamper learning according to Mayer's (2009) limited capacity assumption. While clause separation in subtitles may be classified as influencing cohesion within the textual metafunction, other systems improving cohesion within the textual metafunction for both modes, such as Theme and Information Value, may influence the highlighting of Participants, Processes and Circumstances. This may result in improvement in the learning of these constituents. This is discussed in section 6.6.5.

While Video 4 has the least instances of Modality (see 6.4.2), Video 1 has eight instances of Modality, however they are four separate occurrences of the same two constructions and modal markers (see 6.1.2). This is in the stative verb form with *-s-* affix, e.g., "Ayisabonakali" (It cannot be seen anywhere) and ideophone construction with *-man* affix, e.g., "Waman' ukuyithi krwaqu le mpukane" (He kept glancing at the fly). These are two types of modality, usuality and probability. Video 4 has three instances of *-nga-*, two within *ukuba* clauses, which create further complexity, and an instance of *-za-* and *-ku-* (see 6.4.2). Modality in relation to Mayer's (2009) Personalization Principle for Video 4 is discussed in 6.6.7. Thus, Video 1 may present the least complexity and distinct instances of modality. However, Video 5 would be tied with Video 1 for the least amount of modality, as while there are seven instances, they are separate occurrences of the same modality words *-soloko* (always) and *ngamanye amaxesha* (sometimes) (see 6.5.2). There is only one type of modality in this video, usuality, thus if it was the frequency of different types that created complexity, arguably Video 5 may present the least complexity and distinct instances of modality. Further research in this area in this context is required.

Video 2 and 3 have the most complex instances of Modality. While Video 3 has a total of fifteen instances of Modality and three clauses with multiple modal markers (see 6.3.2), which result in ten modal clauses. It has three types of Modality, which are probability, usuality and obligation. Words that realise Modality in Video 3 include *-za -ku*, *-yaku-*, *ukuba*, *yokuba*, *xa*, *-soloko* (always) and *ngamanye amaxesha* (sometimes) (see 1.2.1.1). Video 2 has eleven instances of Modality and two clauses with multiple modal markers (see 6.2.2), which result nine modal clauses. It has all four types of Modality, which are probability, usuality, obligation and inclination. Words that realise Modality in Video 2 include *-vumela* (to be allowed to),

qhele (usual/unusual), *ukuba*, *xa*, and *-soloko* (always) (see 1.2.1.1). Thus, while Video 3 has more instances of Modality and modal clauses, Video 2 has more types of Modality than Video 3. Further research in this area in this context is required.

Video 3 seems to be more linguistically complex than other videos. Videos 3 and 4 are highly linguistically complex, linguistic complexity both have agents more than objects of verbs and non-standardised conventions from ellipsis, modality and ideophones in Video 3. Furthermore, Video 5, which contains a spelling error, has the most instances of intramodal and intermodal linguistic discrepancies, followed by Videos 3 and 4. However, Video 1 and 2 have similar linguistic complexity distinguished by a single factor, for Video 1 it is ideophones, for Video 2 it is complex modality. In addition, however, Video 2 has the most frequent clause separations between subtitles than any of the other videos. Further research is required to ascertain which of these aspects of linguistic complexity causes a greater challenge in L1 isiXhosa literacy development.

The Pre-Training Principle stipulates that if learners have prior knowledge of content, they will learn easier (Mayer 2009). This relates to the learning intervention discussed in Chapter 4.2.3 and lengthy exposure to the videos that the children received. The videos were selected based on the vocabulary in the videos, as they would have been familiar to Grade 2 and 3 learners. While their reading improved in the three months of the intervention, the statistics show the videos themselves made no significant difference to reading scores post-intervention. Due to the speed of the subtitles being too fast for the average reader at this age and lack of prior subtitle exposure, the addition of pre-training could potentially assist learners in identifying features and indexing locations (Liao *et al.* 2021, see 3.2) that may assist in L1 isiXhosa literacy development. Furthermore, pre-training could assist in alleviating learners' cognitive load and allow them to navigate videos' linguistic complexity. Further research is required to investigate whether videos integrated into lessons are more influential on L1 isiXhosa literacy development, which corresponds to Mayer's (2009) Pre-Training Principle.

Mayer's (2009) Pre-Training Principle was followed to an extent in terms of lexical vocabulary, but literacy development may have improved, if the video was integrated into a lesson and specific words and phrases were trained directly prior to watching the videos. This study did not do that, however, as it was designed to reveal the influence of sole exposures to the videos without additional factors on literacy development. A study with a longer intervention time and

greater sample size might indicate a difference in results. While a three-month exposure period for videos like this may nevertheless be too short to show a large significant difference in this context, as shown by previous longitudinal studies in other contexts (see Kothari 2008), the subtitle rate would have to be adjusted for learners to comfortably follow them. The grammar would also have to be appropriate for the grade, and this and the subtitle rate would have to be adequate before the duration of the exposure would be a significant factor. Furthermore, one would have to test for the optimal amount of time for exposure before this is redundant and learning would become tedious, which would decrease learner attention and any positive effects for learning.

Mayer's (2009) Multimedia Principle stipulates that the combination of pictures and words (either written or spoken, not both) causes easier learning than through words alone. The children demonstrated that they could relate linguistic meaning to visual meaning in the Picture-Word Match test discussed in 5.1.3. However, the paired sample statistics on Long Passage Reading and Word List tests in 5.1.4 demonstrate that the videos did not assist in learners' literacy improvements during the three-month intervention. Thus, while this principle could potentially be true, the videos did not assist learners to improve their reading as the subtitle rate was too fast for the subtitles to be read.

6.6.2 Syndromes and the typical genre of children's narratives

All the videos presented a high frequency of Circumstance of Location: Place and Location: Time. Often, Circumstances of Location: Time further realise usuality types of Modality, e.g., -*soloko* (always) and *ngamanye amaxesha* (sometimes). A high frequency of Circumstances of Location combined with many emphasised deictic words, e.g. "**nalo** ulwimi" (There is [this] tongue) as a Theme in Video 1 and "ukuba ikhona" (that it is there) in Video 4 seem to assist in focusing and maintaining young children's attention (Levinson 2004, Peña 2020). Existential Processes like "nalo" act as deixis of space which refers to a specific Participant in relation to its space and in relation to any similar Participants around it. This does allow for all the videos, depending on words and concepts used, to further be good reinforcement for Circumstances of Location: Time and Circumstances of Location: Place for literacy development, e.g., "ngamanye amaxesha" (sometimes) from Video 3 and "esikolweni" (at school) from Video 5. The couplings (see 3.1) of Material Processes and Participants, Action Processes and Participants and these Circumstances of Location occur throughout all five videos and

represent a syndrome (see 3.1) that is typical for children's narratives internationally (Guijarro and Sanz 2008). The videos' adherence to a broader children's narrative genre may derive from the translation of these videos from English stories. However, it also raises the question as to whether these narrative patterns are typical for children's narratives in isiXhosa and whether these universal narrative patterns are suitable for learners in a South African context. More research is required in this area.

6.6.3 Intersemiotic Additive Relations, Similarity system choices and Mayer's (2009) Coherence Principle

The presence of Intersemiotic Additive Relations allows for the visual mode to provide information that the language does not. For example, in the subtitle "Ndinomhlobo wam oyedwa (omnye) ophuma ebusuku kuphela" (I have one friend of mine who comes out only at night). The image reveals the friend's identity (i.e., the moon). The relationship of meanings between modes is not as clear or similar between modes as Similarity in Intersemiotic Comparative Relations. However, it may be more beneficial for literacy development to add information (e.g., Participants) in the visual mode to add more depth of description to information in the linguistic mode, than if information was added in the visual mode that had no correspondence in the linguistic mode or vice versa. Intersemiotic Additive Relations may allow for better adherence to Mayer's (2009) Coherence Principle than the addition of Participants, Processes and Circumstances in one mode without any corresponding Participants, Processes and Circumstances in other modes. This is due to Mayer's (2009) Coherence Principle describing extraneous material as detrimental to a learner's cognitive capacity. Since cognitive capacity is assumed as limited, an increase of cognitive load could hamper learning. Thus, Intersemiotic Additive Relations may improve literacy development and cause fewer distractions than simply adding information in one mode and no correspondence in the others.

Examples of additions in the visual mode that have no correspondence or are not creating depth in the linguistic mode are the animated lily pads in Video 1, the insects animated in Videos 3 and 4, the additional visual animal Participants in Video 4 and the animated stuffed rabbit in Video 5. These visual Participants might shift the focus of the viewer onto Participant movement and take their eyes off the subtitles, particularly when the subtitles are at a rate that

is too fast to read. This may also cause issues for learners to integrate meanings between modes to build a model of the situation, potentially posing challenges for comprehension (see 3.2 and discussions of multimodal literacy). In this way, the videos may adhere less to Mayer's (2009) Coherence Principle. Further research would be needed in relation to potential visual distractions in the videos, and this could be examined with eye-tracking software. If these additional visual elements were static and not animated, or animated during moments without subtitles, it could allow less focus on them, and their addition may not pose as much of a distraction. Thus, there needs to be a mediation in video design between the visuals adding information to aid the understanding of the language and the addition of visuals that do not relate to the language in the videos that could distract learners. If visual Participants are added, this should be within Intersemiotic Additive Relations or they could be present when there are no subtitles or spoken language in the video.

Further research is required to confirm that Similarity of Correspondence assists literacy development more than other system choices for Similarity, such as Hyponymy and Collocation. An example of Similarity of Hyponymy is in Video 2, the word "izilwanyana" (animals) is only represented in the visual mode by certain types of animals (e.g., dogs, cats, birds and squirrels). An example of Similarity of Collocation also occurs in Video 3, in which "I" and "my father" in the linguistic mode are shown as lions in the visual mode. While this is more similar in meaning than added visual Participants, which have no corresponding linguistic Participant, it may still require more semiotic awareness skills to integrate meanings between modes than if these were Similarity of Correspondence.

6.6.4 Subtitle rate and Mayer's (2009) Redundancy and Segmenting Principles

Mayer's (2009) Redundancy Principle implies that even if on-screen text and narration are the same, people learn better without on-screen text with narration and visuals. Mayer (2009) was taking into account that if similar information was produced in different modes, it may cause learning to be tedious and may result in boredom. This is mentioned as a concern for repeated information by Guijarro and Sanz (2008). Further research is needed to examine the optimal amount of repetition in this context, enough to maximise the benefits for literacy development and minimise its challenges (i.e., boredom or disinterest). Nevertheless, Mayer's (2009)

Cognitive Theory of Multimedia Learning, as mentioned in 3.2.1, proposes principles for a context where linguistic proficiency was assumed and not for literacy development, despite the presence of the dual channel assumption in literacy development theories. Due to Mayer's (2009) use of this assumption and applying it further to other semiotic modes outside of language for literacy, I chose to adapt his principles for the context of literacy development (see 3.2.1). To aid clarity for learners beginning to read, the presence of similar information in both the written and spoken mode is essential for learning. Thus, in this context, Mayer's (2009) Redundancy Principle must be violated in order for learning to occur. In this context, when this thesis' expanded definition of Mayer's (2009) Segmenting Principle is followed, which stipulates that narrated animation and subtitles should be at the learner's pace, then the Redundancy Principle would be considered violated, as learners can follow the subtitles. If they could not follow the subtitles, due to a fast subtitle speed and a violation of this thesis' expanded definition of Mayer's (2009) Segmenting Principle, the presence of the subtitles would be redundant, and this would result in an adherence to Mayer's (2009) Redundancy Principle. Thus, in a literacy development context, Mayer's (2009) Redundancy Principle is the only principle that has to be violated, whereas other Principles require adherence. Conditions that need to be met to violate Mayer's (2009) Redundancy Principle in this context for improved literacy development include the adherence to this thesis' expanded definition of Mayer's (2009) Segmenting Principle and enough repetition of information that is conducive to learning.

In terms of a potential adherence to this thesis' expanded definition of Mayer's (2009) Segmenting Principle, all five videos had a subtitle rate far greater than the ability of the learners to read. Thus, this may have caused this thesis' expanded definition of Mayer's (2009) Segmenting Principle to not be followed. The subtitles are not readable for Foundation Phase learners, which implies that learning from the content would have been better with subtitles omitted than these videos being resources to aid in literacy development. If the subtitles are not readable or removed from the videos, the videos would only be useful in improving visual literacy, general spoken language comprehension and language learning that does not include literacy or written areas of language learning. Therefore, this may realise Mayer's (2009) Redundancy Principle. The only remediation to this would have been if there were fewer words in the video per minute. This can be solved with fewer words in the video overall and less grammatically complex clauses. This is not necessarily the case that the subtitles are unreadable to any isiXhosa reader outside of a context of Foundation Phase L1 literacy development (see

3.2). However, the subtitle rate would pose far too great a challenge specifically to Grade 2 and 3 beginner L1 isiXhosa readers based on the current findings of the intervention discussed in 5.1.

When comparing the videos to see which video had the lowest subtitle rate and potentially greater learning potential according to this thesis' expanded definition of Mayer's (2009) Segmenting Principle, Video 1 had the second-lowest subtitle rate speed, with Video 4 having the lowest rate. There is a correlation between the number of words in the video and subtitle rate as Video 4 had the lowest subtitle rate but was also the videos with the least number of total words. While Video 4 had the lowest subtitle rate, it was still not low enough for learners to follow and thus, may not have adhered to this thesis' expanded definition of Mayer's (2009) Segmenting Principle**.** Furthermore, despite the low subtitle rate, Video 4 may not have been more conducive for learning than other videos as it had many aspects in video design that may have caused less adherence to the other principles (Mayer 2009). These aspects include additional visual Participants (6.6.3), lack of relevant Participant highlighting in the interpersonal and textual metafunctions (6.6.5) and the lowest amount of Similarity and highest amount of Contrast in proportion to the number of words of all the videos (6.6.6). A video that has a slower subtitle speed and a higher Similarity rate may be better for learning than a video like Video 4, which has the slowest subtitle rate, but a potential lack of adherence to the other principles of design. These principles include Mayer's (2009) Coherence, Signaling, Spatial Contiguity and Temporal Contiguity Principles.

A limitation to the study is that learners read subtitles silently while the tests measure ORF rates in this intervention. However, learners' silent reading rates were not collected as Probert (2016) confirms, there is no way to objectively measure a learner's silent reading rate (see 2.1.1). Even with eye-tracking software examining eye movement, an average silent reading score could be deduced if the text was static. When it includes visuals and moving subtitles, there would be too much eye movement to deduce an accurate object reading rate per minute. Even if, hypothetically, the silent reading rate were double the WCPM rates collected in the study, 51 WCPM, which was the highest potential WCPM score in the intervention by Grade 3 readers, was comprised of conjoint (short form) words and not disjoint (long form) words. The subtitles were thus more linguistically complex and had longer word lengths than the ORF tests, which indicates the learners would have read the subtitles slower than the results in the ORF tests. Furthermore, the maximum reading score is a minority score and, therefore, the

majority of learners that can read in the Foundation Phase would not be able to read the subtitles in this video. The discrepancy between ORF scores and subtitle rate is so large that even if there had been a way to collect data for silent reading scores for these learners, it would have likely reflected the same problem of the subtitles being too fast for learners to decode and comprehend.

The learners in this context did not improve in ORF at the word-level or long reading passage level because of the videos. The subtitles were too fast for all the videos and the tests illustrate that there was potentially not enough time for learners to use their semiotic awareness skills to connect repeated words to visuals (see discussions on horizontal reading by Kruger *et al.* (2022) in 2.2.31, 3.1.2 and 3.2). Alternatively, the lack of repetition and emphasis of enough words during the videos could have resulted in the learners being unable to follow the subtitles, even if they were slightly faster than their ORF reading rate levels. All the videos have a high Similarity rate, but they had nuances in how they presented the information that could have caused challenges for learners, depending on the video (see 6.6.1). In terms of reading speed at the word and sentence level, decoding and ORF skills, the videos did not assist. One must decode the linguistic text to comprehend the linguistic text (Wilsenach 2013). For learners to be able to decode the linguistic text and then focus on comprehension aspects, the subtitle rate had to be at a speed that they follow (see discussions on horizontal reading by Kruger *et al.* (2022) in 2.2.3.1, 3.1.2 and 3.2). At the high objective subtitle rate of the videos, learners did improve in their reading in this intervention, even if they could have caught isolated words (see 3.2). Thus, while the videos had a high Similarity rate, even if learners could see a word and map it to corresponding meaning, thus decoding and understanding a word better, this was not done to an extent that overall automaticity of word and passage reading could be improved.

This thesis proposes that if a subtitle rate is too fast, the Similarity rate alone could not improve reading speed, decoding and ORF skills in literacy. Thus, the DMDA highlights Intersemiotic aspects of video design in this context, which depend on the subtitles to be at a rate that learners can follow in order for these aspects to make a difference in literacy development. If the subtitle rate is decreased to a manageable level, but the Similarity rate is less, there would be alternative challenges in reading and comprehension. Both an adequate subtitle rate and a high Similarity rate is necessary for ideal literacy development to occur (see discussions on horizontal subtitle reading and vertical reading in 2.2.31, 3.1.2 and 3.2). If all the videos had slower subtitles, the high Similarity rate present in Video 5 would result in this video being better for literacy

development in terms of design. However, if one takes into account the presence of linguistic complexity, and less utilisation of interpersonal and textual metafunctional systems for emphasis, Video 1 may have been a better video for literacy development. This is due to the misspelt word in Video 5 and large discrepancies of frequency and duration in the linguistic mode intramodally and intermodally. While Tseng and Djonov (2023) indicate that co-patterning may be more significant for learning than frequency and duration, this was learning about time and not literacy development, as well as another context. Research in multimodal literacy (see 3.2) indicate Similarity between modes is integral for integration of information between modes and reducing learners' cognitive capacity for easier comprehension (vertical reading aspect discussed in 3.2). However, frequency and duration are not discussed in relation to meaning relations between modes. Further research should examine these aspects of highlighting and linguistic complexity in relation to one another to confirm which influences L1 isiXhosa literacy development more in this context. Highlighting in the interpersonal and textual metafunctions is discussed in 6.6.5 and Similarity and Contrast in relation to Intersemiotic Texture is discussed in 6.6.6.

6.6.5 Systems in the interpersonal and textual metafunctions and Mayer's (2009) Signaling Principle

Mayer's (2009) Signaling Principle states that highlighting essential words guides learners when processing knowledge and leads to more efficient learning. Types of highlighting are discussed in 3.2.1. Furthermore, I argue that highlighting is not restricted to the auditory channel alone. In the videos, this may be visible in the highlighted information in the visual mode in terms of Social Distance, Zoom, Camera Movement and Salience in the interpersonal metafunction, as well as Framing and Information Value in the textual metafunction. Mood and Theme also highlights essential words and specific information in the linguistic mode. Fully utilising as many of these systems as possible to highlight frequent Participants, Processes and Circumstances may allow for enhanced literacy development, as this guides learners' attention on these words and phrases.

Mood in the interpersonal metafunction and Themes in the linguistic textual metafunction could have been used as to better emphasise specific concepts and actions in the videos. The only video to employ clauses in Imperative and Interrogative Moods is Video 4, and the video

only had two clauses, even though they also presented instances of Modality. Further research on the influence of Mood structures on literacy development in this context is required. As seen earlier in this chapter, the only videos to employ Themes to its best potential were Videos 1 and 2, with frequent recurring Themes not being present in the other videos. Like other systems of the interpersonal and textual metafunctions, Themes could complement the meaning choices in the ideational metafunction by emphasising specific words and phrases. This could improve literacy development, as it could assist in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle.

In relation to visual system choices for these metafunctions, only Video 2 has no Zoom effects, while all the videos use Zoom on predominant Participants in both modes. All videos have Camera Movement, but Videos 1, 2 and 4 have Pan and Tilt, Video 3 has Dolly and Pan and Video 5 just has Pan. Further research is required to examine which Camera Movements would be better to maintain learners' attention and facilitate literacy in this context. Nevertheless, Videos 1, 2 and 3 utilise more than one Camera Movement system and highlight specific predominant Participants. In contrast, non-predominant Participants in the linguistic mode are highlighted by Camera Movements in Video 4, which may not be desirable for literacy development. Furthermore, Framing is used in all videos except Video 4. All videos had their predominant Participants as Centre at some point in the video, but Video 4 had Participants as Centre that were not predominant Participants in the linguistic mode, rather additional visual Participants. The only video that used Salience was Video 1, which included a Sharpness of Focus on Sele-sele and placing insects in the foreground to Sele-sele. Consequently, of these systems and Theme in the linguistic mode, Video 4 utilised the fewest systems to highlight specific predominant Participants.

Long shots and Front, Eye Level, Stationary shots are typical for the genre and conducive to maintain a child's attention (Guijarro and Sanz 2008). However Close and Medium shots may assist in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle. Close and Medium shots add a layer of intimacy and identification between the viewer and the characters (See Smith and Adendorff 2021; Guijarro and Sanz 2008). Only Videos 1, 2 and 3 use Close and Medium shots on predominant Participants. Video 4 and 5 only use Medium shots on predominant Participants. While all the videos use Medium shots, only Videos 1, 2 and 3 fully utilise all types of shots in the video, which may improve learning.

The system of Social Distance assists in adherence to both Mayer's (2009) Signaling and Personalization Principles. Mayer's (2009) Personalization Principle is discussed in 6.6.7.

Interpersonal and textual metafunction systems may assist in adhering to Mayer's (2009) Signaling Principle. Gaze and Horizontal and Vertical Perspective may not assist with this thesis' expanded definition of Mayer's (2009) Signaling Principle as much as other systems in the interpersonal metafunction. Gaze and Horizontal and Vertical Perspective may assist in adherence to this thesis' expanded definition of Mayer's (2009) Personalization Principle.

If there is a lack of consistency in what is highlighted by these systems, it might cause a lack of coherence and disrupt Mayer's (2009) Coherence Principle. Furthermore, if there is a consistency of highlighting, this might assist a high Similarity rate and low Contrast in meanings between modes. This may cause a better adherence to Mayer's (2009) Principles of Spatial and Temporal Contiguity, which could assist in improving literacy development for learners and cause less distraction and confusion.

Intersemiotic Temporal and Consequential Relations may assist in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle. An example of Intersemiotic Temporal Relations in the videos is illustrated in Video 1. When the subtitle "LETSHE!" (Shoot!), an onomatopoeia, appears, Sele-sele's tongue in the visual mode shoots out, latches on to the insect, retracts and the insect is swallowed. The visual mode here illustrates a sequence of complex actions associated with the word "LETSHE!" Thus, the onomatopoeia here appears at the beginning of a sequence of complex visual actions that are associated with it. This may emphasise the visual Processes, which may foster literacy development. Intersemiotic Contingency within Intersemiotic Consequential Relations determines a possibility, with the aftereffect not a guarantee (Liu and O'Halloran 2009:382). The only example of this is in Video 4 (see 3.1.3). In terms of potentially fostering learning, it could make learning more interactive for learners, which could maintain their attention and assist in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle. However, more research is needed on the different systems of Intersemiotic Texture and their influence on literacy development.

6.6.6 Intersemiotic Texture and Mayer's (2009) Principles of Spatial and Temporal Contiguity

The system of Comparative Relations may be the best for L1 isiXhosa literacy development, including for the improvement of skills such as word recognition and comprehension in the design of these videos, in comparison to other types of meaning relation systems such as Additive, Consequential and Temporal Relations. Discussions of this in multimodal literacy theories are in 3.2. Similarity and Contrast of meanings between modes may relate to Mayer's (2009) Principles of Spatial and Temporal Contiguity, which states that if visuals and words are presented simultaneously or closer to one another, learning is improved. This is supported by Liao *et al.* (2021) and their multimodal-integrated language framework for multimodal reading, as well as vertical reading by Kruger *et al* (2022) in 3.2.

As mentioned in 6.6.3, Intersemiotic Additive Relations could be a manner of adding visual Participants, Processes and Circumstances that may not detract from Similarity in Comparative Relations. Temporal and Consequence Relations may assist more in the adherence to this thesis' expanded definition of Mayer's (2009) Signaling Principle, as discussed in 6.6.5. There may be significantly more skills required from learners to connect meanings in these other types of Intersemiotic Relations. The meanings between the visual and linguistic modes may be more complex to interpret in Intersemiotic Temporal, Consequence and Additive Relations than Comparative Relations, where there is a more direct semantic relationship. Further research is required in these areas.

All the videos contain more frequent instances of Similarity than Contrast, and thus their design in terms of the relationship between modes is adequate, potentially even excellent, for literacy development from this resource**.** High scores in the Picture-Word Match tests illustrate that learners have the semiotic awareness skills to connect corresponding language to visuals, thus no significant improvement was found in this skill due to the intervention. However, the learners' proven proficiency to be able to perform this semiotic awareness skill means that they would be able to benefit from videos designed with greater Similarity between modes. The true extent of this effect on their learning, however, cannot be truly determined as the subtitle rate of these videos was too fast, thus the learners could not truly engage with the subtitles in these videos.

The dropping of initial prefixes on nouns and using root verbs without any morphs attached on verbs in the Picture-Word Match test did not impede their ability to connect their comprehension of the word to the corresponding picture. This is unsurprising because versions of these units of meaning exist in isolation in the language, even if there are changes to the form of these units in the video. There are instances when the prefix drops on a noun, and there are instances when the root verbs are used in an imperative form without any affixes in the language.

The 41% of learners that performed poorly in the ORF Word Level 1 pre-test, scored moderately to exceptionally high in the Picture-Word Match test pre-intervention. This suggests that these learners may have used alternative reading strategies at the visual and semantic levels in the Picture-Word Match test, rather than phonological. This is due to their ORF reading results suggesting that they lacked the phonological awareness skills to decode words. The other 59% of the learners that performed poorly in the ORF Word Level 1 test also scored poorly in the Picture-Word Match test. These learners may demonstrate a lack of the skills to deduce or infer the domains of phonological, visual and semantic and connect words and images that are similar in meaning (semiotic awareness). This indicates that, if semiotic awareness skills aside from word recognition for literacy are present in a multimodal literacy context, poor reading skills can be remediated, and this may assist in literacy development. Further research with a battery of various Picture-Word Match tests is required.

While there could have been some guesswork and processes of elimination from words learners knew and words they did not in the Picture-Word Match test, some learners demonstrated that they could connect meanings between modes accurately. They could connect these meanings despite their inability to decode and comprehend linguistic words. This implies skills relating to feature binding and location indexing (see Liao *et al.* 2021) that learners are incorporating to connect the meanings between modes in the Picture-Word Match test, which would be a skill within semiotic awareness (see 1.2.3). This test was timed for 20 minutes, compared to the ORF tests at a minute, however, which implies that restricted time was a variable in decoding and using semiotic awareness skills that would possibly influence comprehension.

Restricted time as a variable when comprehending meaning in the videos, indicates that a subtitle speed that learners could follow may not accommodate semiotic awareness skills to be incorporated in reading, even if learners could not decode or comprehend the full subtitle on

the screen. While this seems to suggest that learners could still comprehend the video and the story in its entirety, Kruger *et al*. (2022) also argue that a fast subtitle speed would cause less time for viewers to connect meanings between the subtitles and the visuals, to understand the story. Furthermore, the speed would result in less time for words to be reread for reinforcement of words and other written linguistic components that are important for literacy development. If decoding is not achieved, comprehension cannot occur. Although accurate decoding does not mean comprehension occurred, it allows for potential comprehension in comparison to an inability to decode. This is supported by Kruger *et al*. (2022), who describe how subtitle speed may influence meaning integration between modes, thus semiotic awareness.

Similarity of Correspondence is the majority instance of Similarity found in the videos and usually contains meanings that have a closer semantic relationship between modes in comparison to other types of Similarity (see 6.6.3). When Similarity is discussed in relation to improved learning based on Mayer's (2009) Principles of Spatial and Temporal Contiguity, this is potentially the type of Similarity discussed. While there are other types of Similarity, namely Hyponymy (Video 2), Polysemy (Video 5) and Collocation (Video 3), the meanings between the visual and linguistic modes may be more complex to interpret for learners in these than in Similarity of Correspondence, where there is a more direct semantic relationship (see 6.6.3). Further research is required in these areas.

Video	Spoken Linguistic Content	Written Linguistic Content	Spoken Language and Visual Similarity	Spoken Language and Visual Contrast	Written Language and Visual Similarity	Written Language and Visual Contrast
Video 1	59%	60%	52%	7%	54%	6%
Video 2	36%	69%	38%	8%	61%	8%
Video 3	33%	45%	30%	3%	39%	6%
Video 4	24%	28%	18%	6%	23%	5%
Video 5	48%	60%	45%	3%	56%	4%

Table 25 Percentages of Spoken and Written Linguistic Content and Each Similarity and Contrast rate in the spoken and written modes per video

In order to compare and contrast the frequency of Similarity and Contrast in proportion to the amount of language present in each of the videos, the percentage values for spoken and written Similarity and Contrast in each of the videos was divided by the total linguistic content

percentages. This was first conducted in the spoken and then in the written mode. The percentages of total spoken linguistic and written linguistic content were calculated based on the total percentages of no language in the Intersemiotic Texture STMs that were annotated and the corresponding percentage that would equal 100%. The percentages were all rounded off to not include decimal places. Table 25 illustrates the videos, the percentage of linguistic content in each video, as well as the percentages within this percentage that are comprised of overall Similarity and Contrast.

Video	Spoken Similarity	Spoken Contrast	Written Similarity	Written Contrast
Video 1	88.14%	11.86%	90.00%	10.00%
Video 2	82.60%	17.40%	88.41%	11.60%
Video 3	90.91%	9.09%	86.67%	13.33%
Video 4	75.00%	25.00%	82.14%	17.86%
Video 5	95.74%	6.38%	93.33%	6.67%

Table 26 Total Values of Similarity and Contrast in proportion to linguistic content

The total values of Similarity and Contrast in proportion to the linguistic content are better represented in Table 26**.** According to Table 26, Video 4 has the lowest rate of Similarity and the highest rate of Contrast in comparison to the other videos in both spoken and written modes. This is followed by Video 3 and Video 2 in the written mode and Videos 1 and 2 in the spoken mode. As seen in the analysis of these separate videos and the metafunctions, it is apparent that there are fewer corresponding Participants, Processes and Circumstances in these modes than in Videos 5 and 1. Video 5 may have the highest Similarity rate and lowest amount of Contrast in both spoken and written modes. It also does not have as many instances of complex grammatical constructions in comparison to the other videos. However, due to the inclusion of spelling mistakes, discrepancies in duration and frequency and less utilisation of the systems in the interpersonal and textual metafunctions, Video 1 may have the most optimal video design for learning. Of all the videos, Video 1. *The Icky Sticky Frog*, employs the most repetition and frequency of systems in addition to a high Similarity rate and thus, may be the most effective for literacy development in terms of Mayer's (2009) principles and optimal learning design.

Ultimately, the subtitle rate in all the videos was too fast for learners to read. This indicates this thesis' expanded definition of Mayer's (2009) Segmenting Principle takes priority above any

other Principles relating to design, aside from Mayer's (2009) Coherence Principle, to which adherence or violation of the other Principles in this context could contribute. This may cause Mayer's (2009) Coherence Principle to be the topmost principle in a hierarchy that includes adapted Coherence, Segmenting, Signaling and Spatial and Temporal Contiguity Principles. A potential hierarchy of principles is established for this context in 7.9.1.

Furthermore, Video 5 also alternates in language between *utitshala* (male teacher) and *utitshalakazi* (female teacher) to refer to the same female teacher in the visual mode. This can cause confusion for learners reading these words. In the ORF Long Passage Reading tests, learners would often read *utitshalakazi* as *utitshala*, indicating there can be a confusion regarding the addition or omission of the feminine suffix. This causes Contrast which may lead to a lack of adherence to Mayer's (2009) Principles of Spatial and Temporal Contiguity. While this rarely occurs in the video in comparison to the high Similarity rate in the video overall, it may nevertheless cause confusion for learners in a way that is not present in the other videos.

6.6.7 Mayer's (2009) Personalization Principle – contextual visuals, Social Distance, Horizontal and Vertical Viewing Perspective and Gaze

Mayer's (2009) Personalization Principle states that, if a text is less formal, learning is easier. This thesis expands this definition for this context to examine personalisation in the sense of a closer relationship between the viewer and the text. Videos that have more Direct Gaze, different Moods and corresponding Modality (e.g., Video 4), the different angles of Horizontal and Vertical Perspectives, more Medium and Close shots and relatable content for their target audience (e.g., Video 5 and Video 2) could assist in the adherence to this principle. In all the videos that include human Participants, there is a lack of Participants of people of colour, as the animations were originally created for English-language videos. The voice-over was later translated into isiXhosa and isiXhosa subtitles added. Considering these visuals are from English-speaking videos, this thesis' expanded definition of Mayer's Personalization Principle could have been adhered to more closely in the visual mode with more relatable visuals for such a target audience.

Due to the instance of Intersemiotic Collocation with the visual lions in Video 3 (see 6.6.3), this video is more useful for family vocabulary than animal vocabulary, as the word lion was

not used, but words associated with family were used. If a visual human family had been illustrated in the visual mode, it would have been better suited to learn this specific set of vocabulary. For general improved learning, however, might have been better to illustrate the family in Video 3 as human, preferably one closer in resemblance to the target culture, than the lion family (Zohrabi et al. 2019). The further into an imaginary space one goes to engage the child, there is a danger that the Personalization Principle adapted from Mayer (2009) and described in this thesis might not be followed as the world becomes less relatable for a child. Video 3 is the closest video in terms of visuals to a particular African-style context with the rondavel hut as a home, but even then, it is mono-dimensional and does not really have enough connections to the African context outside of the visual metafunction and the language being used in the video. Thus, there has to be a balance between creating learner engagement or interest and making the videos relatable to the learners for literacy development. Further research would be required to investigate how video design could best achieve this balance.

An additional factor is that these human Participants in the videos, unrelatable in appearance, are presented in Long shots. This results in less intimacy between viewer and Participants in terms of the interpersonal metafunction, more in Video 2 than in Video 5. This illustrates more potential for this video to be less effective than a video like Video 1. Video 1 has animal Participants but utilises the interpersonal and textual metafunctions more frequently and with more repetition to emphasise these Participants and Processes. This would be an excellent focus for further research with this framework and tests to determine whether animal participants are indeed less effective than human participants. This thesis, however, argues that even if human Participants are more effective for literacy development for children than animal participants. The usage of the systems in the interpersonal and textual metafunctions that highlight Participants and Processes according to Mayer's (2009) Signaling Principle are more frequently used in the videos with animal Participants than with the human Participants. This may cause more engagement between the viewer and the meanings being viewed and comprehended, which may remedy any Social Distance from a non-human Participant. Further research could examine this hypothesis with quantitative tools.

While specific visual Participants, Processes and Circumstances may not align with this thesis' expanded definition of Mayer's (2009) Personalization Principle, the use of Social Distance, Horizontal and Vertical Perspective, as well as Gaze could assist in the adherence of this principle. Social Distance and Horizontal and Vertical Perspective are system choices in the

interpersonal metafunction. Front and Eye Level shots are predominant in the videos, which places viewers in equal relationship with Participants. This may assist in engaging learners' attention, as this might lead them to feel that they are a part of the imaginary space created by the stories. The frequent use of High angles in Video 3 for learners to relate to the moon's perspective further may assist in a closer relationship between viewer and Participants. This may result in better literacy development according to this thesis' expanded definition of Mayer's (2009) Personalization Principle. This is further a possibility with the frequent Close shots in Videos 1 and 2, and Medium shots in the other videos. Close shots would create a closer dynamic between viewer and Participant than Medium shots, thus Video 1, including Direct Gaze and Social Distance may adhere well to this thesis' expanded definition of Mayer's (2009) Personalization Principle, which may improve literacy development. Video 3, including visuals and High angle shots may also adhere well to this thesis' expanded definition of Mayer's (2009) Personalization Principle, which may improve literacy development. However, it adheres less to the other principles in comparison to Video 1, which thus may cause Video 1 to contain the best video design for literacy development of all the videos.

In terms of consistent use of Direct Gaze in the interpersonal metafunction, Video 1 is the best of all the videos, as most of the insects use Direct Gaze, which may elicit sympathy from the audience as they are swallowed by Sele-sele. Video 2 also has instances of this with Sue, however she only gazes directly when with the cats and not the dogs, indicating an imbalance on the animals that Sue loves, thus it may adhere less to this principle in this system than Video 1. The other videos do not contain predominant instances of Direct Gaze. This further indicates Video 1 may utilise this system more efficiently for literacy development than the other videos according to Mayer's (2009) Personalization Principle.

When Indirect Gaze is present, it mostly corresponds with the predominance in offers of information and the Declarative Mood. Similarly, when it relates to Interrogative and Imperative Moods in Video 4, Modality may cause adherence to Mayer's (2009) Personalization Principle, discussed in 3.2.1. The presence of Interrogative and Imperative Mood, discussed in 6.6.5, may highlight information for learners according to Mayer's (2009) Signaling Principle, which may lead to better learner engagement with the video. Due to differing intensities of obligation between modal constructions in the Imperative clause and Interrogative clause in Video 4, there is a conflict of power relations established between the speaker and viewers. In the Interrogative clause, there is less obligation than in the Imperative

clause, thus the learner is made to feel part of the scene with less imposed power restrictions by the speaker (see Zohrabi *et al.* 2019), resulting in a narrower social distance between the learner and the video. In the Imperative clause, the learner has less choice, and the speaker has more power when presenting information, thus widening the social distance between the learner and the video. This causes a dimension of conflict of adherence in this video to Mayer's (2009) Personalization Principle. Based on Mayer's (2009) Principle, more clauses in the Interrogative Mood and low intensity modal constructions of obligation may be better for literacy development. This is, however, beyond the scope of the research in this thesis and more research is required to confirm whether clauses in this Mood and with this type of Modality are more effective in this context.

A DMDA framework that can be used for analysing the effectiveness of SLS videos in promoting literacy in triangulation with literacy test results should include subtitle rate as well as Intersemiotic Texture and SF-MDA systems. These are vital in determining the Similarity and Contrast and relationship between linguistic and visual modes. They provide an understanding of how Mayer's (2009) Principles are being followed in terms of adequate video design for learning. However, the more systems in the metafunctional framework that the video harnesses to create engagement from the audience, the more likely a better learning environment could be created from the videos in order to foster literacy development in this context. This may assist in creating more tools to incorporate in lessons, such as in test design to complete with viewing for better learner engagement and active viewing of the videos, rather than passive viewing (Dal 2012). Passive viewing according to this study has shown to potentially lead to no significant improvement in learner literacy. This suggests the incorporation of Mayer's (2009) Pre-Training Principle to mediate literacy development in this context. Suggestions on better video design are provided in 7.8 and 7.9. The next chapter, which concludes this thesis, summarises further areas of research and responds to the research questions of this thesis.

Chapter 7 – Conclusion

This chapter concludes this thesis by discussing the limitations and opportunities for future research from this thesis. It also provides the answers to research questions, and each of these research questions are discussed below:

1. How is meaning construed in terms of the systems and sub-systems of the mode of spoken language in the videos? (7.1).
2. How is meaning construed in terms of the systems and sub-systems of the mode of written language (subtitles) in the videos? (7.2)
3. How is meaning construed in terms of the systems and sub-systems of the visual mode in the videos? (7.3)
4. What is the relationship between the various modes of language and image within these videos according to the DMDA systems and sub-systems of these modes? (7.4)
5. How does the combination of the modes influence learner literacy levels? (7.5)
6. What effect does repeated exposure to the videos have on literacy scores? (7.6)
7. What are the opportunities and challenges associated with learning word recognition, oral reading fluency and semiotic awareness with these videos? (7.7)
8. What recommendations can be made for the video use and design for future educational videos? (7.8)
9. How can DMDA be used in combination with psycholinguistic approaches to literacy to investigate and foster literacy development in schools? (7.9)

This section concludes in 7.10 with limitations and opportunities for future research and 7.11 concludes with final remarks and reflections of this research.

7.1 How is meaning construed in terms of the systems and sub-systems of the mode of spoken language in the videos?

Spoken language appears less in the videos as it naturally unfolds in all the videos in comparison to its written counterpart. The written language (subtitles) maintains a longer screen time than the spoken language, thus subtitles appear for a longer duration in the videos than spoken language. Potential challenges of intramodal discrepancies between duration and

frequency of systems and their choices are predominantly noted in the mode of spoken language (see 6.6.1). This may result in a greater cognitive load for learners in relation to Mayer's (2009) limited capacity assumption. Video 2 presented no intramodal discrepancies between duration and frequency of systems and their choices. This may result in a greater cognitive capacity for learners in this aspect for this video, which may be more conducive for literacy development. Intramodal discrepancies between duration and frequency were identified within one system and its choices for Video 1, two systems for Videos 3 and 4 and three systems for Video 5. This results in Video 5 having the most discrepancies between systems and their choices intramodally, which may cause learners to experience a greater cognitive load in this aspect of video design for this video than the other videos.

Naturally, one cannot learn literacy skills from spoken language alone, but the presence of the spoken language may be useful to help children decode unknown written words (a skill within word recognition and Oral Reading Fluency) without external assistance by a teacher. Furthermore, it could assist in horizontal reading of subtitles (see Kruger *et al.* 2022). The addition of spoken language would particularly be useful if those words are repeated often within a video. The spoken language had pauses and emphasis, but due to the subtitles following the spoken rate of the narrator, it could have potentially been a factor as to why the subtitle rate was too fast for learners to comprehend (Kruger *et al.* 2022). While the spoken language could potentially have adhered to Mayer's (2009) Segmenting Principle, which highlights that learners learn better with information presented in a segmented way, according to the learners' pace, the written language did not. This indicates that even if the spoken language was at a pace for L1 learners to understand, its written counterpart not being at a pace that the learners could understand could lead to an adherence to Mayer's (2008) Redundancy Principle. In this case, the subtitles were too fast for learners to read, and it would have been better for learning purposes in general if the written text was not present. Naturally, if literacy is to be learned or influenced from the videos, the video designers would have to prioritise catering the language around the subtitles and not the subtitles being an addition to the rate of the spoken language.

Mayer's (2009) Modality Principle posits spoken language and visuals are more effective for learning than written language and visuals, and this is assumed due to the cognitive load of decoding written language and the visuals along the same channel (dual-channel assumption, see 3.2.1). Nevertheless, as mentioned, the written language and mode is crucial for literacy

development, therefore, if a subtitled video (or even a book with pictures) is presented to learners, the Modality Principle in terms of literacy is always violated, and for this context, simply cannot be followed. It can, however, be remedied, if the spoken language is presented in combination with the other two modes (and then, this would violate the Redundancy Principle). A discussion of the exact hierarchy of principles for the purpose of SLS isiXhosa L1 literacy video design is in 7.9.1.

The subtitles generally followed the words and sentences in the spoken language with very little difference, which allowed for less cognitive overload on the viewers' part as there were fewer differences between linguistic units within both modes. Both spoken and written language could thus either align with the visual mode or not without too many differences between them. This is why this thesis ultimately focused on meanings conveyed through the written linguistic mode (subtitles) and how they aligned with the visuals or not, as the spoken language was identical. Since spoken and written modes are identical in terms of the content, I discuss patterns of system choices and their learning opportunities and challenges more in 7.2.

7.2 How is meaning construed in terms of the systems and sub-systems of the mode of written language (subtitles) in the videos?

The longer duration of SLS than spoken language allows for children to have longer exposure to linguistic concepts. Intersemiotic Relations between SLS and visuals are thus present for longer in the videos than the Intersemiotic Relations between spoken language and visuals. As mentioned in section 7.1, this is a benefit for learning in comparison to the shorter amount of time in the videos that the spoken language occurs. It is, however, useless for learning and literacy purposes if the rate of the written language is too fast for the learners to understand, even when the written language is presented for a longer period in the videos overall. As the videos are used for literacy purposes, this requires the written mode to be prioritised in the videos and the rate of SLS supersedes the importance of the speech rate. If the written rate has been adjusted to the spoken rate, one could either slow down the spoken language to adjust the speed of the written rate, or one could simplify the language used. If the spoken rate is decreased at a rate that is far slower than natural speech, this may create boredom and lessen

learner motivation in literacy development. A potential solution would be to simplify the speech for a SLS video that is designed to assist or influence literacy development for isiXhosa.

A major challenge in the act of construing meaning in the written language (subtitles) within the videos is their fast subtitle rate. As mentioned in 7.1, the subtitles in the videos did not adhere to Mayer's (2009) Segmenting Principle and are redundant, which adheres to Mayer's (2009) Redundancy Principle. This was illustrated in Chapter 5 as the videos were not significant for learning in the intervention. Although learning did occur, it was not significantly attributed to the videos. Furthermore, a comparison of the video subtitle rate in Chapter 6 and learners reading rates in Chapter 5 indicated that the learners could not read these subtitles in the videos, as the rate would have been too fast for learners to read. However, as studies have shown (see 2.2), literacy can be improved with SLS and they are not redundant in many other contexts. In this context, they are redundant as they are at a rate that learners could not process and are thus more likely to be extraneous material that presents additional cognitive load for the learner. This means that the subtitles and the videos did not entirely adhere to Mayer's (2009) Coherence Principle, as the more extraneous material is present in the video, the less cognitive capacity the learner has when learning. When developing literacy, however, subtitles should not be extraneous material. The written mode is the core mode for literacy development. This means that the videos should be designed in a way that the subtitles are integral information, scaffolded to learner ability, and the spoken and visual modes are the additional material that aligns with the written mode. Furthermore, it is conducive to literacy development for the videos to maintain their high Similarity rate between modes, to prevent any of the modes, particularly the additional spoken and visual modes, from becoming extraneous information.

Meaning potentials may result in most of the videos assisting in the learning of words and phrases relating to animals. If Video 5, *I Can Go To School*, were to be used for this learning goal, however, it may not be useful for literacy development in this area. There are various ways to learn vocabulary from the videos, but this is subject to the intended literacy development goals in the classroom. This has illustrated that better literacy development is dependent on what meaning potentials the videos have for literacy goals for the classroom, even if a high Similarity rate is present in the videos. A topic of further research would be to investigate exactly how the sub-types of Similarity influence learning, see 6.6.3.

A high frequency of syndromes of Material Processes and their Participants and Circumstances of Location are typical for children's narrative patterns (Guijarro and Sanz 2008). This may be beneficial for literacy development universally, but more research is required to determine whether this is similar specifically in this research context. A high frequency of Circumstance of Location: Place and Circumstance of Location: Time are seen in the videos, see section 6.6.2. This may assist in focusing a child's attention to specific vocabulary within Circumstance of Location: Place and Circumstance of Location: Time, and thus may reinforce literacy development in this manner. "Nalo" as an Existential Process refers to a specific Participant and this acts as a deixis of space (Levinson 2004).

The use of the disjoint (long) form in the videos results in greater word lengths and more linguistically complex words, particularly for verbs, than if the conjoint (short) form were used. This would result in more complex meanings construed in the videos, as well as difficulties in decoding and comprehension for Foundation Phase readers. Therefore, for Foundation Phase learners, it may be recommended that language in stories is presented in conjoint form for easier decoding. Other challenges in linguistic complexity that this could assist in alleviating include clause separations between subtitles, particularly in Video 2, as these clause separations occur due to long clause lengths and a resulting long subtitle length. Non-standard writing conventions in isiXhosa, such as incorrect uses of ellipses in Videos 3, 4 and 5, spelling errors such as in Video 5 and incorrect prefixes used in the title of Video 2 might also negatively influence the perception and comprehension of construed meanings in the written mode.

A greater omission of nouns, particularly in a written disjoint form, would be more natural for Actors and agent nouns of Processes for the isiXhosa language. Videos 3 and 4 have fewer Goals than Actor Participants durationally (see 6.6.1 for analysis discussion), which may be an indication of the videos' translation from English into isiXhosa. Videos 1, 2 and 5 have more Goals than Actor Participants, or in Video 2, more Phenomenon than Sensor Participants durationally, which may be more beneficial for isiXhosa L1 literacy development.

Furthermore, there are also multiple instances of complex verb forms in the videos, i.e., passive, stative, causative, applicative and reciprocal, which are difficult grammatical constructions in any language. Complex ideophones are common Themes and are thus emphasised within the design of the videos. To determine exactly how difficult, it would be for Foundation Phase learners to read complex verb forms and ideophones for their L1 literacy level, further research

is required. There are two questions posed for further research from the repetition of complex grammatical constructions in the videos. The first question is "When in a L1 isiXhosa child's literacy development can they read and comprehend complex verb form constructions and complex ideophone predicates?" This would require a range of tests to provide information on these areas of research. The second question is what ramifications ideophone predicates have for SFL classification. In terms of SFL classification, this type of grammatical construction is classified according to the semantic properties of the ideophone to the right of the -thi verb. Further analysis of this is discussed in 6.6.1.

As seen in 6.6.3, a potential challenge for literacy development is the discrepancy of frequency and duration within modes of meaning systems in the ideational metafunctions. This applies to intermodal linguistic discrepancies and intermodal linguistic-visual discrepancies. Videos 1 and 2 were better in this aspect than Videos 3, 4 and 5, with Video 5 containing both intramodal and intermodal linguistic discrepancies. This may negatively influence construed meanings and their perception in this context. Furthermore, between the linguistic and visual mode, for example, Mental and other Processes in the linguistic mode are not often presented concurrently in the visual Mode. A high Similarity rate of meaning between linguistic and visual modes should be maintained in order to adhere to Mayer's (2009) principles of design to reduce cognitive overload.

Not all the videos fully utilised Mood and corresponding Modality, specifically Interrogative, and Themes to their full potential to highlight specific clauses and smaller structures for L1 literacy development. Themes and other metafunctional systems assist in emphasising meaning choices within the ideational metafunction. Modality may further create complexity in construed meanings; however, further research is required to examine whether this type of meaning system is too complex for a Foundation Phase context.

7.3 How is meaning construed in terms of the systems and sub-systems of the visual mode in the videos?

Most visual Participants and Processes in the ideational metafunction for videos are Actors within Actions and often Goals that assist in creating the high Similarity rate in meaning between Material Processes and Actor and Goal Participants in the language. This allows the

videos to adhere to Principles of Spatial and Temporal Contiguity and the Coherence Principle. Most of the videos have a higher frequency of these types of Processes and Participants in the video, particularly videos like Video 1, however even in Video 1 there is a Mental Process in the visual Mode but a Material Process in the language. This is an instance of Contrast, which may cause such videos to not adhere to Mayer's (2009) Principles of Coherence, Temporal and Spatial Contiguity. More representation of Mental linguistic Processes and other Processes such as Existential or Verbal Processes in the visual mode may have been more conducive for Similarity of meanings and literacy development in this context.

The video that used the most meaning potentials within the interpersonal and textual metafunctions to emphasise particular relevant Participants, Processes and Circumstances was Video 1, particularly with *USele-sele* in terms of Zoom effects, Direct Gaze, Salience and Close shots often Centred when not presented as Given information. Often in videos like Video 2 or Video 5, there is no use of Salience and less effective use of Zoom Effects, Gaze and Information Value to highlight particular Processes and Participants.

Similarity of Correspondence is the most frequent instance of Similarity in the system of Comparative Relations found in the videos. Similarity of Correspondence usually contains meanings that have a closer semantic relationship between modes in comparison to other types of Similarity (see 6.6.3). The meanings between the visual and linguistic modes are more complex to interpret in Similarity of Hyponymy, Polysemy and Collocation than in Correspondence, where there is a more direct semantic relationship. Furthermore, it may be easier to interpret meanings in the systems of Comparative Relations for the purposes of L1 isiXhosa literacy development than in Intersemiotic Temporal and Additive Relations. Further research is required in these areas mentioned in 6.6.3. Video 3 could have been useful to learn animal vocabulary if the word lion was used. Since the vocabulary for family is present in this video, if a human family had been illustrated in the visual mode, it would have better suited to learn this specific set of vocabulary. Furthermore, the closer the video can resemble the target culture visually, the higher the chances of better literacy development, as this could also create a type of adherence to Mayer's (2009) Personalization Principle (see 6.6.7). An example in Video 2 is discussed in 6.2. If the child could relate to the videos more due to their familiarity with visual components, there is a greater possibility an improvement in learning.

The Personalization Principle states the less formal a resource is, the easier it is to learn from the resource, i.e. the more informal a video is presented, the more suitable it is for learning. This thesis expands this principle to define "informal" in this context as being closer to the culture of the viewer, or rather include concepts and language that are more familiar to the viewer. This includes both the incorporation of human Participants over animal Participants, but also more incorporation of visuals that are closer to the cultures of the target audience. Animal Participants may not align with Mayer's (2009) Personalization Principle, as they are not human and might be less relatable to the learner. These aspects of learning and the presentation of visuals that are closer to the target culture of the learner require further research in this context.

7.4 What is the relationship between the various modes of language and image within these videos according to the DMDA systems and sub-systems of these modes?

Video 1 has the highest rate of Correspondence between visuals and language. There is also the instance where, in Video 4, Actor Participants and Action Processes and Material Processes do not necessarily correspond with the same Actors, which causes less Correspondence than if they did. The most frequent Participants do not always correspond between modes, again due to the noun-omitting nature of the isiXhosa language.

Salience as a system has often been omitted in the videos, where it and other systems within the interpersonal and textual systems could be used to further emphasise meanings conveyed through the ideational metafunction. Of the videos analysed, Video 1 is one of the few videos that use both Salience and Themes alongside Given information or Centre in the visual Mode to emphasise specific integral Participants to the story (see Chapter 6, particularly 6.1). This highlights how through the ideational metafunction, content is conveyed that is useful for L1 isiXhosa literacy development while the interpersonal and textual metafunctions may potentially be used to emphasise specific structures in that content. These may need to be aligned between linguistic and visual Modes for ideal learning conditions (see 6.6.5).

The presence of Intersemiotic Additive Relations allows for the visual mode to provide information that the language does not, however its relationship of meanings between modes

is not as clear or similar between modes as Similarity in Intersemiotic Comparative Relations. Nevertheless, it may be more beneficial for literacy development to add information (e.g., Participants) in the visual mode to add more depth of description to information in the linguistic mode, than if information was added in the visual mode that had no correspondence in the linguistic mode or vice versa. All videos have a high Similarity Rate between visual and linguistic modes, although in various ways, as discussed in 6.6.3. (e.g., Correspondence for Video 1, Hyponymy in Video 2, Collocation for Video 3 and Polysemy in Video 5). More research would determine the best for learning, but Correspondence is the most predominant in the videos and it also is the closest in a type of Similarity of meaning that would potentially cause fewer alternative interpretations that may result in less confusion for learners. This relates to Mayer's (2009) Principles of Spatial and Temporal Contiguity, as well as the model of multimodal-integrated language framework for multimodal reading (Liao *et al.* 2021).

Nevertheless, the conducive Intersemiotic Relations for literacy development almost become extraneous in this thesis, due to the subtitle rate being too fast for learners to decode and comprehend and for literacy this is integral (discussed in Chapter 5 and 6.6.4). Therefore, difficulties in horizontal reading may create difficulties in vertical reading and semiotic awareness (Kruger *et al.* 2022). Thus, in a literacy development context, Mayer's (2009) Segmenting Principle is more integral to adhere to than Principles of Spatial and Temporal Contiguity for conducive learning. A hierarchy of Mayer's (2009) Principles of video design for this context is discussed in 7.9.

7.5 How does the combination of the modes influence learner literacy levels?

Since learners already had a high basic semiotic awareness ability to begin with, based on the ceiling effects present in the Picture-Word Match test, the combination of modes did not necessarily improve their skills in this area. It could have had the potential to improve their decoding and comprehension of subtitles, had the subtitles been at a pace they could follow. The speed of subtitles in multimodal texts may cause difficulties for meaning integration between modes to create an accurate comprehension of the text (Liao *et al.* 2021; Kruger *et al.* 2022).

It was identified that the written mode is central for literacy development and if it is compromised, the three modes together are redundant, potentially causing cognitive overload and hindering learning. If Mayer's (2009) Segmenting Principle is followed in the written mode, it could allow for Mayer's (2009) Redundancy Principle to not be followed, which would create learning potential for a combination of spoken and written linguistic modes and visual modes in one text. The combination between the written, spoken and visuals could have been drawn upon in test design for a deeper understanding of exactly how the high Similarity rate between modes would assist learners in furthering their semiotic awareness. This would present a good understanding of mapping meaning between modes that would delve further into the Picture-Word Match test results and other areas of semiotic awareness. This is an area for further study and research. This was not researched further in this thesis, due to the subtitle rate and complexity being a barrier for children to learn effectively from the videos and probably impeding learning from Intersemiotic Relations. This indicates that the importance of modes, in this context, the written one, can impede any benefits of the presentation of modes as a whole and the meanings that they construe together. A change to one semiotic system, either in one mode or the multimodal text, could create a domino or ripple effect across the entire "meta-system" (O'Halloran 2023:174), or a "butterfly effect" (Kothari 2002:55) in the perception and literacy development of children.

7.6 What effect does repeated exposure to the videos have on literacy scores?

Learning potentials could be realised from the DMDA but the subtitle rate (see 7.5) became a barrier. This caused children to not learn literacy skills from the videos before meanings could be explored in-depth with the learners as in the analysis. While literacy scores improved in ORF at the word and longer reading passage levels during the intervention, the videos alone made no significant contribution to this improvement. The average reading scores of the children in this thesis did not come close to the objective subtitle rate in the videos, thus this thesis concludes that learners may not have been able to decode subtitles sufficiently to improve their literacy skills from the videos.

There is, however, a need for further research on how videos integrated in a specific syllabus for literacy improvement, i.e., as an additional resource integrated in the curriculum and not as a stand-alone resource, could improve literacy scores and in specific literacy skills. This is because passive viewing of videos without learning outcomes may have also resulted in the videos not making as great an impact in literacy scores as they could have. Furthermore, the learners' lack of prior subtitle exposure and the lack of cultural relevance to South Africa of the videos' content may also have influenced the challenges learners face when engaging with this semiotic resource.

In addition, there was no significant difference found between pre- and post-tests for Picture-Word Match, which mostly had ceiling effects. This illustrated the type of semiotic awareness, which the children required to complete this type of test was already at a high enough level initially to relate meanings between modes when presented with them in the videos. Since the visuals there were dynamic and animated, further tests that accommodate that movement would have to be used. This could then investigate how these extra semiotic resources of rate, speed and movement influence semiotic awareness in children as they identify and compare meanings in the comprehension of dynamic texts. In addition, the nuances of semiotic awareness as a skill in this context need to be explored and researched further.

7.7 What are the opportunities and challenges associated with learning word recognition, oral reading fluency and semiotic awareness with these videos?

Learner reading rates in comparison to the subtitle rates of the videos indicated that they could not follow the subtitles in the videos themselves. This may be the primary reason, along with challenges in linguistic complexity of the subtitles, that resulted in video exposure having no significant difference for automaticity and faster word recognition and ORF scores. When presented with static images from the videos and related lexical units on a page, children could relate these meanings intermodally to one another. This indicates that the isiXhosa L1 learners in this study have a type of advanced semiotic awareness (see 1.2.3). They are able to use various modes and understand the semiotic resources with which they are presented and compare these meanings in a way that indicates their perception of Similarity and Contrast of

meanings between modes. Additional tests would need to be designed and undertaken to understand the animated nature of the visuals in the video in comparison to the static images used in the text. Nevertheless, a proposition of this research is that when there is ambiguity of meaning in one mode, we use other semiotic resources to navigate through the web of meanings to comprehend multimodal texts. Without adequate understanding of modes and the semiotic resources they employ to create different layers of meaning, it can lead to a text for a particular context either being misinterpreted or being rendered completely illegible, incomprehensible, and not as useful as it could be for the purposes of furthering literacy in the language. This was demonstrated by the rate of the subtitles (an affordance of a semiotic resource within multiple modes) being potentially too fast for learners to follow.

This thesis primarily examined word recognition, Oral Reading Fluency skills and semiotic awareness skills in terms of grasping similar or contrasting meanings presented between modes. These skills of course are dependent on other awarenesses outlined in 2.1 for literacy in all languages including isiXhosa, e.g., phonetic and phonological awareness skills, morphological awareness skills, orthographic awareness skills, and vocabulary. Types of vocabulary that could have been learned through these videos are mostly animal words and words relating to home and school. Other vocabulary that could have been enhanced or learned through these videos include Action Processes and words relating to Circumstances of Location: Place and Circumstances of Location: Time. It is encouraged to design videos that align with learning objectives for specific learning goals to be achieved in these areas. Certain vocabulary content incorporated in the video design will naturally assist learners with specific vocabulary and reading learning goals. These videos were not graded readers and an opportunity to learn with videos like this would be to have graded videos designed for specific learning outcomes, which would have the potential to make them more effective as resources in this context. This would further allow active engagement around the videos with other resources for the learner to fully learn these concepts, with the videos as an additional resource in the curriculum rather than necessarily a primary and stand-alone resource.

In multimodal texts used for literacy and learning, a delicate balance needs to be taken when designing pedagogical resources as with too much information at this level, there is a risk of cognitive load and Mayer's (2009) Redundancy Principle being followed. With the subtitle rate of these videos being too fast, the spoken, written and visual modes were not effectively combined for learning in this context. The use of the isiXhosa language with linguistic

complexities in the videos may be potentially too challenging for learners at this level to fully grasp or cause additional cognitive strain in the process of decoding and comprehending the written language. Challenges of linguistic complexity are described in relation to disjoint form, non-standardised conventions of written isiXhosa, lack of noun omission, complex verb forms such as passive, stative, applicative, causative and reciprocal, frequency and duration (intramodal and intermodal linguistic discrepancies), system choice co-occurrences in the subtitles, ideophone constructions, clause separation in subtitles and modality. These in combination with any additional emphasis on extraneous information could lead to Mayer's (2009) Coherence Principle being violated due to the cognitive load required, causing challenges in literacy development. However, if the language is oversimplified, with simple sentences or even words with corresponding Participants and Processes and Circumstances on the screen, it runs the risk of not being as engaging for learners and could be less motivating. This could also impact its opportunity as a literacy development resource. Thus, when videos for this context are designed, a delicate balance needs to be made between ensuring there is enough repetition, simplicity, and Similarity for learning, but also ensuring the text is still engaging and not boring for learners. More instances of emphasis from intersemiotic systems in the interpersonal and textual metafunctions in video design could be used to ensure learner engagement, as well as assist in establishing an emphasis on Similarity. This may improve literacy development.

7.8 What recommendations can be made for the design and use of future educational videos in line with the policies of the DBE?

Videos for this level of learning in isiXhosa (Grade 2 and Grade 3) for L1 literacy objectives require simpler language to be used, i.e., conjoint form and non-standardised conventions of written isiXhosa. However, the EGRAs take this into consideration, while the CAPS documents do not consider difficulties in disjoint form for Foundation Phase learners. If incorporating SLS videos for L1 isiXhosa literacy development, the Department of Basic Education (DBE) would have to take these forms into consideration. Linguistic complexities such as noun omission, complex verb forms, ideophones, discrepancies in frequency and duration, co-occurrences of system choices, and modality should also be considered in

designing videos for these grades. These complexities may further assist in the challenge of alleviating clause separations between subtitles, another potential challenge for learners' literacy development that may result in cognitive overload. The videos could be designed in a style of animated flashcards with more systems from the interpersonal and textual metafunctions to assist in focusing on specific language, visuals and concepts conveyed in the ideational metafunction. As outlined in 7.7, this needs to be balanced to ensure learner engagement and that the design is not too simple and causes learners to be bored and less motivated while developing literacy skills from them. Future video design will need to be tested in the classroom to gauge what this balance would look like for this context.

The average learner at the research site met the reading benchmark of 20 WCPM for Grade 2, but the average Grade 3 learner could not reach the 35 WCPM benchmark for Grade 3. Furthermore, there were many learners who could not read a single word. The maximum subtitle rate per minute should be 41 WPM for Grade 2 Foundation Phase learners and 50 WPM for Grade 3 Foundation Phase learners as this seems to be the maximum reading rates for each grade. Even if the silent reading rate was double the oral reading rate, then a 41 WPM and 50 WPM subtitle rate would account for a learner oral reading rate that is on average 20 WCPM and 30 WCPM. A limitation of this study is the inability to measure silent reading rate, as this is entirely independent on the learner self-reporting this information, which is not necessarily accurate. Furthermore, eye-tracking does not necessarily state whether decoding took place, even if it tracks the movement of the eye along the subtitle. This requires further testing on subtitles that follow this rate to confirm. The objective subtitle rate should be catered towards these ranges with further tests done to explore the ideal rate for subtitles in the language for these grades.

The "five finger strategy" in CAPS mentioned in 1.2.3 provides five reading strategies for new and challenging words based on the fingers of a child's hand (DBE 2011). This strategy indicates that children should use meanings conveyed from other modes, such as a static image, to help them understand meanings conveyed from the written linguistic mode that they do not understand. This is a strategy that should be used when teaching with video, but it will only work if the subtitles and animation are set at the pace of the learner and adhere to Mayer's (2009) Segmenting Principle. A method to alleviate the introduction of complex videos would be to utilise Mayer's (2009) Pre-Training Principle, which would also ensure active engagement with the videos and not passive viewing by the learners.

When examining literacy using this resource and potentially implementing it into any curriculum for L1 isiXhosa teaching and learning in schools, the videos would need to be adequately designed as described above where Mayer's (2009) Redundancy Principle is violated, and all other principles are followed. Grade learning goals, objectives, aims, and requirements are necessary to incorporate when designing videos for specific areas of the curriculum to ensure those areas are focused on adequately. A video such as Video 1, *The Icky Sticky Frog*, that has repeated semiotic choices and utilises a wide range of systems within the interpersonal and textual metafunction to emphasise specific Participants, Processes and Circumstances would be a good video to model in video design. However, this video would still require simpler language for the Foundation Phase and less linguistic content for there to be an adequate subtitle rate for learners to follow.

Ultimately, these videos should be designed with the assistance of an L1 isiXhosa speaker and be scripted in isiXhosa rather than translated. It requires an isiXhosa L1 designer who would ideally also be an educator or have education experience. If one person possessing all these qualifications cannot be found, then definitely a trans-disciplinary team is required. Videos with human Participants should be in a South African context to bring the context of videos closer to its intended audience. This may improve literacy development. It would also assist in social transformation. This type of work would also eventually lead to not only a normalisation of this type of video, but also lead to subtitling companies for the language that would then employ translators and a normalised isiXhosa subtitle standard. This would create a plethora of job opportunities for the language and increase the power of the isiXhosa language, but also increase the likelihood for such videos to assist learners in this context. This is not, however, an instant process and would require multiple videos to be designed, these videos to be tested and particularly in terms of the curriculum before they could be useful and benefit learners in their literacy.

SLS isiXhosa videos, such as the ones examined in this study, may be useful as an additional resource for teachers to use to teach the skills outlined in the CAPS documents and incorporate within their lessons. This could only occur if the rate of the written language is at a rate that the learners could follow. There is potential for these videos to assist in fulfilling literacy objectives for each grade. However, as described above, this would require SLS isiXhosa videos to be analysed in greater detail to determine with which of the specific skills in the

CAPS policies and school curriculum they would align. This research has started this process and further research is required.

7.9 How can DMDA be used in combination with psycholinguistic approaches to literacy to investigate and foster literacy development in schools?

As observed with this literacy intervention and video analysis, the results can give us insight into if and how this resource can assist in this context, particularly for young children who may not be able to provide in-depth feedback into their own literacy development. The ORF Long Passage and word recognition test scores, as well as Picture-Word Match test scores are triangulated with the Comparative Relations Similarity and Contrast rate. These methods are used to illustrate how meanings are similar and contrast in the different modes, as well as the objective subtitle rate score to determine its legibility to learners. These literacy tests were implemented in this study, as, due to the nature of not knowing for sure whether learners could decode effectively if they read silently, the ORF score would be different to a silent reading score. More studies are required in other research sites in this context, as well as with additional and alternative test methods to confirm these findings. A Picture-Word Match test that is more dynamic and takes other variables into account such as animation would also be beneficial in fully understanding how learners navigate the meanings in terms of Similarity and Contrast. A potential additional test, for example, would include learners being presented with similar images and corresponding language and observing if they could distinguish between explicit and implicit meanings. This type of research could also be done with eye-tracking methodology at this level to further confirm what Matthew (2020) found in a different context, as well as the findings in this thesis.

The nature of multimodal analysis is that systems are flexible and context dependent. While it may be difficult to use a universal set of systems for analysis in all contexts, it is nevertheless possible that Intersemiotic Comparative Relations and Subtitle Rate, and the written mode are extremely important systems in a literacy context such as this one. A multimedia theory of learning, such as Mayer's (2009) principles, should be utilised as a lens to connect quantitative literacy tests and scores with DMDA. On its own, DMDA only examines how meaning is

portrayed and not how it influences learning, which is an area that requires additional theory and systems of analysis.

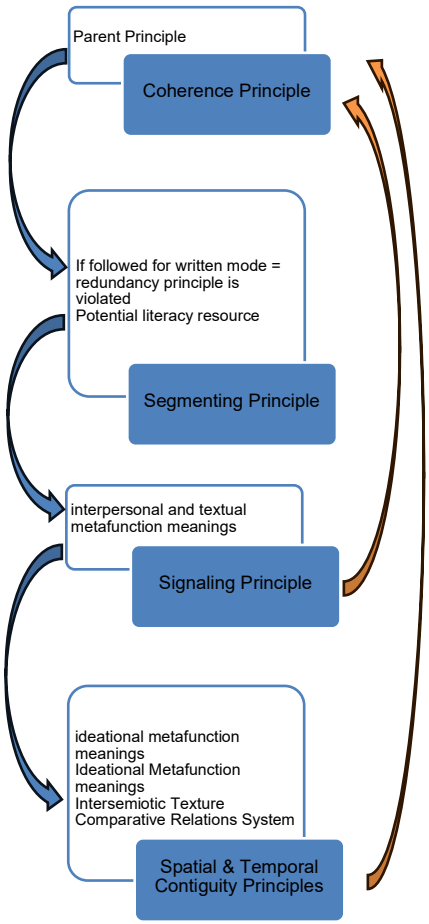
Fine-grained phonological systems in DMDA could be used to further investigate texts for phonological awareness. Other systems in DMDA could be tied into examining more metalinguistic skills for a finer-grained analysis between psycholinguistic literacy skills and the way meanings are presented in text. Since subtitle rate is an essential system in this context, other factors of subtitles, such as colour and size, could be incorporated into a system for the analysis of subtitles that further align to the context and language.

Multimodal Analysis Video as a software was innovative for the context as it could represent State Transition Machine diagrams from annotated texts. There were, however, various bugs that made annotation challenging, including having to create duplicate systems when annotations overlapped. This was due to the STMs not incorporating overlapping annotated data in their final percentages. More software of this kind is needed to fully allow for easier annotation and various graph representations that also incorporate editable systems such as subtitle rate calculations. In this study, subtitle rate had to be calculated separately. Furthermore, for these kinds of analyses that require specific time measurements, the annotation timestamps must be fine-grained, which would ensure more precision in annotation. Similar software to *Multimodal Analysis Video* or adaptations thereof could be more inclusive in terms of the types of systems incorporated, which allow for more trans-disciplinary types of analysis to be conducted.

This thesis has illustrated that Mayer's (2009) principles have a hierarchy in this context. The following section describes this hierarchy in more detail.

7.9.1 Mayer's (2009) Principles of Learning for the isiXhosa L1 SLS video literacy education context

The hierarchy of Mayer's (2009) basic principles are suggested for this context based on the findings of this study. The proposed hierarchy is as follows: Coherence Principle, and if the Segmenting Principle occurs, then the Redundancy Principle is violated. If the Redundancy Principle is violated, subsequently Signaling, Spatial and Temporal Contiguity Principles could be realised (see Figure 99).



Hierarchy of Mayer's (2009) Principles of isiXhosa L1 SLS Literacy Development.

Figure 99 Hierarchy of Mayer's (2009) Principles of isiXhosa L1 SLS Literacy Development

Videos use a combination of words and pictures, thus adhering to Mayer's (2009) Multimedia Principle, which may improve literacy development, but Mayer's (2009) Modality Principle indicates that spoken language and pictures are better than written text and graphics. Print literacy development requires written text, therefore, while this principle could assist in learning information, the use of written text for literacy development is non-negotiable. Thus, for all literacy resources, Mayer's (2009) Modality Principle will be violated unless the spoken language is used simultaneously and as a complement to the written language.

If Mayer's (2009) Segmenting Principle for the written text is followed, then Mayer's (2009) Redundancy Principle is violated and a multimodal text like this could be useful for literacy education. If Mayer's (2009) Segmenting Principle is violated, Mayer's (2009) Redundancy and Coherence Principles are not violated, which also means the videos cannot be used for

literacy education in this context. Mayer's (2009) Coherence Principle would also be violated in this context, as the written mode would qualify as extraneous material, which may cause the entire text to be rendered useless for the learning goals.

Mayer's (2009) Coherence Principle, along with Spatial and Temporal Contiguity Principles could also be violated if there is a high frequency of Contrast. However, the findings suggest that Mayer's (2009) Segmenting Principle, and consequently Redundancy Principle, have a rank higher than Mayer's (2009) Coherence Principle. Both Segmenting Principle and Redundancy Principle may assist in alignment of Mayer's (2009) Coherence Principle, as well as Spatial and Temporal Contiguity Principles. Therefore, the suggested parent of the hierarchy structure is Mayer's (2009) Coherence Principle (see Figure 99), followed by Segmenting Principle, Redundancy Principle, Signaling Principle, and Spatial and Temporal Contiguity Principles. Mayer's (2009) Signaling, and Spatial and Temporal Contiguity Principles, when not followed, may create a video design that produces extraneous material that could create challenges in literacy development. Mayer's (2009) Segmenting Principle, however, would simply cause the information communicated in modes to be illegible to learners. Thus, it is suggested that Mayer's (2009) Signaling, Spatial and Temporal Contiguity Principles may influence the adherence to Mayer's (2009) Coherence Principle in video design.

Adherence to Mayer's (2009) Signaling Principle may be determined by analysing the interpersonal and textual metafunctions. This principle could assist in adherence to Mayer's (2009) Coherence Principle when more extraneous material is present and when Comparative Relations do not align. It may assist in directing the attention of viewers to specific Participants, Processes and Circumstances in the video. Consequently, it may assist in establishing literacy goals in this context, and may influence learner attention, motivation and engagement. Based on the findings in this study, the rate of information presented could have caused the inability of learners to comprehend the content in the videos. It suggests that Mayer's (2009) Signaling Principle is potentially higher in priority for video design in this context than the alignment of meanings within Intersemiotic Parallel Structures. This may further include meanings conveyed by the ideational metafunction between modes. Meanings conveyed by the interpersonal and textual metafunction may be used to direct attention rather than necessarily conveying new learning content to the ideational metafunction. Less highlighting may cause videos to be less effective in their conveyance of ideational meaning, which may impede the

high Similarity rate in Comparative Relations and Mayer's (2009) Principles of Spatial and Temporal Contiguity.

Mayer's (2009) Personalization Principle may also contribute to the different realisations of metafunctions, for example, with the visual Participant presented (ideational metafunction) and Gaze (interpersonal metafunction). Further research could illustrate its potential position in the hierarchy and investigate this principle further in terms of literacy education. As discussed in 7.3, this thesis has illustrated how this principle could be expanded for this context to include the presentation of visuals that are more familiar to the culture of the target audience. This was both in the presentation of the type of Participant (animal or human) and in this context, presenting visuals that are closely aligned to amaXhosa culture for children that speak isiXhosa as their L1. Mayer's (2009) Pre-Training Principle would also need further research for such a resource, but this principle and Mayer's (2009) Personalization Principle seem to be additional principles that are more independent than the others. They may assist in literacy development in this context in terms of the principles in the hierarchy, but this was beyond the scope of this research.

Therefore, in terms of the meanings related to literacy development, content is conveyed within the ideational metafunction, but since it is presented between modes, it is visible in Parallel and Cohesive structures in Comparative Relations. This is a type of Relations within Intersemiotic Texture and other types of meaning relations and their influence in L1 isiXhosa literacy development in this context requires further research. Analysing these systems for videos as literacy resources may be useful in determining what meanings are conveyed to understand the potentials for literacy development. As mentioned in 7.8, this thesis informs teachers of potential benefits and challenges in using the videos for isiXhosa L1 literacy development. It also informs potential curriculum design and how learners may develop literacy skills with this resource in alignment with DBE policies. Some potential opportunities for literacy development in this context with these videos include learning specific vocabulary and grammatical constructions. Systems within the interpersonal and textual metafunction allow for emphasis on these specific constructions and lexical units that adheres to Mayer's (2009) Signaling Principle. This may enhance literacy development. It could also potentially aid in increasing attention, motivation, and semiotic awareness skills. These could be realised if the subtitle rate is at a pace which the learner could follow.

In this thesis, I have redefined the Coherence, Signaling, Segmenting and Personalization Principles for this context based on the findings. According to Mayer (2009), the Coherence Principle states the more extraneous material (unnecessary information for the instructional goal) present in multimedia, the less cognitive capacity a learner has for integral information. This may relate to structure and content Similarity between modes and metafunctional systems, as content that is extraneous and contrasts to construed meaning within a particular mode may pose a greater cognitive load for the learner. This depends on the learning aims of the video. Thus, in this context, the Coherence Principle would state that the more Contrast and Additive meanings present in the videos that are unnecessary information for the instructional goal, the less cognitive capacity a learner may have for integral information. It is therefore potentially more difficult for learners to focus on integral information for literacy development.

The Signaling Principle states that highlighting essential words guide learners when processing knowledge and lead to more efficient learning (Mayer 2009). Highlighting can be in vocal emphasis and textual headings (Mayer 2009), therefore is not restricted to the auditory channel but is intermodal. In terms of DMDA, this may be visible in Salience, Framing and Information Value, which focus on highlighting specific information in the visual mode. In the linguistic mode, this may be visible in Mood and Theme. Thus, a new definition of this principle for this context would be that the Signaling Principle may utilise meanings conveyed by multiple systems of the interpersonal and textual metafunctions to highlight specific and essential units of meaning. These units of meaning could be groups or clauses, which may lead to more efficient literacy development in this context.

The Segmenting Principle stipulates that narrated animation should be at the learner's pace (Mayer 2009). This thesis expands on this principle in this context. For L1 isiXhosa literacy development, the Segmenting Principle should not be restricted to narrated animation, but to further include the subtitle rate. This, as well as the animation, should be at the pace of the learner for potentially better literacy development with this resource. Thus, the Segmenting Principle in this context should include the rate of all types of narrated semiotic resources in the visual mode, as well as written language such as subtitles. This rate of all types of narrated semiotic resources in the visual mode should be at the learner's pace.

The Personalization Principle states that, if a text is less formal, learning is easier (Mayer 2009). The videos in this research are translations of storybooks and arguably are more formal than informal, but also are not entirely formal texts. This thesis further postulates that in addition to

the above definition, the Personalization Principle should be expanded for this context to include that if a text is closer to a learner's socio-cultural context, then easier literacy development may occur. This is a type of personalisation of the video that could allow a video to be more relatable to the learner, thus potentially allowing for better and more efficient literacy development to occur. There has to be a balance between creating learner engagement or interest and making the videos relatable to the learners for literacy development with this principle. Further research would be required to investigate how video design could best achieve this balance.

7.10 Limitations and opportunities for future research

A large limitation to this research is my positionality as a white, fourth-language speaker of the language. Despite my years of study of the language and its linguistics, I do not possess the intimate understanding nor the cultural background to understand the meanings within this context that a L1 speaker would have brought to this research. I had a team of isiXhosa L1 speakers to advise and assist in data collection and analysis, and isiXhosa L1 linguist advisors for the translation and annotation phases of this project to ensure linguistic accuracy. There is, however, a need for primary isiXhosa L1 researchers to expand on this research and bring to it additional knowledge and further insight, particular in the design of a video and tests. This would show the positive learning benefits for SLS isiXhosa videos seen in other studies and contexts like Kothari *et al.* (2002).

Despite the lack of a significant difference in this study, there is potential for SLS videos to assist literacy development if learners are exposed to subtitles early and if the subtitles are at a rate that the learners could follow. Further research could highlight optimal learning conditions and best video design required for successful literacy development with this resource. This research acts as a guide for the next generation of researchers and a gauntlet for isiXhosa L1 researchers to pick up and carry forward to truly highlight the impact such a resource could have for this language and for literacy in L1 isiXhosa and other African languages. There is very little understanding of isiXhosa subtitles as there is no industry for their creation and more research in this area could assist in literacy and even create employment for translators and transliterators. This research has added to prior research in the field of African language subtitles and has shed some light on how this operates in the context of L1 isiXhosa Foundation

Phase learners. It suggests how videos could be better designed to be better suited to this context.

Only one research site was used in this thesis and while the sample was representative of the site, it was small and only a single site. More research is required on multiple sites to see if this phenomenon is not isolated to this one site, as well as other languages to see if there are similar challenges. Further research is needed on Intermediate Phase and Senior Phase and whether they could read subtitles at this speed and the subtitles of these specific videos. When simpler videos are designed, they should also be researched to determine whether simpler videos would be less effective as children age and progress in reading ability. More sites would also lead to stronger statistical evidence to further support the findings in this thesis.

Different tests designed around semiotic awareness to gauge the nuances of this skill are required, as the Picture-Word Match test in this study reached ceiling effects. This test thus demonstrated that learners have basic semiotic awareness skills. In addition, they demonstrated a type of advanced semiotic awareness skill for static basic concepts and actions when presented with similar meanings in multiple modes. Ultimately, this research serves as a guide to a simpler video design. The reading tests could have incorporated video content. However, there was a concern the exposure to the vocabulary in reading tests before the intervention would be a factor that could make understanding of the video exposure alone more challenging to determine.

Passive viewing of videos (Dal 2012) in this case and context is not effective for literacy development in this context and there a few reasons why that could be:

1. The speed of the objective subtitle rate is too fast for learners to decode.
2. The subtitles are too linguistically complex for Foundation Phase learners to decode.
3. There is a lack of subtitle exposure and culture of reading subtitles in the country for this language.
4. There is a lack of prior exposure to the videos and this kind of video in this context.
5. The length of the subtitles and the challenges of too many shots per subtitle could often create difficulty for literacy development with this resource.
6. Literacy research has not confirmed whether learners are comfortable with the complexity of language constructions in the subtitles, such as verb forms and

ideophones. These videos are not graded, and this should be considered when videos are designed.

Due to the learners being unable to read the subtitles, and ceiling effects on the Picture-Word Match test, learner feedback on how the meanings conveyed through the metafunctions and how they could assist literacy development would have been beyond the scope of this research. Only with linguistically simplified videos could this be further explored, as well as the principles in this context from a mixed-method approach. Other systems in DMDA could be tied into examining more metalinguistic skills for finer-grained analysis between psycholinguistic literacy skills and the way in which meanings are presented in text.

7.11 Final Remarks: Towards finding a "butterfly for literacy"[24] for isiXhosa

While the field of multimodality moves more towards technology and computation, this project has shown that context is key in this area and technology is not always a solution for every problem and every context. There is a lack of African language subtitle culture in South Africa, and a lack of knowledge on how subtitles work for isiXhosa. The question remains if a movement is needed by the people to create subtitles in the language and normalise it for this to be more effective in the language as a literacy resource. "The butterfly effect" for literacy in this context (Kothari *et al.* 2002) unfortunately has not extended to South Africa yet, and there needs to be discussions of why that is the case. We should provide more technological resources to empower this language on a global stage, where predominant global languages have been subtitled and have industries in subtitling. Yet, as we have also observed by examining isiXhosa subtitles, technology itself is not a solution and a miracle of teaching and learning, as often observed in early research of FL learning contexts (see Schafli 2020). This research, however, proposes that we cannot assume subtitles in these languages will never be useful, rather it raises a question of how we create opportunities for development in the field of isiXhosa subtitles for them to be useful. IsiXhosa deserves subtitles like many global languages to create opportunities for access, and we can only arrive at this point by increasing subtitle industries. These are influenced by each other, and it requires vast amounts of effort to begin this process,

[24]Kothari *et al.* 2002

not only in a creation of the subtitle industry for translators that would create employment and access opportunities, but also further insight as to best practices for learning how to read in isiXhosa with this resource. It would also create further insight into how learners decode and comprehend complex constructions in isiXhosa and work towards a scaffolded curriculum for the language that, when designed simply enough linguistically, could still include videos like this.

This research has described a potential hierarchy of Mayer's (2009) principles in context. It shows us that one mode, that is the written linguistic mode, may disrupt any learning potential from other modes in this context, even as a multimodal text, thus creating a domino or ripple effect, where a change to one semiotic system can impact the entire "meta-system" (O'Halloran 2023:174). This hierarchy is context dependent, as the written language may not always be a vital mode for a context, but certainly in the case of literacy development for a language, whether this be subtitles or static written text, it is the primary mode. Depending on how the resource is used in literacy development, whether for vocabulary, comprehension or other specific areas, video could be used as an additional resource instead of a primary resource. Longer interventions are needed and with more schools to determine whether this would be effective, along with more tests which require L1 language insight. The end of this project serves as a call for L1 isiXhosa researchers to further expand and extend this research, finding a South African "butterfly for literacy" (Kothari *et al.* 2002:55) for isiXhosa.

Reference List

Abrams, Z.I. 2016 “Possibilities and challenges of learning German in a multimodal environment: A case study”. *ReCALL*, 28(3), 343–363.

Adams, M.J. 1990. *Beginning To Read: Thinking and Learning about Print*. Cambridge, MA: MIT Press.

Adetomokun, I.J. 2018. “Exploring Semiotic Remediation in Performances of Stand-up Comedians in Post-Apartheid South Africa and Post-Colonial Nigeria”. Unpublished PhD thesis. Cape Town: University of the Western Cape. http://etd.uwc.ac.za/xmlui/handle/11394/6684. 15 July 2022.

Adobe Corporate Communications. 2019. “The Future of Adobe Air”. https://blog.adobe.com/en/publish/2019/05/30/the-future-of-adobe-air. 19 December 2019

Amundrud, T.M. 2017. “Analyzing classroom teacher-student consultations: A systemic multimodal perspective”. Unpublished PhD thesis. Sydney: Macquarie University. https://www.academia.edu/36343207/Analyzing_classroom_teacher_student_consultations_A_systemic_multimodal_perspective?email_work_card=view-paper. 27 May 2020.

ANA. 2013. “Report on the Annual National Assessment of 2013: Grades 1 to 6 & 9”. Department of Education (ed). http://superonnies.co.za/ANA%20Report%202013.pdf. 7 July 2020.

Andrason, A. 2017. “The “exotic” nature of ideophones–from Khoekhoe to Xhosa”. *Stellenbosch Papers in Linguistics*, 48(1), 139-150.

Andrason, A. 2020. “Ideophones as Limguistic 'Rebels': The Extra-Systematicity of Ideophones in Xhosa (Part 1)”. *Asian & African Studies*, 29(2).

Archer, A. 1997. “The function of the visual in teaching English as an additional language: The case of video”. Unpublished MA research report. Johannesburg: University of the Witwatersrand. http://hdl.handle.net/10539/21355. 27 May 2020.

Archer, A. 2004. “Access to academic practices in an engineering curriculum: drawing on students’ representational resources through a multimodal pedagogy”. Unpublished PhD thesis. Cape Town: University of Cape Town. http://hdl.handle.net/11427/23682. 27 May 2020.

Archer, A. 2017. “Using multimodal pedagogies in writing centres to improve student writing”. *Stellenbosch Papers in Linguistics Plus*, 53(1), 1–12.

Archer, A., and Newfield, D. 2014. *Multimodal Approaches to Research and Pedagogy: Recognition, Resources, and Access*. Routledge, New York & Oxon.

Ardington, C., Wills, G., Pretorius, E., Deghaye, N., Menendez, A., Mohohlwane, N., Mtsatse, N. and Van der Berg, S. 2020. “Technical Report: Benchmarking early grade reading skills in Nguni languages”. Department of Basic Education.

Baldry, A. (ed) 2000. *Multimodality and Multimediality in the Distance Learning Age*. Campobasso: Palladino Editore.

Banda, F. and Oketch, O. 2011. “Localizing HIV/AIDS discourse in a rural Kenyan community”. *Journal of Asian and African Studies*, 46(1), 19–37.

Bezemer, J. and Jewitt, C. 2010. “Multimodal Analysis: Key issues”. In: L. Litosseliti (ed), *Research Methods in Linguistics*. London: Continuum. 180-197.

Bhaskar, Roy. 1975. *A realist theory of science*. New York: Routledge.

Bianchi, F. and Ciabattoni, T. 2008. Captions and subtitles in EFL learning: An investigative study in a comprehensive computer environment. In *From DIDACTAS to eCoLingua,* 69-90. EUT Edizioni Università di Trieste.

Bird, S.A. and Williams, J.N. 2002. “The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling”. *Applied Psycholinguistics*. 23(4), 509–533.

Black, S. 2021. “The potential benefits of subtitles for enhancing language acquisition and literacy in children: An integrative review of experimental research”. *Translation, Cognition & Behavior*, 4(1), 74-97.

Bloor, T. and Bloor, M. 2004. *The Functional Analysis of English*. 2nd ed. New York: Oxford University Press.

Bolowana, A. “Department of Education plans to open online schools to reduce pressure on system”. Palech. https://palech.org/department-of-education-plans-to-open-online-schools-to- reduce-pressure-on-system/. 16 July 2021.

Brady, C.K. 2015. “A Multimodal Discourse Analysis of Female K-Pop Music Videos”. Unpublished MA thesis. Birmingham: University of Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/cels/essays/appliedlinguistics/Dissertation-Multimodal-Analysis-84-Distinction.pdf. 15 July 2022.

Brown, Q. 2012. “Exploring the scalar equivalence of the Picture Vocabulary Scale of the Woodcock Munoz Language Survey across rural and urban isiXhosa-speaking learners”. Master's thesis. University of the Western Cape, Cape Town

Browner, W.S., Newman, T.B. and Hulley, S.B. 2007. “Estimating sample size and power: applications and examples”. *Designing clinical research*, 3, 66-85.

Bryman A. 2006. “Integrating quantitative and qualitative research: How is it done?” *Qualitative Research*, 6, 97-113.

Buffalo City Metropolitan Municipality. 2020a. “RFQ/SPD/2020-21/67 - FEASIBILITY STUDY FOR MSOBOMVU SETTLEMENT, VARIOUS PORTIONS OF FARM 270, NEWLANDS” https://www.buffalocity.gov.za/CM/uploads/tenders/20200210101601627636RFQ-SPD-2020-21-67-FEASIBILITYSTUDYFORMSOBOMVUSETTLEMENTVARIOUSPORTIONSOFFARM270NEWLANDS.pdf. 24 October 2022

Buffalo City Metropolitan Municipality. 2020/2021. “INTEGRATED DEVELOPMENT PLAN REVIEW OF BUFFALO CITY METROPOLITAN MUNICIPALITY” https://lg.treasury.gov.za/supportingdocs/BUF/BUF_IDP%20Final_2021_Y_20220225T091602Z_sandiswal.pdf. 24 October 2022

Buffalo City Metropolitan Municipality. 2020b. “Newlands LSDF (2015)” in Review of Spatial Policy. https://buffalocity.gov.za/CM/uploads/documents/20201409091600118732Annexure1.ReviewofLocalSpatialDevelopmentFrameworks-2.pdf. 24 October 2022

Buffalo City Metropolitan Municipality. 2021. “SFD Report Buffalo City Metropolitan Municipality Eastern Cape, South Africa”. https://www.susana.org/_resources/documents/default/3-4394-7-1633084732.pdf. 24 October 2022

Bączkowska, A. 2011. “Some remarks on a multimodal approach”. *Linguistics Applied*, 4, 47-65.

Caimi, A. 2006. “Audiovisual Translation and Language Learning: The Promotion of Intralingual Subtitles”. *The Journal of Specialised Translation*, 6, 85–98.

Carlisle, J.F. 2003. “Morphology matters in learning to read: A commentary”. *Reading Psychology*, 24(3-4), 291-322.

Chai, H. 2007. “Automatic Annotation for Korean--Approach Based on the Contextual Exploration Method”. In 18th International Workshop on Database and Expert Systems Applications (DEXA 2007). 278-282. IEEE.

Childs, G.T. 2003. *An introduction to African languages.* 1. John Benjamins Publishing.

Christie, C.R. 2019. “Khoekhoe Lexical Borrowing in Namaqualand Afrikaans”. Doctoral thesis, Rhodes University.

Coltheart, M., Curtis, B., Atkins, P. and Haller, M. 1993. “Models of reading aloud: Dual-route and parallel-distributed-processing approaches”. *Psychological review*, 100(4), 589.

Combrinck, C. and Mtsatse, N. 2019. “Reading on paper or reading digitally? Reflections and implications of ePIRLS 2016 in South Africa”. *South African Journal of Education*, 39(2)

Commissioning Protocol. 2010. “COMMISSIONING PROTOCOL FOR INDEPENDENTLY PRODUCED SOUTH AFRICAN PROGRAMMING”. https://www.etv.co.za/sites/default/files/Commissioning_protocol-final_July2010.pdf. 28 May 2020

Conley, C.M., Derby, K.M., Roberts-Gwinn, M., Weber, K.P. and McLaughlin, T.F. 2004. “An analysis of initial acquisition and maintenance of sight words following picture matching and copy, cover, and compare teaching methods”. *Journal of Applied Behavior Analysis*, 37(3), 339-349.

Cotrim, H., Granja, C., Carvalho, A. S., Cotrim, C., and Martins, R. 2021. “Children's Understanding of Informed Assents in Research Studies”. *Healthcare (Basel, Switzerland)*, 9(7), 871.

Creissels, D. 2021. “The mismamed “participial” forms of Tswana and other Southern Bantu languages”. *Handout from Bantu*, 8.

Creswell, J.W. 2014. *A concise introduction to mixed methods research*. SAGE publications.

Creswell, J.W. and Clark, V.P. 2011. *Mixed methods research*. SAGE Publications.

D'Ydewalle, G. and Van de Poel, M. 1999. “Incidental Foreign-Language Acquisition by Children Watching Subtitled Television Programs”. *Journal of Psycholinguistic Research*, 28(3), 227–244.

Dal, M. 2012 “Digital Video Production and Foreign Language Learning.” In Conference: *World Conference on Educational Multimedia, Hypermedia and Telecommunications*. 2012. 1–13.

https://www.researchgate.net/publication/277984755_Digital_Video_Production_and_Foreign_Language_Learning. 27 May 2021.

Danan, M. 2004. "Captioning and Subtitling: Undervalued Language Learning Strategies". *Meta*, 49(1), 67–77.

Daries, M.A. and Probert, T.N. 2020. "A linguistic analysis of spelling errors in Grade 3 isiXhosa home-language learners". *Reading & Writing*, 11(1), 1-10.

Davies, H.J. "Lights, camera, caption! Why subtitles are no longer just for the hard of hearing". https://www.theguardian.com/tv-and-radio/2019/jul/21/subtitles-tv-hearing-no-context-twitter-captions. 14 June 2020.

De Linde, Z. and Kay, N. 2014. *The Semiotics of Subtitling*. London & New York.

De Vos, M., Van Der Merwe, K., and Van der Mescht, C. 2015. "A Linguistic Research Programme for Reading in African Languages to underpin CAPS". *Journal for Language Teaching*, 48(2), 149-177.

Demuth, K. 2003. "The acquisition of Bantu languages". In K. Demuth. *The Bantu languages*, 209-222.

DBE. 2011. Department of Basic Education "Curriculum and Assessment Policy Statement: Foundation Phase R-3". https://www.education.gov.za/Portals/0/CD/National%20Curriculum%20Statements%20and%20Vocational/CAPS%20isiXhosa%20HL%20%20GRADES%20R-3%20FS.pdf?ver=2015-01-27-154618-763. 7 May 2021.

Department of Education. 1997. "Language in education policy", *Section 3(4)(m) of the National Education Policy Act, 1996 (Act 27 of 1996)*. Pretoria, RSA.

Department of Higher Education and Training. 2021. "Fact Sheet. Adult Illiteracy in South Africa". https://www.dhet.gov.za/Planning%20Monitoring%20and%20Evaluation%20Coordination/Fact%20Sheet%20on%20Adult%20Illiteracy%20in%20South%20Africa%20-%20March%202021.pdf. 7 May 2021.

Diab, P., Matthews, M. and Gokool, R. 2016. "Medical students' views on the use of video technology in the teaching of isiZulu communication, language skills and cultural competence," *African Journal of Health Professions Education*, 8(1), 11-14.

Diemer, M.N. 2015. "The contributions of phonological awareness and naming speed to the reading fluency, accuracy, comprehension and spelling of Grade 3 isiXhosa readers". Master's thesis, Rhodes University, Grahamstown.

Dobakhti, L., Zohrabi, M. and Tadayyon, P. 2020. “An Investigation into the Impact of Rote and Mnemonic Strategies on Vocabulary Learning of Iranian Elementary EFL Learners”. *Teaching English Language*, 14(2), 323-344.

DSTV Now. 2022. “Can I watch NOW with subtitles or closed captions?”. https://help.nowtv.com/article/can-i-watch-with-subtitles. 12 Feb 2022.

Du Plessis, T. 2023. “The emergence of South African Sign Language as South Africa's 12th official language”. *Tydskrif vir Geesteswetenskappe*, 63(3), 585-614.

Dwivedi, A.K., Mallawaarachchi, I. and Alvarado, L.A. 2017. “Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method”. *Statistics in medicine*, 36(14), 2187-2205.

d’Ydewalle, G., and de Bruycker, W. 2007. “Eye movements of children and adults while reading television subtitles”. *European Psychologist,* 12, 196-205.

d’Ydewalle, G., and Gielen, I. 1992. “Attention allocation with overlapping sound, image, and text”. In K. Rayner (Ed.): *Eye movements and visual cognition: Scene perception and reading*, 415-427. New York, NY: Springer.

Eren, F. 2016. “Re-specification of Concepts in the Morphogenetic Approach for Property Market Research”. *International Journal of Architecture & Planning*, 4(2), 15-34.

Fagerland, M.W. and Sandvik, L. 2009. “The wilcoxon–mann–whitney test under scrutiny”. *Statistics in medicine*, 28(10), 1487-1497.

Fei, L.V. 2011 A Systemic Functional Multimodal Discourse Analysis Approach to Pedagogic Discourse. Doctoral Thesis. National University of Signapore

Ferris, F.S. and Banda, F. 2015. “'Poof! a’m heppily saving the Lord…': multimodality and evaluative discourses in male toilet graffiti at the University of the Western Cape”. *African Identities*, 13(4), 243–261.

Fleetwood, S. 2014. Bhaskar and critical realism. *Oxford handbook of sociology, social theory, and organization studies: Contemporary currents*, 182-219.

Fletcher, J. 2010. “The prosody of speech: Timing and rhythm”. *The handbook of phonetic sciences*, 521-602.

Fong, G.C.F. and Au, K.K. eds. 2009. *Dubbing and subtitling in a world context*. Chinese University Press.

Fossett, B. and Mirenda, P. 2006. “Sight word reading in children with developmental disabilities: A comparison of paired associate and picture-to-text matching instruction”. *Research in developmental disabilities*, 27(4), 411-429.

Frehan, P. 1999. “Beyond the sentence: Finding a balance between bottom-up and top-down reading approaches”. *LANGUAGE TEACHER-KYOTO-JALT-*, 23, 11-14.

Fuchs, L.S., Fuchs, D., Hosp, M.K. and Jenkins, J.R. 2001. “Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis”. *Scientific studies of reading*, 5(3), 239-256.

Gachago, D., Cronje, F., Ivala, E., Condy, J. and Chigona, A. 2014. “Using Digital Counterstories as Multimodal Pedagogy among South African Pre-service Student Educators to produce Stories of Resistance”. *Electronic Journal of e-Learning*, 12(1), 29–42.

Gallien, C., Hotho, S. and Staines, H. 2000. “The impact of input modifications on listening comprehension: A study of learner perceptions”. *JALT journal*, 22(2), 271-295.

Garza, T J. 1991. “Evaluating the use of captioned video materials in advanced foreign language learning”. *Foreign Language Annals*, 24(3), 239-258.

Gee, J.P. 2015. “Discourse, small d, big D”. *The international encyclopedia of language and social interaction*, 3, 1-5.

Gernsbacher, M.A. 2015. “Video Captions Benefit Everyone”. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 195-202.

Godwin, K.E., Almeda, M.V., Seltman, H., Kai, S., Skerbetz, M.D., Baker, R.S. and Fisher, A.V. 2016. “Off-task behavior in elementary school children”. *Learning and Instruction*, 44, 128-143.

Gonzalez Peña, P. 2020. “Spatial deixis in child development”. Doctoral thesis, University of East Anglia.

Goswami, U. 2010. “A psycholinguistic grain size view of reading acquisition across languages”. In *Reading and dyslexia in different orthographies*. 41-60. Psychology Press.

Gove, A., and Wetterberg, A. (eds.) 2011. *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy*. Research Triangle Park, NC: RTI Press.

Guichon, N., and McLornan, S. 2008. “The Effects of Multimodality on L2 Learners: Implications for CALL Resource Design”. *System*, 36(1), 85-93.

Guijarro, J.M. and Sanz, M.J.P. 2008. “Compositional, interpersonal and representational meanings in a children's narrrative: A multimodal discourse analysis”. *Journal of Pragmatics*, 40(9), 1601-1619.

Guthrie, M. 1971. “The western Bantu languages”. *Current trends in linguistics*, 7, 357-366.

Gxilishe, S. 2004. “The acquisition of clicks by Xhosa-speaking children”. *Per Linguam: a Journal of Language Learning=Per Linguam: Tydskrif vir Taalaanleer*, 20(2), 1-12.

Gxilishe, S., Smouse, M.R., Xhalisa, T. and de Villiers, J. 2009. “Children’s insensitivity to information from the target of agreement: The case of Xhosa.” in J. Crawford et al. (eds.), *Generative Approaches to Language Acquisition North America (GALANA)*, 46–53, Cascadilla Proceedings Project, Somerville, MA.

Hakkani-Tür, D.Z., Oflazer, K. and Tür, G. 2002. “Statistical morphological disambiguation for agglutinative languages”. *Computers and the Humanities*, 36, 381-410.

Halliday, M.A.K. 1976. *System and Function in Language*. Kress, G.R. (ed.). London: Oxford University Press.

Halliday, M.A.K. 1985 *Spoken and Written Language*. Victoria: Deakin University Press.

Halliday, M.A.K. 1994 “Spoken and Written Modes of Meaning”, in D. Graddol D. and 0. Boyd-Barett (eds) *Media texts: Authors and Readers*. Clevedon, England: Multilingual Matters & The Open University, 51-73.

Halliday, M.A.K. 2004. *An introduction to functional grammar*. 3rd ed. Revised by C.M.I.M. Matthiessen. London: Arnold.

Halliday, M.A.K. and Matthiessen, C.M. 2014. *Halliday's introduction to functional grammar*. 4th ed. Routledge.

Hazamy, A.A. 2009. “Influence of pictures on word recognition”. Master’s thesis, Florida Atlantic University, Statesboro, Georgia

Hefer, E. 2013a “Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English”, *Southern African Linguistics and Applied Language Studies*, 31(3), 359-373.

Hefer, E. 2013b. “Television subtitles and literacy: where do we go from here?”. *Journal of Multilingual and Multicultural Development*, 34(7), 636-652.

Heinemann, T. 2019. “Energy crisis in South Africa–a barrier to higher economic growth”. *KfW Research Economics in Brief*, 178.

Hodge, R. and Kress, G. 1988. *Social Semiotics*. Cornwall: Cornell University Press.

Howie, S., Combrinck, C., Roux, K., Tshele, M., Mokoena, G., Palane, N.M. 2016. *Progress in International Reading Literacy Study 2016: South African Children’s Reading Literacy Achievement*. Pretoria: Centre for Evaluation and Assessment.

Howie, S., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit, C., Scherman, V., Archer, E. 2006. “PIRLS 2006 Summary Report: South African Children’s Reading Literacy

Achievement". https://nicspaull.files.wordpress.com/2011/04/howie-et-al-pirls-2006-sa-summary-report.pdf. 7 July 2020.

Huang, H.C. and Eskey, D.E. 1999. "The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students". *Journal of Educational Technology Systems*, 28(1), 75-96.

Huettig, F. and McQueen, J.M. 2007. "The tug of war between phonological, semantic and shape information in language-mediated visual search". *Journal of memory and language*, 57(4), 460-482.

Iedema, R. 2001. "Analysing film and television: a social semiotic account of Hospital: an Unhealthy Business" In van Leeuwen, T. and Jewitt, C. (eds.) *Handbook of Visual Analysis*. London, California, New Delhi: SAGE Publications Ltd. 183–206.

Jewitt, C. 2006 *Technology, Literacy, Learning*. Oxon, New York: Routledge.

Jewitt, C. 2009 "Different Approaches to Multimodality". In *The Routledge Handbook of Multimodal Analysis*. London: Routledge. 28–39.

Jewitt, C. 2014, "An Introduction to Multimodality". In Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis, 2nd ed.*, London: Routledge, 15-30.

Jewitt, C. 2012. "An introduction to using video for research". NCRM Working Paper. https://eprints.ncrm.ac.uk/id/eprint/2259/. 13 August 2021

Jewitt, C., Bezemer, J. and O'Halloran, K. 2016. *Introducing Multimodality*. Oxon: Routledge.

Jilingisi, N. 2013. "Gendered Roles and Social Behaviour Towards Women in Marginalised Communities: The Case of Newlands Location in East London". Doctoral dissertation, Nelson Mandela Metropolitan University.

Jones, M. 2007. "Presence as External versus Internal Experience: How Form, User, Style, and Content Factors Produce Presence from the Inside". *Proceedings of the Tenth Annual International Meeting of the Presence Workshop*, Barcelona, Spain, 115–126.

Jordan, A.C. 1966. *A practical course in Xhosa 1st ed.* Wisconsin: Longmans Southern Africa (PTY) limited.

Kaltenbacher, M. 2004. "Multimodality in language teaching CD-ROMs". In Ventola, E., Charles, C. and Kaltenbacher, M. (eds.) *Perspectives on Multimodality*. Amsterdam, Philadelphia: John Benjamins Publishing Company. 119–136.

Karakaş, A. and Sariçoban, A. 2012. "The Impact of Watching Subtitled Animated Cartoons on Incidental Vocabulary Learning of ELT Students". *Teaching English with Technology*, 12, 3-15.

Kehe, J. 2018. “The Real Reason You Use Closed Captions for Everything Now”. https://www.wired.com/story/closed-captions-everywhere/. 6 September 2021.

Khuluvhe, M. 2021. *Adult illiteracy in South Africa*. Pretoria: South African Department of Higher Education and Training.

Khumalo, L. 2009. “The passive and stative constructions in Ndebele: A comparative analysis”. *Nordic Journal of African Studies*, 18(2), 21-21.

Kim, T.K. 2015. “T test as a parametric statistic”. *Korean journal of anesthesiology*, 68(6), 540-546.

Knox, J.S., Patpong, P. and Piriyasilpa, Y. 2011. “ขาวหน าหน ◌่ ◌่ ◌่ ◌่ ง (Khao naa nung): A Multimodal Analysis of Thai-language Newspaper Front Pages”. In Martin, J.R. and Bednarek, M. (eds.) *New Discourse on Language: Functional Perspectives on Multimodality, Identity, and Affiliation*. London & New York: Continuum International Publishing Group. 80–110.

Koda, K. 2007. “Reading and language learning: Cross-linguistic constraints on second language reading development”. *Language learning,* 57(suppl.1), 1–44.

Kothari, B. 2008. “Let a billion readers bloom: Same language subtitling (SLS) on television for mass literacy”. *International Review of Education*, 54(5-6), 773-80.

Kothari, B., and Bandyopadhyay, T. 2014. “Same language subtitling of Bollywood film songs on TV: Effects on literacy”. *Information Technologies & International Development*, 10(4), 31-47.

Kothari, B., Pandey, A. and Chudgar, A. 2004. “Reading Out of the ‘Idiot Box’: Same-Language Subtitling on Television in India”. *Information Technologies and International Development*, 2(1), 23-44.

Kothari, B., Takeda, J., Joshi, A., and Pandey, A. 2002. “Same Language Subtitling: A Butterfly for Literacy?”. *Information Technologies & International Development*, 21(1), 55-66

Kress, G. and T. van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Oxford UK: Oxford University Press.

Kress, G. and T. van Leeuwen. 2006 [1996]. *Reading Images*. London: Routledge.

Kress, G. 2015. “Semiotic work: Applied Linguistics and a social semiotic account of Multimodality”. *AILA Review*, 28, 49–71.

Kress, G. and van Leeuwen, T. 2001, *Multimodal Discourse: The Modes and Media of Contemporary Communication*, London: Edward Arnold.

Kruger, J.L. 2012. “Ideology and Subtitling: South African Soap Operas”. *Meta*, 57(2), 496–509.

Kruger, J.L., Hefer, E. and Matthew, G. 2013. August. “Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts”. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, 62-66.

Kruger, J.L., Hefer, E. and Matthew, G. 2014. “Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2”. *Journal of Eye Movement Research*, *7*(5).

Kruger, J.L., Wisniewska, N. and Liao, S. 2022. “Why subtitle speed matters: Evidence from word skipping and rereading”. *Applied psycholinguistics*, 43(1), 211-236.

Kuosmanen, J. 2020. “Getting Your Daily Flix: A Reception Study of Netflix and Its Subtitles”. Master's thesis, Itä-Suomen yliopisto.

Land, S.J. 2015. “Reading isiZulu: Reading processes in an agglutinative language with a transparent orthography”. Doctoral thesis, 1–195.

Lantz, M.E., Cui, A.X., and Cuddy, L.L. 2020. “The role of duration and frequency of occurrence in perceived pitch structure”. *PloS one*, 15(9)

Lappalainen, M. 2021. “When does domestication go too far? Viewers’ reception of domesticated subtitles”. Master's thesis, University of Helsinki

Lavaur, J-M., and Bairstow, D. 2011. “Languages on the screen: Is film comprehension related to the viewers’ fluency level and to the language in subtitles?”. *International Journal of Psychology*, 46(6), 455-462.

Lavrakas, P.J. 2008. *Encyclopedia of survey research methods*. Sage publications.

Le Doeuff, H. 2020. “A sociolinguistic, phonetic and phonological overview of isiXhosa clicks”. Master's thesis. New Sorbonne University, Paris. https://www.researchgate.net/publication/343239485_A_sociolinguistic_phonetic_and_phonological_overview_of_isiXhosa_clicks. 29 October 2022.

Lenninger, S. 2006. “Piaget and Vygotsky on the child becoming sign-minded”. Heterogénesis: Revista de Artes Visuales.

Lepola, J., Poskiparta, E., Laakkonen, E. and Niemi, P. 2005. “Development of and relationship between phonological and motivational processes and naming speed in predicting word recognition in grade 1”. *Scientific studies of reading*, 9(4), 367-399.

Levinson, S.C. 2004. *Deixis. The handbook of pragmatics*. Horn & G. Ward (eds.).

Liao S., Yu L., Reichle, E.D. and Kruger J.L. 2021. “Using Eye Movements to Study the Reading of Subtitles in Video”. *Scientific Studies of Reading*, 25(5), 417-435.

Lim, F.V. and Toh, W. 2020. “How to Teach Digital Reading?”. *Journal of Information Literacy*, 14(2), 24-43.

Linebarger, D., Piotrowski, J.T. and Greenwood, C.R. 2010. “On-screen print: the role of captions as a supplemental literacy tool”. *Journal of Research in Reading*, 33(2), 148–167.

Linebarger, D.L. and Walker, D. 2005. “Infants’ and toddlers’ television viewing and language outcomes”. *American behavioral scientist*, 48(5), 624-645.

Liu, D. 2014. “On the classification of subtitling”. *Journal of language Teaching and Research*, 5(5), 1103.

Liu, Y. and O’Halloran, K. 2009. “Intersemiotic Texture: analyzing cohesive devices between language and images”. *Social Semiotics*, 19(4), 367–388.

Louwrens, L.J. and Poulos, G. 2006. "The status of the word in selected conventional writing systems—the case of disjunctive writing". *Southern African linguistics and applied language studies*, 24(3), 389-401.

Lum, S.A. and Tiokou, C.D. 2015. “Analysis of the prospects and challenges of subtitling as a mode of audiovisual translation in Cameroon”. *Journal of Languages and Culture*, 6(7), 61-70.

Ma, X. 2017. “Ellipsis in the vP domain in Mandarin and Xhosa". Doctoral thesis. Rhodes University, Grahamstown.

Mackenzie, N. & Knipe, S. 2006. “Research dilemmas: paradigms, methods and methodology”. *Issues In Educational Research*, 16, 1-15.

Made, Z.J. 2010. “An investigation into implementation of language policy in the Eastern Cape with specific reference to isiXhosa”. Doctoral thesis, Nelson Mandela Metropolitan University.

Mafofo, L. and Banda, F. 2014. “Accentuating institutional brands: A multimodal analysis of the homepages of selected South African universities”. *Southern African Linguistics and Applied Language Studies*, 32(4), 417–432.

Makalela, L. and Fakude, P.F. 2014. “'Barking' at texts in Sepedi oral reading fluency: Implications for edumetric interventions in African languages”. *South African Journal of African Languages*, 34(1), 71-78.

Makaure, Z.P. 2016. “Phonological processing and reading development in Northern Sotho-English bilingual children”. Master’s thesis, University of South Africa, Pretoria.

Malda, M., Nel, C. and Van de Vijver, F.J. 2014. "The road to reading for South African learners: The role of orthographic depth". *Learning and Individual Differences*, 30, 34-45.

Malinga, S. 2020. "Where to go online for digital learning in SA" https://www.itweb.co.za/content/lLn14Mmj1peqJ6Aa. 14 March 2021.

Markham, P. and Peter, L. 2003. "The influence of English language and Spanish language captions on foreign language listening/reading comprehension". *Journal of Educational Technology Systems*, 31(3), 331-341.

Markham, P.L. 1999. "Captioned videotapes and second-language listening word recognition". *Foreign Language Annals*, 32(3), 321-328.

Markham, P.L., Peter, L., & Mccarthy, T. 2001. "The effects of native language vs. target language captions on foreign language students' DVD video comprehension". *Foreign Language Annals*, 34(5), 439-445.

Martin, J.R. and Rose, D. 2007. "Interacting with text: The role of dialogue in learning to read and write". *Foreign Languages in China*, 4(5), 66-80.

Maseko, P. 2017. "Exploring the History of the Writing of isiXhosa: An Organic or an Engineered Process?". *International Journal of African Renaissance Studies - Multi-, Inter- and Transdisciplinarity*, 12(2), 81-96.

Matielo, M., D'Ely, R.C.S.F, Baretta, L. 2015. "The effects of interlingual and intralingual subtitles on second language learning/acquisition: a state-of-the-art review". *Trabalhos em Linguística Aplicada*, 54(1), 161-182.

Matjila, D.S., & Pretorius, E.J. 2004. "Bilingual and biliterate? An exploratory study of grade 8 reading skills in Setswana and English". *Per Linguam*, 20, 121.

Matthew, G. 2020. "The effect of adding same-language subtitles to recorded lectures for non-native, English speakers in e-learning environments". *Research in Learning Technology*, 28, 1-16.

Mayer, R. E., and Moreno, R. 2003. "Nine ways to reduce cognitive load in multimedia learning". *Educational Psychologist*, 38(1), 43–52.

Mayer, R.E. 2008. "Applying the science of learning: evidence-based principles for the design of multimedia instruction". *American psychologist*, 63(8), 760.

Mayer, R.E. ed. 2005. *The Cambridge handbook of multimedia learning*. Cambridge university press.

Mayer, R.E. ed. 2009. *The Cambridge handbook of multimedia learning*. Cambridge university press.

McBride-Chang, C, Wagner R.K., Muse, A., Chow, B.W.Y. and Shu, H. 2005. “The role of morphological awareness in children's vocabulary acquisition in English”. *Applied Psycholinguistics*, 26(3), 415-435.

McConachy, T. 2018. “Critically engaging with cultural representations in foreign language textbooks”. *Intercultural Education*, 29(1), 77-88.

Meiring, C. 2020. “Impact evaluation of Funda Wande in-service teacher coaching intervention”. Master's thesis, Stellenbosch University, Stellenbosch.

Mihajlik, P., Fegyó, T., Tüske, Z. and Ircing, P. 2007, August. “A morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages-like Hungarian”. In *INTERSPEECH*, 1497-1500.

Minucci, M.V. and Cárnio, M.S. 2010. “Movie subtitles reading skills of elementary school children”. *Pró-Fono Revista de Atualização Científica* 22(3), 227–232.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A. 2019. “Descriptive statistics and normality tests for statistical data”. *Annals of cardiac anaesthesia*, 22(1), 67.

Mkabile, H. 2019. “The morphotactic constraints of verbal extensions in isiXhosa”. Master's Thesis. Rhodes University

Mkhize, D.N. 2013. “The nested contexts of language use and literacy learning in a South African fourth grade class: Understanding the dynamics of language and literacy practices”. Doctoral thesis, University of Illinois at Urbana-Champaign.

Moropa, K. 2007. “Analysing the English-Xhosa parallel corpus of technical texts with Paraconc: a case study of term formation processes”. *Southern African linguistics and applied language studies*, 25(2), 183-205.

Morse J.M., and Niehaus, L. 2009. *Mixed method design: Principles and procedures*. Walnut Creek: Left Coast Press

Mossberger, K., Tolbert, C.J. and Stansbury, M. 2003. *Virtual inequality: Beyond the digital divide*. Georgetown University Press.

Mowrer, D.E. and Burger, S. 1991. “A comparative analysis of phonological acquisition of consonants in the speech of 2½-6-year-old Xhosa-and English-speaking children”. *Clinical Linguistics & Phonetics*, 5(2), 139-164.

Mtsatse, N. and van Staden, S. 2021. “Exploring differential item functioning on reading achievement: A case for isiXhosa”. *South African Journal of African Languages*, 41(1), 1-11.

Mubenga, K.S. 2020. “Film Subtitling: A New Research Genre in Africa”. https://www.researchgate.net/profile/Kajingulu-Mubenga/publication/349716294_Film_Subtitling_A_New_Research_Genre_in_Africa/links/604c117992851c2b23c5731a/Film-Subtitling-A-New-Research-Genre-in-Africa.pdf. 7 May 2021

Mullis I.V.S., Martin M.O., Kennedy A., Foy P. 2007. *PIRLS 2006 international report: IEA’s Progress in International Reading Literacy Study in 113 primary schools in 40 countries*. Chestnut Hill: PIRLS International Study Center, Boston College.

Mullis, I.V., Martin, M.O. and Sainsbury, M. 2016. *PIRLS 2016 reading framework*. PIRLS, 11-29.

Mzamo, L., Helberg, A. and Bosch, S. 2019. January. “Towards an unsupervised morphological segmenter for isiXhosa”. In *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 166-170. IEEE.

Neubert, C. 2021. “Language variation in South Africa: A sociophonetic study of the vowel system of black South African English”. Doctoral dissertation. Universität Regensburg

Neuman, S.B., and Koskinen, P. 1992. “Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students”. *Reading Research Quarterly*, 27, 94-106.

Nurse, D. “A persistive aspect”. In Nurse, D. 2008. *Tense and aspect in Bantu*. Oxford University Press, USA. 24

Nurse, D., and Philippson, G. 2003. “Introduction”. In D Nurse & G Philippson (Eds.), *The Bantu Languages*. London: Routledge.

Office of Communications. 2006. “Television access services Review of the Code and guidance”. Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0016/42442/access.pdf. 30 July 2020.

Olivier, J. 2011. “Acknowledging and protecting language rights on SABC TV through the use of subtitles”. *Communicatio*, 37(2), 225-241.

Oosthuysen, J.C. 2015. “Extricating the Description of the Grammar of isiXhosa From a Eurocentric Approach.” *South African Journal of African Languages*, 35, 83–92.

Oosthuysen, J.C. 2016. *The grammar of isiXhosa*. African Sun Media.

O'Halloran, K. 2008. "Systemic functional-multimodal discourse analysis (SF-MDA): constructing ideational meaning using language and visual imagery". *Visual Communication*, 7(4), 443–475.

O'Halloran, K. and Lim Fei, V. 2014. "Systemic functional multimodal discourse analysis". In Norris, S. and Maier, C. (eds.) *Texts, Images and Interactions: A Reader in Multimodality*. Berlin: Mouton de Gruyter, 137–153.

O'Halloran, K.L. 2005. *Mathematical Discourse: Language, Symbolism and Visual Images*, London and New York: Continuum.

O'Halloran, K.L. 2023. "Matter, meaning and semiotics". *Visual Communication*, 22(1), 174-201.

O'Halloran, K.L., Tan, S., Wiebrands, M., Sheffield, R., Wignell, P. and Turner, P. 2018. "The multimodal classroom in the digital age: The use of 360 degree videos for online teaching and learning". In *Multimodality Across Classrooms* (84-102). Routledge.

O'Toole, M. 1994. *The Language of Displayed Art*. London: Leicester University Press.

Paivio, A., 1969. "Mental imagery in associative learning and memory". *Psychological review*. 76(3), 241.

Pamla, A., Thondhlana, G. and Ruwanza, S. 2021. "Persistent droughts and water scarcity: households' perceptions and practices in Makhanda, South Africa". *Land*, 10(6), 593.

Park, Y.S., Konge, L. and Artino, A.R. 2020. "The positivism paradigm of research". *Academic medicine*, 95(5), 690-694.

Parkhill, F., Johnson, J. and Bates, J. 2011. "Capturing literacy learners: Evaluating a reading programme using popular novels and films with subtitles". *Digital Culture & Education*, 3(2), 140–156.

Pasch, H. 2008. "Competing scripts: the introduction of the Roman alphabet in Africa". *International journal of the sociology of language*, (191), 65-109.

Perego, E., del Missier, F., Porta, M. and Mosconi, M. 2010. "The Cognitive Effectiveness of Subtitle Processing". *Media Psychology*, 13, 243–272.

Perez, M.M., Van Den Noortgate, W. and Desmet, P., 2013. "Captioned video for L2 listening and vocabulary learning: A meta-analysis". *System*, 41(3), 720-739.

Phandle, G. 2021. "More than 1,500 Eastern Cape schools are overcrowded". https://www.dispatchlive.co.za/news/2021-05-14-more-than-1500-eastern-cape-schools-are-overcrowded/. 22 November 2021.

Piaget, J., and Inhelder, B. 1969. *The Psychology of the Child (H. Weaver Trans.)*. New York, NY: Basic Books.

Pihlajakoski, M. 2008. “Alakouluikäiset lapset elokuvakäännösten vastaanottajina”. Master’s thesis, Tampere University.

Piper, B., and Zuilkowski, S.S. 2015. “Assessing reading fluency in Kenya: Oral or silent assessment?”. *International Review of Education*, 61(2), 153-171.

Pretorius, E. & Mokhwesana, M. 2009. “Putting reading in Northern Sotho on track in the early years: Changing resources, expectations and practices in a High Poverty School”, *South African Journal of African Languages*, 1, 54–73.

Pretorius, E.J. 2010. “Issues of complexity in reading: Putting Occam’s razor aside for now”. *Southern African Linguistics and Applied Language Studies*, 28(4), 339-358.

Pretorius, E.J. 2012. “Butterfly effects in reading? The relationship between decoding and comprehension in Grade 6 high poverty schools”. *Journal for Language Teaching*, 46(2), 74-95.

Pretorius, E.J. 2015. “Failure to launch: Matching language policy with literacy accomplishment in South African schools”. *International Journal for the Sociology of Language*, 2015(234), 47-76

Pretorius, E.J., and Mokhwesana, M.M. 2009. “Putting reading in Northern Sotho on track in the early years: Changing resources, expectations and practices in a high poverty school”. *South African Journal of African Languages*, 29(1), 54-73.

Price, K. 1983. “Closed-captioned TV: An untapped resource”. *MATESOL Newsletter*, 12, 1-8

Prinsloo, M. and Stein, P. 2004. “What's inside the box? Children's early encounters with literacy in South African classrooms”. *Perspectives in education*, 22(2), 67-84.

Probert, T. and De Vos, M. 2016. “Word recognition strategies amongst IsiXhosa/English Bilingual Learners: The interaction of orthography and language of learning and teaching”. *Reading & Writing*, 7(1), a84.

Probert, T. 2016. “A comparative study of syllables and morphemes as literacy processing units in word recognition: IsiXhosa and Setswana”. Master's thesis, Rhodes University, Grahamstown.

Probert, T.N. 2019. “A comparison of the early reading strategies of isiXhosa and Setswana first language learners”. *South African Journal of Childhood Education*, 9(1), 1-12.

Psyridou, M., Tolvanen, A., Lerkkanen, M.K., Poikkeus, A.M. and Torppa, M. 2020. “Longitudinal stability of reading difficulties: Examining the effects of measurement error, cut-offs, and buffer zones in identification”. *Frontiers in psychology*, 10, 2841.

Ramrathan, L. 2020. “School curriculum in South Africa in the Covid-19 context: An opportunity for education for relevance”. *Prospects*, 51(1-3), 383-392.

Rao, V.C. 2018. “English spelling and pronunciation: a brief study”. *J. Res. Sch. Profess. Eng. Lang. Teach*, 2, 1-10.

Rapp, B. 2002. “Uncovering the cognitive architecture of spelling”. In Hillis, A. (ed.) *The handbook of adult language disorders*. New York: Psychology Press. 47–62.

Rees, S.A. 2016. “Morphological awareness in readers of isiXhosa”. Master's thesis, Rhodes, University, Grahamstown.

Royce, T.D. and Bowcher, W.L. 2007. “Intersemiotic complementarity: a framework for multimodal”. *New directions in the analysis of multimodal discourse*, 63, 109.

Rudolph, M. 2017. “Cognitive theory of multimedia learning”. *Journal of Online Higher Education*, 1(2).

Satyo, S. 1985. “Topics in Xhosa verbal extension”. Doctoral thesis. University of South Africa

Saul, Z.W. 2013. “Significance of Accuracy in the Orthographical Development of IsiXhosa in a Post-democratic South Africa”. Doctoral dissertation, University of Fort Hare.

Savić, S. 2020. “Tense and aspect in Xhosa”. Doctoral thesis. Rhodes University, Makhanda.

Schaefer, M. and Kotzé, J. 2019. “Early reading skills related to Grade 1 English Second Language literacy in rural South African schools”. *South African Journal of Childhood Education*, 9(1).

Schaefer, M., Probert, T. and Rees, S. 2020. “The roles of phonological awareness, rapid automatised naming and morphological awareness in isiXhosa”. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 36(1), 90-111.

Schafli, S.L. 2020. “Pedagogic videos as a foreign language learning resource in textbooks used in the German studies section of a South African university: A digital multimodal discourse perspective”. Master's thesis, Rhodes University, Grahamstown.

Schoonenboom J, Johnson R.B. 2017. “How to Construct a Mixed Methods Research Design”. *Kolner Zeitschrift fur Soziologie und Sozialpsychologie*. 69(2), 107-131.

Scollon, R., and Scollon, S.W. 2003. *Discourses in Place: Language in the material world*. London & New York: Routledge.

Scott HK, Cogburn M. 2023 "Piaget". In: *StatPearls. Treasure Island (FL)*: StatPearls Publishing; https://www.ncbi.nlm.nih.gov/books/NBK448206/. 22 March 2023

Sefotho, M.P. 2019. “Strategies for reading development among Sesotho-English bilinguals: efficacy of translanguaging”. Doctoral dissertation, University of the Witwatersrand.

Sibanda, G. 2009. “Vowel processes in Nguni: Resolving the problem of unacceptable VV sequences”. In *Selected Proceedings of the 38th Annual Conference on African Linguistics*. 38-55. Cascadilla Proceedings Project Somerville, MA.

Simelane-Kalumba, P.I. 2014. “The use of proverbial names among the Xhosa society: Socio-cultural approach”. Doctoral thesis, University of Western Cape.

Slemmons, K., Anyanwu, K., Hames, J., Grabski, D., Mlsna, J., Simkins, E. and Cook, P. 2018. “The impact of video length on learning in a middle-level flipped science setting: Implications for diversity inclusion”. *Journal of Science Education and Technology*, 27, 469-479.

Smith, J. and Adendorff, R. 2021. “Reading parents: Parody and paradox in Go the Fuck to Sleep”. *Language & Communication*, 77, 81-92.

South African Broadcasting Corporation. 2019. “The SABC Submission on the Code for Persons with Disabilities Draft Regulations”. https://www.icasa.org.za/uploads/files/SABC-2018-submission-on-the-Draft-Code-for-Persons-with-Disabilities.pdf. 9 December 2021.

South African Broadcasting Corporation. 2020. *Annual report*. Auckland Park: SABC Corporate Publications.

South African Broadcasting Corporation. 2014. “Subtitles and Signing”. https://static.pmg.org.za/131010subtitling.pdf. 9 December 2021.

South African Social Attitudes Study. 2001-2018. https://hsrc.ac.za/special-projects/sasas/. 7 May 2021

Spaull, N., Mohohlwane, N. and Pretorius, E. 2020. “Investigating the comprehension iceberg: Developing empirical benchmarks for early-grade reading in agglutinating African languages”. *South African Journal of Childhood Education*, 10(1), 1-14.

Specker, E.A. 2008. “L1/L2 Eye movement reading of closed captioning: A multimodal analysis of multimodal use”. Doctoral thesis, The University of Arizona.

Stanovich, K.E. 1980. “Toward an interactive-compensatory model of individual differences in the development of reading fluency”. *Reading research quarterly*, 32-71.

Statistics South Africa, 2011. “South African Population Census”. https://www.statssa.gov.za/?page_id=3839. 15 July 2021.

Statistics South Africa, 2018. “Provincial profile: Eastern Cape Community Survey 2016” https://cs2016.statssa.gov.za/wp-content/uploads/2018/07/EasternCape.pdf. 24 October 2022

Statistics Kingdom. 2017. "Wilcoxon Signed Rank test calculator". https://www.statskingdom.com/index.html. 7 July 2022

Stein, P. 2008. *Multimodal Pedagogies in Diverse Classrooms: Representation, Rights and Resources*. USA & Canada: Routledge.

Stein, P. and Newfield, D. 2006. “Multiliteracies and multimodality in English in Education in Africa: Mapping the terrain”. *English Studies in Africa*, 49 (1), 1–22.

Stewart, M.A., and Pertusa, I. 2004. “Gains to language learners from viewing target language closed-captioned films”. *Foreign Language Annals*, 37(3), 438-447.

Stutchbury, K. 2021. *CRITICAL REALIST RESEARCH. Thinking Critically and Ethically about Research for Education: Engaging with Voice and Empowerment in International Contexts*. London & New York: Routledge

Stöckl, H. 2004 “In between Modes: Language and Image in Printed Media”. In Ventola, E., Charles, C. and Kaltenbacher, M. (eds.) *Perspectives on Multimodality*. Philadelphia & Amsterdam: Benjamins. 119–136.

Szarkowska, A. 2016. *Report on the results of an online survey on subtitle presentation times and line breaks in interlingual subtitling. Part 1: Subtitlers.* London

Szarkowska, A. and Gerber-Morón, O. 2018. “Viewers can keep up with fast subtitles: Evidence from eye movements”. *PloS one*, 13(6), p.e0199331.

Tamm, M. 1990. “The semiotic function: studies in children's representations”. Doctoral dissertation, Umeå Universitet.

Tan, L., and Guo, L. 2014. “Multiliteracies in an Outcome-Driven Curriculum: Where Is Its Fit?”. *The Asia-Pacific Education Researcher*, 23(1), 29–36.

Tan, S., O'Halloran, K.L., Wignell, P., Chai, K. and Lange, R. 2018. “A multimodal mixed methods approach for examining recontextualisation patterns of violent extremist images in online media”. *Discourse, Context & Media*, 21, 18-35.

Tan, S., O’Halloran, K. and Wignell, P. 2016. “Multimodal research: Addressing the complexity of multimodal environments and the challenges for CALL”. *ReCALL*, 28(3), 253–273.

Tan, S., Smith, B.A. and O’Halloran, K. 2015. “Online leadership discourse in higher education: A digital multimodal discourse perspective”. *Discourse & Communication*, 9(5), 559–584

Taylor, C.J. 2003. “Multimodal Transcription in the Analysis, Translation and Subtitling of Italian Films”. *The Translator*, 9(2), 191-205.

Taylor, G. 2005. “Perceived processing strategies of students watching captioned video”. *Foreign Language Annals*, 38(3), 422-27.

The Human Development Teaching & Learning Group. 2020. “Human Development”. https://pdx.pressbooks.pub/humandevelopment/. 25 October 2022

Thibault, P. 2000. “The Multimodal Transcription of a Television Advertisement: Theory and Practice”. In Baldry, A. (ed) *Multimodality and Multimediality in the Distance Learning Age*, Campobasso: Palladino Editore, 311-385.

Tiittula, L.M., Kurimo, M., Mansikkaniemi, A. and Rainò, P. 2018. “Ohjelmatekstityksen toimivuus eri kohderyhmien näkökulmasta”. *MikaEL Kääntämisen ja tulkkauksen tutkimuksen symposiumin verkkojulkaisu*.

Tomczak, M. and Tomczak, E. 2014. “The need to report effect size estimates revisited. An overview of some recommended measures of effect size”. *Trends in Sport Sciences*, *21*(1).

Tseng, C.I. and Djonov, E. 2023. “Children’s comprehension of time in audiovisual narratives: A multimodal discourse and empirical approach”. *Linguistics and Education*, 73, 101144.

Uhmann, S. 1992. “Contextualizing relevance: on some forms and functions of speech rate changes in everyday conversation”. *The contextualization of language*, 297-336.

Van Leeuwen, T. 1999. *Speech, Music, Sound*. Macmillan, London.

Van Leeuwen, T. 2005. *Introducing social semiotics*. Psychology Press.

Van Lommel, S., Laenen, A. and d'Ydewalle, G. 2006. “Foreign-grammar acquisition while watching subtitled television programmes”. *British Journal of Educational Psychology*, 76(2), 243-258.

Van Rooy, B. and Pretorius, E.J. 2014. “Is reading in an agglutinating language different from an analytic language? An analysis of Zulu and English reading based on eye movements”. *Southern African Linguistics and Applied Language Studies*, 31(3), 281-297.

Vanderplank, R. 1990. “Paying attention to the words: Practical and theoretical problems in watching television programmes with unilingualceefax sub-titles”. *System*, 18, 221-234.

Veii, K. and Everatt, J. 2005. “Predictors of reading among Herero-English bilingual Namibian school children”. *Bilingualism*, 8(3), 239.

Voeltz, F.E. and Kilian-Hatz, C. “Introduction”. In Voeltz, F.E. and Kilian-Hatz, C. (eds.) *Ideophones* (Vol. 44). John Benjamins Publishing. 1-9

Wa Kivilu, M., Morrow, S. 2006. “What do South Africans think about education?”. https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/325/download/4835. 18 January 2023.

Walsh, M. 2010. “Multimodal literacy: What does it mean for classroom practice?”. *Australian Journal of Language and Literacy*, 33(3), 211–239.

Webb V, Lafon M, Pare P. 2010. “Bantu languages in education in South Africa: An overview. Ongekho akekho! - the absentee owner”. *The Language Learning Journal*, 38(3), 273–292.

Wilmot, P.D. 1999. “Graphicacy as a Form of Communication”. *South African Geographical Journal*, 81(2), 91-95.

Wilsenach, C. 2013. “Phonological skills as predictor of reading success: An investigation of emergent bilingual Northern Sotho/English learners”. *Per Linguam*, 29(2), 17-32.

Wilsenach, C. 2015, “Receptive vocabulary and early literacy skills in bilingual Northern Sotho-English bilinguals”. *Reading and Writing*, 6(1), Article 77, 11.

Winke, P., Gass, S., and Sydorenko, T. 2010. “The effects of captioning videos used for foreign language listening activities”. *Language Learning & Technology*, 14(1), 65–86.

Winke, P., Gass, S., and Sydorenko, T. 2013. “Factors influencing the use of captions by foreign language learners: An eye-tracking study”. *The Modern Language Journal*, 97(1), 254–275.

Wolf, M. 2012. *Proust and the Squid*. London: Harper Perennial.

Xhosa Kids. 2019. “Qongqothwane | Knock-Knock Beetle Rhyme in Xhosa | Abantwana iingoma | kwimfundo yasesikolweni”. https://www.youtube.com/watch?v=7pRLNCmjces&ab_channel=XhosaKids. 15 December 2019.

Zanon, N. T. 2007. “Learning Vocabulary through Authentic Video and Subtitles”. *Tesol-Spain Newsletter*, 31, 5-8.

Zappavigna, M. 2011. “Ambient affiliation: A linguistic perspective on Twitter”. *New Media & Society*, 13(5), 788–806.

Zappavigna, M., Dwyer, P. and Martin, J.R. 2008. “Syndromes of meaning: exploring patterned coupling in a NSW Youth Justice Conference”. In Mahboob, A. and Knight,

N. (eds.) *Questioning Linguistics*. Cambridge: Cambridge Scholars Publishing. 164–184.

Ziegler, J.C. and Goswami, U. 2005. “Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory”. *Psychological bulletin*, 131(1), 3-29.

Zohrabi, M., Dobakhti, L. and Pour, E.M. 2019. “Interpersonal Meanings in Children's Storybooks”. *Iranian Journal of Language Teaching Research*, 7(2), 39-64.

Appendix A – Video Transcripts

Video 1[25]

The Icky Sticky Frog – ISele eLibi eliNcangathi

NguDawn Bentley

Imifanekiso nguSalina Yoon

Kwelinye ichityana elihle elizuba, phezu kwesikhondo esimdaka, kwakuthe ngcu usele-sele oluhlaza, ethe cwaka.

In some little blue lake, on a brown trunk, perched green Sele-sele, who remained silent.

Nantso impukane ibhabhela kufuphi. Sh-h-h!

Then came a fly flying nearby. Sh-h-h!

Wathula Sh-h-h! Cwaka usele-sele.

He kept quiet Sh-h-h! Silence Sele-sele!

Waman' ukuyithi krwaqu le mpukane, ibhabhela kufuphi.

He kept glancing at this fly, flying nearby.

LETSHE! Nalo ulwimi lukaSele-sele, luncangathi kwaye lulude.

SHOOT! [There is] that tongue of Sele-sele, sticky and long.

SHWAKA! Ayisabonakali ndawo impukane!

GONE! The fly can no longer be seen anywhere!

Njengokuba usele-sele eginya impukane, gqi ibhungane elimbejembeje lirhubuluza lisondela.

As Sele-sele swallows the fly, a colourful beetle appears, crawling coming nearby.

Sh-h-h! Cwaka usele-sele.

Sh-h-h! Sele-sele kept quiet.

Waman' ukulithi krwaqu eli bhungane lirhubuluza kufuphi.

He kept glancing at the beetle crawling nearby.

LETSHE! Nalo ulwimi lukasele-sele, luncangathi kwaye lulude.

SHOOT! [There is] that tongue of Sele-sele, sticky and long.

SHWAKA! Alisabonakali ndawo ibhungane.

[25] Hyperlinks to each video on Youtube are in each Video heading

GONE! The beetle can no longer be seen anywhere!

Usele-sele waliginya ibhungane, ngendlela efanayo naginya ngayo impukane, nantso ke nentethe eluhlaza isiza itsiba-tsibela kufuphi.

Sele-sele swallowed the beetle just as he swallowed the fly, now there comes the grasshopper jumping nearby.

Sh-h-h! Cwaka usele-sele.

Sh-h-h! Sele-sele kept quiet.

Waman' ukuyithi krwaqu le ntethe itsiba-tsibela kufuphi.

He kept glancing at this grasshopper jumping nearby.

LETSHE! Nalo ulwimi lukasele-sele, luncangathi kwaye lulude.

SHOOT! [There is] that tongue of Sele-sele, sticky and long.

SHWAKA! Ayisabonakali ndawo intethe!

GONE! The grasshopper can no longer be seen anywhere!

Usele-sele wayiginya intethe ebitsiba-tsibela kufuphi, ngendlela efanayo nale naginye ngayo ibhungane, naginye ngayo impukane, emva koko wabona, ibhabhathane elihle.

Sele-sele swallowed the grasshopper jumping nearby in the same way he swallowed the beetle, and the same way he swallowed the fly, after that he saw a beautiful butterfly.

Sh-h-h! Cwaka usele-sele.

Sh-h-h! Sele-sele kept quiet.

Waman' ukulithi krwaqu eli bhabhathane libhabhela kufuphi.

He kept glancing at this butterfly flying nearby.

LETSHE! Nalo ulwimi lukaSele-sele, luncangathi kwaye lulude.

SHOOT! [There is] that tongue of Sele-sele, sticky and long.

BIMBILILI! Akasabonakali ndawo usele-sele!

GULP! Sele-sele can no longer be seen anywhere!

Isiphelo.

The end.

Video 2

They all loved Sue - Bonke babemthanda USue (Since the story is about animals and not people, the appropriate translation would be Zonke zazimthanda uSue) 8

Ibhalwe nguCynthia Haslam Arnholt 11

Imifanekiso nguKathryn Vingi Gage 13

USue wayezithanda izilwanyana. 12

Sue loved animals.

Wayezithanda zonke iintlobo zezilwanyana, kodwa wayethanda iikati nezinja kakhulu. 30

She loved all kinds of animals, but she loved cats and dogs more.

USue wayekhathazekile kuba wayehlala kwigumbi eliqeshiswayo apho kwakungavumelekanga kubekho naziphi na izilwanyana zasekhaya. 50

Sue was sad/hurt because she lived in a rented room where pets were not allowed.

Iikati, izinja, amafudo, iintaka, izilwanyana zalo, naluphi na uhlobo zazingavumelekanga kweso sakhiwo sakhe. 42

Cats, dogs, tortoises, birds, all kinds of animals were not allowed in her building.

Kwakukho uphawu olukhulu olwalubekwe apho olufundeka ngolu hlobo; 26

IZILWANYANA ZASEKHAYA AZIVUMELEKANGA! 16

There was a big sign put up there that says/reads NO PETS ALLOWED!

Xa esiya emsebenzini, uSue wayesoloko ebonakalisa ububele kuzo zonke izilwanyana azibonayo. 40

On her way to work, Sue always showed kindness to all the animals she would see.

Ubekuthanda kakhulu ukuphulula izinja. 19

She loved patting/stroking dogs a lot.

Kwakukho izinja ezinkulu, ezincinane, ezingaqhelekanga, ezintle kunye nezinja ezimbi 34.

There were big, small, unusual, beautiful, and ugly dogs.

Ezinye izinja bezikhonkotha, ezinye zingakhonkothi.20

Other dogs barked, others didn't/ did not.

USue ubezithanda zonke. 9

Sue loved them all.

Kungekudala uSue waqalisa ukuphatha izimuncumuncu zezinja emsebenzini. 31

Soon Sue started carrying dog snacks to work.

Zonke izinja zazisuke zize kuye zibaleka ukuza kumbulisa xa zimbona. 28

All the dogs would just run to greet her when they saw her.

Kungekudala zonke izinja zasebumelwaneni bezimthanda uSue. 23

Soon all the neighbourhood dogs loved Sue.

Emva kwemini uSue uyewaqonda ukuba neekati ziyafuna ukukhathalelwa nazo. 31

In the afternoon Sue felt that the cats also needed to be cared for.

Ubekuthanda ukuwola iikati. 12

She loved cuddling cats.

Kungekudala waqalisa ukuphatha izimuncumuncu zeekati. 24

Soon she started carrying cat snacks.

Kwisithuba seeveki ezimbalwa, zonke iikati zabamelwane zamthanda uSue. 28

In a few weeks, all the neighbours' cats loved Sue.

Kusasa zonke izinja zabamelwane zazimlandela uSue xa esiya emsebenzini. 29

In the morning all the neighbours' dogs would follow Sue to work.

USue wayonwabile kakhulu. 10

Sue was very happy.

Emva kwemini, zonke iikati bezimjikeleza, zikwenze kubenzima ukuya ekhaya. 29

In the afternoon all the cats would surround her making it hard for her to go home.

USue wayonwabile kakhulu! 10

Sue was very happy!

Iindaba zavakala kuzo zonke izinja nakuzo zonke iikati ngoSue, kungekudala zonke izilwanyana zesixeko zazimazi uSue. 46

The news reached all the dogs and all the cats, soon all the animals in the city knew Sue.

Apho aya khona uSue, izinja neekati zazifuna ukuba lapho akhoyo uSue. 29

Wherever Sue went, dogs and cats wanted to be where Sue was.

Ngoko, nangona uSue ebengenazo izilwanyana zasekhaya ezizezakhe, ubenamawaka amaninzi ezilwanyana ezimthandayo.47

So, although Sue didn't/did not have pets of her own, she had thousands of animals that loved her.

USue wayonwabe kakhulu, kakhulu! 12

Sue was very very happy!

Nanjengoko uSue ubethandwa kakhulu zizo zonke izilwanyana zasekhaya esixekweni, ibiyinto eqhelekileyo ukuba achithe ixesha elininzi kunye nazo. 59

Since Sue was loved by all the pets in the city, it was a usual thing for her to spend a lot of time with them.

Ngenxa yoko uSue, wavula elakhe ishishini, elithi “IShishini leeNkonzo zokuGcinwa kweZilwanyana likaSue.” 39

Because of that, Sue opened her own business called “Sue’s Animal Keeping Services.”

USue wayenabathengi abaninzi. 13

Sue had many customers.

USue ubengoyena mntu ugcina izilwanyana zasekhaya kakuhle kuso sonke eso sixeko. 35

Sue was the most outstanding pet keeper in that whole city.

USue wayonwabe ngaphezu kwento yonke! 13

Sue was extremely happy!

Isiphelo. 4

The end.

Video 3

The Moon Followed me Home

INyanga iNdilandele ukuya eKhaya

Ibhalwe ngu-Elizabeth Bewley Imizobo nguMasumi Furukawa

Ndinomhlobo wam oyedwa (omnye) ophuma ebusuku kuphela.

I have one friend of mine who comes out only at night.

Akazange alahlekane nam yaye ukhangeleka emhle kakhulu.

He has never lost me/We have never lost each other, and he looks very beautiful.

Namhlanje, mna notata siyagoduka, ngelixa ilanga lisiya kutshona.

Today, my father and I are going home, as the sun sets.

Inyanga yam ithi molo kum; ayilibali.

My moon says hello to me; it does not forget.

Ngobunye ubusuku uliceba, into encinci.

On one night he is a small piece.

Apha ebengezelayo esibhaka-bhakeni.

Where he shines in the sky.

Nokuba ufihlakele kuloo mafu asibekeleyo, asiphosani.

Even when he is hidden under the clouds, we do not miss each other.

Ngamanye amaxesha inyanga yam inkulu, njengokuba injalo ngobu busuku.

Sometimes my moon is big, as it is tonight/this night.

Uyakhanya, uyabengezela kwaye uyamenyezela, uphanyazisa isibhakabhaka ngokukhanya.

You shine, you gleam and sparkle, you blind the sky with brightness.

Siye sithi naxa sihambela kude okanye kufutshane nekhaya, umhlobo wam unyanga usoloko endifumana nokuba sele sibhadula phi na.

Even when we are walking far or close to home, my friend, moon, always finds me wherever we wander.

Sithi xa sifika ngasekhaya, akukho mfuneko yokuba ndijonge ukuze ndazi ukuba umhlobo wam unyanga uhamba ngqo emveni kwethu.

When we get close to home, there is no need for me to look behind us in order for me to know if my friend, moon, is directly behind us.

Kum uyakusoloko eqaqamba.

To me, he will always shine bright.

Ngoku sele siza kufika ekhaya, utata undibalisela amabali apha endleleni.

Now we are about to get home, my father tells me stories on the way.

Usoloko endenza ndihleke, ndincume, ngalo lonke ixesha sincokola.

He always makes me laugh and smile every time we chat.

Kodwa umhlobo wam unyanga akathethi kangako.

But my friend, moon, does not speak much.

Umgumhlobo othuleyo.

He is a silent friend.

Ndicinga ukuba wonwabela nje ukukhanya.

I think he is just happy to shine.

Ndithembele ekukhanyeni kwakhe.

I trust in his light.

Ngaphaya kwefestile ndibona umama wam.

Through the window I see my mother.

Utata uyandiphakamisa ahambe nam ukungena ngaphakathi eyadini.

My father picks me up and walks with me to get inside the yard.

Konke kuzolile kuthe cwaka.

Everything is calm and quiet.

Ngelixa sivula ucango, ndiyazi ukuba asisodwa.

When we open the door, I know we are not alone.

Ndijonga phezulu esibhakabhakeni ndibone…

I look up at the sky and see…

Ukuba inyanga indilandele ukuya ekhaya.

That the moon followed me home.

Isiphelo.

The end.

Video 4

Iintshukumo Zasendle – Jungle Tumble

Ibhalwe nguPatrick Glendening

Imifanekiso nguRebecca Elliot

Izikhwenene zizifihla phezulu emithini.

Parrots hide themselves high up in the trees.

Tsala isebe lomthi…Uza kubona!

Pull a tree branch… You will see!

Isele liluhlaza okwegqabi.

The frog is as green as a leaf.

Yiyo lento lizifihla…ngaphantsi!

Therefore it hides itself… below!

Amachaphaza engwe ayayinceda ukuba ingabonakali.

The spots of a leopard help it to not be seen.

Kodwa ungabona…ukuba ikhona!

But you could see… that it is there!

Ilovane lingafana nawo nawuphina umbala womnyama, ukususela kumasebe amdaka kakhulu… ukuya kwingca eluhlaza yaka.

A chameleon can look like any colour of the rainbow, ranging from very dark brown branches… to very green grass.

Inyoka izisonga emagqabini.

The snake wraps itself around the leaves.

Xa imvula isina…ikhwela phezulu emithini.

When it rains… it climbs high up on the tree.

Ookrebe baphumla emigxobhozweni.

Sharks rest in swamps.

Imilomo yabo ivala…ngo "Kram!"

Their mouths close with a "Kram!" sound.

Amatye ayatsha lilanga elifudumeleyo.

Rocks/stones are burnt by the warm sun.

Kuphuma…Ufudwazana eziva kamnandi.

Little tortoise comes out feeling good.

Ingaba yinyoka emdaka le ihleka emthini?

Is that a brown snake laughing in the tree?

Hayi. Ngumsila we…nkawu.

No. It's a monkey tail.

Iintyatyambo ezimibala-bala zingafihla ibhabhathane… ixesha elide.

Multi-coloured flowers can hide butterflies… for a long time.

Umgqumo wehlosi uvakala okweendudumo.

The roar of a leopard sounds like thunder.

Lithi, "Ndiyakuthanda ukubukela iiNtshukumo zaseNdle!"

It says, "I like watching Jungle Tumble!"

Video 5

I can go to school – Ndingaya esikolweni

Ibhalwe nguDouglas Segal

Imifanekiso nguKay Widdowson

Ndiya esikolweni!

I am going to school!

Kusasa utata wam undisa eklasini yam.

In the morning my father takes me to my class.

Ndiziphathela ngokwam ityesi yesidlo sasemini!

I carry my own lunch box!

Sibulisa sithi, "molo" kutitshalakazi wam.

We greet and say "hello" to my teacher.

Ze ndibulise nabahlobo bam ndithi, "molweni!"

And then I also greet my friends and say, "hello all!"

Ndisoloko ndonwabile kukubabona.

I am always happy to see them.

Phambi kokuba ahambe utata wam uyandanga.

Before my father leaves, he hugs/kisses me.

"Ube nemini emnandi" atsho utata.

"Have a good day" father says.

Kusasa sihlala emethini simamele amabali.

In the morning we sit on the mat and listen to stories.

Ndiyathanda ukuhlala kwicala lemethi elinemibala yomnyama.

I like to sit on the side of the mat with rainbow colours.

Utitshala wam usixelela ngemisebenzi yemini, asifundele nebali.

My teacher tells us about the activities of the day, and she reads a story for us.

Senza nezinye izinto ezonwabisayo ezifana… nokuthetha ngemo yezulu, nokuba loluphi usuku esikulo evekini.

We do other fun things… such as talking about the weather and which day of the week we're in.

Ngamanye amaxesha sicula iingoma, okanye sabelane ngamabali obomi bethu.

Sometimes we sing songs, or we share stories about our lives.

Emva kwesidlo sasemini, sidlala phandle.

After lunch, we play outside.

Kukho oojingi, imityibilizi nendawo ethe gabalala yokudlala.

There are swings, slides and an open playing field.

Ndithanda ukudlala icekwa okanye ibhola nabahlobo bam.

I like to play tag or soccer with my friends.

Okanye ngamanye amaxesha ndithanda ukunyova ibhayisekile

Or sometimes I like to ride a bicycle.

Ndiyakwazi ukuzoba emgangathweni ngetshokhwe.

I know how to draw on the ground with chalk.

Emva kwemini siba nexesha lokusebenza ngokuzolileyo.

In the afternoon we have time to work calmly.

Ukudlala iphazile yeyona nto ndiyithandayo.

Building a puzzle is the thing I love the most.

Ngamanye amaxesha ndihlala esitulweni esitofo-tofo ndifune incwadi.

Sometimes I sit on the soft chair and read a book.

Ngamanye amaxesha ndizoba imifanekiso yosapho lwam.

Sometimes I draw pictures of my family.

Xa ndidiniwe ndingalala emthini otofotofo.

When I am tired, I can sleep on the soft tree.

Ukuphuma kwesikolo, umama uyandilanda.

After school, my mother picks me up.

Ndisoloko ndivuya ukumbona.

I am always happy to see her.

Ndithi, “usale kakuhle” kutitshalakazi nakubahlobo bam bonke.

I say “stay well” to my teacher and all my friends.

Ndifika ndibonise umama wam omnye umsebenzi endiwenzileyo.

I always get to my mom and show her some of the work I did.

Usoloko ezingca ngam.

She is always proud of me.

Ndiyakuthanda ukumbalisela ngemini yam, nangabahlobo bam nezinto ezonwabisayo endizenzileyo.

I like telling her about my day, my friends and the things that are enjoyable that I did.

Ndiyayilangazelela imini elandelayo esikolweni!

I am looking forward to the next day in school!

Isiphelo.

The end.

Appendix B – Tests

Section A

FIHLA

VALA

IKATI

IKHAYA

INJA

HLALA

HAMBA

INTETHE

CULA

ISELE

Section B

1.

mama	tata	atsho	cula	kuba
kati	thetha	senza	funa	cinci
bona	apha	hlala	jonga	atsho
ndiya	thanda	inja	hamba	kwethu
siza	kodwa	khaya	engwe	cinga
hayi	vala	intethe	tsala	ibali
wabona	izinto	imini	umhlobo	fihla
zuba	phezu	kakhulu	sila	phumla
nantso	kusasa	isele	fika	sithi
thula	cwaka	naxa	isikolo	funda

2.

umfo	bhuti	sisi	sana	zathi
kanti	vuka	lila	Iti	wabo
vasa	luma	bomvu	mali	luma
moto	sela	bisi	yena	jezi
wona	fama	cawa	wona	jika
qala	zisa	imoto	lilela	ipali
phunga	wavuka	udade	njani	igusha
ubisi	imali	kuqala	ihagu	isofa
ikofu	wabeleka	qoqa	sonka	vota
buza	xoxa	zola	phila	blowu

	Igama kumgca ngamnye
ULolo wayeza kuqala ukuya kwisikolo esikhulu.	***6***
Lwalungathi alufiki olu suku yimivuyo.	***5***
Ngomhla wokuvulwa kwezikolo wavuka	***4***
ekuseni. Wahlamba waza watya isidlo	***5***
sakusasa.	***1***
Emva koko uLolo wabeleka ubhaka wakhe	***6***
waya esikolweni nomama wakhe. Esikolweni	***5***
waboniswa utitshalakazi wakhe.	***3***
Utitshalakazi wabalisa ibali likaSidumo.	***4***
USidumo libali elidlala kumabonakude. ULolo	***5***
wahleka izinto ezenziwa nguSidumo. Wabona	***5***
ukuba kumnandi esikolweni esikhulu.	***4***
	(***53 amagama***)

Amagama achazwe kakuhle ngemizuzu 60 ☐