Proceedings of the International Conference of the Digital Humanities Association of Southern Africa (DHASA) 2021

Virtual Conference

29th of November - 3rd of December 2021

With special thanks to our sponsors

Peta sponsor

Tera sponsor and best paper award sponsor

Videolectures.net and audio-visual sponsor

Swearing in South Africa: Multidisciplinary research on language taboos

Van Huyssteen, Gerhard B
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Abstract

Research on swearing (used here as a hypernym to include other phenomena and/or synonyms like cursing, profanity, taboo language, etc.) has been prevalent for many years internationally, also from a variety of scientific disciplines. Most of the research literature, however, is on swearing in English, although studies have also been conducted on some other languages. By contrast, very little to no research has been done on swearing within the South African context, which is quite surprising, given that using certain swearwords (i.e., racial slurs) is punishable by law.

To address this void, we established a multidisciplinary project with its primary roots in the digital humanities, and with inputs from and implications to (amongst others) linguistics, literary studies, communication studies, neurology, psychology, sociology, computer sciences, and law. This project (and specifically the topic of swearing) holds the potential to provide insights in human cognition and social interaction, while situating it broadly within the scope of the Fourth Industrial Revolution. The project commenced in July 2019, and is currently ongoing.

In this paper, we firstly provide a rationale for the project, before introducing each of the five subprojects. These subprojects pertain to swearing and the law; a swearing constructicon (a kind of online dictionary) for Afrikaans; swearing in the entertainment world and in the media; swearing as a linguistic innovation; and an end-user facing project website. We also report on some of the outputs from the project that are already available, and others that are still being developed and investigated. We conclude with a brief overview of some of the potential impacts of the project.

Keywords: censorship, computational linguistics, cursing, language change, taboo

1 Introduction

Swearing is a fascinating phenomenon that not only gives us deep insights in human cognition and neurophysiology, but also in social interactions and power dynamics. However, very little multidisciplinary research has been done on swearing in the South African context – a lacuna that the project *What the Swearword?!* (WTS) aims to fill with insights from the digital humanities, and with inputs from and implications to linguistics, literary studies, communication studies, psychology, neurology, sociology, computer sciences, and law. The project commenced in July 2019 with a three-year set-up and exploratory phase (focusing only on Afrikaans, and other languages in its ecosystem – including other Germanic languages), ending in June 2022. Thereafter, the project will continue in directions determined by the interests of the multidisciplinary team members, and depending on the availability of funding.

The following types of (popular) questions are of interest to researchers in the project:

- If a website contains swearing, what legal obligations does the owner/developer have?
- Should parents protect their children from hearing swear words?
- What is the best way to determine objective offensiveness ratings for swearwords, e.g., to determine advisories for films and/or books?
- How does it happen that an Afrikaans word like *be·fok* (a verbalized form of *fuck*) can mean, among others, both 'good' (as in *Dit was nou befok gewees!* 'That was really fucking A'), and 'angry' (as in *Hy is al weer befok!* 'He is once again fucked off!')?
- How is swearing used as a linguistic innovation that causes short-term and/or rapid language change?
- What are the views on swearing of writers, dramatists, poets, TV and film makers,

producers, directors, actors, musicians, editors, journalists, podcasters, bloggers?

- How and why do these content creators apply self-censorship with regards to swearing? What is the impact of cancel culture on their language usage in the content they create?
- What is the interaction between swearing and societal change?
- What is the neurological impact when someone hears a racial, homophobic, or sexist slur?

In addition to the primary focus on swearing, the project also has a secondary, subjacent aim, namely, to investigate alternative, contemporary opportunities of scholarly communication, specifically focusing on podcasts, blogs, videos, and webinars. Traditional main-stream outlets for communicating research results, i.e., monographs, edited books, journal articles, conference publications, and presented talks and posters, are by and large still the only research outputs that carry weight in academic appointments and promotions, and in the national and international evaluations of universities. This is especially true for the humanities and social sciences, and even more so in the South African context. A fundamental (albeit radical) presupposition of this project is that these main-stream outlets for communicating research results are already outdated and will become even more outdated and less appropriate in a technologically revolutionized society [1]. We therefore aim to experiment with how to incorporate and integrate peer-reviewing in new communication channels (to ensure quality); how to optimize such means to stimulate multidisciplinary interest and foster new collaborations; and how to use these channels to enable and fast-track research (e.g., increasing respondent participation).

The aim of this paper is twofold: (1) To provide rationales for each of the subprojects; and (2) To report on some of the outputs and milestones of the project after two years of research and development. The overarching theme is that the digital humanities afford one with even more opportunities to stimulate multidisciplinarity in and outside the humanities. In the next section, we give a brief overview of previous research on the topic, indicating that there is a lacuna in knowledge on, and understanding of swearing in the South African context. In Section 3, each of the five subprojects are introduced, while we report on some of the outputs in these subprojects in Section 4. We conclude with a brief perspective on some of the other benefits and impact of the project.

2 Multidisciplinary research on swearing

For many decades, swearing (used here as a hypernym to include other phenomena and/or synonyms like cursing/cussing, profanity, blasphemy, obscenity, vulgarity, verbal abuse, verbal sparring, (racial) slurs, terms of abuse, insults, dirty language, and taboo language) has been researched internationally from various disciplines, including literary studies, journalism and communication studies, psychology, sociology, law, philosophy and ethics, cultural anthropology and history, pediatrics, neurology and other neurosciences. In linguistics specifically, studies range from comparative etymology, lexicology and lexicography, typology, and grammar, to first- and second-language acquisition, variation studies and dialectology, and sign-language, gestures and kinesics. Interdisciplinary research is often conducted within the fields of sociolinguistics, psycholinguistics, computational linguistics, and neurolinguistics. It is true that most of the literature is on swearing in English, although studies have also been conducted on many other languages, such as Cantonese, Danish, Dutch, Finnish, French, Italian, Japanese, Latin, and Russian, amongst many others. The titles of a few seminal and/or recent books serve to illustrate: *The Oxford Handbook of Taboo Words and Language* (Allan 2019); *Advances in swearing research: New languages and new contexts* (Beers Fägersten & Stapleton 2017); *What the F – What swearing reveals about our language, our brains, and ourselves* (Bergen 2016); *Why we curse: A neuro-psycho-social theory of speech* (Jay 2000); *Nine nasty words: English in the gutter: then, now, and forever* (McWhorter 2021); *Offensive Language: Taboo, offence and social control* (O'Driscoll 2020); *Linguistic Taboo Revisited: Novel Insights from Cognitive Perspectives* (Pizarro Pedraza

2018); and *Rot lekker zelf op: Over politiek incorrect en ander ongepast taalgebruik* (Van Sterkenburg 2019).

By contrast, very little to no research has been done on swearing within the South African context, which is quite surprising, given that using certain swearwords (i.e., racial slurs) is punishable by law. Most of the linguistic research has focused on the lexicographic treatment of swearing (e.g., Dekker 1991; Van Huyssteen 1998), while only a handful of studies focused on grammatical aspects of swearing (e.g., Calitz 1979; Feinauer 1981; Van Huyssteen 1996). Most recently, Van der Walt's (2019) MA dissertation at the North-West University (NWU), had a section on swearing as part of her analysis of Zefrikaans (an informal variety of Afrikaans). In other fields, research has also been sparse; for example, in Coetzee's 2018 article on children's swearing in multilingual contexts, there are only three references to other (socio)linguistic research that has been conducted in the South African context.

To address the lacuna in knowledge on, and understanding of swearing in the South African context, we conceptualized five initial subprojects; the rationale for these is discussed in the next section.

3 Subprojects

3.1 A: Swearing and the law

The South African Film and Publication Board (FPB) regulates age restrictions on films, computer games, and publications that don't fall under the jurisdiction of the Press Ombudsman, which are released/published in South Africa. One of their criteria relates to what they call "strong language", which is defined as "crude words, threats, abuse, profanity or language that amounts to prejudice" (Republic of South Africa 2019). They will add the label "L" to a film, computer game or publication to alert users that there is use of strong language "of a mild, moderate, strong or very strong impact". However, this offensiveness scale is nowhere operationalized.

Following from this, several questions arise (to mention but a few):

- Can these categories of the FPB be predicted automatically (e.g., through machine learning algorithms)?
- Should adults and children be treated differently regarding swearing? Is swearing considered "adult/mature content", or simply as "explicit content"?
- These guidelines refer specifically to films and computer games, but what about other media, such as websites, literary texts, memes, songs/lyrics, and podcasts with swear words? Should these also carry content advisories? What are end-users' (e.g., parents) expectations about such advisories?
- What about swear words/text linked to images, videos and/or sound? For example, what about swearwords in lyrics and music videos?
- Given the history of censorship in South Africa (Van Rooyen 2012), how should we balance freedom of speech and freedom of choice, vs. protecting the citizens (e.g., children) of South Africa?

3.2 B: *Vloekepedia*: An encyclopaedic constructicon of Afrikaans swearing

Dictionaries and encyclopedias of swearing in English, Dutch, Spanish, Cantonese, Russian, etc. abound, none exists for Afrikaans and/or other indigenous South African languages. In addition, many of the dictionaries and encyclopedias for other languages are not authoritative, but mainly presented as popular entertainment (with two notable exceptions: Hughes (1991), and Sheidlower (2009), with the latter restricted to only the word *fuck* and its compounds and derivations). To address this lacuna for Afrikaans, we commenced to compile an encyclopaedic constructicon of Afrikaans swearing, called *Vloekepedia*.

Theoretically, the *Vloekepedia* will be underpinned by cognitive construction grammar, specifically as a constructicon, which is "a theoretical conception of language as a structured inventory of constructions, and … a collection of construction descriptions, essentially a practical instantiation of the former concept" (Lyngfelt *et al.* 2018:1). The

idea of a constructicon as a "dictionary of constructions" was first suggested by Fillmore *et al.* (2008), which subsequently lead to constructicon projects for Brazilian Portuguese, German, Japanese, Russian, and Swedish. Constructicography is a blend between construction grammar and lexicography, with the aim to compile a practically usable descriptive resource of lexical, morphological and/or syntactic constructions.

A central tenet of cognitive construction grammar is that it is usage-based, i.e., the view that constructions are generalisations over specific, real-world instances, based on, among others, frequency and salience. To identify and describe constructions, methods from corpus linguistic and/or psycholinguistic are most often used; this approach is therefore in its very essence suitable for multidisciplinary research.

Another important principle of cognitive construction grammar is its view of semantics being encyclopaedic, i.e., that meaning cannot be captured by means of a (lexical) definition only. Instead, usage patterns, pragmatics, associations, inferred knowledge, cultural importance, etc. are all part of the conceptual "meaning" of words and expressions. It is admittedly difficult (if not impossible) to capture such vast knowledge of constructions in the form of a (linear, linguistic) dictionary, but one could at least attempt to include elements such as real-world examples, frequency-based collocations, extensive pragmatic tags, mixed media, related information from other languages in the ecosystem, etymological information, etc.

One particular type of encyclopedic information that we are focusing on (also in relation to subproject A), is the rating of swear words and expressions on a taboo scale. To obtain offensiveness ratings for words has been done for a few languages (see Beers Fägersten (2007; 2012) for an overview), but never before for Afrikaans.

3.3 C: Swearing in the entertainment world and media

One of the landmark cases in censorship in the South African context, was the banning of *Magersfontein, O Magersfontein!* (Leroux 1976) in 1977. The main arguments for banning the book were based on the language in the book: "… excessive foul language, excessive vain use of the Name of the Lord, vulgar references to defecation, masturbation, loss of virginity, prevention of conception by rinsing with soap, menstruation, genitals and prostate trouble …" [translated] (Leroux 1990). Much has changed since the fierce grip that the Film and Publication Board had on South African entertainment and media in the 1970s and 1980s.

To our knowledge, no focused research has ever been done (also not recently) on why and how content creators use swearing in entertainment and the media. In this subproject, we therefore investigate the views on swearing of content creators in entertainment and the media (e.g., writers, dramatists, poets, TV and film makers, producers, directors, actors, musicians, editors, journalists, podcasters, bloggers). Of special interest, is how they are potentially impacted by the current cancel culture (as a form of social censorship).

3.4 D: Swearing, linguistic innovation, constructionalisation, and language change

Linguistic innovation (a.k.a. linguistic creativity) as an instigator of language change has been studied widely in linguistics. With regard to linguistic innovation in the 21st century, Paradowski & Jonak (2012) note that "[e]rstwhile research on language evolution and change focused on large timescales, typically spanning at least several decades. Nowadays, observable changes are taking place much faster. According to the Global Language Monitor (2009) a new English word is born roughly every 98 minutes …" Analyses of linguistic data from so-called Web 2.0 sources (e.g. blogs, microblogs, social media, and comments on websites) potentially provide us with insight into complex, dynamic systems, including "society, variations and typology, the rise of new grammatical constructions, semantic bleaching, language evolution in general, and the spread and competition of both individual expressions, and entire languages …" (Paradowski & Jonak 2012).

For example, in a post on Facebook on 7 April 2019 the user Don Dapper commented on a

Figure 1: The "what in the X fuck" construction

fashion photo of a person wearing accordion-like attire: "What in the accordion FUCK is this???" (See Figure 1). Two days later, someone in a WhatsApp group commented on a picture of the elevation profile of a half-marathon (see Figure 1): "What in the steep cliff FUCK are they talking about!!". The first expression could be considered a syntactic extension of the expression *what the fuck*, which in itself could be considered a syntactic innovation – i.e., *what the X* is only used in contexts where X could be filled with a swear word (or euphemism). Similar swearing-specific constructions could be observed in Afrikaans (e.g., *de X in*, as in *de moer in*, *de bliksem in*, *de fok in*, etc.), Dutch (e.g. *krijg de X*, as in *krijg de tyfus*, *krijg de rambam*, *krijg de pokken*, etc.), or English (e.g. *by X!*, as in *by God!*, extended to *by Toutatis!* or *by Jupiter!* in the Asterix comic book series).

This subproject has the strongest linguistic focus of all subprojects, since we investigate morphological and syntactic constructions that are specific to the domain of swearing, i.e., part of a swearing constructicon. One of our main interests is how new constructions are continuously added to the constructicon via the process of constructionalisation (Traugott & Trousdale 2013). In this regard we also focus on the role of cross-linguistic constructionalisation (Höder 2018), specifically focusing on Afrikaans, English, Dutch (and potentially French with relation to Flemish Dutch). For example, is the above-mentioned *de X in* construction the source for Afrikaans *wat de fok!*, or is it rather the result of transfer from English *what the fuck!*? Or is it a combination of both? This part of the research is not only relevant to the swearing domain, but also more generally to language change (in contact situations).

The importance of traditional social networks as a determining factor in language change has been accepted widely in linguistics (see Labov (2001), for instance). In recent years, the role of modern social networks (in the form of social media) has gained prominence in research on rapid linguistic change (e.g., Goel *et al.* 2016). The basic idea is that linguistic innovations can potentially gain momentum in speech communities more rapidly and widespread through social media, than is the case in traditional social networks and through traditional media. By analyzing unedited linguistic data from social media (e.g., Twitter, Facebook, Reddit, or comments on blogs, newspaper articles, etc.), we can therefore potentially observe language change "as it happens".

3.5 E: *Vloekcoza*: project website and social media presence

It is not uncommon for research projects to have independent websites with unique, easy-to-remember URLs. We have therefore set up a secure, technology-rich, end-user facing project website, vloek.co.za, as a means to create awareness of and cultivate new collaborations on the project, to publish outputs from the project, and to create a platform where registered users can participate in the above-mentioned surveys. In order to create a wide awareness of the website, we have also created project pages on Facebook, Twitter, Instagram, and Pinterest. Our main focus, however, is on Facebook, where an additional group, *Vloek*, has been established. This group serves as the first stop to gather information and data from end-users, as well as to disseminate information.

4 Progress and outputs

4.1 Subproject A (law)

Since the one of overarching questions of this subproject is how to classify Afrikaans swearwords according to the categories identified by the South African FPB, the main output of this subproject is the *Vloekmeter* ('swearing meter'; see

Figure 2: Vloekmeter showing results for "fokken" ('fucking') and "frieken" ('fricking')

vloek.co.za/vloekmeter). The *Vloekmeter* is purely data-driven: Based on data from single word surveys (SWSs), statistics are presented on an interactive dashboard on the website (see Figure 2). In each SWS, only one swearword is presented to registered participants. The aim with SWSs is to keep each one as short as possible, in order to prevent respondent fatigue (Lavrakas 2008). The assumption is that one would cover more words over a period of time, than if one were to present the same number of words to participants in a single session. As of 15 August 2021, 51 such SWSs have been posted, with a total of 6 243 responses (an average of 122.4 responses per SWS). These results have already been used for research on statistic modelling in the digital humanities ([REFERENCE 1 REMOVED]), as well as lexicology studies ([REFERENCE 2 REMOVED]).

This subproject also provided the impetus for two master's degree students currently working on their dissertations. One of the students (Mart-Mari van der Merwe; University of Pretoria (UP)) is identifying the 50 most prototypical Afrikaans swearwords, in order to obtain offensiveness ratings / taboo values for them. Another student (Colette Combrink; (NWU)) focuses on cancel culture as a form of social censorship, and how it impacts on a variety of writers and authors.

At the beginning of 2021, we (in collaboration with Maroela Media, the largest Afrikaans online news publication, and WatKykJy, a very free-thinking Afrikaans blog site) have conducted a large-scale survey to determine what the attitudes of adult Afrikaans speakers are towards content advisories for films and books (e.g., indication of suitability for certain age groups, themes covered, etc.). These results are currently (as of 15 August 2021) being processed and interpreted, and will be published during 2021/2.

4.2 Subproject B (*Vloekepedia*)

Until now, this subproject has focused only on data collection, and more specifically on lexical items (i.e., words, rather than phrases and expressions). A core lexicon of 711 words has been compiled in 2019/20, mainly based on data from WatKykJy. It was supplemented with data crawled from UrbanDictionary, resulting in 131 additional usable entries. Both datasets were manually curated by a student assistant.

In 2021, the *Woordeboek van die Afrikaanse Taal* (WAT), *Handwoordeboek van die Afrikaanse Taal* (HAT) and Centre for Text Technology (CTexT) of the NWU agreed generously to supply the project with relevant material from their respective databases. This data was amalgamated with the above-mentioned data, to construct a single

database consisting of 3,858 entries (as of 15 August 2021). Subsequently, one of the computational linguists on the project (Jaco du Toit) wrote a complex script to not only retrieve frequencies for all entries from all available corpora on VivA's Corpus Portal (VivA 2021), but also to extract all examples where these entries occur. This resulted in a database of 273 MB, containing more than 3,5 million sentences. This database needs to be curated, which in itself will be a gigantic task. Currently a master's student (Mart-Mari van der Merwe) is working on solutions to clean-up at least a portion of the data. Work will continue into the foreseeable future.

4.3 Subproject C (entertainment/media)

In addition to work already mentioned under subproject A, work in this subproject has focused by and large on the production and release of the podcast series *Wat de Vloekwoord?!*. This is a podcast series that explores the views and attitudes of content creators in the entertainment world and media on swearwords and taboo topics. Through interviews with well-known (Afrikaans) writers, TV and filmmakers, directors, actors, musicians, editors, journalists, podcasters and bloggers, we explore censorship in South Africa, what the function of swearing is, how viewers and listeners respond to swearwords, and so on.

The first episode of the first season was launched on 4 September 2020; the fourteenth (and last) episode of the first season was published on 18 December 2020. The series was co-hosted by psychologist me and Elmarie Claassens (clinical psychologist), and was technically produced by Gifford Peché (Decibel Studios). The first season consisted of interviews with prominent figures in the South African entertainment and media industry, including Anton Goosen, Amanda Strydom, Claire Johnson, Neil Sandilands, and Hunter Kennedy (to name but a few).

On Anchor.fm (the platform where the podcast is hosted), these fourteen episodes have been played a total number of 2,704 times (an average of 195 times per episode, as of 15 August 2021). Planning for a second season of fourteen episodes in a different format has commenced. The second season should launch in September 2021.

In collaboration with Afrikaans.com, a campaign related to this project was run from August till November 2020. This project not only created awareness of the project (with a significant increase in the number of registered users), but also promoted three questionnaires related to swearing in entertainment and the media. In addition, five blogs by renowned journalists (and one student) have been published on vloek.co.za. All of these blogs centered around the theme of swearing in the media, including Afrikaans music, radio, and newspapers.

4.4 Subproject D (linguistics)

Being one of the central subprojects of this project (since the main project's focus is on a phenomenon that manifests in language usage), and since many of the members are trained and/or practicing linguists, it is expected that this subproject will be the long-term focus of the main project. Hence, this is also the subproject where most of the fundamental "thinking" about directions for the other subprojects happens. Despite its central role, it is however the subproject with the least number of outputs to date, but the one with the most important outputs (in my personal opinion).

The first important output that this subproject directly lead to, is the establishment of an honors-level course in linguistics, called *Pornolinguistics: Swearing and other language taboos in cognitive neurosciences*. This course was conceptualized as a collaboration initiative between the departments of Afrikaans (and Dutch) at NWU and UP, re-utilizing existing course modules at both institutions. Virtual teaching and learning – due to the Covid-19 pandemic – played a central role in establishing the course, since it became more "natural" for students from two universities to be in the same virtual classroom, while it also afforded the opportunity to involve many other experts to teach specialized sections of the course.

The course was designed around four disciplines, with specific themes in each of these (see Table 1). Aside from lecturers from the above-mentioned departments of Afrikaans (and Dutch), lecturers also included computational linguists from the NWU's CTexT; a pediatric neurologist,

Table 1: Honors module

Discipline	Theme
Introduction	• What is swearing and language taboos? • What is cognitive neurosciences?
Linguistics	• The constructicon and constructicography • Construction grammar • Constructionalization and subjectification • Methodology: Sociolinguistics • Methodology: Corpus linguistics
Computer sciences	• Artificial intelligence • Sentiment analysis • Hate speech recognition
Neurology	• Neuroanatomy and language • Coprolalia and other disorders • Neuro-imaging
Psychology	• Emotion • Language acquisition

psychologist, and speech therapist from NWU's Centre for Health and Human Performance (CHHP); and a clinical psychologist in private practice.

Five students enrolled for the course in 2021, while a number of guests also joined the classes per occasion. As part of the course outcomes, students have to write blogs, popular articles, a research proposal, and a conference presentation (among others). Many of these outputs will be submitted for publication towards the end of 2021.

The course will continue in 2022, but with two additional opportunities:

(1) All lectures will be presented as public symposia, in order to enable external people to also attend these lectures.

(2) The linguistics section will also be attended by students from the University of Leiden (The Netherlands) who are enrolled for a postgraduate course in Afrikaans linguistics.

The second output of this subproject, is the establishment of a research discussion group on construction grammar and constructionalization, consisting of linguistics researchers from NWU and UP. Members of the group meet virtually for a weekly discussion session of 90 minutes, where they attend online courses together, discuss recent publication, and work together on research outputs. Two presentations at international symposia during October 2021 have already been accepted, while the first scholarly publications from this group is scheduled for submission before the end of 2021.

4.5 Subproject E (Vloekcoza)

Setting-up, designing, developing and implementing the project website and associated social media pages, took up most of the financial and other resources during the first twelve to eighteen months of the project. Since the project website is meant to be a fully functional, secure, technology-rich, end-user facing product, it was of utmost importance to ensure that it is a secure platform, is able to handle traffic, can accommodate various kinds of posts, is easy to maintain by non-technical people, can work well on mobile devices, etc.

In addition to the main functionalities of the website, a complete end-user facing, online booking system for the podcast series have been developed by BlueTek Computers. This system enables the interviewers to interact in a professional and systematic way with podcast guests, specifically to make bookings for online or personal interviews, obtain official permission for release of podcasts from guests, etc.

In our assessment, the initial investment in this subproject was well-worth the time and money. One of the best dividends is that the project has a dedicated platform to host a variety of surveys for data collection; as was mentioned in 4.1, we have already been able to publish more than 50 surveys, with more than 120 responses on average per survey (also see Table 2). In some disciplines this

Table 2: End-user interactions of Vloekcoza

Interactions	2019/20	% increase	2020/21	% increase
Registered users on website	1 434	100%	2 075	31%
Questionnaires	30	100%	51	41%
Responses to questionnaires	4 801	100%	6 243	23%
Facebook group: Members	553	100%	708	22%
Facebook page: Likes	264	100%	359	27%
Facebook page: Followers	281	100%	385	27%
Instagram: Followers	158	100%	218	28%
Twitter: Followers	13	100%	21	33%
Pinterest: Followers	4	100%	25	84%
Pinterest: Engaged audience	0	0%	659	100%
Anchor.fm: Plays	0	0%	2 704	100%

might be deemed a small response rate, but for the kind of data we are collecting, it is quite substantial. For the sake of comparison, for similar research conducted by Beers Fägersten (2012), she only used 60 respondents from one university campus.

Table 2 also presents some of the other interactions with end-users. The percentage of increase for the first period (September 2019 to August 2020) is a 100% in all cases, since the project started with zero interactions. A slow, but steady growth can be observed for the second period of reporting (September 2020 to August 2021). We are confident that this trend will continue in the years to come.

5 Conclusion

In addition to the above tangible outputs, the project also have (potential) impact in other ways:

- In addition to creating opportunities for post-graduate students, the project has also created part-time job opportunities for student assistants (one per year), and a web and social media editor (one per year).
- Since one of our secondary aims is to foster collaboration outside our "usual" disciplines and networks, the project has already shown its potential to create such new opportunities. We hope that this will increase substantially in future, with collaboration with even more disciplines, more institutions, and more countries.
- All the data and corpora that have been, are being, and will be developed during the course of this project, will be made available for distribution under an open-source license by the South African Centre for Digital Language Resources (SADiLaR), so that it could eventually be utilized in many other follow-up or competing projects.
- Given the priority of the development and integration of new and emerging indigenous ICTs, as well as an exponential rise in interest in artificial intelligence research and development, this project stimulates the conceptualization, design, development and implementation of new resources and technologies (at least for Afrikaans, until now). We believe that it holds the potential to also attract, expand, and support research in the digital humanities as part of the process of building South Africa's information society.

Notes

[1] To illustrate this presupposition with two examples: (1) South African's National Skills Fund (NSF) CEO Mvuyisi Macikama said in October 2018 that the NSF has a target of training 30,000 artisans a year by 2030, and that students in the social sciences and humanities do not contribute to the job market. He argued that funding should therefore be channeled away from "soft degrees" offered by universities (MyBroadband 2018). Although we do not subscribe to this view in any

possible way, it does illustrate the perception that social sciences and humanities are irrelevant in the South African context. Perhaps if research results from these disciplines have been more visible (and "digestible"), perceptions like these could be changed over time.

(2) In his now widely known (albeit controversial) article, Meho (2007) stated: "It is a sobering fact that some 90% of papers that have been published in academic journals are never cited. Indeed, as many as 50% of papers are never read by anyone other than their authors, referees and journal editors." These statistics have been challenged and re-evaluated by numerous other scholars, but Remler (2014) still concludes that "[a]cademic publishing needs fixing", especially since more than 80% of articles in humanities are never cited. This trend to "fix" academic publishing is seen in numerous other forms, including a strong drive towards open-access publication. For example, in March 2019 the University of California cancelled its subscription to Elsevier, the world's largest publisher of academic journals, as part of their crusade to transform scholarly communication.

If one looks more closely at linguistics, and specifically at linguistics in (South) Africa, one could for example note that on the renowned Scimago Lab's list of journals (www.scimagojr.com), only six linguistics journals from the continent appear with a Scimago Journal Rank (SJR) indicator, with *Lexikos* rated highest (SJR=0.280 for 2020), and *Stellenbosch Papers in Linguistics Plus* lowest (SJR=0.104 for 2020). To take the latter for illustration purposes, it means that 48 articles were published between 2017-2019, and these articles were only cited 8 times up to 2020. One of the most renowned international linguistics journals, *Language*, published 139 articles between 2017-2019, and these were cited only 240 times up to 2020.

In all honesty, one should take cognizance of the fact that, in contrast to publications in the natural sciences with relatively quick turn-over times, humanities journal articles are typically cited over a longer period of time. In addition, citations in books are often counted to a limited degree (or not at all), and this potentially has some effect on impact and evaluations. Nonetheless, these quoted figures illustrate a general tendency regarding publications in linguistics and the humanities specifically.

Acknowledgements

This research is partially funded by the Suid-Afrikaans Akademie vir Wetenskap en Kuns, and partially made possible through barter agreements with BlueTek Computers, Afrikaans.com, and WatKykJy.co.za. The *Woordeboek van die Afrikaanse Taal* (WAT), *Handwoordeboek van die Afrikaanse Taal* (HAT) and Centre for Text Technology (CTexT) of the North-West University are hereby also acknowledge for generously supplying the project with material from their respective databases.

Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the NWU (ethics number: NWU-00632-19-A7). Additional ethics clearance for one of the master's students was obtained from the Faculty of Humanities (UP), with reference number 16002360 (HUM017/0920).

The author is a director of the not-for-profit company Viridevert NPC (CIPC registration number: 2016/411799/08), who owns and manages the website vloek.co.za. This website was developed specifically for this project, and this conflict of interest has been approved by NWU.

I would like to acknowledge the inputs of Liesbeth Augustinus and Peter Dirix (Catholic University of Leuven, Belgium) in the initial conceptualization of this project, as well as Suléne Pilon (UP) in the ongoing re-conceptualization of the project. Thanks also to all the coworkers, collaborators, and students on the project; the list is too long to publish here, and it is therefore published on vloek.co.za/oor-ons.

None of the results and/or opinions in this paper can be ascribed to any of the people or organizations mentioned above.

References

Allan, K (ed.) 2019, *The Oxford Handbook of Taboo Words and Language*, Oxford University Press, Oxford.

Beers Fägersten, K 2007, *A sociolinguistic analysis of swearword offensiveness*, Universität des Saarlands, Saarbrücken, view 22 August 2021, <https://www.researchgate.net/publication/265009714_A_sociolinguistic_analysis_of_swearword_offensiveness>.

Beers Fägersten, K 2012, *Who's Swearing Now? The Social Aspects of Conversational Swearing*, Cambridge Scholars Publishing, Newcastle upon Tyne.

Beers Fägersten, K & Stapleton, K (eds.) 2017, *Advances in Swearing Research: New Languages and New Contexts*, John Benjamins, Amsterdam.

Bergen, BK 2016, *What the F: What Swearing Reveals About Our Language, Our Brains, and Ourselves*, Basic Books, New York.

Calitz, FC 1979, Spot, skel en verwante verskynsels in Afrikaans [Mockery, swearing and related phenomena in Afrikaans], PhD thesis, Stellenbosch University, Stellenbosch.

Coetzee, F 2018, 'Hy leer dit nie hier nie ('He doesn't learn it here'): talking about children's swearing in extended families in multilingual South Africa', *International Journal of Multilingualism*, vol. 15, no. 3, pp. 291-305.

Dekker, L 1991, 'Vloek, skel en vulgariteit: Hantering van sosiolinguisties aanstootlike leksikale items', *Lexikos*, vol. 1, pp. 52-60.

Eiselen, ER & Van Huyssteen, GB 2021, 'Using ordinal logistic regression to analyse self-reported usage of, and attitudes towards swearwords', *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*, 29 November to 3 December, DHASA, South Africa.

Feinauer, AE 1981, Die taalkundige gedrag van vloekwoorde in Afrikaans [The linguistic behaviour of swearwords in Afrikaans], MA dissertation, Stellenbosch University, Stellenbosch.

Fillmore, CJ, Bernal, E & DeCesaris, J 2008, 'Border Conflicts: FrameNet Meets Construction Grammar', *Proceedings of the XIII EURALEX International Congress*, Universitat Pompeu Fabra, Barcelona.

Global Language Monitor 2009, 'Death of Michael Jackson', view 22 August 2021, <http://www.languagemonitor.com/news/death-of-michael-jackson/>.

Goel, R, Soni, S, Goyal, N, Paparrizos, J, Wallach, H, Diaz, F & Eisenstein, J 2016, 'The Social Dynamics of Language Change in Online Networks', International Conference on Social Informatics (SocInfo16), view 22 August 2021, < https://arxiv.org/abs/1609.02075>.

Höder, S 2018, 'Grammar is community-specific. Background and basic concepts of Diasystematic Construction Grammar', in Boas, HC and Höder, S (eds.) *Constructions in Contact: Constructional perspectives on contact phenomena in Germanic languages*, Benjamins, Amsterdam, pp. 37–70.

Hughes, G 2006, *An encyclopedia of swearing: the social history of oaths, profanity, foul language, and ethnic slurs in the English-speaking world*, M.E. Sharpe, Armonk.

Jay, T 2000, *Why we curse: A neuro-psycho-social theory of speech*, John Benjamins, Amsterdam.

Labov, W 2001, *Principles of Linguistic Change, Volume 2: Social Factors*, Language in Society, Wiley-Blackwell, London.

Lavrakas, PJ 2008, *Encyclopedia of survey research methods* (Vols. 1-0), Sage Publications, Thousand Oaks, doi: 10.4135/9781412963947.

Leroux, E 1976, *Magersfontein, O Magersfontein!*, Human & Rousseau, Cape Town.

Leroux, E 1990, *Magersfontein: Die dokumente*, Human & Rousseau, Cape Town.

Lyngfelt, B, Borin, L, Ohara, K & Torrent, TT (eds.) 2018, *Constructicography: Constructicon development across languages*, John Benjamins, Amsterdam.

McWhorter, JH 2021, *Nine nasty words: English in the gutter: then, now, and forever*, Kindle edn, Penguin, New York.

Meho, LI 2007, 'The rise and rise of citation analysis', *Physics World*, vol. 20, no. 1, pp. 32-36.

MyBroadband 2018, 'South Africa wasting money on funding unemployable humanities and social science university students', view 22 August 2021,

<https://mybroadband.co.za/news/government/278877-south-africa-wasting-money-on-funding-unemployable-humanities-and-social-science-university-students.html>.

O'Driscoll, J 2020, *Offensive Language: Taboo, offence and social control*, Bloomsbury, London.

Paradowski, MB & Jonak, Ł 2012, 'Diffusion of Linguistic Innovation as Social Coordination', *Psychology of Language and Communication*, vol. 16, no. 2, pp. 131-142.

Pizarro Pedraza, A 2018, *Linguistic Taboo Revisited: Novel Insights from Cognitive Perspectives*, Cognitive Linguistics Research [CLR], De Gruyter Mouton, Berlin.

Remler, D 2014, 'Are 90% of academic papers really never cited? Searching citations about academic citations reveals the good, the bad and the ugly', view 22 August 2021, <https://dahliaremler.com/ 2014/04/09/are-90-of-academic-papers-really-never-cited-searching-citations-about-academic-citations-reveals-the-good-the-bad-and-the-ugly/>.

Republic of South Africa, 2019, 'Films and Publications Act (65/1996): Classification guidelines for the classification of films, interactive computer games and certain publications', Film and Publication Board, Department of Communications, Government gazette, no. 42380, notice 539, 5 April 2019, pp. 21-59, view 22 August 2021, <http://www.gpwonline.co.za/Gazettes/Gazettes/42380_05-4_NationalGovernment.pdf>.

Sheidlower, J 2009, *The F-Word*, Kindle edn., Oxford University Press, Oxford.

Traugott, EC & Trousdale, G 2013, *Constructionalization and constructional changes*, Oxford University Press, Oxford.

Van der Walt, A 2019, Linguistiese eienskappe en konvensionalisering in Zefrikaans op die WatKykJy?-blog: 'n korpuslinguistiese ondersoek [Linguistic features and conventionalisation in Zefrikaans on the WatKykJy? blog: a corpus linguistic study], MA dissertation, North-West University, Vanderbijlpark.

Van Huyssteen, GB 1996, 'The sexist nature of sexual expressions in Afrikaans', *Literator*, vol. 17, no. 3, pp. 119-135.

Van Huyssteen, GB 1998, 'Die leksikografiese hantering van seksuele uitdrukkings in Afrikaans [The lexicographic treatment of sexual expressions in Afrikaans]', *South African Journal of Linguistics*, vol. 16, no. 2, pp. 63-71.

Van Huyssteen, GB & Eiselen, R 2021, 'Oor feekse en helleveë [On shrews and harridans]', *Tydskrif vir Geesteswetenskappe*.

Van Rooyen, K 2012, *A South African Censor's Tale*, Protea Boekhuis, Pretoria.

Van Sterkenburg, PGJ 2019, *Rot lekker zelf op: Over politiek incorrect en ander ongepast taalgebruik*, Scriptum, Schiedam.

VivA 2021, Virtual Institute for Afrikaans: Corpus Portal, view 19 August 2021, <http://viva-afrikaans.org>.

Morphology-based investigation of differences between spoken and written isiZulu

Marais, Laurette
CSIR
laurette.p@gmail.com

Wilken, Ilana
CSIR
iwilken@csir.co.za

Abstract

Research attempting to describe and quantify the differences between spoken and written language has been done for languages such as English, but not for isiZulu. In this paper, we present a quantitative investigation into such differences by considering the morphology of tokens in a transcribed spoken isiZulu corpus and a written isiZulu corpus. We use morpheme tags as a proxy for features that typically differ between spoken and written language, and calculate relative differences of the occurrence of specific morpheme tags from analyses produced by ZulMorph, a finite-state morphological analyser for isiZulu. This analysis presents information that could inform the development of voice-enabled computer applications for isiZulu.

Keywords: spoken language, written language, voice computing, isiZulu

1 Introduction

Studies investigating the differences between speech and writing have been conducted by researchers from various fields for a variety of reasons. From an anthropological perspective, understanding such differences contribute to the study of cultural evolution and the role that writing and literacy play in human culture. Educators and psychologists have studied the differences in order to understand the cognitive factors affecting acquisition of both modalities, while an understanding of the lexical and grammatical differences of the two modalities has been the focus of linguists and language teachers (Akinnaso 1982, Olson 1996, Hung 2017).

In this work, we study the differences between the spoken and written modalities with a different aim: to inform design choices in the development of spoken language applications for isiZulu, especially given its resource scarce context.

When developing voice-enabled computer applications for a given language, it is important to have an understanding of the typical features of the spoken form of the language. Moreover, since corpora used for language modelling are often based on written text, it is useful to have an understanding of the differences between the spoken and written forms of the language. Features that are known to occur more frequently in spoken language could be considered during development, whether by engineering rules to deal with them appropriately or by ensuring that systems are trained on corpora that exhibit the desired features in a balanced way. This is especially important in a resource scarce context, where existing data may not perfectly fit the intended use case and where informed decisions must be made in order to utilise the data most effectively.

Research attempting to describe and quantify the differences between spoken and written language has been done for languages such as English, but not for isiZulu. In this paper, we present a quantitative investigation into such differences by considering the morphology of tokens in a transcribed spoken isiZulu corpus and a written isiZulu corpus. We use morpheme tags as a proxy for features that typically differ between spoken and written language, and calculate relative differences of the occurrence of specific morpheme tags from analyses produced by ZulMorph (Pretorius & Bosch 2003), a state-of-the-art finite-state morphological analyser for isiZulu.

2 Spoken and written language

One of the prominent themes in studies of differences between spoken and written language has been "disentangling the numerous factors that codetermine differences between spoken and writ-

ten language" (Redeker 1984), of which the most important are "the amount of planning, the conventionally expected level of formality in the situation, the nature and size of the audience, and the subject matter". In order to study specific differences, researchers have often opted to control for these codetermining factors in various ways: for example, a study of lexical differences in Dutch by Drieman (1962) was based on the assumption that topic, participants and the circumstances of obtaining data from participants should not vary (Akinnaso 1982), while Redeker (1984) studied the differences in degree of involvement/detachment as well as fragmentation/integration by keeping plannedness, formality and audience constant.

In this work, our aim is not to study features that differ between written and spoken isiZulu in a general way, but to understand the nature of the differences between the kind of language data for isiZulu that is readily available (namely written corpora) and the kind of isiZulu that voice-enabled applications would be expected to model. This reduces the need to control for various codetermining factors, since the goal of the work is not primarily a linguistic or discourse analytic result, but a characterisation of *required* resources in relation to *available* resources.

What language modelling resources would be ideal for the development of voice-enabled applications for isiZulu? To answer this, we need to understand typical use cases for such applications.

While it is almost impossible to predict the ways in which technology may be applied to improve the lives of people, a useful starting point is to consider where written and spoken language are typically used. As Akinnaso (1982) notes, the two modalities are often found in "complementary distribution" in society: "natural conversations are always carried out in spoken language, whereas, in modern industrial societies, speech is inappropriate for much bureaucratic communication such as applying for a job, requesting social services, filling out tax and credit application forms, and so on." From this description it is clear that the "modern industrial societies" in view are assumed to have high levels of literacy in the language in question. In South Africa, however, literacy rates are low and home language literacy rates even more so (Posel 2011), which seems to indicate that spoken isiZulu is used beyond the "natural conversations" mentioned by Akinnaso. Presumably, therefore, voice-enabled applications for isiZulu could prove useful in a larger variety of domains than might be the case for the languages of societies with high levels of literacy. This conclusion does not point to the requirement of a very specific kind of spoken language modelling resource, and therefore, presumably, any data comprising spontaneous spoken isiZulu, and perhaps especially spoken dialogue, would be suitable.

3 Resources and methodology

The basic requirements for performing an investigation into the difference between spoken and written isiZulu are, in the first place, suitable corpora that exhibit the features of the two modalities, and secondly, in the case where the identified corpora are not annotated in some way, a natural language processing tool that could enable a form of quantitative analysis. For a morphologically rich language, such as isiZulu, where many grammatical features are marked in the morphology, a morphological analyser provides a suitable instance of the latter. The South African NCHLT project delivered both written (Eiselen & Puttkammer 2014) and spoken (De Vries et al. 2014) corpora for isiZulu, although the spoken corpora do not exhibit spontaneous speech. It was compiled by recording written prompts and hence cannot be assumed to exhibit typical features of spoken isiZulu. In contrast, van der Westhuizen & Niesler (2018) compiled a corpus from transcribed South African soap opera data, mainly for the purposes of studying code-switching between various South African languages. The complete corpus contains five languages, namely English, isiXhosa, isiZulu, Setswana and Sesotho, and includes many code-switched segments, along with a few thousand monolingual isiZulu utterances. The authors note that a comparison of the transcriptions with the

original scripts for the episodes shows "a strong tendency in the actors to ad-lib", and they therefore conclude that the corpus can be considered as spontaneous speech.

Having identified suitable corpora, our methodology can be summarised as follows:

1. From available literature, compile a list of features that characterise the difference between spoken and written English.
2. Identify, where possible, concrete measures of these features (or related features) for isiZulu that can be achieved by analysis of the surface forms of the text or morphology-based analysis.
3. Perform the analysis on the spoken and written corpora and compare the results.

3.1 Features to be investigated

Table 1 lists a number of features compiled from the literature on spoken and written English (Akinnaso 1982, Redeker 1984, Cornbleet & Carter 2001, Zhang 2019, Tottie 1991) and Dutch to a lesser degree (Drieman 1962). For each feature, we indicate which kind of analysis was performed, namely either a simple textual analysis of the surface forms or an analysis of morpheme tags. For a number a features, such as eg. false starts, it was determined that this method would not be sufficient to shed light on the feature - syntactic or even semantic information would be necessary - and hence these features were not investigated.

3.2 Corpus preparation

The spoken corpus was extracted from transcriptions of South African soap opera episodes (van der Westhuizen & Niesler 2018). In total, 4 362 entirely monolingual isiZulu utterances were extracted, and this served as the spoken isiZulu corpus. The number of tokens contained in the monolingual isiZulu corpus was 13 929.

The written corpus was extracted from the NCHLT isiZulu text corpus (Eiselen & Puttkammer 2014), which consists mostly of government related texts. A corpus "equivalent" in size to the spoken corpus could be composed in at least two ways: either by including an equal number of utterances, or an equal number of tokens. As discussed in Section 4, the analysis was done on the token level, and so extracting a subset of the NCHLT corpus was done by selecting complete sentences from the corpus at random until the same number of tokens was reached as the spoken corpus. In the end, the written corpus contained 712 utterances and 13 943 tokens.

For the purposes of this work, these two corpora were assumed to represent the two modalities of isiZulu with regards to, in the case of the written corpus, what is typically available to developers of natural language processing applications, and in the case of the spoken corpus, spontaneous isiZulu dialogue, which is the kind of language voice-enabled isiZulu applications would typically have to model.

4 Morphology-based analysis

The ZulMorph analyser represents the state-of-the-art in isiZulu morphological analysis. It also has a substantial lexicon with over 20 000 roots and stems (Pretorius & Bosch 2009). A known effect of morphological analysis is the possibility of multiple analyses per token, and this is also the case with ZulMorph, which might produce as much as 20 possible analyses for some tokens. The applicable analysis for a token occurring in the context of a specific utterance would typically be determined via some disambiguation process, perhaps via a constraint grammar. In the absence of such a resource, it is not a simple task to determine which of the possible analyses for any given token is the correct one. The use of any other heuristic for performing disambiguation is likely to introduce unpredictable errors and biases, especially if the goal is to count the occurrences of specific morpheme tags.

One way of overcoming this problem is simply to consider all analyses. Admittedly, the absolute counts of specific morphemes in such sets would

Table 1: Typically different features of spoken and written English

Feature	**Surface analysis**	**Morphology-based analysis**
Length of text	✓	
Length of words	✓	
Monosyllabic words	✓	
Variety in vocabulary	✓	
Number of attributive adjectives		✓
Number of verbs		✓
Subordinate vs coordinate constructions		
Declaratives and subjunctives vs imperatives, interrogatives, and exclamations		✓
Passive vs active voice		✓
Definite articles vs demonstratives		✓
Gerunds		✓
Participles		
Modal and perfective auxiliaries		✓
Deliberate organization of ideas		
False starts, repetitions, digressions		
Negation		✓
Time relationships		✓
Personal discourse markers		✓

not be indicative of anything. However, the relative counts of the all possible analyses from the two corpora would still be significant. For example, suppose we wanted to investigate the occurrence of negation in two distinct corpora of 100 tokens each, and suppose the analyser returned about 500 analyses in total for both corpora. This would mean that the "overgeneration" of analyses on the two corpora were more or less equal, which implies similar patterns of overgeneration in both corpora. If we then found that the first set of analyses contained 81 tokens with negative prefix morphemes and the second set of analyses contained only 43, we could not conclude that about 16% of tokens in the first corpus exhibited negation in comparison to about 8% in the second corpus, because we do not know which kinds of tokens contributed relatively more possible analyses. However, we might reasonably conclude that the first corpus exhibits about twice as much negation as the second corpus.

As it happens, the effect of applying the ZulMorph analyser to the spoken and written isiZulu corpora did result in sets of analyses of similar size. Specifically, of the 13 929 tokens in the spoken corpus, the analyser produced analyses for 12 073 of the tokens, while for the written corpus of 13 943 tokens, the analyser produced analyses for 12 129 of the tokens. In total, the analyser produced 67 199 analyses for the spoken corpus and 70 345 analyses for the written corpus, giving a ratio of 1 to 1.05. We deem this to be sufficiently similar to assume that relative counts in the two corpora are indicative of relative occurrences of specific morpheme tags. Essentially, our assumption is that the context provided by existing results for English, combined with a reasonable relative measure for isiZulu, provides a useful indication of the differences between the two isiZulu corpora in question.

Specific morpheme tags were identified as representing or relating to specific features, such as negative prefixes representing negation. Appendix A contains a table that shows the mapping from feature to tags in the first two columns, followed by absolute counts and their relative difference in the fol-

lowing columns. Features that could be investigated via simpler means were approached in the following way:

Length of text The spoken language text exhibited shorter utterances than the written text, which is contrary to what was found in some of the literature (Drieman 1962). We expect this to be due to the nature of the spoken corpus, which is typically dialogue, and hence may exhibit a degree of interruptions not included in Drieman's data.

Length of words For this feature, we calculated the average lengths of the words in the corpora. We found an average length of 6.5 characters per word for the spoken corpus compared to 8.2 characters per word in the written corpus, consistent with the literature.

Monosyllabic words A naïve definition of monosyllabic words was used to make this comparison, namely that they are words consisting either of a vowel, or a vowel preceded and/or succeeded only by consonants. This yielded 138 such words in the written corpus compared to 949 in the spoken corpus, which is consistent with the literature.

Variety in vocabulary For this feature, we first considered the number of unique tokens. In the spoken corpus, 4 670 unique tokens appear in the set of 13 929 tokens, while in the written corpus, 7 920 unique tokens appear in the set of 13 943 tokens, giving a ratio of 1 to 1.6. Then, we counted unique verb roots and noun stems, with the spoken corpus containing 1194 and the written corpus containing 1586, giving a ratio of 1 to 1.33. Hence, this feature is also consistent with the literature, and the results additionally suggest that the written corpus contains more morphological variety.

5 Discussion

In order to improve the readability of this section, all numbers mentioned refer to the frequency of some morpheme tag in the spoken corpus relative to the written corpus. For example, a relative frequency of 10 means that the tag in question appeared 10 times more frequently in the spoken corpus than in the written corpus.

The first result to note is that of verbs and copulatives. While the spoken corpus contains 4 362 utterances, the written corpus contains 712, which is a ratio of about 6 spoken utterances to every written sentence. However, two typical kinds of verb phrases, namely verb based and copulative based verb phrases, occur only 2 and 3 times as often in the spoken corpus. This is surprisingly low, and seems to indicate that the utterances in the spoken corpus tend to lack verb phrases. This may be because of interruptions that occur during a dialogue, or it may be some form of ellipsis.

A feature that stands out, however, is the imperative, as suggested by the relative frequencies of the imperative prefix (about 10) and imperative suffix (almost 7). This is consistent with the summary provided by Akinnaso (1982), who mentions imperatives alongside interrogatives. In our experiment, both interrogative tags in the ZulMorph tagset had a relative frequency of about 4. This is especially intuitive considering the nature of the spoken corpus, which typically takes the form of a dialogue between characters in a soap opera. It is therefore also unsurprising that the relative frequency of the first person singular morpheme tag is 7.5, while the second person singular tag has a relative frequency of almost 3. We note that the first and second person plural tags have significantly lower relative frequencies, namely 1.5 and 0.9, respectively. In fact, the second person plural is one of only two features to have relative frequencies below 1, indicating that the feature occurs more frequently in the written corpus. However, in this case, the number is very close to 1, and therefore rather indicates that the feature occurs equally frequently in both corpora.

The other feature occurring more frequently in the written corpus is the passive voice, which again accords with the literature for English. Here, the passive voice is almost twice as frequent in the written corpus as in the spoken corpus.

We note that the negative prefix has a relative frequency of about 2.5, consistent with the literature for English. isiZulu does not have an explicit definite or indefinite article, but demonstratives have a

relative frequency of about 2. Cornbleet & Carter (2001) state that "various differences" can be found between written and spoken English with regards to time relationships, and in this work we see that especially the use of the future tense is more frequent in the spoken corpus, although the past tense also occurs slightly more frequently.

One instance where a clear confirmation was not found was in the case of gerunds, which we approximated for isiZulu by counting noun stems from class 15, the class of infinitive nouns (Poulos & Msimang 1998). Contrary to gerunds in English, the spoken isiZulu corpus did not exhibit fewer infinitive nouns than the written corpus. This is likely due to the fact that infinitive nouns in isiZulu are not sufficiently equivalent to gerunds in English: indeed, infinitive nouns have a "dual nature" (Poulos & Msimang 1998), and a more syntactically informed investigation would be required to differentiate their nominal and verbal usage in the two modalities.

Our investigation has shown a basic similarity between isiZulu and more well-studied languages, such as English, for features that can be identified morphologically. The similarities found on the morphological level would suggest that other relative differences between spoken and written language at the syntactic and semantic levels, may also be exhibited by isiZulu.

6 Conclusion and future work

In this study, we performed a quantitative comparison between a corpus of written isiZulu and a corpus of spontaneous spoken isiZulu. The comparison was mainly done on morphological analyses of the corpora obtained via a finite-state morphological analyser, and the methodology followed allowed for estimates of relative occurrences of morpheme tags in the corpora. The morpheme tags were chosen to represent or relate to features that are known to differ between written and spoken English. Broadly speaking, it was found that isiZulu exhibits many of the differences in its spoken and written modalities that languages such as English (and Dutch) exhibit. Our results also provide a quantitative characterisation of these differences, which could inform the development of voice-enabled applications for isiZulu in a resource scarce context.

One aspect of the resource scarcity of isiZulu is the available tools for analysing corpora. While the ZulMorph analyser was able to provide reliable morphological analyses of tokens in the corpora, no disambiguation tool currently exists, and this had a significant impact on the methodology and the kinds of conclusions that could be drawn, namely that we had to express the differences between the corpora in relative rather than absolute terms. Additionally, as evidenced by the results obtained by the approximation of gerunds in English by infinitive nouns in isiZulu, a purely morphological approach is not sufficient to investigate some grammatical features, and hence a syntactically informed tool, such as a parser, would enable more complete and accurate results.

Currently, however, morphological analysers exist for some of the other Nguni languages, including isiXhosa (Pretorius & Bosch 2009), as well as Setswana (Pretorius et al. 2005), both of which are also included in the multilingual soap opera corpus, and so similar morphology-based investigations could also be performed for these languages.

Another possibility would be to investigate social media text in isiZulu, in order to compare it with both the written corpus and the spontaneous spoken corpus used in this work. In his doctoral thesis, Wikström (2017) investigates "talk-like tweeting" in English as part of a study of "linguistic and metalinguistic practices in everyday Twitter discourse in relation to aspects of speech and writing". A comparison of social media text to corpora that represent the speech and writing modalities of in a more traditional way, could shed light on the extent to which social media text corpora could provide useful data for language modelling in voice-enabled applications for the resource scarce languages of South Africa.

References

Akinnaso, F. N. (1982), 'On the differences between spoken and written language', *Language and speech* **25**(2), 97–125.

Cornbleet, S. & Carter, R. (2001), *The language of speech and writing*, Routledge London.

De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E. & De Waal, A. (2014), 'A smartphone-based asr data collection tool for under-resourced languages', *Speech communication* **56**, 119–131.

Drieman, G. H. (1962), 'Differences between written and spoken language: An exploratory study', *Acta Psychologica* **20**, 36–57.

Eiselen, R. & Puttkammer, M. J. (2014), Developing text resources for ten south african languages., *in* 'LREC', pp. 3698–3703.

Hung, R. (2017), *Education between speech and writing: Crossing the boundaries of Dao and Deconstruction*, Routledge.

Olson, D. (1996), 'Towards a psychology of literacy: on the relations between speech and writing', *Cognition* **60**(1), 83–104.

Posel, D. (2011), 'Adult literacy rates in south africa: A comparison of different measures', *Language Matters* **42**(1), 39–49.
URL: *https://doi.org/10.1080/10228195.2011.571703*

Poulos, G. & Msimang, C. (1998), *A linguistic analysis of Zulu*, Vol. 1, Via Afrika, Pretoria.

Pretorius, L. & Bosch, S. (2009), Exploiting cross-linguistic similarities in zulu and xhosa computational morphology, *in* 'EACL Workshop on Language Technologies for African Languages', Association for Computational Linguistics.

Pretorius, L. & Bosch, S. E. (2003), 'Finite-state computational morphology: An analyzer prototype for zulu', *Machine Translation* **18**(3), 195–216.

Pretorius, R., Viljoen, B. & Pretorius, L. (2005), 'A finite-state morphological analysis of tswana nouns', *South African Journal of African Languages* **25**(1), 48–58.
URL: *https://doi.org/10.1080/02572117.2005.10587248*

Redeker, G. (1984), 'On differences between spoken and written language', *Discourse processes* **7**(1), 43–55.

Tottie, G. (1991), *Negation in English speech and writing: A study in variation*, San Diego: Academic Press.

van der Westhuizen, E. & Niesler, T. (2018), A first south african corpus of multilingual code-switched soap opera speech., *in* 'LREC'.

Wikström, P. (2017), I tweet like I talk: Aspects of speech and writing on Twitter, PhD thesis, Karlstads universitet.

Zhang, M. (2019), 'Exploring personal metadiscourse markers across speech and writing using cluster analysis', *Journal of Quantitative Linguistics* **26**(4), 267–286.
URL: *https://doi.org/10.1080/09296174.2018.1480856*

Appendix A: Feature counts

Feature	Tag	Number of occurrences		Relative diff.
		Written analyses	Spoken analyses	
Number of attributive adjectives	AdjStem	3344	2717	0,8125
	PC	26192	31674	1,2093
	RC	12695	10980	0,8649
	RelStem	780	1483	1,9013
	RelSuf	774	549	0,7093
Number of verbs	VRoot	39643	84305	2,1266
	CopPre	659	1979	3,0030
Declaratives, subjunctives,/ imperatives, interrogatives, and exclamations	ImpPre	29	314	10,8276
	ImpSuf	6	40	6,6667
	Interrog	966	4055	4,1977
	InterrogSuf	941	3648	3,8767
Passive/active voice	PassExt	5955	3553	0,5966
Definite articles/demonstratives	Dem	1110	2262	2,0378
Gerunds	15 + NStem	36039	41381	1,1482
Modal and perfective auxiliaries	Pot	776	2758	3,5541
	AuxVStem	477	1982	4,1551
Negation	NegPre	3519	8752	2,4871
	PotNeg	266	980	3,6842
Time relationships	Fut	3044	6826	2,2424
	FutNeg	15	96	6,4000
	SCPT	9759	16060	1,6457
	RCPT	2894	3590	1,2405
	VTPerf	11883	16306	1,3722
Personal discourse markers	1ps	2025	15256	7,5338
	2ps	5975	17697	2,9618
	1pp	1997	3003	1,5038
	2pp	2151	2096	0,9744

Using ordinal logistic regression to analyse self-reported usage of, and attitudes towards swearwords

Eiselen, Roald, and Van Huyssteen, Gerhard B
Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
{roald.eiselen/gerhard.vanhuyssteen}@nwu.ac.za

Abstract

Likert-type data is commonly used in many research fields in humanities: from gauging the usability of different user-interface designs, to determining users' likeliness to vote for a particular political party, to evaluation of course materials – to name but a few examples. Despite its prevalence, there is still some disagreement within the statistics community on whether Likert-type scales are true ordinal variables, and by implication whether parametric tests are legitimate to be used in such cases (Endresen & Janda 2017).

In this paper, we explore one parametric statistical test, viz. cumulative odds ordinal logistic regression (OLR), as an analysis method for self-reported data in the humanities. For illustration purposes, our focus is specifically on data of users' self-reported usage of, and attitudes towards swearwords, with the aim of identifying demographic attributes that are predictive of their usage and/or attitudes.

After a brief description of the data we're using, including how the data is being collected, we give a layman's overview of OLR. Since one of our aims is to demonstrate the usability of OLR, we apply our discussion practically to a step-by-step procedure (based on Laerd Statistics 2015) that could be followed easily. We demonstrate the usefulness of the results in reporting on the usage of, and attitude towards two near synonymous Afrikaans swearwords. We show, amongst others, that the odds ratios that are generated as part of the modelling procedure can be used to draw direct conclusions about specific demographic groups.

Keywords: Likert scale, linguistics, offensiveness, ordinal logistic regression, statistical modelling

1 Introduction

Over the last several decades, the use of statistical methods in linguistic investigations have become increasingly common, even the norm in many sub-fields of linguistics (Gries 2015). Deciding on which statistical method to use can be a somewhat daunting task, as the nature of the test, as well as the assumptions associated with the statistical test, can limit the types of tests available to a researcher. These factors, of course, also have a direct impact on the types of analysis and interpretation of the results that can be done.

Several types of analysis are commonly used in linguistic analysis, including the use of descriptive statistics, goodness-of-fit tests, monofactorial designs, and linear modelling (see, amongst others, Baayen 2019; Eddington 2015; Gries 2013). However, the use of generalised (i.e., mixed effect) logistic modelling, which take into account multiple predictor (i.e., independent) variables to predict the value of an outcome (i.e., dependent) variable, has been less prevalent. Given the fact that aspects of language production (speak/write) and perception (hear/read), as well as attitudes such as offensiveness of a word, perceived prominence of a word, etc., can be the result of a combination of factors, it is expected that the use of generalised models could be a valuable statistical tool for the analysis and interpretation of linguistic phenomena (Baayen & Linke 2020; Gries 2021). This would however not be applicable to linguistics only, but also more broadly in other fields of digital humanities. With this in mind, we investigate the use of one particular type of generalised logistic model, viz. cumulative odds ordinal logistic regression (OLR).

OLR is a parametric statistical test which describes the relationship between an ordinal outcome variable (i.e., ordered categorical data), and one or more ordinal, categorical or continuous predictor variables. OLR lets you determine which of your predictor variables have a statistically significant effect on an outcome variable, as well as determining how well the OLR model predicts the outcome variable, given a set of predictor variables. In addition to determining variable interaction and prediction, OLR can easily be interpreted as an odds ratio, which provides an

additional interpretation possibility for applying the results of OLR models in real-world contexts (Friendly *et al*. 2015; Harrel 2015).

To investigate the applicability of OLR for linguistic research, we use data collected from the *What The Swearword?!* (WTS) project [1]. One of the aims of this project is to determine offensiveness ratings for Afrikaans swearwords (i.e., any word or expression that could be offensive to some users in some contexts), which could be relevant for content developers, such as authors, publishers, film producers, etc.

The aim of this paper is to demonstrate the usefulness of OLR for this kind of inquiry. For this exploratory study and for illustrative purposes, we determine for only two near-synonymous swearwords, viz. *feeks* and *helleveeg* ('shrew, vixen, harridan'), the relationship between demographic information, and self-reported usage and attitudes ratings. We specifically want to answer the following questions:

- Can OLR be used to predict the usage of, and attitudes towards swearwords?
- Which predictor variables have a statistically significant effect on the usage of, and attitudes towards these two swearwords?
- Are the predictor variables with a statistically significant effect on a particular outcome variable the same for near synonyms?
- Can the interpretation of odds ratios be used to provide practical advice for content developers regarding swearwords?

To answer these questions, we commence with a brief overview of the data that we are using for purposes of this paper, including discussions on our sampling and collection procedures. Section 3 provides an overview of the four assumptions of OLR, as well as the procedure to follow for OLR modelling. This procedure is then illustrated extensively in 4.1, before we also provide more concise ways of presenting results in 4.2. We conclude with a brief discussion of our conclusions, as well as ideas for future work.

2 Swearword data

The WTS project website (vloek.co.za) was designed and developed with the main purpose to collect data from users, while experimenting with a variety of surveys, polls, questionnaires, and other data collection tools. Volunteer respondents, recruited through opportunistic and snow-ball sampling (i.e., via social media), have to register as users to participate as (self-selected) respondents. As of 21 August 2021, there are 2,088 registered users on the website, who are all eligible to participate in the surveys.

2.1 Demographics

During the registration process, participants provide demographic information, as well as self-reported information on their religious, political and world views. The selection of these questions and their categories is based on similar psycho-sociolinguistic studies (e.g. Beers Fägersten 2007; Dewaele 2016; Janschewitz 2008; Jay 2000, 2020; Van Sterkenburg 2001; Vingerhoets *et al*. 2013) where statistical relationships between one or more of these factors have been correlated with usage of and attitudes to swearwords. The following information, amongst others, is available for all participants in the study (with options for "other" or "don't want to answer" in some cases):

- Age group (three categories; ordinal)
- Sex (four categories; nominal)
- Gender (three categories; nominal) [2]
- Race (five categories; nominal)
- Length group (eight categories; ordinal)
- Highest qualification (12 categories, nominal)
- Income group (eight categories; ordinal)
- Religious view (five-point scale, from Not religious at all, to Very religious)
- Political view (five-point scale, from Very liberal, to Very conservative)
- World view (pertaining to moral and social issues; five-point scale, from Very liberal, to Very conservative)

Due to our sampling method and mode (social media, and a website), we assumed a priori that our sampling population will not be representative of the general Afrikaans population, since there are some inherent assumptions about this population. These include that they:

- have regular access to a computer/mobile device, and an internet connection;
- are technologically savvy (e.g., they are using social media platforms);
- have an interest in language, and specifically swearword or other taboos;
- are therefore probably less easily offended by such words and taboos (even though they may not use and/or approve of such words); and
- thus perceive themselves as rather enlightened/liberal.

These assumptions are confirmed when we look at the descriptive statistics of the groups that responded to the questionnaires for the two words under consideration (for *feeks* n=133; for *helleveeg* n=90). Only a small percentage of the respondents are 60 or older (21.1% for *feeks* and 18.9% for *helleveeg*); for both questions the entire population is white, and there are more males than females (unlike in the general Afrikaans population [3]); and the population is highly educated (64.7% of the respondents for *feeks* and 73.3% for *helleveeg* have a university degree). Although the entire population for both questions is mostly religious to some degree, only 7.5% (*feeks*) and 8.8% (*helleveeg*) of the respondents identify as conservative or very conservative.

When interpreting any of the results in this project, one should therefore be aware of the fact that the sample population is not representative of the Afrikaans community. Such results should therefore be preferably seen as individual pieces of empirical evidence that should be corroborated with other evidence, to get the full picture of a bigger puzzle.

2.2 Collection of self-reported data

One of the project's main types of short surveys, is the single word survey (SWS), where only one swearword per survey is presented to registered participants. The aim with SWSs is to keep each one as short as possible, in order to prevent respondent fatigue – "a well-documented phenomenon that occurs when survey participants become tired of the survey task and the quality of the data they provide begins to deteriorate" (Lavrakas 2008). The assumption is that one would cover more words over a period of time, than if one were to present the same number of words to participants in a single session.

Participants are therefore not required to complete questionnaires on all words, but only those ones that they want to participate in, and/or that they have time for. The implication of this way of sampling is that we cannot assume that (a) the data per word is independent (because some of the respondents might have answered all the SWSs); or (b) the responses are from the same sampling group (because some of the respondents might not have answered all the SWSs). Evaluating the pros and cons of this sampling method is, however, not the focus of this paper, but will be addressed in future research.

To make it as easy as possible for participants, they must self-report their usage of, and attitudes towards a given word on Likert-type scales, which are typically used to collect qualitative data in a way that provides quantitative values, thereby making statistical analysis of the data possible (Dubois 2013). For this study, a 9-point scale was used, where only the scores at the two extreme ends are descriptively categorised; this reinforces the notion that there are equal distances between each point on the scale (Endresen & Janda 2017). Respondents are asked to report their judgments on each of the following eight questions:

1. How often do you *say* or *write* the word? (Never … Very often)
2. How often do you *hear* or *read* the word? (Never … Very often)
3. How *offensive* do you find the word personally? (Not at all … Very)
4. How *taboo* or socially unacceptable is the word for people in general? (Not at all … Very)

5. To what extent is the word *emotionally charged* for you? (Very negative … Very positive)
6. How *prominent* is the word? (Not at all … Very)
7. How well do you know what the word *means*? (Not at all … Very well)
8. Is the word used pertaining to *men only*, *men and women*, or *women only*? (Men only … Women only)

The responses to each of these questions are considered as the outcome variables, while the demographic data are considered as the predictor variables. The hypothesis is that one or more demographic factors (such as age, or political view) will have a statistical effect on the usage of, or attitudes towards the swearwords (see Beers Fägersten 2007; Dewaele 2016; Janschewitz 2008; Jay 2000, 2020; Van Sterkenburg 2001; Vingerhoets *et al.* 2013).

3 OLR modelling

OLR modelling is a parametric statistical test to determine whether one or more predictor variables have a statistically significant effect on an outcome variable, and how well the model can predict the value of the outcome variable, given a set of predictor variables (Friendly *et al.* 2015; Harrel 2015; Laerd Statistics 2015). OLR has four assumptions that need to be considered in order to determine if it is a valid statistical approach for a particular study.

The first two assumptions are related to the design of the study and the measurements taken. *Assumption one* requires that you have a single ordinal outcome variable. *Assumption two* states that you should have one or more predictor variable(s) that are continuous, categorical, or ordinal. It should be noted that ordinal predictor variables are treated as categorical (i.e., they lose any internal ordering distinctions as part of the modelling procedure).

The last two assumptions relate to how the data fits the OLR model to provide valid test results. *Assumption three* states that there should be no multicollinearity between two or more continuous predictor variables. This means that if two continuous predictor variables are highly correlated, the results cannot be interpreted accurately, since it will not be possible to determine which one of the two predictor variables contribute to the explanation of the outcome variable. *Assumption four*, which is the fundamental assumption of OLR, states that you must have proportional odds, which means that each predictor variable has an identical effect at each cumulative split in the ordinal outcome variable.

Informed by the procedure suggested by Laerd Statistics (2015), the first step of the OLR modelling procedure is to ensure that the data adheres to the assumptions of the test. The first assumption requires an outcome variable that is ordinal. Although parametric tests, such as OLR, have been applied widely to Likert-type data in various other studies (e.g. Zhou *et al.* 2009), there is some disagreement within the community on whether Likert-type scales are true ordinal variables, and by implication whether parametric tests such as OLR are legitimate to use in such cases (Endresen & Janda 2017). However, Endresen & Janda (2017) show that for Likert-type data, the results for parametric and non-parametric tests have comparable results. With this in mind, we assume that Likert-type data is indeed ordinal, and that this type of parametric analysis is valid. Adherence to the second and third assumption is more easily confirmed, since all the predictor variables (i.e., the demographic information) are ordinal or categorical.

Verifying adherence to the fourth assumption is relatively easily tested in a statistical package such as SPSS by using "Test for parallel lines". This test (also known as a full likelihood ratio test) compares the fit of the proportional odds model to a cumulative odds model without the proportional odds assumption. If the assumption is met, the Chi-square value of the model will be small and not statistically significant ($p > 0.05$). Any variables that violate this assumption must be excluded from an OLR model.

After removing all predictor variables that violate any of the assumptions, the OLR is run, using an appropriate statistical package (SPSS in our case). The OLR test produces three important test results that should be reviewed before

investigating the full set of model parameter estimates:

1. a deviance goodness-of-fit test, which indicates if the model is a good fit for the data, where larger values are more indicative of a good fit;
2. an omnibus test (the likelihood ratio test [4]), which indicates whether the model predicts the outcome variable statistically significantly better than an intercept-only model (i.e. a model that does not take predictor variables into account); and
3. the effects of the different predictor variables, by looking at the Wald χ^2 test statistic and associated statistical significance (where $p < 0.05$).

Next, depending on the statistical significance of the model fit, and the effect of the different predictor variables in the model, additional predictor variables that clearly do not have an effect on the outcome variable, could be removed – both to simplify the model, and to improve the fit of the model. Therefore, given the results, one can either report the model as is, or try to improve the model by only selecting a subset of the predictor variables to see if there is any improvement in the overall fit of the model. However, care should be taken, since there are often intervariable effects, which might mean that a combination of predictor variables (e.g. gender plus age) could create a better model fit, even though one of these predictor variables does not have a statistically significant effect on the outcome variable.

The final step in the procedure is to interpret the predictor variable parameter estimates for each category of the predictor variables, and their significance. This interpretation should provide insight into the specific effect of each category of that predictor variable on the outcome variable.

4 Examples

4.1 Extensive example of OLR procedure

For the purposes of illustrating the procedure described in the previous section, we select one of the words, *feeks*, and one outcome variable, Tabooness ("How taboo or socially unacceptable is the word for people in general?"), as an application example for the full procedure. Additional, more concise examples of results are presented in section 4.2.

Table 1: Test of model effects: Tabooness of "feeks"

Predictor variable	*Wald* χ^2	*df*	*Significance*
Age	3.340	2	.188
Gender	11.153	1	.001
Length	11.507	5	.042
Country	14.961	8	.060
Political view	11.101	4	.025

Step 1: Determine if the data adheres to the assumptions of the OLR test

Given that the outcome variable is ordinal (i.e., data on a 9-point Likert scale), and all predictor variables are categorical, the first three assumptions of OLR are adhered to. For the fourth assumption, all predictor variables are tested for violation of the proportional odds assumption. For the word *feeks* and the Tabooness outcome variable, four of the predictor variables violate the assumption of proportional odds, viz. Qualification, Income, Religious view, and World view. Five variables do not violate this assumption, and will therefore remain in the initial OLR model.

Step 2: Run OLR and review results

For the Tabooness outcome variable and five predictor variables, the deviance goodness-of-fit test indicated a good fit of the observed data $\chi^2(748)=408.662$, $p=.546$, and the likelihood ratio test does statistically significantly predict the outcome variable over and above the intercept-only model, $\chi^2(20)=39.284$, $p=.006$. The model effects produced by OLR, presented in Table 1, show that three of the variables have a statistically significant effect on the outcome variable, Gender ($p=.001$), Length ($p=.042$) and Political view ($p=.025$). Age ($p=.188$) and Country ($p=.060$) do not show statistical significant effect on the outcome variable, although Country does account for the most data.

Step 3 (optional): Exclude uncorrelated predictor variables to simplify the model, and improve its fit

Table 2: Parameter estimates table (condensed): Tabooness of "feeks"

Parameter	*Beta*	*Wald* χ^2	*Sign.*	*Exp(B)*	*Odds*
Gender: Male	-1.209	11.191	.001	.299	1:3.34
Gender: Female	0	.	.	1	1:1
Political: Very conservative	-2.891	4.337	.037	.056	**1:17.85**
Political: Very liberal	-1.346	8.769	.003	.260	**1:3.85**
Political: Moderate	-.461	1.359	.244	.631	**1:1.58**
Political: Conservative	.446	.408	.523	1.562	1.56:1
Political: Liberal	0	.	.	1	1:1

In the case of Tabooness of *feeks*, we removed the variables that do not show statistically significant effect (Age and Country), but this decreased the fit of the model (p=.010). Therefore, keeping all five variables produced the most statistical significant fit for the observed data. This is most likely due to the fact that there are interactions between predictor variable groups, e.g. including both Age and Gender, that contribute to the overall model fit.

Step 4: Interpret results by reviewing parameter estimates, and determining the odds-ratios for values of specific predictor variables

Table 2 provides a condensed view of the most important information from the parameter estimates table for two variable groups, viz. Gender and Political view. The table includes the beta, Wald χ^2, significance, and odds-ratio (Exp(B)) values.

Since we have established that the OLR model for Tabooness of *feeks* fits the observed data, we can now interpret the information in the parameter estimates table for effects between specific groups of respondents and the outcome variable.

Keep in mind that OLR expresses parameter estimates in terms of one reference group (i.e., one of the categories under a variable). For each predictor variable, one category is selected as the reference group, and no beta or significance values are calculated for such a selected category; in Table 2 these are Gender: Female, and Political: Liberal. The Exp(B) value represents the odds ratio, i.e., the odds that that group will either assign a higher score (values larger than 1), or a lower score (values smaller than 1). As an example: For the word *feeks*, the odds that a man will assign a lower Likert score than a woman, are 3.34 (1/.299) times, which is a statistically significant effect, $\chi^2(1)=11.191$, p=.001. In other words, we could expect that women are more likely to rate *feeks* with a higher taboo score than men.

Another example: The odds of people who are Very liberal to assign a lower Likert score than people who are Liberal, are 3.85 times, also a statistically significant effect, $\chi^2(1)=8.769$, p=.003. In contrast, politically conservative respondents are 1.56 time more likely to assign a higher score than liberal respondents, but this is not a statistically significant effect (p=.523).

In the following section we apply the same procedure to two outcome variables for both *feeks* and *helleveeg*, to show how results can be more concisely reported. We also illustrate further interpretations of the results.

4.2 Concise examples of results

Given two words, *feeks* and *helleveeg*, and eight outcome variables, a total of 16 OLR models are possible. Since the aim of this paper is to demonstrate the applicability of OLR models to this type of inquiry, and for the sake of brevity, we report on the OLR tests and procedures for only two outcome variables, namely:

- How often do you *hear* or *read* the word? (Hear/Read)
- How *prominent* is the word? (Prominence)

As discussed in the previous section, the first three assumptions of OLR are not violated, since the outcome variables are all Likert-type data, and all

the predictor variables are categorical in nature. For the fourth assumption, all variables are tested for violation of the proportional odds assumption. For both outcome variables, across both words, many of the variables violated the assumption of proportional odds, and therefore cannot be included in the remainder of the procedure. A summary of the remaining predictor variables for each outcome variable for both words is provided in Table 3.

Table 3: Remaining predictor variables after testing for assumption of proportional odds

	Feeks (n=133)	*Helleveeg (n=90)*
Hear/Read	Age Gender Length Country Qualification Religious view Political view	-
Prominence	Income Religious view Political view World view	Age Gender Length Country Income Political view

The first thing to note from this table, is that there is no overarching set of predictor variables that adhere to the proportional odds assumption for both words across the two outcome variables. Separate models and variable selection are therefore necessary for each swearword, and for each outcome variable.

Also note that all the predictor variables for *helleveeg* violate the proportional odds assumption for the Hear/Read outcome variable. This stems from the fact that the distribution of assigned scores is very skewed, and 74.4% assigned either a 1 or 2 on the scale, indicating that they never or very rarely read or hear the word. [5] For *feeks,* on the other hand, there is a much more equal distribution across the various scale scores, with between 9% and 15.3% of responses in 7 of the 9 scale scores. Although there is no inherent assumption about the distribution of data for OLR, in cases where the distribution is highly skewed on the outcome variable, it is likely that either all the predictor variables will violate the proportional odds assumption, or that the resultant model will not be significantly better than an intercept-only model.

Given these remaining predictor variables, we firstly create OLR models that include all of the predictor variables that are valid for the OLR test. We then review the first three statistical tests to determine (a) the fit; (b) whether the model performs statistically significantly better than an intercept-only model; and (c) what the effects of the different predictor variables on the outcome variables are.

The following subsections provide the results for the words *feeks* and *helleveeg* for the two outcome variables, where only the best model for each outcome variable is described and interpreted. The aim is to illustrate that the entire statistical procedure can be expressed much more succinctly for each set of outcome and predictor variables.

Feeks

An OLR was run to determine the effect of Length, Qualification, Religious view, and Political view on how often participants Hear/Read the word *feeks*. There were proportional odds as assessed by a full likelihood ratio test comparing the model with varying location parameters, $\chi^2(154)=175.326$, $p=.115$. Although the deviance goodness-of-fit test indicated that the model was a good fit of the observed data, $\chi^2(866)=478.641$, $p=1.00$, the final model did not statistically significantly predict the outcome variable over and above the intercept-only model, most likely due to the high rate of empty cells for combinations of predictor variables (> 50%) [6]. Various models with fewer variables, which decrease the empty cell rate, also did not improve the fit of the overall model significantly.

For the Prominence of *feeks*, an OLR was run to determine the effects of Religious view [7]. The full likelihood ratio test indicated that there were proportional odds, $\chi^2(28)=17.670$, $p=.934$, while the deviance goodness-of-fit test also indicated that the model was a good fit of the observed data,

$\chi^2(28)=18.601$, $p=.664$. The overall model statistically significantly predicted the outcome variable over and above the intercept-only model, $\chi^2(4)=18.049$, $p=.001$. The odds of respondents scoring prominence lower than Religious respondents are statistically significant for two categories: Average religious, 2.81 times, ($p=.030$), and Not at all religious, 4.30 times ($p=.001$). The odds that Very religious participants would rate *feeks* higher on the Likert scale, is 1.41 times, but it is not statistically significant ($p=.425$). The results from this model indicate that more religious people are more likely to find the word *feeks* prominent when compared to people who are less religious.

Helleveeg

Since no variable adhered to the assumption of proportional odds for the Hear/Read variable, an OLR was only run for Prominence to determine the effects of the variables listed in Table 3. The first model, which included all six variables, did not predict the outcome variable statistically significantly over and above the intercept-only model, $\chi^2(26)=38.909$ and $p=.05$. By removing the predictor variable with the least effect, Gender, the model did improve, $\chi^2(25)=38.895$, $p=.038$, and statistically significantly predicted the outcome variable over and above the intercept-only model. Of the predictor variables, Age accounted for the largest proportion of the data: respondents between the ages of 40 and 59 were 4.54 times more likely to find *helleveeg* prominent, than people over the age of 60, a statistically significant effect ($p=.019$). Although the odds of people under the age of 40 is 1.47 times more likely to find the word more prominent, this effect is not statistically significant ($p=.539$).

5 Conclusion

In this paper, we have demonstrated that OLR is able to generate models that, in some instances, statistically significantly predict the outcome variable over and above an intercept-only model. Using data from three questions for two words, OLR was able to identify demographic variables that have a statistically significant effect on the Likert scale for the three questions. However, the demographic variables that have a statistically significant effect varies for both the different questions, and the different words. This is partly due to the fact that different predictor variables violate the primary assumption of proportional odds and therefore cannot be included in the OLR model. We concluded that it is essential to do rigorous testing of adherence to OLR's four assumptions for all predictor variables, in order to ensure that the OLR model is valid.

Beyond the differences in the predictor variables for the different questions and words, we also found that different numbers of variables are required to find the best fit for the data. In some cases, such as the prominence of *feeks*, a single predictor variable created the best model, while five variables were necessary for the Tabooness model of *feeks*.

Given the fact that the same variables do not have significant effects for the different words, we concluded that our kind of data and sampling methods do not allow to directly compare the OLR results or models of different words – at least at this stage of our research. This could possibly be due to two aspects:

1. Data for the two words were collected from two different, but potentially overlapping sampling groups. We expect intuitively that there should be larger overlaps of predictor variables (e.g. we might expect that very religious people will rate most swearwords more offensive than people who are perhaps less religious).
2. The semantic fields of different swearwords might also play a role. Near synonymous swearwords come from the same semantic domain (e.g. RELIGION) and we might expect that their tabooness ratings will all depend on similar predictor variables.

Since we have not observed these expectations in the results above, we will need to investigate how to deal with these anomalies in future studies. Other or additional statistical tests will most probably be needed to allow direct comparisons between the outcome variables for different words (see Van Huyssteen & Eiselen, 2021).

Based on the OLR models that have been created for the respective questions and words, we have shown that the odds ratios that have been generated as part of the modelling procedure, can be used to draw direct conclusions about specific demographic groups. For example, what are the odds that women will rate *feeks* as more Taboo than men, or that people under the age of 60 will find *helleveeg* more prominent than older respondents.

Although these are encouraging results for using OLR to investigate the usage of, and attitudes towards swearwords, several outstanding issues need to be addressed to determine how well this type of modelling works for this kind of data. In addition to matters already mentioned above, these include:

- the applicability of OLR modelling to other swearwords that are used more often and are more well-known;
- how sample size and demographic distribution affect the model's descriptive quality;
- how data distribution affects the ability of the models to identify variable effects;
- model visualisations that make the data and results more accessible to publishers and writers; and
- whether the models will be more or less useful indicators of variable effect on smaller Likert scales, such as a 3- or 5 point scale.

Notes

[1] A comprehensive overview of this project is provided in another paper (submitted for presentation) at this conference. See Van Huyssteen, 2021.

[2] The question is: "Do you identify with one or more specific gender groups?", with options Yes, No, Don't want to answer. If a respondent choose Yes, they can specify which group(s).

[3] The ratio male:female for both words was 57:43. For the general South African population, the ratio in the 2011 Census was 49:51. Based on data in Centre for Risk Analysis (2020), we can calculate that of the total white population in South Africa, 29.8% males and 31.5% females consider Afrikaans their first language.

[4] Two separate likelihood tests are performed as part of the OLR procedure, and they should not be confused with one another. The first, referred to as the full likelihood ratio test, is an assumption test for proportional odds; the second determines the fit of the full model.

[5] This is corroborated by data from all the corpora on VivA-KPO (2021): the distribution *feeks*:*helleveeg* is 93:7 per hundred examples.

[6] Empty cells in this context refers to a combination of predictor variables with no respondents, e.g. a person who is taller than 199cm (Length), has a doctorate (Qualification), is very conservative (Religious), and is very liberal (Political view). High rates of these empty cells, about which no statistical information is available, can be detrimental to the quality of the model and usually occurs if the sample group is relatively small, and a large number of variables, with a large number of categories are included in the model.

[7] The categories for Religious view are: Not religious at all; Not particularly religious; Average religious; Religious; Very religious. Respondents also had the option to specify something else, or to choose not to answer the question.

Acknowledgements

This research is partially funded by the Suid-Afrikaans Akademie vir Wetenskap en Kuns, and partially made possible through barter agreements with BlueTek Computers, Afrikaans.com, and WatKykJy.co.za.

Ethical clearance for the research project was obtained through the Language Matters Ethics Committee of the North-West University (ethics number: NWU-00632-19-A7).

The second author is a director of the not-for-profit company Viridevert NPC (CIPC registration number: 2016/411799/08), who owns and manages the website vloek.co.za. This website was developed specifically for this project, and this conflict of interest has been approved by North-West University.

We would like to thank Jaco du Toit (NWU) for his help with data processing. Thank you also to Adam Lund (Laerd Statistics) for suggesting OLR for our kind of data.

None of the results and/or opinions in this paper can be ascribed to any of the people or organisations mentioned above.

References

Baayen, RH 2008, *Analyzing linguistic data: A practical introduction to statistics using* R, Cambridge University Press, Cambridge

Baayen, RH & Linke, M 2020, 'Generalized Additive Mixed Models', in Paquot, M & Gries, ST (eds.) *A Practical Handbook of Corpus Linguistics*, Springer Nature, Cham, pp. 563-591.

Beers Fägersten, K 2007, A sociolinguistic analysis of swearword offensiveness, Universität des Saarlands, Saarbrücken, view 16 August 2021, <https://www.researchgate.net/publication/265009714_A_sociolinguistic_analysis_of_swearword_offensiveness>.

Centre for Risk Analysis 2020, *Socio-Economic Survey of South Africa*, Melville, view 16 August 2021, <https://cra-sa.com/products/socio-economic-survey/2020>.

Dewaele, J-M 2016, 'Self-reported frequency of swearing in English: do situational, psychological and sociobiographical variables have similar effects on first and foreign language users?', *Journal of Multilingual and Multicultural Development*, vol. 38, no. 4, pp. 330-345.

Dubois, D 2013, 'Statistical reasoning with set-valued information: Ontic vs. epistemic views', in Borgelt, C, Gil, MA, Sousa, JMC, & Verleysen, M (eds.) *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, Springer, Berlin, pp. 119–136.

Eddington, D 2015, *Statistics for linguists: a step-by-step guide for novices*, Cambridge Scholars Publishing, Newcastle upon Tyne.

Endresen, A & Janda, LA 2017, 'Five statistical models for Likert-type experimental data on acceptability judgments', *Journal of Research Design and Statistics in Linguistics and Communication Science,* vol. 3, no. 2, pp. 217–250. https://doi.org/10.1558/jrds.30822.

Friendly, M, Meyer, D & Zeileis, A 2015, *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*, 1st edn, Chapman and Hall, Boca Raton.

Gries, ST 2013, *Statistics for Linguistics with R: A Practical Introduction*, 2nd edn, De Gruyter, Berlin.

Gries, ST 2015, 'Quantitative linguistics', in Wright, J. (ed.), *International Encyclopedia of the Social and Behavioral Sciences*, 2nd edn, Vol. 19, Elsevier, Oxford, pp. 725–732.

Gries, ST 2021, '(Generalized Linear) Mixed-Effects Modeling: A Learner Corpus Example', *Language Learning*, vol. 71, no. 3, pp. 757-798.

Harrel, FE 2015, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression and survival analysis*, 2nd edn, Springer, Heidelberg.

Janschewitz, K 2008, 'Taboo, emotionally valenced, and emotionally neutral word norms', *Behavior Research Methods*, vol. 40, no. 4, pp. 1065-74.

Jay, T 2000, *Why we curse: A neuro-psycho-social theory of speech*, John Benjamins, Amsterdam.

Jay, T 2020, 'Ten issues facing taboo word scholars', in Nassenstein, N and Storch, A (eds.) *Swearing and Cursing*, De Gruyter Mouton, Berlin, pp. 37-52.

Laerd Statistics 2015, 'Ordinal logistic regression using SPSS Statistics', *Statistical tutorials and software guides*, view 16 August 2021, <https://statistics.laerd.com/>.

Lavrakas, PJ 2008, *Encyclopedia of survey research methods* (Vols. 1-0), Sage Publications, Thousand Oaks, doi: 10.4135/9781412963947.

Van Sterkenburg, PGJ 2001, *Vloeken. Een cultuurbepaalde reactie op woede, irritatie en frustratie*, 2nd edn, Sdu Uitgevers, The Hague.

Vingerhoets, AJJM, Bylsma, LM & De Vlam, C 2013, 'Swearing: A Biopsychosocial Perspective', *Psychological Topics*, vol. 22, no. 2, pp. 287-304.

VivA-KPO 2021, Virtual Institute for Afrikaans: Corpus Portal Comprehensive, view 19 August 2021, <http://viva-afrikaans.org>.

Van Huyssteen, GB 2021, 'Swearing in South Africa: Multidisciplinary research and scientific communication on language taboos', *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa 2021*, 29 November to 3 December, DHASA, South Africa.

Van Huyssteen, GB, Eiselen, ER 2021, (In print), 'Oor feekse en helleveë [On shrews and harridans]', *Tydskrif vir Geesteswetenskappe*.

Zhou, F, Wu, D, Yang, X & Jiao, J 2008, 'Ordinal logistic regression for affective product design', IEEE International Conference on Industrial Engineering and Engineering Management, IEEE, pp. 1986-1990.

UPLOrc: A Networked, Live Coding Laptop Orchestra based in South Africa

Melandri Laubscher
N.Cert. (Sound Technology), BMus (UP)
m.laubscher@tuks.co.za

Abstract

In this article I report on the current and emerging practices of *UPLOrc (University of Pretoria Laptop Orchestra)*, a networked live coding laptop orchestra based in Southern Africa. Since its establishment in 2019, the ensemble has performed live coded network music using the TidalCycles live coding environment at various conferences and live streamed events. The development of these practices is owed to, among other aspects, the fieldwork experience I obtained with the trans-continental network ensemble *SuperContinent*. I describe how this knowledge has been implemented into the activities of UPLOrc, alongside some of our own emerging practices. Particular problems that emerged during the performance preparation process is also highlighted, as well as the strategies that could be implemented to address some of these problems.

Keywords: Digital Humanities, Networked Communities, Network Music Performance, Laptop Orchestra Pedagogy, Live Coding

1 Introduction

Contemporary music performance, particularly network-based performance, has recently experienced an increase in popularity for a number of reasons. First, increased interaction has become commonplace for people who use technology to maintain relationships across long distances and "political borders" (Schrooten 2016), whether professional or personal in nature. These online relationships increased on an unprecedented scale after the outbreak of the Coronavirus pandemic in late 2019. Consequently, many collaborative activities were forced to move to online platforms, including the activities of some laptop orchestras (Fasciani 2020). Network music, a corollary to telematic music (Oliveros et al. 2009), is performed using an internet connection, where collaborators are often, but not always located in the same country or region (although it is also common to work collaboratively across continents) (Carôt et al. 2006). For example, collaborators of the trans-continental network ensemble SuperContinent are located, at minimum, 500 kilometers apart. The group performs regularly at various events and conferences (Betancur et al. 2021), and is a sub-project of ongoing research at McMaster University's collaborative research center the Networked Imagination Laboratory or NIL [1]. An opportunity to join SuperContinent in 2020 became available when my research supervisor could not participate due to other academic and creative commitments. I viewed this as an opportunity to gain experience as a performer, and to learn how others approach collaborative laptop ensemble performance. In the most general sense collaborative laptop ensemble performance involves the staging and performance of contemporary art forms through the use of computers, or more accurately, laptops. These contemporary art forms, and the technologies that make them possible, may vary widely depending on the context. For instance, some ensembles such as Princeton University's PLOrk (Princeton Laptop Orchestra) make use of self-contained stations consisting of a laptop and a hemispherical speaker (Trueman 2007), or in some cases Digital Musical Instruments (DMI's) (Ferguson & Wanderley 2010, Berdahl et al. 2018). However, due to the limitations of this type of setup, many ensembles make exclusive use of software to perform and improvise music collaboratively (Freeman & Troyer 2011). In some contexts, as is the case with SuperContinent and UPLOrc, software similar but not limited to Estuary [2], is further combined with networks to enable collaborators to perform together across long distances (Knotts 2015, Ogborn et al. 2017, Carôt et al. 2006).

My interactions on the Estuary platform included attending weekly SuperContinent rehearsals alongside NIL-related activities, both of which introduced me to new concepts, practices, and ways of performing music collaboratively. In particular, NIL hosted weekly *fromZero* workshops (Ogborn et al. 2015) which I was attending in addition to performing regularly with the members of SuperContinent [3]. Most of my time was spent observing and interacting with a number of new processes I had never experienced elsewhere, working with a wide variety of technology such as the intuitively-designed Estuary platform (Ogborn et al. 2017). SuperContinent is one of many ensembles that use Estuary to perform live coded (Collins et al. 2003, Ogborn 2016, Nilson 2007) network music, a performance practice which forms a central part of current pedagogical strategies in formal academic contexts, in particular STEM education [4] (Soon & Knotts 2018).

Through engagements with the members of SuperContinent, I was able to note the requirements involved with planning and coordinating the activities of a typical live coding laptop ensemble (Betancur et al. 2021). As the coordinator of the then newly established *University of Pretoria Laptop Orchestra* (UPLOrc), I was tasked with developing and facilitating ensemble activities. This article presents the process of using such techniques alongside the progress made by UPLOrc thus far. I begin with a brief background of how UPLOrc came to be and where we currently find ourselves as a group. I then attempt to address some of the questions surrounding collaborative musical creativity, an idea put forward by Bishop (2018), in the context of a network-based live coding ensemble. Further, I present the planning and coordination that went into three UPLOrc performance cycles in addition to the communication involved in network music performance. The article then concludes with a discussion detailing the current problems we have encountered, as well as the lessons we learned during the process.

2 UPLOrc, so far

UPLOrc was established in May 2019 by my research supervisor and artistic director of UPLOrc, Dr. Miles Warrington [5]. Our debut showcase was held at the annual University of Pretoria Music Festival (UPMF) in 2019 [6] and although a wonderful experience and opportunity, I was only beginning to familiarise myself with collaborative laptop performance practices and felt that I had much to learn in this area. Since then I have spent time interacting with other network musicians and used various technologies, attempting to absorb as much information as possible. The knowledge I acquired during this time has therefore greatly impacted the development of UPLOrc activities.

In obtaining this knowledge my goal was to understand how other individuals were able to collaborate, particularly focusing on the ways in which others would approach performing live coded music in real-time (Collins et al. 2003). I would spend hours watching and deconstructing the content of TidalCycles video tutorials presented by its creator McLean (2014). I further observed the ways in which SuperContinent members approached live coding with MiniTidal, a version of TidalCycles available to use on Estuary. TidalCycles and MiniTidal, often referred to as Tidal to include both versions, is an audio programming language environment used to perform live coded music.

My initial objective entering into SuperContinent was to improve my own skills as a live coder, then incorporating this knowledge and performance experience as a reference point for developing the practices and objectives of UPLOrc. What I did not expect to experience, was the the complete musical freedom afforded to me by the other members of SuperContinent. I experienced new forms of interaction that would otherwise be impossible without the technology facilitating those interactions. Encountering the work of Bishop (2018) allowed me to identify a shift in within my own identity as a musician. She states that in order to understand this shift further research should be conducted into de-

termining how the individual's mind is constrained by their imagination, and whether they are able to transform their frame of reference to make room for new sonic structures (Boden 2004). Evidently, the tools used to express these musical ideas, in this case MiniTidal and Estuary, should be evaluated in further detail to determine whether these technologies either facilitate or constrain the individual's ability to express their individual musical ideas in a collaborative setting (Bishop 2018, Knotts & Collins 2014, Knotts 2015).

Experiencing this sort of musical freedom has enabled me to embrace a similar openness to the perspectives and tastes of others. It is my primary objective as the manager of UPLOrc activities to facilitate a similar kind of approach to collaborative performance. My duties as the manager involves, among other things, planning and presenting educational content focusing on the technical aspects and logistics of live coded network music. While I am responsible for presenting this information to the members of the ensemble, I acknowledge that it is crucial to maintain these relationships in a manner that encourages freedom of expression from all involved in this project. Moving forward, our collective aim is to develop a fundamental understanding of the ways in which musical ideas can be generated in collaboration with others, where the performer is often required to monitor multiple actions in real-time (Xambó et al. 2016, Xambó 2017). Learning how to live code, and observing others who do, has become a fundamental part of the process of drawing closer to developing my understanding of how UPLOrc is able express musical ideas as an ensemble.

2.1 The network orchestra

The initial months of coordinating UPLOrc included a great deal of experimentation until I was confident that I had developed an efficient approach to preparing myself and my fellow ensemble members for upcoming performances. We initially intended for UPLOrc to perform live concerts in halls and venues and, before the Covid-19 pandemic, we had already decided that a portion of our activities would be held online. I had moved to another province in South Africa and was to travel to Pretoria when I needed to be there for UPLOrc events. Since that did not materialise we were forced to, like many other ensembles, make use of additional technological tools that would allow us to perform collaboratively from the safety of our homes.

UPLOrc currently has six members located in all corners of Southern Africa, including the Western Cape, Gauteng, the Free State, Mpumalanga and Namibia. Members comprise of undergraduate students, post-graduate students and University of Pretoria faculty, as is the case with many other laptop ensembles in higher education, for example SLOrk (Stanford Laptop Orchestra) (Wang et al. 2009). A new challenge presented itself to UPLOrc in 2020. In navigating our activities as a "new" network ensemble, we needed to explore other modes of communication. Communicating our ideas became possible using tools including, but not limited to, Estuary, MiniTidal, Slack [7] and Discord [8]. These have provided us with the most efficient, no-cost option for meeting twice a week to rehearse and attend workshops. Communications between ensemble members are discussed in further detail in section 3.3 below.

3 Hardware and software tools

Between the members of UPLOrc we have three MacBooks and three Windows laptops. Since Estuary requires no installation of additional software, all members of UPLOrc require is a computer that is able to run the Google Chrome browser. It is safe to assume that most university-attending individuals have some computing device enabling them to attend online academic-related events, therefore having the ability to at least access Estuary (Feerrar 2019, Ogborn et al. 2017). Making efficient use of Estuary may be a challenge for some however, especially if their device does not meet minimum system requirements needed to run Estuary. I elaborate on the relationship between our laptops and navigating a rehearsal or performance in Estuary in section

3 below. Our current devices are, for the most part, entry-level devices which most would agree are best used for long-term administrative use. We have been fortunate to have access to at least two devices that could easily stream a live performance on YouTube, as has been common practice for some SuperContinent events. Software, as opposed to hardware, is our primary form of technology that we use to perform. Our laptops are currently the only form of hardware we interact with, meaning that any musical gesture we generate originates solely from typing code (Salazar 2017). For UPLOrc, the benefit of making exclusive use of software for performance means a low entry-level if a participant is not able to purchase additional equipment. More so if that software is completely free to use, and optimised in such a way that anyone with little to no live coding experience will be able to perform simple, yet interesting, musical ideas with a few lines of code (Ogborn 2012).

3.1 Estuary

The Estuary platform is a browser-based and multilingual live coding platform, providing instant access to a collaborative gathering space for novices and experienced programmers through the browser (Ogborn et al. 2017). According to Estuary's GitHub repository [9], it is recommended that users access Estuary using either Google Chrome or any browser that is based on the Chromium browser project [10]. Attempting to access Estuary from browsers such as Safari and Firefox, whose architecture does not use Chromium, is not currently permitted. The majority of UPLOrc'ers (the name we use to refer to our members) have had the best experience with Estuary using Microsoft Edge, another browser platform that uses Chromium. Some experienced Edge as performing better than Chrome on their older devices. When one member recommended I use Edge on my 2011 MacBook Pro, I immediately noticed a significant difference in the way Estuary was performing.

Getting to know the platform is simple, even for those who are not so comfortable using technology. The overall layout and design of Estuary assists workshop instructors like myself in customising some features of the platform. For example, adjusting what is displayed on screen, adding and removing an ensemble, and adjusting tempo is among a long list of available commands. Most of these features, called terminal view commands, can be performed using one-word commands which are accessible by clicking on the question mark in the top right corner of the Estuary screen (see Number 3: Figure 1: Estuary login screen). When accessing Estuary for the first time, the user is presented with *solo mode* and *collaborate mode* (number 1 and 2 on Figure 1: Estuary login screen). Collaborating in an ensemble requires that collaborate mode is used. The following screen displays a list of all the current ensembles active on the platform. Once the correct ensemble is selected, the user is prompted to provide their username, and optionally, their location. The ensemble password is entered, taking the participant to a final screen where they are then able to collaborate with the entire ensemble (see Figure 2: UPLOrc screen layout).

Other useful tools include a terminal chat window (Number 2: Figure 2) used to communicate during activities, space to enter your name and your code (Number 1: Figure 2), a list of participants letting everyone else know who has logged in (Number 3: Figure 2), and a useful information bar used to monitor CPU usage or what is loosely referred to as "glitching" [11]. "Glitching" has become a regular term used among the members of UPLOrc, and is used to describe the point at which one of our laptops cannot process the current code running on Estuary, presenting us with a glitching effect of the audio. While glitching can be interesting at times, it can severely affect the audio at times where the glitching becomes a hindrance to the performance.

3.2 MiniTidal

MiniTidal is among an extensive list of audio and visual programming languages available to use use on Estuary [12]. Some of its features excludes some TidalCycles functionality, although new ones are

constantly being added by researchers and developers at NIL. Developed by McLean (2014) in collaboration with a growing community of developers and users, TidalCycles is currently one of the most prevalent live coding environments and music programming languages being used for live coded performance. This is true in individual and collaborative settings, but also for teaching live coding in a variety of educational contexts (Ogborn et al. 2021, Soon & Knotts 2018) across the globe. TidalCycles is a programming language written in Haskell, and is specifically designed for live coded music performance. SuperContinent mainly performs with MiniTidal and Punctual (Betancur et al. 2021), an audio and visual live coding language developed by David Ogborn [13]. Presently, UPLOrc is live coding exclusively with MiniTidal, due to the ease with which novice live coders are able to participate in coding exercises. Before moving online UPLOrc was using SuperCollider to run TidalCycles using the Atom IDE [14]. Various problems would emerge when installing software on some devices and it became challenging to assist newcomers with the installation of these tools. The simplicity of logging into a platform that is Estuary is extremely appealing to novice and experienced network music performers alike. An additional benefit of developing experience as a live coder and network ensemble performer, is the ease with which those skills can be attained and extended (Ogborn 2012). Provided of course that sufficient time is spent cultivating those skills, particularly in areas that develop musical expression.

3.3 Ensemble communications

In the context of network music performance communication between ensemble members becomes challenging when members are not physically located in the same room. (Freeman & Troyer 2011). Bishop (2018, p. 6) describes communication as "the transfer of information that occurs between members of a group" and identifies various forms of communication. In the case of UPLOrc, communication occurs in a number of different ways. Our primary mode of communication, and perhaps the most challenging to master as a new member, is that which occurs when live coding with MiniTidal. Similar to the communication of an instrumental ensemble, the audio transmitted from Estuary while UPLOrc is live coding is interpreted by each member in real-time. Since almost all of our live coding activities are improvised, it is impossible for anyone to predict what the outcome of a live coding performance will be, and therefore members are required to adjust to what is heard in real-time. Marie et al. (Forthcoming, p. 6) refer to this as a "layer of unpredictability." UPLOrc, and so too SuperContinent, are required to deal with these unpredictabilities as they occur.

Another, which points to the limitations or restrictions of the technologies UPLOrc and SuperContinent use, has to do with what Marie et al. (Forthcoming, p. 6) further refer to as "layers of unpredictability between human and machine." For example, if one member of the ensemble unknowingly makes a change in their code that initiates the aforementioned audio glitches, almost all [15] members of the ensemble will experience the unwanted effect. This will be a direct result of a process that no one in the ensemble has control over, due to some combination of events that are simply incompatible in that moment. A simple readjustment or removal of a piece of code should quickly resolve the issue. This continuous readjustment of code in real-time, which is essentially the act of live coding, is centred around the notion of emergence and group flow. Bishop (2018) defines this as performing "in a way that cannot be attributed to any one individual." Being aware of one's position in and amongst the other voices who would like to be heard, is essential to maintain the balances of power and freedom of expression in collaborative performance contexts Collins (2003), Knotts & Collins (2014), Knotts (2018).

Our second mode of communication is in the form textual communications. UPLOrc engages in these interactions using the Estuary terminal chat window, allowing members to communicate during a

rehearsal or performance. Further communication takes place during post-rehearsal discussions on Discord - a practice I initially observed as a member of SuperContinent. I incorporated this into our own practices as I observed the benefits of post-rehearsal reflection. This provides members the opportunity to voice their opinions and make suggestions, or simply to reflect on the shared experience. Discord's voice hangout functionality, originally intended for use with streaming and playing online games, has recently become a tool UPLOrc has been using during performances (Laubscher et al. 2021*a*). Some of the the newest members have only performed as a network ensemble once prior, and therefore Discord was a useful tool in directing and prompting other members during specific stages of our performance. During an earlier performance (Laubscher et al. 2021*b*), we opted to perform without using Discord as an extra line of communication. This is sufficient in certain situations (with SuperContinent for example, where constraints and ensemble goals are different from that of UPLOrc) and may become unnecessary when members become more comfortable in their ability to perform.

4 UPLOrc performance cycles

UPLOrc performance cycles consist of three related activities; workshops, rehearsals and concert performances. Workshops and rehearsals are structured to prepare the entire ensemble for upcoming scheduled performances. Our first cycle, approximately six months long, was the longer of the two, with cycle two (Laubscher et al. 2021*a*,*b*) lasting four months. During this time I attempted to develop members' skills as quickly as possible, hence the difference in cycle length. Our debut online performance was held at Estuary's *five year since commit* [16] event in December 2020 (Laubscher et al. 2020). The circumstances and my experience as an instructor were much different in both cycles, allowing me to learn from previous errors and correcting them where possible. I present some notable approaches and problems that emerged from the preparation of my fellow ensemble members for performance.

4.1 UPLOrcShops

UPLOrc workshops (UPLOrcShops), like all of our other ensemble activities, are scheduled according to the times that best suit the majority of the group. These are held every Wednesday for one hour and involves prepared content that is presented and demonstrated during the session. Members listen in using Discord's voice channel capability, where I am able to display my screen directly in the application. Depending on the prepared content I may either opt to have TidalCycles running in Atom, since it would be ideal for members to fully grasp the language and all its capabilities. Only when I demonstrate code that requires audible output, do I open Estuary in Microsoft Edge. The second portion of the session is usually dedicated to collectively experimenting with some of the content covered in the workshop, thus reinforcing some of the concepts discussed.

All UPLOrcShops are recorded using screen capturing software, mainly so that they can be reviewed and improved upon, but also so that members can view missed content. I reflect on all activities in weekly UPLOrcShop documents, briefly discussing an overview of the particular workshop, preparations for rehearsals and workshops and, problems in the session and possible solutions. These documents are distributed to members of the ensemble so they may be informed of what they should prepare for the next session, thereby also becoming familiar with the content covered during the workshop. All videos that are made available to ensemble members are archived, unlisted videos that are uploaded to YouTube. These are only accessible to individuals who have access to the URL link.

To maintain engagement with other members of the ensemble, I have developed strategies to provide them with a wide range of instructional and educational tools to learn and experiment with TidalCycles. Some of these include content in the form of instructional videos where I discuss a chosen topic and dissect it from a technical point of view. These videos follow a similar format and approach such

as the multitude of tutorials available on YouTube. Alex McLean himself has a series of freely available TidalClub Tutorials I include as supplemental content [17]. This typically involves a demonstration of how I would approach a completely improvised line of code, explaining the thought process and decisions that were made. Further tasks are given to ensemble members in the form of problem sets, a set of instructions provided in plain English. This is deliberately done so that members will eventually be able to compile TidalCycles code drawing from what they have learned in workshops and tutorials. Should members choose to complete this task, it should assist them in preparing sufficiently for rehearsals.

Conducting UPLOrcShops in this manner, in accordance with many other university music departments (Cheng 2019), is purely pedagogically motivated and attempts to facilitate the development of essential twenty-first century skills (Feerrar 2019). At the time of writing it is not clear whether any skills have been developed by the current members of UPLOrc (not including myself). I, on the other hand, have experienced increased awareness of my sonic surroundings when I collaborate with others (Cheng 2019). This is futher explored in Laubscher (Forthcoming).

4.2 Rehearsals

UPLOrc rehearsals are currently held on Fridays for one hour, divided into two parts. The first, is a thirty minute "jam" [18] in which we improvise the entire performance. As our end-of-cycle performance/s draw near we move to rehearsing pre-planned improvised or "comprovised" content, an idea put forward in the work of Dudas (2010), Tsabary (2012) and Tsabary & Woollard (2014). We attempt to extend this notion that live coding in laptop performance can be approached from either a compositional or improvisational perspective, or a combination of the two. For instance, Albert (2012), reports on a similar approach taken by the Laptop Orchestra of Louisiana (LOLs), where performers are improvising within a structured, pre-planned arrangement of musical events. These events are often organised in terms of their duration, density, and gestural structure.

The second portion of UPLOrc rehearsals involve a post-rehearsal discussion on our Discord server channel with the same name. I use this as an opportunity to determine how members are responding to the content. A general question is given to members each week to determine whether anyone had any issues or problems that they'd like to bring to everyone's attention, or sometimes whether anything in particular stood out to them. I pose this general question to prompt thoughts and extract ideas from my collaborators, seeking to promote a collaborative atmosphere where everyone has the opportunity to express their musical thoughts and ideas freely. Of course anyone is free to prompt ideas or ask questions, as it is intended to be a collaborative project where we interact with, and explore each others point of view.

4.3 Performance preparation

UPLOrc first performed online in December 2020, as part of performance cycle one. Together with our performance on 24 June this year, we felt that we needed to re-examine how we plan and execute our performances. While these performances each had several interesting moments, at times it became challenging to hear ourselves. It was clear that we needed to incorporate a wider range of sampled sounds. Additionally, we sought to include more rhythmic, harmonic and melodic material. I decided to compile a new strategy that would best suit our current situation and subsequently distributed a package of documents and tutorial content to ensemble members. In this I describe multiple strategies on how to express various musical elements using TidalCycles functions.

Some useful strategies have emerged from planning our second performance cycle. First, I explicitly list each action that needs to be performed in a step-by-step manner. Instructions are provided, asking the performer to make small changes, in addition to

computing their code more often. In an attempt to incorporate a wider range of musical phrasing UPLOrc turned to two applicable TidalCycles functions. The *struct* and *up* functions enabled UPLOrc to incorporate rhythmic and harmonic transformations in combination with our chosen sound palette for the performance. This, and all our performance preparation documents, are accessible from a dedicated GitHub repository. [19].

UPLOrc members are also provided with an additional document referred to as a "Cheat Sheet." This has been a useful tool for UPLOrc, especially for the beginning live coder. Taking a quick glance at the sheet often prompts ideas for improvising, and acts as a quick reference guide for some essential TidalCycles functions. With the wide range of functions that TidalCycles offers, I have noticed that novices tend to struggle retaining all the functionalities of the environment. This sheet was compiled to assist members with this problem. An essential advantage of working with TidalCycles is that users are, usually within a short amount of time, able to compose complex combinations of TidalCycles functions [20]. More often than not, these function combinations produce equally complex musical material. For that, essential functional programming knowledge (McLean 2014) has become of great important when it comes to understanding the behaviour of a particular TidalCycles function. For example, when examining a function such as *struct*, one can deduce that it accepts a boolean pattern (true and false values), expressed in binary numbers (zero's and one's). *struct* therefore, is useful for compiling rhythmic patterns with TidalCycles code, by simply assigning a pattern of binary numbers to a *struct* function:

```
struct "0 1 1 0 1 0" $ s "bleep:2" #up "1 3"
```

Studying and analysing TidalCycles functions in terms of their behaviour and construction, has stimulated a greater understanding of the environment and what is musically possible within the constraints we have set for ourselves. Design constraints are essential for defining the limitations of "musical expression" (Magnusson 2010, p. 69), and while developing an understanding of the technology UPLOrc uses to perform, I would argue that the most important aspect of our preparation is owed to a combination of these two perspectives. Magnusson (2010) further points to the importance of time spent experimenting and discovering the constraints within which an entire ensemble is able to perform using a programming language similar to, but not excluding, TidalCycles.

5 Lessons learned

Throughout this article I have pointed to the majority of the problems that UPLOrc has encountered thus far. An additional problem we have yet to address is the manner in which we express our musical ideas. This is an avenue I feel needs to be explored further in my research. Combining our collective experience and knowledge from a variety of fields in musicology such as composition, performance, technology and education, is our greatest advantage. With this combined expertise we aim to further develop and cultivate our identity as an ensemble, in the process of experimentation, exploration and presentation of ourselves in the form of live streamed YouTube content. Due to the limitations of some of our current equipment UPLOrc will, for the foreseeable future, continue experimenting and performing with TidalCycles. Though completely sufficient for UPLOrc at this time, in the near future we aim to incorporate other technology into our performances. This would require additional invested time and funding to learn how to navigate these tools.

The aforementioned problems some of us experience with "glitching," is a continuous issue for UPLOrc'ers. Not only do we need to monitor the constraints of the MiniTidal language in Estuary, but so too do we have to monitor Estuary to ensure that our audio output remains without any unintended glitched effects. We did manage to maintain a steady audio output during our YouTube event on 31 July, at the cost of having to reduce the amount of func-

tions we were able to include in our performance. The way in which I planned and compiled instructions for our final performance ultimately included the perspective of one person - my own (Laubscher et al. 2021*a*). For this project to be truly considered collaborative, where participants are equally responsible for making decisions, more invested time from our members is required. Controlling and performing musical gestures in this purely instructional manner becomes problematic in that control over the musical output can only be performed in a memorised and sequential manner, as suggested by Salazar (2017) and Ogborn (2012). This ultimately limits and, to some extent, removes the control a participant exerts on their overall musical output, thereby depriving them of their individual musical agency (Knotts 2015, Bishop 2018).

6 Conclusion and future research

As the principal researcher of my forthcoming research titled, *Establishing a laptop orchestra in South Africa: An emic-centred inquiry into computer music performance* (Laubscher Forthcoming), I aim to further understand the interactions between myself and other network musicians I encounter. This article has provided me with an opportunity to reflect on the work I have completed thus far. In this process of writing this reflective piece, and as a member of the live coded and network music communities, I have been able to reach some initial assumptions about my work. At the time of writing, my current research examines whether and to which extent a novice live coder is able to develop a musical identity as a network performer within a pre-determined set of constraints (Bishop 2018). The study will be conducted with my own progress as a live coding performer in mind, and as such is presented from an insider's perspective (Morey & Luthans 1984). Through interaction, communication, observation and experimentation, UPLOrc is closer to establishing a distinct musical identity - an identity that is in constant flux. Similar to the exploration of the musical possibilities of new modes of connectivity and communication through the use of MiniTidal and Estuary, UPLOrc is constantly redefined through the development of our individual identities as performers of network music (MacDonald et al. 2002).

While pre-determined musical parameters and technological constraints may limit the possibilities of musical expression, the musical decisions and actions of members of an ensemble should not. The question "does technology facilitate or constrain creativity" posed by (Bishop 2018, p. 13), and placed in the context of collaborative live coded performance, remains unanswered at this time. I hope my forthcoming research will provide more information that extends to a more complete answer of this question. Myself and the other members of UPLOrc recognise that we have much to learn as an ensemble and as individuals within the current constraints and limitations we currently face. We intend to extend and expand on the practices we have developed thus far, with particular attention to restoring performer agency through increased engagement and development as live coding musicians.

Notes

[1] Located in Ontario, Canada, Research at NIL is focused on developing media and software for collaborative network music performance and is funded by the Social Sciences and Humanities Research Council of Canada (SSHRC). Visit https://nil.mcmaster.ca for more information.

[2] Accessible from `https://estuary.mcmaster.ca/`

[3] Six performances to be exact. See `https://www.youtube.com/playlist?list=PLroSCmh5yBWAHsSjTMY3hXtNoVB1I8Snh`

[4] Science, Technology, Engineering and Mathematics education.

[5] See `https://orcid.org/0000-0003-1947-7055`

[6] See `https://www.up.ac.za/school-of-the-arts/article/2821812/public-lectures-`

[7] https://slack.com/intl/en-au/

[8] https://discord.com/

[9] Accessible at https://github.com/dktr0/estuary

[10] https://www.chromium.org/

[11] For more in-depth information about how to access and use Estuary, see Ogborn (2019, June 11) and Ogborn (2020, December 3).

[12] Other live coding environments hosted on Estuary include CQenze, LaCalle, Sucixxx, Togo, BlackBox, Punctual, CineCero, TimeNot, Seis8s and Hydra.

[13] https://github.com/dktro/Punctual

[14] Integrated Development Environment

[15] I say "almost all" because this is also dependent on the computing abilities of the device a particular member is using

[16] Five years since Estuary was first released.

[17] See https://www.youtube.com/watch?v=M-Y5pAEBXXQ&list=PL2lW1zNIIwj3bDkh-Y3LUGDuRcoUigoDs

[18] Another common term used by live coders on the Estuary platform.

[19] See https://github.com/djmelan3/Academic_Articles/tree/main/DHASA_2021

[20] See https://tidalcycles.org/ for detailed information concerning the capabilities of TidalCycles

Acknowledgements

I'd like to thank Dr. Miles Warrington, faculty member at the University of Pretoria School of the Arts, for giving me the opportunity to manage UPLOrc for the past three years, but most importantly for introducing me to the world of live coding. It has opened my mind in ways I never thought possible. A special thank you to David Ogborn and my co-members of SuperContinent, with whom I have created a close bond despite being separated by such great distances. Thank you for all you have shared and contributed to my experiences as a member of SuperContinent and early-career academic. I am truly grateful. An enormous thank you to my mom and dad, without whom this article and my forthcoming dissertation would not be possible. To my Nicola, thank you for your kindness, love and patience (and for proofreading my work). Your support means everything.

References

Albert, J. (2012), 'Improvisation as tool and intention: Organizational practices in laptop orchestras and their effect on personal musical approaches', *Critical Studies in Improvisation/Études critiques en improvisation* **8**(1).

Berdahl, E., Pfalz, A., Blandino, M. & Beck, S. D. (2018), Force-feedback instruments for the laptop orchestra of louisiana, *in* 'Musical haptics', Springer, Cham, pp. 171–191.

Betancur, C., Khoparzi, A., Knotts, S., Laubscher, M., Marie, M., Ogborn, D., Oka, C. & Tsabary, E. (2021), 'Supercontinent: Global, collective live coding improvisation', *International Conference on New Interfaces for Musical Expression* . https://nime.pubpub.org/pub/f3hum4he. **URL:** *https://nime.pubpub.org/pub/f3hum4he*

Bishop, L. (2018), 'Collaborative musical creativity: How ensembles coordinate spontaneity', *Frontiers in psychology* **9**, 1285.

Boden, M. A. (2004), *The creative mind: Myths and mechanisms*, Routledge.

Carôt, A., Krämer, U. & Schuller, G. (2006), Network music performance (nmp) in narrow band networks, *in* 'Audio Engineering Society Convention 120', Audio Engineering Society.

Cheng, L. (2019), 'Musical competency development in a laptop ensemble', *Research Studies in Music Education* **41**(1), 117–131.

Collins, N. (2003), 'Generative music and lap-

top performance', *Contemporary Music Review* **22**(4), 67–79.

Collins, N., McLean, A., Rohrhuber, J. & Ward, A. (2003), 'Live coding in laptop performance', *Organised sound* **8**(3), 321–330.

Dudas, R. (2010), "Comprovisation": The various facets of composed improvisation within interactive performance systems', *Leonardo Music Journal* **20**, 29–31.

Fasciani, S. (2020), 'Network-based collaborative music making'.
URL: *https://www.youtube.com/watch?v=GZCueJeg168*

Feerrar, J. (2019), 'Development of a framework for digital literacy', *Reference Services Review* .

Ferguson, S. & Wanderley, M. M. (2010), 'The mcgill digital orchestra: An interdisciplinary project on digital musical instruments.', *Journal of Interdisciplinary Music Studies* **4**(2).

Freeman, J. & Troyer, A. V. (2011), 'Collaborative textual improvisation in a laptop ensemble', *Computer Music Journal* **35**(2), 8–21.

Knotts, S. (2015), 'Changing music's constitution: Network music and radical democratization', *Leonardo Music Journal* **25**, 47–52.

Knotts, S. (2018), Social Systems for Improvisation in Live Computer Music, PhD thesis, Durham University.

Knotts, S. & Collins, N. (2014), The politics of laptop ensembles: A survey of 160 laptop ensembles and their organisational structures., *in* 'NIME', Citeseer, pp. 191–194.

Laubscher, M. (Forthcoming), Establishing a laptop orchestra in south africa: An emic-centred inquiry into computer music performance, Master's thesis, University of Pretoria: School of the Arts.

Laubscher, M., Warrington, M. S., Hannay, B. M. & Allen, G. (2020), 'University of Pretoria Laptop Orchestra (UPLOrc) Performance: Estuary 5th Birthday Live Stream'.
URL: *https://youtu.be/fd83R6gTgxY?list=PLroSCmh5yBWCwxQ6jnR4Ott1Dxt2kB7l* = 1655

Laubscher, M., Warrington, M. S., Matthysen, D., Hannay, B. M., Annandale, L. F. & Lobo Tabora, D. A. (2021*a*), 'University of Pretoria Laptop Orchestra (UPLOrc): Mid-Year Concert 2021.'.
URL: *https://youtu.be/O8Y6fhg3xLs*

Laubscher, M., Warrington, M. S., Matthysen, D., Hannay, B. M., Annandale, L. F. & Lobo Tabora, D. A. (2021*b*), University of Pretoria Laptop Orchestra (UPLOrc) Performance: Siyaphumelela Day 3: Afternoon session, *in* 'Siyaphumelela Conference 2021'.
URL: *https://youtu.be/m6DJj$_G$AOcw?t* = 8646

MacDonald, R. A., Hargreaves, D. J. & Miell, D. (2002), *Musical identities*, Oxford University Press.

Magnusson, T. (2010), 'Designing constraints: Composing and performing with digital musical systems', *Computer Music Journal* **34**(4), 62–73.

Marie, M., Knotts, S., Tsabary, E. & Laubscher, M. (Forthcoming), Layers of unpredictability: Developing the aesthetic and identity of a network-based live coding ensemble, *in* 'International Conference on Live Coding (ICLC)'.

McLean, A. (2014), Making programming languages to dance to: live coding with tidal, *in* 'International workshop on Functional art, music, modeling & design (FARM), pages=63-70'.

Morey, N. C. & Luthans, F. (1984), 'An emic perspective and ethnoscience methods for organizational research', *Academy of Management Review* **9**(1), 27–36.

Nilson, C. (2007), Live coding practice, *in* 'Proceedings of the 7th international conference on New interfaces for musical expression', pp. 112–117.

Ogborn, D. (2012), 'Composing for a networked, pulse-based, laptop orchestra', *Organised Sound* **17**(1), 56–61.

Ogborn, D. (2016), 'Live coding together: Three potentials of collective live coding', *Journal of Music, Technology & Education* **9**(1), 17–31.

Ogborn, D. (2019, June 11), 'NIL Live Coding Intensive 2019: Using Estuary in different situations workshop'.
URL: *https://www.youtube.com/watch?v=9rOi8bpC9XYt=145s*

Ogborn, D. (2020, December 3), 'Estuary 5th birthday livestream'.
URL: *https://youtu.be/fd83R6gTgxY?t=24010*

Ogborn, D., Beverley, J., del Angel, L. N., Tsabary, E. & McLean, A. (2017), Estuary: Browser-based collaborative projectional live coding of musical patterns, *in* 'International Conference on Live Coding (ICLC) 2017'.

Ogborn, D., Hattwick, I., Herrou, A., Khoparzi, A., Littler, C., Oduro, K., Roberts, A., Stewart, D. A., Tsabary, E. & van der Walt, J. S. (2021), Educational applications of the estuary live coding platform, *in* 'Digital Humanities Summer Institute (DHSI) Conference 2021'.
URL: *https://www.youtube.com/watch?v=no9TNTdGf4t=1867s*

Ogborn, D., Tsabary, E., Jarvis, I., Cárdenas, A. & McLean, A. (2015), Extramuros: making music in a browser-based, language-neutral collaborative live coding environment, *in* 'International Conference on Live Coding', pp. 163–69.

Oliveros, P., Weaver, S., Dresser, M., Pitcher, J., Braasch, J. & Chafe, C. (2009), 'Telematic music: six perspectives', *Leonardo Music Journal* **19**(1), 95–96.

Salazar, S. (2017), Searching for gesture and embodiment in live coding, *in* 'Proceedings of the International Conference on Live Coding'.

Schrooten, M. (2016), Writing efieldnotes: Some ethical considerations, *in* 'eFieldnotes', University of Pennsylvania Press, pp. 78–93.

Soon, W. & Knotts, S. (2018), Aesthetic coding: Exploring computational culture beyond creative coding, *in* 'Art Machines: International Symposium on Computational Media Art', p. 87.

Trueman, D. (2007), 'Why a laptop orchestra?', *Organised Sound* **12**(2), 171–179.

Tsabary, E. (2012), 'Comprovisation for laptop orchestra'.
URL: *https://www.youtube.com/watch?v=R2WeIMHE-Lo*

Tsabary, E. & Woollard, J. (2014), '"Whatever Works": An action-centred approach to creation and mediation in designing laptop orchestra performances', *Gli spazi della musica* **3**(2).

Wang, G., Bryan, N. J., Oh, J. & Hamilton, R. (2009), Stanford laptop orchestra (slork), *in* 'International Computer Music Conference', Citeseer.

Xambó, A. (2017), Embodied music interaction: creative design synergies between music performance and hci, *in* 'Digital Bodies', Springer, pp. 207–220.

Xambó, A., Freeman, J., Magerko, B. & Shah, P. (2016), Challenges and new directions for collaborative live coding in the classroom, *in* 'International Conference of Live Interfaces (ICLI 2016). Brighton, UK'.

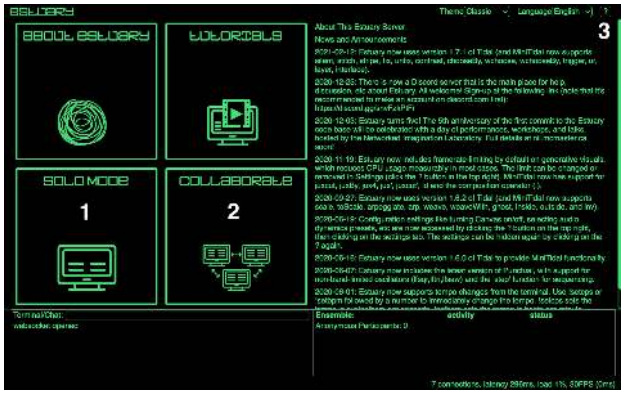

Figure 1: Estuary login screen

Figure 2: UPLOrc screen layout

Finding topic boundaries in literary text

Heyns, Nuette
North-West University, South Africa
nuette.heyns@gmail.com

van Zaanen, Menno
South African Centre for Digital Language Resources, South Africa
menno.vanzaanen@nwu.ac.za

Abstract

When performing a distant reading analysis of large amounts of literary texts, we would like to be able to automatically identify the high level structure or story lines of these texts. Story lines are not always linear, but contain transitions, such as flashbacks or changes of scenery. While working towards our goal of identifying story lines in text, we first start by identifying topic transitions. We propose a system that aims to identify a boundary describing a topic transition in the text. First, we split the text in short snippets. Next, topics are assigned to each of the snippets using LDA, a topic modelling approach. Based on this sequence of LDA topics, potential transition boundaries between snippets are identified. Potential transitions occur between snippets with the smallest intersection of the LDA topics that occur on either side of the potential transition. If multiple potential transitions are available, the system selects one at random. To evaluate this system, we apply it to the concatenation of two texts such that the real boundary is known. We provide results of this system with respect to a random baseline and an oracle system that always selects the best transition when more than one possible transition is available. The system consistently outperforms the baseline. Future work will focus on extending this system to allow for the identification of multiple transitions.

Keywords: topic modelling, LDA, boundary identification

1 Introduction

With the availability of huge amounts of texts, in depth literary analysis of all texts using manual close reading approaches is infeasible. Distant reading approaches (Moretti 2013) that rely on the automatic analysis of the texts should be considered instead. The idea of distant reading is that the computer can perform large scale and objective analyses of the texts, in contrast to the more time consuming and subjective manual analyses. (However, it is generally assumed that close reading approaches can provide a more fine-grained analysis compared to the distant reading approaches.)

One type of literary analysis deals with the identification of story lines, that can be found, for instance, in literary texts. Structuralist theorist Gérard Genette discerns four important levels of a literary text; order, duration, frequency and mood (Genette et al. 1980). We focus on the first level, order, where the sequence of events is viewed in relation to the order of narration. Many literary texts do not follow a linear story line, but apply literary techniques such as the use of different perspectives, different locations, or variations in the time line (e.g., flashbacks or flashforwards). In particular, we are interested in the transitions that occur in the story lines throughout a literary text. This allows for high level comparisons, for instance, of writing styles of different authors or structural differences in texts from different genres.

Transitions in the story line can be seen as boundaries, separating the text into parts of the text that have different properties. How these parts are different depends on the type of transition, but because the text before the transition and that after will be different in some aspect(s), we may assume that such transitions can be automatically identified based on the differences between properties of the part of the text before and after the transition.

In this article, we propose a method that aims to automatically identify a topic transition in a text. This method assumes that transitions can be identified by considering changes that can be described by

topics (as identified using the Latent Dirichlet Allocation, or LDA (Blei et al. 2003*a*), topic model). In particular, we subdivide the text into smaller snippets and apply LDA to these snippets to determine their topics. The method then analyses the sequence of LDA topics to identify potential transitions between snippets. These potential transitions occur at all positions where the intersection of the sets of LDA topics that occur before and the LDA topics that occur after the potential transition is the smallest. In other words, the system finds boundaries such that the topics occurring "on the left" of the boundary is maximally different from the topics that occur "on the right" of the boundary. These are positions where the text before and after the transition is different on the basis of LDA topics.

To evaluate the method, we construct a text by concatenating two different texts, such that the position of the real transition is known. We then apply our method, which proposes the location of a transition. The proposed transition is then compared against the real transition. To measure how well the proposed transition fits the real transition, the root mean squared error (RMSE) is computed, which takes distance into account. Lower values for RMSE are better. This system is evaluated against a baseline (which does not use the LDA model) and an oracle system (which always selects the best possible transition, in contrast to the proposed system which makes a selection from all possible transitions at random).

The LDA topic modelling system has a parameter that indicates how many topics LDA may assign to the snippets. As the system relies on the differences between the topics on both sides of the potential transition, we may expect that the number of LDA topics will have an influence of the performance of the transition identification system. In fact, in order to apply the system to the snippets of the text, we need to define the number of LDA topics beforehand, so it is useful to know more about the influence of the number of LDA topics on the performance of the system to make an informed choice when applying the system to a new text.

In this article we will focus on the following research questions.

1. Can a system that identifies a transition in a text based on LDA topics of snippets outperform a random baseline?
2. What is the influence of the random selection of the possible boundaries on the performance of the LDA based system?
3. What is the influence of the number of LDA topics on the performance of the LDA based system?

To answer the first question, we will apply the system and the random baseline to a text with a known transition and compare the results. For the second question, we compare the results of the system against an oracle system, which always selects the best of the possible boundaries. We also run the system with several values for the number of LDA topics and evaluate their performance to better understand how to answer the third question.

2 Background

The system proposed in this article depends heavily on the performance of LDA. Fortunately, there has already been research on the performance of LDA in different settings. In particular, the length of the documents given to LDA has a direct influence of the performance of LDA. We will look at this research first. Next, we briefly discuss different ways of evaluating the performance of LDA, mostly focusing on the limitations of evaluating LDA directly.

With respect to the automatic identification of transitions in literary text, unfortunately, to our knowledge there is not much previous research. Aurnhammer et al. (2019) performed a comparison between a close reading approach, which relied on manually annotated Reddit posts and a distant reading approach which relied on the identification of topics using LDA. Here the texts were already separated (as they were individual posts), but this work showed that there is a relationship between manually annotated texts and LDA texts. Similarly,

for instance, Huang & Huant (2013) and Zhou et al. (2016) genereated story lines given a collection of news articles. Gupta et al. (2009) aimed to visualise story lines (from video data), but rely on weakly labelled data.

2.1 Length of LDA documents

Sbalchiero & Eder (2020) focused on the model fitting process of topic modelling when applied to long texts. In their study, they examined the performance of LDA on literature text by splitting the text using six different sample sizes (500, 1000, 5000, 10000, 20000, and 50000). Based on this, they found that there is a relationship between the length of text chunks and the number of topics. Intuitively an extremely short text chunk and a large number of topics will provide many very specific topics, which causes the model to overfit. Reversely, an extremely large text chunk combined with too few topics will result in very broad, general topics causing the model to underfit. Sbalchiero & Eder (2020) state that given a corpus, the optimal number of topics is inversely proportional to the length of text chunks. Therefore, the larger the size of the text chunk, the lower the optimal number of topics will be. However, they also mention that the extreme cases where there are too many topics combined with a very short sample chunk will overfit the model and too few topics combined with an extremely large text chunk will underfit the model. From this statement, we can derive that the optimal number of topics to size of the text chunk should be in equilibrium. Sbalchiero & Eder (2020) conclude that the best number of topics for different sizes of the samples should be evaluated using, for example, the elbow method suggested by Kodinariya & Makwana (2013).

According to Sbalchiero & Eder (2020), previous studies have already demonstrated that LDA performs well when applied to short texts, but there is a lack of empirical evidence to show that LDA also performs well on longer texts. Syed & Spruit (2017) argue that longer text are less affected by noise in the topic-word distributions, resulting in more coherent topics. However, limited research has been done on this subject.

Jockers & Mimno (2013) indicate that the ideal size of the sample texts should be large enough to allow for the proper measurement of word cooccurences, but small enough that it can reasonably be assumed to contain a small number of themes. They found that applying LDA to full texts typically results in vague topics. However, splitting texts into approximately 1000 word samples, breaking at the nearest sentence boundary, results in more highly interpretable topics. Studies like Syed & Spruit (2017), Blei et al. (2003*b*) and others suggest using abstracts as a suitable size of sample texts.

2.2 Evaluation of LDA

LDA models can be evaluated using either extrinsic or intrinsic methods. Extrinsic evaluation methods measure LDA models' performance on a secondary task, such as document classification or information retrieval (Wallach et al. 2009). Intrinsic methods include measurements that help distinguish between topics that are semantically interpretable and topics that are artefacts of statistical inference. Usually an intrinsic method rely on the estimation of the probability of an unseen held-out data set given the trained model (Wallach et al. 2009). Popular intrinsic methods are log-likelihood and perplexity measures, as well as topic coherence.

The log-likelihood approach measures how well an LDA model fits the data. The probability of a held-out data set, not used during training, can be estimated in several ways, such as importance sampling methods, harmonic mean, annealed importance sampling, a Chib-style estimator, or a left-to-right evaluation algorithm (Wallach et al. 2009). Perplexity can also be used to measure the quality of the LDA model. Perplexity describes how well an LDA model predicts a topic for a sample by computing the normalised log-likelihood of a held-out test set. A model will be considered good when it has a high log-likelihood and, hence, a low perplexity score. Chang et al. (2009) have, however, shown that the log-likelihood and perplexity scores

have poor correlation to human judgement and are sometimes even slightly anti-correlated.

Another approach deals with the top words found in a topic. For each word, a vector representation can be created based on occurrences in large amounts of texts. Based on these vectors, the LDA topics can be evaluated by measuring the cosine distance between the words that describe the topic. If words from the same topic are closer together, the coherence is considered high. The underlining idea behind topic coherence is the distributional hypothesis of linguistics. The distributional hypothesis states that words with similar meaning tend to occur in similar contexts (Harris 1954).

Note that, ideally, human topic rankings should be available, to compare the LDA topic coherence scores to the human topic rankings. Unfortunately, in most cases topics identified by humans are not available and researchers have to rely on some automatically computed coherence score alone.

3 Methodology

3.1 Systems

The transition identification system that we propose in this article consists of four steps. First, it takes the input text T and subdivides the text into n snippets: $S = \langle s_1, \ldots, s_n \rangle$, where $T = s_1 \oplus s_2 \oplus \cdots \oplus s_n$ with $\oplus$ the concatenation operator. Second, this sequence of snippets (S) is given to the LDA system, which essentially provides a mapping LDA, which results in a sequence of LDA topics: $LDA(S) = \langle LDA(s_1), LDA(s_2), \ldots, LDA(s_n) \rangle$. Third, potential transitions are identified. Each position between two snippets, (s_x, s_{x+1}) in the sequence (with $x = 1 \ldots n-1$) is considered. For each of these positions, the size of the intersection of the set of LDA topics before this position and the set of LDA topics after the position is computed. The minimum value of all of these intersections indicates the best potential transition and there may be several positions that have the same minimum intersection sizes: $\arg\min_{x=1}^{n} |\bigcup_{i=1}^{x} LDA(s_i) \cap \bigcup_{j=x+1}^{n} LDA(s_j)|$. Finally, the system selects one of the potential transitions. If there are multiple potential transitions, it selects one at random.

The transition identification system is compared to two other systems: a baseline and an oracle system. The baseline system does not use any LDA information, but selects a transition at random from all possible positions between the snippets. This system serves as a lower limit. In contrast, the oracle system follows the regular transition identification system with one difference: when multiple potential transitions are identified, this system selects the best of these potential transitions. In other words, it makes use of information of where the real transition can be found. This method serves as an upper limit.

3.2 Data

In order to properly evaluate the performance of the system, we need to apply the system to a text in which the transition is known. For this, we create a text by concatenating two source texts that we know discuss different topics. Here, we used two books as source texts: Utilitarianism (Mill 1861) and Hide and Seek (Collins 1861). Straightforward preprocessing is applied to these text: stopwords are removed using NLTK[1], as these words occur so frequently that they do not help in identifying LDA topics of the snippets (but they do have an impact on the size of the snippets). Additionally, the text is lower cased, lemmatised, and punctuation is removed using spaCy[2].

From these two books, we selected the first 25 snippets of 500 words each, resulting in a list of 50 snippets in total with the known transition after 25 snippets. Table 1 shows a sample from both of the source texts.

3.3 Experimental settings

As mentioned before, the transition identification system relies on LDA to identify topics for each of the snippets. LDA has a parameter that sets the number of topics that LDA is allowed to assign to the snippets. As this is a manually assigned param-

Table 1: Sample from each of the two source texts.

Source text	Sentence:
Mill (1861)	desire different thing desire happiness love music desire health They included happiness They elements desire happiness made Happiness abstract idea concrete whole parts And utilitarian standard sanctions approves Life would poor thing ill provided sources happiness provision nature things originally indifferent conducive otherwise associated satisfaction primitive desires…
Collins (1861)	ruddy face suddenly turned pale left circus determined find really going behind red curtain He walked round outside building wasting time found door apply admission At last came sort passage tattered horse-cloths hanging outer entrance You can't come said shabby lad suddenly appearing inside shirt sleeves Mr. Blyth took half-a-crown I want see deaf dumb child directly Oh right go muttered lad pocketing money greedily Valentine hastily entered passage As soon inside sound reached ears heart sickened turned faint No words describe horror helplessness moan pain dumb human creature…

eter, we can vary this parameter in the experiments. In the experiments described in this article, we varied the number of LDA topics from two to 30 in steps of two. For each of the number of LDA topics, the system is run 100 times (as LDA may lead to slightly different results due to a random factor.) We provide the median, average, and standard deviation results for each of these settings.

3.4 Evaluation

To measure how well the different systems perform, we need to decide on an evaluation metric. We are interested in finding a transition that is as close as possible to the real transition (the real transition is known as we have essentially created a text by concatenating two different texts). In other words, we would like to have an evaluation metric that takes into account the distance between the proposed and real transition. For this, we use the root mean squared error, which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(p_i - r\right)^2}{n}}$$

where n is the number of runs, p_i is the position of the proposed transition position (which can range from one to 49) in run i (which ranges from one to 100, as we run the system 100 times due to the random factor of LDA and the random selection in case of multiple possible transitions) and r is the position of the real transition (at position 25). The scikit-learn Python package[3] was used to calculate the RMSE.

Note that this approach does not directly evaluate the performance of LDA, but instead focuses on how well the overall system identifies the boundaries. In other words, we perform an extrinsic evaluation.

4 Results

To investigate the performance of the transition identification system, we provide the RMSE results of the system as well as the random baseline and oracle system in Table 2. This table also shows this information for each of the settings for the number of LDA topics.

From these results we see that the RMSE of the baseline is around 15. Note that the baseline always selects a random position for the transition, which may range from one to 49, with the real transition at position 25.

Our system performs perfectly with two LDA topics as can be seen by the RMSE of 0.0 and a standard

Table 2: RMSE results of each system for the range of LDA topics. Note that the baseline does not rely on LDA and hence has no # LDA topics provided.

system	# topics	median	mean	sd
Baseline		15.0	14.603	1.067
Our	2	0.0	0.000	0.000
Our	4	0.0	0.957	2.941
Our	6	0.0	2.943	5.357
Our	8	0.0	5.814	7.437
Our	10	6.0	7.229	7.390
Our	12	6.5	9.543	8.576
Our	14	6.0	8.371	9.305
Our	16	9.0	10.057	8.485
Our	18	9.0	10.671	8.759
Our	20	12.0	12.914	8.165
Our	22	9.5	12.114	9.454
Our	24	10.5	12.086	7.910
Our	26	10.5	11.014	8.893
Our	28	10.0	12.886	8.596
Our	30	10.0	12.129	9.119
Oracle	2	0.0	0.000	0.000
Oracle	4	0.0	0.471	2.263
Oracle	6	0.0	0.843	3.242
Oracle	8	0.0	1.843	4.652
Oracle	10	0.0	3.429	6.788
Oracle	12	0.0	3.000	6.347
Oracle	14	0.0	2.514	6.611
Oracle	16	0.0	1.243	4.206
Oracle	18	0.0	1.543	4.989
Oracle	20	0.0	2.429	6.135
Oracle	22	0.0	1.700	5.176
Oracle	24	0.0	0.714	2.649
Oracle	26	0.0	1.714	5.491
Oracle	28	0.0	1.229	4.304
Oracle	30	0.0	0.914	3.202

deviation of 0.0 (remember, lower values of RMSE are better as they relate to the distance of the position of the proposed transition compared to the position of the real transition). In each run, exactly the right position for the transition is identified. Effectively, LDA identifies that there are two main topics that can be identified in the complete text and these correspond to the two original texts that were concatenated.

The performance of our system gradually deteriorates when more LDA topics are made available. When four LDA topics are available, the performance is still quite good with a RMSE of 0.957, but the standard deviation is already 2.941, which indicates that if a wrong transition is identified it may be relatively far away from the real position.

Increasing the number of LDA topics generally decreases the performance. Overall, the mean RMSE becomes larger, indicating that more often incorrect positions for the transition are proposed. The standard deviation also becomes relatively large, which again indicates the spread of proposed transitions. Note that the median also becomes larger which emphasises the larger spread. The slight improvement of the system at 22 LDA topics is probably due to the random factors of LDA and the selection of the proper transition. The standard deviation is relatively large, so it is unlikely to be a real improvement.

If we now consider the performance of the oracle system, we see that the oracle system, like our system, performs well with low number of available LDA topics. The fact that our system already performed perfectly with two LDA topics means that the oracle system cannot improve as our system already always selects the best position for the transition. However, with four available LDA topics, sometimes the oracle system leads to runs that do not contain the correct transition. Here we can see the impact of the random factor of LDA as the oracle system always selects the best transition position. Incorrect possible transitions are also found, which leads to a lower score for our system with four LDA topics. The performance of the oracle system

also deteriorates with larger number of LDA topics, which means that the correct transition cannot be found in any of the proposed transitions according to the sequence of LDA topics. However, the median remains at 0, indicating that often the correct transition is proposed. There are runs in which the random factor of LDA leads to sets of possible transitions where the correct transition is not proposed. In other words, the lower performance of the oracle system with larger number of LDA topics is the result of the random factor in the LDA system, whereas the difference between the oracle system and our system can be attributed to the random selection of transitions when multiple possible transitions are identified.

To support the idea that the number of potential transitions increases with the number of LDA topics, we can look at the correlation between the number of LDA topics and the number of possible boundaries. Computing Pearson's product-moment correlation results in a moderate significant ($p < .0001$) correlation between the number of LDA topics and the number of possible boundaries with $r = .698$ (an $r > .70$ is considered a strong correlation). Figure 1 shows the relationship between the number of topics and the number of possible boundaries the system identifies. The x-axis of the graph shows the number of LDA topics and the y-axis the number of potential transitions identified by the system. Note that the transparency of the points in the graph indicate how frequently that situation occurs, with darker points having higher frequency. We see that when increasing the number of LDA topics indeed increases the number of possible positions for the transitions. This results in situations where our transition identification system has a harder time as there are more possible transitions to choose from. The line is computed using the local polynomial regression fitting and the shaded area indicates the 95% confidence interval.

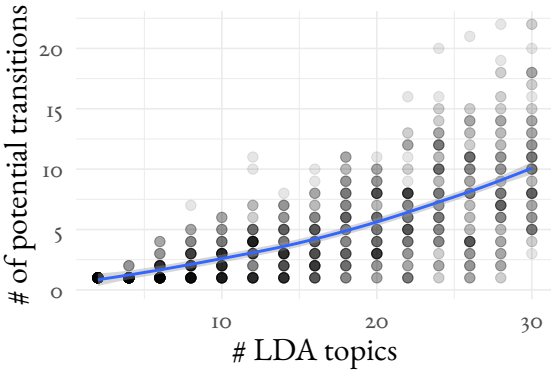

Figure 1: The relationship between the number of LDA topics and the number of possible transitions proposed by the system. Darker points indicate higher frequency of that situation. The line indicates the local polynomial regression fitting and the shaded area around the line represents the 95% confidence interval.

5 Discussion

Ultimately, we are interested in the transitions that occur in the story lines throughout a literary text. However, given the nature of story lines, this task is difficult to achieve. A more contained problem is that of identifying transitions in the topics of literary text. There should be no argument of where the topic transitions occur in the text, therefore the algorithm can be evaluated against a clear answer. We propose to build on this algorithm in the future so that it can also identify story line transitions.

In this article, we proposed a system that aims to identify topic transitions in the story line in a text by first subdividing the text into smaller snippets. This sequence of snippets is used as the input to LDA, which assigns topics to each of the snippets. Next, based on the size of the intersection of the LDA classes of the snippets "to the left" and "to the right" of each of the positions between snippets, the best potential transitions are identified. If multiple potential transitions are found, one is selected at random.

The reasoning behind using the size of the intersection of the LDA topics on both sides of the potential transition is that transitions will show a change of topics. The current system assumes that the topics at one side of the transition will not occur at the other side of the transition (or at least less frequently). This means that the approach described here relies on the performance of LDA in assigning

the correct topics.

We have evaluated the system with snippets from two texts. One may assume that it would be easiest to identify the transition between the texts if only two LDA topics are requested. The results show that this is indeed true. However, it may be the case that the snippets from one source text already contain two or more topics. In that case, LDA may have problems assigning the right topics to the snippets. Essentially, in that case, underfitting will occur. This corresponds to the idea described by Sbalchiero & Eder (2020).

It is interesting, however, that the performance goes down if the number of LDA topics goes up. The system does not directly evaluate the performance of the LDA system, it only relies on the intersection of the topics. From this, we can conclude that increasing the number of LDA topics results in the creation of topics that occur frequently on both sides of the potential transitions, which essentially introduces noise when trying to decide on the best transition. This again, shows that the system is overfitting the data, again following the results from Sbalchiero & Eder (2020).

Based on the results, we see that the proposed system can be used to identify transitions in a text. The system is relatively stable, even if the system tries to assign more LDA topics than are represented in the text, the system still has reasonable performance. Currently, however, several variables have not been evaluated yet, such as the influence of the actual texts, and the length of the snippets. We already know (again, based on Sbalchiero & Eder (2020)) that there is a relationship between these variables.

6 Conclusion

In this article, we aimed to answer three related research questions. The first question focused on the performance of the transition identification system that we introduced in this article. This system subdivides a longer text into smaller snippets, which are the input to LDA. The system then tries to identify possible transitions by considering the size of the intersection of the LDA topics on either side of the possible transition, which may occur between each pair of snippets. The positions that show the smallest intersection are considered possible transitions and if more than one is found, the system selects one at random.

The system consistently outperforms the baseline, indicating that the information that comes from LDA is indeed useful. When more LDA topics are requested, the performance goes down, but perfect results were found when LDA was run with only two topics.

The second question dealt with the influence of the random selection of the system in case multiple transitions were found. We saw that an oracle system, which always selects the best transition, leads to somewhat better results, but even with the oracle system, the performance drops when using more LDA topics. Sometimes the oracle system does lead to perfect results and sometimes it does not, which is the influence of the random factor in the LDA system.

The third question focused on the influence of the number of LDA topics the system used. We see from the result that increasing the number of LDA topics leads to lower results. This means that with higher numbers of LDA topics, additional topics that do not really seem to describe proper topics are assigned to snippets in the text. We can conclude this as they influence the performance of the system as more topics can be found on both sides of the potential transitions. Essentially, this introduces more noise in the LDA topics, due to overfitting.

7 Future work

The research described in this article shows good results, but also raises questions that should be addressed in future work. Specifically, we identify three main areas for future work.

First, the current system only identifies one transition in a text. Future work will need to focus on extending the system to allow for the identification of multiple transitions. The same evaluation strategy can be taken as it is possible to concatenate

three or more texts together. However, the evaluation metric will need to be adjusted to handle multiple boundaries. This scenario, however, is closer to the scenario we would find with a real text. It is yet unclear, exactly how the identification of the transitions will then need to take place. Perhaps a probabilistic approach which assigns probabilities for each of the positions between snippets, combined with a threshold may work. It is also unclear what the influence of the number of LDA topics will be.

Second, the current experiments were only performed on snippets from one pair of texts. Some of the specific results we found (such as the drop in performance around 22 topics) may be attributed to those texts. Experiments on additional pairs of texts, for instance, closer related semantically, may provide more insight in the actual behaviour of the system.

Finally, We may want to investigate the influence of the length of the snippets that are being used when assigning the LDA topics. From previous work, we know that LDA needs texts of a particular length in order to get reasonable probabilities to learn the topic model, but very short snippets (e.g., sentences) allow us to better identify the transitions in the text. Alternatively, we may use paragraphs as snippets, if we assume that no transition will occur within a paragraph.

Notes

[1] `https://www.nltk.org/`

[2] `https://spacy.io/`

[3] `https://scikit-learn.org`

References

Aurnhammer, C., Cuppen, I., van de Ven, I. & van Zaanen, M. (2019), 'Manual annotation of unsupervised models: Close and distant reading of politics on reddit.', *DHQ: Digital Humanities Quarterly* **13**(3).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003*a*), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003*b*), 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**(null), 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. & Blei, D. (2009), Reading tea leaves: How humans interpret topic models, *in* Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams & A. Culotta, eds, 'Advances in Neural Information Processing Systems', Vol. 22, Curran Associates, Inc.

Collins, W. (1861), *Hide and Seek*, Sampson Low.

Genette, G., Lewin, J. E. & Culler, J. D. (1980), 'Narrative discourse : an essay in method', *Comparative Literature* **32**, 413.

Gupta, A., Srinivasan, P., Shi, J. & Davis, L. S. (2009), Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 2012–2019.

Harris, Z. (1954), 'Distributional structure', *Word* **10**(2-3), 146–162.

Huang, L. & Huant, L. (2013), Optimized event storyline generation based on mixture-event-aspect model, *in* 'Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing', pp. 726–735.

Jockers, M. L. & Mimno, D. (2013), 'Significant themes in 19th-century literature', *Poetics* **41**, 750–769.

Kodinariya, T. & Makwana, P. R. (2013), 'Review on determining number of cluster in k-means

clustering', *International Journal of Advance Research in Computer Science and Management Studies* **1**(6), 90–95.

Mill, J. S. (1861), *Utilitarianism*, Oxford University Press UK.

Moretti, F. (2013), *Distant Reading*, Verso, London.

Sbalchiero, S. & Eder, M. (2020), 'Topic modeling, long texts and the best number of topics. some problems and solutions', *Quality & Quantity* **54**, 1095–1108.

Syed, S. & Spruit, M. (2017), Full-text or abstract? examining topic coherence scores using latent dirichlet allocation, *in* '2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)', pp. 165–174.

Wallach, H. M., Murray, I., Salakhutdinov, R. & Mimno, D. (2009), Evaluation methods for topic models, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ICML '09, Association for Computing Machinery, New York, NY, USA, p. 1105–1112.

Zhou, D., Xu, H., Dai, X.-Y. & He, Y. (2016), Unsupervised storyline extraction from news articles., *in* 'IJCAI', pp. 3014–3021.

An analysis of readability metrics on English exam texts

Sibeko Johannes
Nelson Mandela University
Johannes.Sibeko@mandela.ac.za

Menno van Zaanen
South African Centre for Digital Language Resources
Menno.vanZaanen@nwu.ac.za

Abstract

Readability metrics provide information on how difficult a text is to read. This information is relevant, for instance, to identify suitable texts for learner readers. Readability metrics have been developed for several languages, but no such metrics have been developed for the indigenous South African languages. One of the limitations in the development of the metrics is the availability of texts in these languages for which the readability is known. To resolve this issue, we would like to consider texts that are used in final year exams of language subjects at highschool. We expect these texts to have consistent readability throughout the years. Additionally, in South Africa, language subjects may be taught both as home language or first additional language. We expect there to be differences in readability between the exam texts for these subjects. To test these assumptions, in this article, we compute readability scores using nine existing readability metrics for the final year exams of English home language and English first additional language. The results show that indeed the readability of the texts is consistent over the years and significantly different between the two subjects. Generalizing over these results, we expect that we can use final year exam texts of other languages to develop readability metrics for the indigenous South African languages in future work. An analysis of the performance of the readability metrics on the English texts serves as a starting point to identify useful text properties to use for the development of the readability metrics for the indigenous South African languages.

Keywords: English, readability metrics, text readability, highschool exam texts

1 Introduction

The research presented in this article forms part of a bigger project that aims to develop readability metrics for indigenous South African languages. To develop these metrics, we consider using educational texts, such as reading comprehension and summary writing texts used in final year exams, as these are expected to have known or at least consistent readability. Currently, however, it is still unclear whether these exam texts indeed have consistent readability and as such are suitable for the development of readability metrics. The explorative research described here investigates readability of English comprehension and summary exam texts used in South Africa.

We focus on English since tried and tested metrics for measuring text readability in English exist. If the results for the English exam texts are as we expect, then we assume that we can use the same type of exam texts for the indigenous South African languages. The analysis of text readability using English readability metrics may also provide information on what text properties to investigate further when developing text readability metrics for the indigenous South African languages.

In the case of South African official languages, as far as could be ascertained, only Afrikaans has readability metrics. The four Afrikaans readability metrics are based on the English readability metrics (*see* Jansen, Richards and Van Zyl 2017). Fashioned after Afrikaans, we learn from the already established scholarship of text readability in English.

South African schools offer English on three levels (DBE 2012). The English home language (HL) subject is aimed at learners who start school with English competency skills such as listening, speaking, reading, and writing (DBE 2011b). The English as the first additional language (FAL) is proposed for learners who start school with some exposure to English (DBE 2011a; 2012; 2016), whereas English as the second additional

language (SAL) is intended for learners who start school with no exposure or competency skills in English (DBE 2011c). The content, teaching schedule, and the overall curricula for these English subjects are governed by the Curriculum and Assessment Policy Statements (CAPS). Curricula for other official languages are translated from the generic English CAPS (Van der Walt 2010; De Vos, Van der Merwe and Van der Mescht 2014; Tshesane 2014; Probert and De Vos 2016; Van der Walt 2018). In this article, SAL examinations are excluded for two reasons. First, SAL examinations are set at provincial level and there is no certainty that the same rigorous processes followed at National level are followed. Second, SAL examination papers are inconsistently uploaded on the DBE website and many examination papers could not be located.

Given that the FAL subject is aimed at learners with lower proficiency than those in the HL subject (DBE 2011a, p.8; 2011b, p.8), we expect that the texts used for reading comprehension in the FAL examination will be easier to read than texts used in the HL examination. Grade 12 teachers preparing learners for the final examinations are encouraged to source different texts and adapt them to their learners' levels (DBE 2017, p.5). Unfortunately, the guidelines do not specify the text characteristics that teachers can adapt, so the selected texts by the teachers cannot be used reliably in this research. Additionally, examination guidelines do not include any information on whether readability metrics are used to prepare examination papers.

In order to understand the readability of texts used in the English highschool subjects, this article sets out (i) to check whether the readability of the English reading comprehension and summary writing exam texts is consistent (that is, whether there are no differences between the readability of the texts of the different examination opportunities and whether there are differences between the HL and FAL exam texts), and (ii) to investigate whether different readability metrics are consistent with these results in order to get an idea of what text properties (used in the metrics) might be useful for the development of similar metrics for other languages.

2 Background

Measuring text readability can be approached from different perspectives. One perspective depends on readers' characteristics (Nouwens, Groen and Verhoeven 2016; Duff 2019, p.562-3; Kärbla, Uibu and Männamaa 2019; 2020; Phillips Galloway et al. 2020, p.4). From this perspective, the readability of a text depends on how well a reader can either understand the literal meaning of the text, infer meaning from the text, or use evaluative techniques to comprehend the text (Basaraba et al. 2013; Tennent 2014; Kärbla, Uibu and Männamaa 2020). As such, since text readability is viewed in relation to the specific reader, it is used interchangeably with text difficulty and reading difficulty (*see* Collins-Thompson (2014)).

Another perspective, which relates to the readability metrics used in this study, does not view text readability in relation to the reader. Instead, text readability is viewed as a subcategory of text complexity (Amendum, Conradi and Hiebert 2018, p.122), which focuses on independent linguistic factors that can be manipulated (Mesmer, Cunningham and Hiebert 2012, p.235) as opposed to how the text interacts with the reader (McNamara, Louwerse and Graesser 2002; Meyer 2003; Stahl 2003; Stenner et al. 2006; Benjamin 2012; Spencer et al. 2019). Readability metrics are described as mathematical formulas obtained through regression analysis (Mc Laughlin 1969, p.640) that are used to measure readability (Heydari 2012, p.423; Begeny and Greene 2014, p.198). They focus on the style of writing (Courtis 1987, p.20) as manifested, among others, through word and sentence lengths (Stevens, Stevens and Stevens 1992), syllable counts (Kate et al. 2010, p.547), and wordlists (Vajjala and Meurers 2014, p.3). Readability formulas generally output estimated grades or levels of education appropriate for each text, but other numeric values may also be computed.

Table 1: Extracts and summary information from the 2016 November HL and FAL examination texts.

	HL	FAL
Extract	'Hand gestures are really a powerful aspect of communication, from both the speaker's and the listener's end,' says Dr Carol Kinsey Goman, body language expert. Last year, a study analysing human gestures found that the most popular, prolific speakers used an average of 465 hand gestures, which is nearly twice as many as the least popular speakers used. Other research has found that people who 'talk' with their hands tend to be viewed as warm, agreeable and energetic, while those who are less animated are seen as logical, cold and analytical.	South Africa ranks as one of the top thirty driest countries in the world. This knowledge should encourage a new approach towards the way we use our fragile water resources. As South Africans, we have had to change our behaviour to adapt to electricity cuts, so the water crisis demands a change in our habits relating to water usage. South Africa loses billions of Rands annually through leaking taps and water pipes. It is important to repair or replace damaged water connections and washers to stop all leaks.
Sentences	3	5
Tokens	91	89
Syllables	153	142

In the South African context, studies on text readability of health documents (Joubert and Githinji 2014; Leopeng 2019; De Wet 2021) and textbooks evaluations (Sibanda 2013; Wissing, Blignaut and Van Den Berg 2016) using classical readability metrics have been conducted. However, according to our knowledge, there are no empirical studies investigating the readability of reading comprehension and summary writing texts in the domain of South African basic education.

3 Methodology

3.1 Material

The South African Department of Basic Education (DBE) affords grade 12, which is the final grade in the South African basic education schooling system, candidates two examination opportunities. The end-of-year grade 12 examinations are written in November of each year. Until 2018, supplemental examinations were written in February/March. Since 2019, the supplemental examinations are written in May/June. From 2016, low performing learners who could not cope with the grade 12 curriculum were allowed to write three subjects at the end of the academic year (November session) and complete the remaining three subjects in May/June of the following year. The Multiple Examination Opportunities Policy was discontinued after the May/June 2019 examinations (DBE 2019).

Although some exam material is not available on DBE's website for public access, most of the texts used in the exams can be found there.

Our data set comprises 48 exam texts composed of twelve HL and twelve FAL November texts from 2008 to 2019, eight HL and eight FAL February/March texts from 2011 to 2018, and four HL and FAL May/June texts from 2016 to 2019. The exam texts were manually extracted from the PDF documents, which were downloaded from DBE's website. Headings were manually punctuated to ensure the correct identification of sentence boundaries. Footnotes, endnotes, and source references were manually removed from the text. Table 1 provides example extracts from the 2016 November exam texts for the HL and the FAL examinations including some of the textual properties that are used in readability metrics.

Table 2: Classical readability formulas used in the study.

Formula	Calculation
Kincaid	$= 0.39\left(\frac{\#tokens}{\#sentences}\right) + 11.8\left(\frac{\#syllables}{\#tokens}\right) - 15.59$
Flesch	$= 206.835 - 1.015\left(\frac{\#tokens}{\#sentences}\right) + 84.6\left(\frac{\#syllables}{\#tokens}\right)$
SMOG	$= 3.1291 + 1.043\sqrt{\#polysyllabic\ words\ x\ \frac{30}{\#sentences}}$
Fog	$= 0.4[\left(\frac{\#tokens}{\#sentences}\right) + 100\left(\frac{\#complex\ words}{\#words}\right)]$
Coleman-Liau	$= 0.0588\left(\frac{\#letters}{\#samples}\right) - 0.296\left(\frac{\#sentences}{\#samples}\right) - 15.8$
ARI	$= 4.7\left(\frac{\#letters}{\#words}\right) + 0.5\left(\frac{\#words}{\#sentences}\right) - 21.43$
LIX	$= \left(\frac{\#long\ words}{\#words}x\ 100\right) + \left(\frac{\#words}{\#sentences}\right)$
RIX	$= \frac{\#long\ words}{\#sentences}$
Dale-Chall	$= 0.0496\left(\frac{\#words}{\#sentences}\right) + 11.8\left(\frac{\#difficult\ words}{\#words}\right)x\ 0.1579 + 3.6365$

3.2 Procedure

To evaluate the readability of the different texts, we compute the readability according to nine well-known readability metrics, namely, Flesch-Kincaid Grade Level (Kincaid) (Kincaid et al. 1975), Flesch Reading Ease (Flesch) (Flesch 1948), Simple Measure of Gobbledygook (SMOG) (Mc Laughlin 1969), Gunning Fog index (Fog) (Gunning 1952; 1969), lisbarhetindex (LIX) and Rate index (RIX) (Anderson 1983), Automated Readability index (ARI) (Senter and Smith 1967; Kincaid and Delionbach 1973), Coleman-Liau index (Coleman and Liau 1975), and the Dale-Chall index (Dale and Chall 1948). The formulas used in each of the metrics are presented in Table 2.

We have used the Python readability package (version 0.3.1) to compute these. All of these metrics have been developed specifically for English texts. Note that for all metrics, lower scores imply easier to read texts, except for the Flesch metric which shows higher scores for easier to read texts.

'Polysyllabic words' as used in SMOG, and 'complex words' as used in Fog, refer to words with more than two syllables (Eltorai et al. 2015, p. 831; Harden 2018, p. 37). Fog does not count proper nouns and three-syllable words formed by adding suffixes such as -es and -ed.

In SMOG, one uses three samples of ten sentences each, one from the beginning of the text, one from the middle and one from the end of the text (Mc Laughlin 1969, p. 639; Zhou, Jeong and Green 2017, p. 100). The summed results from the samples are then used in the formula. The Coleman-Liau formula divides the text into shorter pieces of 100 words each. The 100-word pieces of text are each analysed individually and the averages are used in the calculations. The LIX and the RIX formulas use 'long words' to signify words with more than six characters. It is suggested that for calculation of both LIX and RIX, ten samples of ten sentences be used for the analysis (Anderson 1983, p. 495). As the exam texts are below 100 sentences each, no sampling was necessary.

Table 3: Mean scores for text properties used in the metrics.

	February		May		November		Overall	
	FAL	HL	FAL	HL	FAL	HL	FAL	EngHL
Tokens	915.25	1173.75	891.00	1157.00	962.42	1134.67	933.70	1151.42
Syllables	1347.25	1726.50	1315.00	1774.75	1373.42	1727.00	1353.87	1734.79
Syllable/ word	1.47	1.47	1.48	1.54	1.43	1.52	1.45	1.51
Sentences	55.63	64.38	54.00	63.75	59.17	59.08	57.13	61.63
Words/ sentence	16.71	18.27	16.53	18.30	16.45	19.48	16.54	18.88
Letters	4400.25	5562.63	4283.50	5648.75	4534.75	5539.50	4443.48	5565.42
Letters/ word	4.81	4.74	4.81	4.89	4.72	4.87	4.77	4.83
Long words	221.13	278.50	204.50	308.25	220.75	296.92	218.13	292.67
Complex words	121.63	166.88	113.75	193.25	108.83	179.58	113.74	177.63
Difficult words	281.75	375.00	269.50	395.25	294.00	388.33	285.09	385.04

Dale-Chall uses 'difficult words' to signify words that do not appear in the wordlist of 3000 frequently used words. Commonly used words are identified as words in the list together with plurals of basic words in the list, -s, -ed, -ing, and -ied verbs, -ly adverbs, names of people and organisations with organisation names only being counted two times per 100-word sample, abbreviations, and compound words if both words appear on the list (Barry and Stevenson 1975, p. 219). For our data sets, we used the basic setting of the readability package which samples four evenly spaced 100-word samples for each text. This type of sampling is recommended (Dale and Chall 1948, p. 37).

3.3 Analysis

To analyze the performance of the different readability metrics on the HL and FAL exam texts, we will first provide mean and standard deviation values of each of the metrics. Additionally, we investigate correlations between the results of the different metrics. As the different metrics aim to describe the same property of the text, we expect there to be relatively high significant correlations.

Once the descriptive statistics are provided and discussed, we create linear regression models for each of the readability metrics. This indicates the relationship between the readability of the HL and FAL texts, including the years and months of the exams. We expect these analyses to identify significant differences between the HL and FAL texts and we expect no significant differences based on the years and months.

4 Results

4.1 Descriptive statistics

To get a better understanding of the behavior of the metrics on the texts, first, an overview of the textual properties used in the metrics discussed in this article is presented in Table 3. Second, we provide mean and standard deviations for the different metrics for both the HL and FAL texts in Table 4.

Table 4: Means and standard deviations (within brackets) of the different readability metrics for the English HL and FAL exam texts.

Metric	HL	FAL
Kincaid	9.53 (1.49)	7.99 (1.10)
Flesch	60.18 (8.60)	67.18 (6.20)
SMOG	12.26 (1.12)	10.74 (0.94)
Fog	13.68 (1.57)	11.51 (1.25)
ARI	10.76 (1.59)	9.28 (1.35)
Coleman-Liau	11.06 (1.30)	10.42 (1.17)
LIX	44.22 (4.30)	39.97 (3.61)
RIX	4.80 (0.92)	3.88 (0.67)
Dale-Chall	9.85 (0.48)	9.30 (0.60)

Table 5: Correlation between the different readability formulas. All correlations are significant (p<.0001).

Metrics	Kincaid	Flesch	SMOG	Fog	ARI	Coleman-Liau	LIX	RIX	Dale-Chall
Kincaid	1.00	-.95	.94	.96	.96	.85	.95	.96	.78
Flesch	-.95	1.00	-.93	-.92	-.87	-.93	-.91	-.88	-.78
SMOG	.94	-.93	1.00	.99	.87	.81	.89	.89	.73
Fog	.96	-.92	.99	1.00	.90	.79	.91	.92	.74
ARI	.96	-.87	.87	.90	1.00	.85	.96	.97	.81
Coleman-Liau	.85	-.93	.81	.79	.85	1.00	.89	.84	.84
LIX	.95	-.91	.89	.91	.9	.89	1.00	.99	.84
RIX	.96	-.88	.89	.92	.97	.84	.99	1.00	.81
Dale-Chall	.78	-.78	.73	.74	.81	.84	.84	.81	1.00

In Table 4, we see consistent differences in the scores where FAL texts are considered to be easier than the HL texts. These results do not consider the influence of the different months or years. This will be investigated in more detail with the linear regression models. The results of Pearson's correlations between the results of the different formulas are presented in Table 5.

All of these correlations are significant (p<.0001). We see that most pairs of metrics show strong positive correlations, except for the Flesch metric, which (in contrast to the other metrics) shows strong negative correlations as higher values mean easier-to-read texts. The lowest absolute correlations are found for Dale-Chall and SMOG (r=.73), and Dale-Chall and Fog (r=.74) metrics (although these are still considered strong correlations). Overall, these results show that the metrics provide very similar behavior.

We also present density plots (see Figure 1) for the different metrics for both HL and FAL texts. This shows that the readability scores are generally normally distributed.

4.2 Linear regression analyses

To investigate the influence of the subject (HL and FAL), and year and month of the exam on the readability, we created linear regression models for each of the readability metrics. For this, we use Subject (HL vs FAL), Year, and Month as independent variables (we also consider the possibility of interaction between the last two variables in the model) and the readability score as the dependent variable.

These results do not consider the influence of the different months or years.

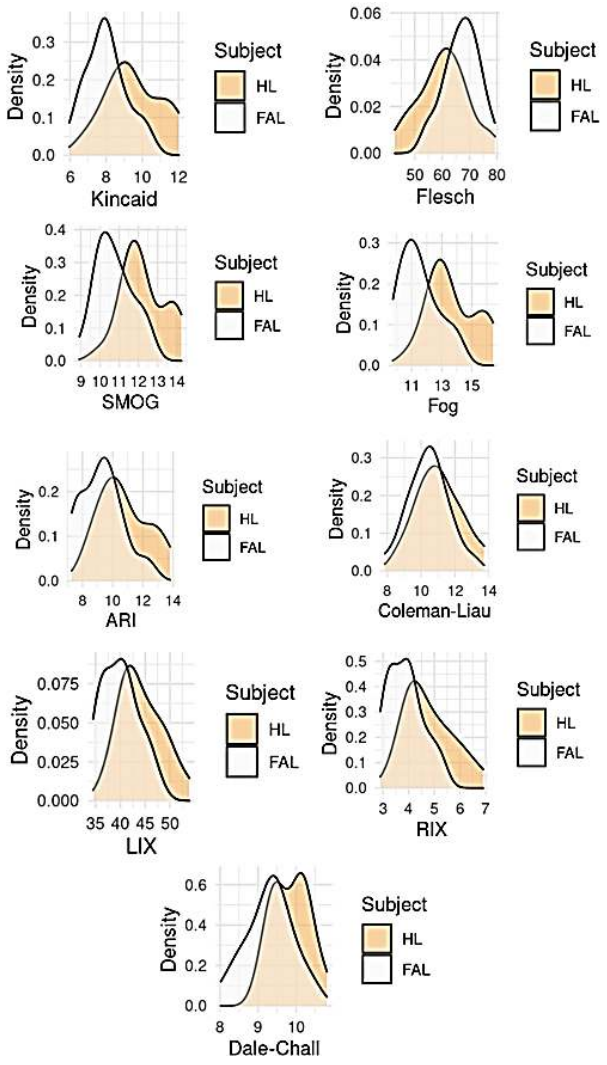

Figure 1: Data distributions density plots for all of the metrics, separated by the HL and FAL texts.

Table 6: Linear regression results. The difference indicates the estimated difference between the HL and FAL values. Error indicates the standard error. F(1, 23) provides the results of the F test with the corresponding p values.

	Difference	error	F(1, 23)	p
Kincaid	-1.54	0.38	16.8471	.0004
Flesch	7.01	2.08	11.3126	.0027
SMOG	-1.53	0.32	23.2077	<.0001
Fog	-2.16	0.44	24.5655	<.0001
ARI	-1.48	0.42	12.4651	.0018
Coleman-Liau	-0.65	0.31	4.3258	.0489
LIX	-4.25	1.13	14.0233	.0011
RIX	-0.92	0.23	15.3211	.0007
Dale-Chall	-0.55	0.12	21.5054	.0001

This will be investigated in more detail with the linear regression models. Although the results are consistent, the grade levels as used in these metrices are not purposed for the South African schooling system. As such, the actual grade levels cannot be determined using these metrics at this point. Both Year and Month are sum-coded, allowing investigation of the influence of these variables with respect to the mean values. Note that although the readability scores are strongly correlated, they are not combined in any model.

For all metrics, the residuals conform fairly well to the normality assumption (according to their histogram and Q-Q plots), although the Kincaid, Coleman-Liau, and Fog metrics show slight deviations from the normal distribution. Similarly, the homoscedasticity for all models is also good (according to the residual plots). The information of all the linear regression models can be found in Table 6.

As you can see, all the linear regression models indicate significant differences between the readability values for both subjects (HL vs FAL) for each of the metrics. Note that the models (according to ANOVA analyses) did not show any significant differences resulting from the year and month variables ($p>.05$), except the model for Coleman-Liau, which showed a significant influence of Year ($F(11, 23) = 2.3303$; $p=.04$) and the model for Dale-Chall, which showed a significant interaction between Year and Month ($F(10, 23) = 3.2075$; $p=.01$). Note that Dale-Chall already showed the largest deviations in correlations with other metrics, which may be due to the significant influence of the interaction between the Year and Month variables.

5 Discussion

Our results indicate that grade 12 examiners for the HL and FAL subjects have selected texts that are consistent over time (no significant differences between year and month) and different for each level (significant differences between HL and FAL). This corresponds to the viewpoint that the HL subject is more complex and caters for learners with higher language competency than those in the FAL subject, as it is supported by the text readability of English exam texts.

From this result, we hope that this will also be the case for the exam texts of the indigenous languages. As mentioned in the background section, the indigenous languages' curricula are often translated from the English curricula. Because of this, we hope that the selections of texts in the indigenous languages also mirror the English practices in as far as readability is concerned.

The nine metrics investigated in this article all show similar behaviour. First, the correlation results indicate no significant differences between any of the metrics. All metrics correlate strongly when considering the readability of the exam texts. Second, the metrics provide similar linear regression models with only minor differences.

There are two unexpected findings. First, Coleman-Liau shows a strong positive correlation with other metrics in the study, but the linear regression model indicates that there are significant differences between the different years. This may be because Coleman-Liau splices texts into pieces of 100 words each and then uses the averages to calculate the overall outcome.

Second, Dale-Chall linear regression model also shows statistically significant differences in terms of Years and Months. Unfortunately, this explorative research does not explore these peculiarities in detail. Nonetheless, if one is to use a metric fashioned after the Dale-Chall index,

a list of frequently used words would need to be generated in the language of choice. At this point, such lists do not yet exist in the indigenous South African official languages. An exception is McKellar's list of frequently used Afrikaans words compiled for the Afrikaans text readability metric (Jansen, Richards & Van Zyl 2017, p.154).

6 Conclusion

In this article, we explored the readability of the English HL and FAL highschool exam texts used in South Africa. We used nine classical readability metrics to investigate the readability of each text. We showed that the nine metrics are significantly and positively correlated to each other. Linear regression models showed that there are consistent significant differences between the HL and the FAL texts. Moreover, the models did not identify significant differences of the HL or FAL texts used at the different examinations.

One of the aims of the exploratory research described in this article was to get a sense of whether exam texts can be used in the development of text readability metrics for South African indigenous languages. Given that the indigenous language curricula are translated from the English generic curricula we may expect similar readability characteristics for the indigenous languages exam texts. That is, the HL and FAL texts for the indigenous South African languages can be expected to be significantly different while texts used at different examinations are expected to show no significant differences in readability.

The results found in this article also indicate areas for future work. For instance, one could explore reasons for the lower correlation between the Dale-Chall index and the SMOG and Fog indexes. Additionally, lists of frequently used words could be compiled for each indigenous language to explore how corpus-based metrics, such as the Dale-Chall index, affect text readability outcomes in the official indigenous South African languages. Furthermore, the syllable-based metrics show the need for the development of computational linguistic tools for the indigenous languages, such as syllabifiers to automatically identify syllables.

Acknowledgements

This research was funded in part by the National Research Foundation (NRF) South Africa (Grant number: 128875). It forms part of a larger study by the first author supervised by the second author. The Grant holder acknowledges that opinions, findings, and conclusions or recommendations expressed in any publication generated by the NRF supported research are those of the authors, and that the sponsors accept no liability whatsoever in this regard.

References

Amendum, SJ, Conradi, K & Hiebert, E 2018, 'Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension', *Educational Psychology Review,* vol. 30, no. 1, pp. 121-151.

Anderson, J 1983, 'Lix and rix: Variations on a little-known readability index', *Journal of Reading,* vol. 26, no. 6, pp 490-496.

Barry, JG & Stevenson, TE 1975, 'Using a computer to calculate the Dale-Chall Formula', *Journal of Reading*, Vol. 19, no. 3, pp. 218-222.

Basaraba, D, Yovanoff, P, Alonzo, J & Tindal, G 2013, 'Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist?', *Reading and Writing,* vol. 26, no. 3, pp. 349-379.

Begeny, JC & Greene, DJ 2014, 'Can readability formulas be used to successfully gauge difficulty of reading materials?', *Psychology in the Schools,* vol. 51, no. 2, pp. 198-215.

Benjamin, RG 2012, 'Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty', *Educational Psychology Review,* vol. 24, no. 1, pp. 63-88.

Coleman, M & Liau, TL 1975, 'A computer readability formula designed for machine scoring', *Journal of Applied Psychology,* vol. 60, no. 2, pp. 283-284.

Collins-Thompson, K 2014, 'Computational assessment of text readability: A survey of current and future research', *ITL-International*

Journal of Applied Linguistics, vol. 165, no. 2, pp. 97-135.

Courtis, JK 1987, 'Fry, smog, lix and rix: Insinuations about corporate business communications', *The Journal of Business Communication,* vol. 24, no. 2, pp. 19-27.

Dale, E & Chall, JS 1948, 'A formula for predicting readability: Instructions', *Educational research bulletin***,** Vol. 27, No. 2, pp. 37-54.

DBE 2011a, *English first additional language, Further Education and Training Phase Grades 10–12: Curriculum and Assessment Policy Statement (CAPS),* Government printing works, Pretoria.

DBE 2011b, *English home language, Further Education and Training Phase Grades 10–12: Curriculum and Assessment Policy Statement (CAPS),* Government Printing Works, Pretoria.

DBE 2011c, *English second additional language, Further Education and Training Phase Grades 10–12: Curriculum and Assessment Policy Statement (CAPS)*, Government Printing Works, Pretoria.

DBE 2012, *National policy pertaining to the programme and promotion requirements of the National Curriculum Statement Grades R – 12,* Government printing works, Pretoria.

DBE 2016*, Curriculum and Assessment Policy Statement Foundation Phase Grades 1-3: English Second Additional Language,* Government printing works, Pretoria.

DBE 2017, *Official Languages: First Additional Language Examination Guidelines,* Department of Basic Education, viewed 13 Aug 2021, https://www.education.gov.za/Portals/0/CD/2017%20NSC%20Exam%20Guidelines/AMENDED%20FAL%20GR%2012%20Exam%20Guidelines%202017.pdf?ver=2017-02-22-095618-000

DBE 2019, *The discontinuation of the multiple examination opportunities (MEO) dispensation with effect from 2020*, Circular E 29 of 2019, Department of Basic Education, viewed 26 Aug 2021, https://www.education.gov.za/Portals/0/Documents/Publications/Circular%20E29%20of%202019.pdf?ver=2019-10-28-142703-507

De Vos, M, Van der Merwe, K & Van der Mescht, C 2014, 'A linguistic research programme for reading in African languages to underpin CAPS.' *Journal for Language Teaching* vol. *48, no.* 2, pp. 149-177.

De Wet, A 2021, The development of a contextually appropriate measure of individual recovery for mental health service users in a South African context, doctoral thesis, Stellenbosch University, Stellenbosch.

Duff, D 2019, 'The effect of vocabulary intervention on text comprehension: Who benefits?', *Language, speech, and hearing services in schools,* vol. 50, no. 4, pp. 562-578.

Eltorai, AE, Naqvi, SS, Ghanian, S, Eberson, CP, Weiss, APC, Born, CT & Daniels, AH 2015, 'Readability of invasive procedure consent forms', *Clinical and translational science*, vol. 8, no. 6, pp. 830-833.

Flesch, R 1948, 'A new readability yardstick', *Journal of applied psychology,* vol. 32, no. 3, pp. 221-233.

Gunning, R 1952, '*The technique of clear writing*', McGraw-Hill: New York.

Gunning, R 1969, 'The fog index after twenty years', *Journal of Business Communication,* vol. 6, no. 2, pp. 3-13.

Harden, S 2018, Comparison of readability indices with grades 1-5 narrative and expository texts, doctoral thesis, Wayne State University, Michigan.

Heydari, P 2012, 'The validity of some popular readability formulas', *Mediterranean Journal of Social Sciences,* vol. 3, no, 2, pp. 423-423.

Jansen, C, Richards, R & Van Zyl, L 2017, 'Evaluating four readability formulas for Afrikaans', *Stellenbosch Papers in Linguistics Plus*, vol. *53*, pp.149-166.

Joubert, K & Githinji, E 2014, 'Quality and readability of information pamphlets on hearing and paediatric hearing loss in the Gauteng Province, South Africa', *International journal of pediatric otorhinolaryngology,* vol. 78, no. 2, pp. 354-358.

Kärbla, T, Uibu, K & Männamaa, M 2019, 'National Estonian-language tests: What is measured in text comprehension tasks?', *New Trends and Issues Proceedings on Humanities and Social Sciences,* vol. 6, no. 5, pp. 8-16.

Kärbla, T, Uibu, K & Männamaa, M 2020, 'Teaching strategies to improve students' vocabulary and text comprehension', *European Journal of Psychology of Education***,** vol. 35, pp. 1-20.

Kate, R, Luo, X, Patwardhan, S, Franz, M, Florian, R, Mooney, R, Roukos, S & Welty, C 2010, 'Learning to predict readability using diverse linguistic features', *23rd International Conference on Computational Linguistics' (Coling 2010)*, August 2010, Beijing, pp. 546-554.

Kincaid, JP & Delionbach, LJ 1973, 'Validation of the Automated Readability Index: A follow-up', *Human Factors,* vol. 15, no. 1, pp. 17-20.

Kincaid, JP, Fishburne Jr. RP, Rogers, RL & Chissom, BS 1975, *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*, vol. 56, pp. i-39.

Leopeng, MT 2019, Translations of informed consent documents for clinical trials in South Africa: are they readable?, masters dissertation, University of Cape Town, Cape Town.

Mc Laughlin, GH 1969, 'SMOG grading-a new readability formula', *Journal of reading,* vol. 12, no.8, pp. 639-646.

Mcnamara, DS, Louwerse, MM & Graesser, AC 2002, Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension, Technical report, Institute for Intelligent Systems, University of Memphis.

Mesmer, HA, Cunningham, JW & Hiebert, EH 2012, 'Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future', *Reading research quarterly,* vol. 47, no. 3, pp. 235-258.

Meyer, BJ 2003, 'Text coherence and readability', *Topics in Language Disorders,* vol. 23, no. 3, pp. 204-224.

Nouwens, S, Groen, MA & Verhoeven, L 2016, 'How storage and executive functions contribute to children's reading comprehension', *Learning and Individual Differences,* vol. 47**,** pp. 96-102.

Nyman, P, Kearl, BE & Powers, RD 1961, 'An attempt to shorten the word list with the Dale-Chall readability formula', *Educational Research Bulletin*, vol. 40, no. 6, pp.150-152.

Phillips Galloway, E, Uccelli, P, Aguilar, G & Barr, CD 2020, 'Exploring the cross-linguistic contribution of Spanish and English academic language skills to English text comprehension for middle-grade dual language learners', *AERA Open,* vol. 6, no. 1, pp. 1-20.

Probert, T, & De Vos, M 2016 'Word recognition strategies amongst isiXhosa/English bilingual learners: The interaction of orthography and language of learning and teaching.' *Reading & Writing-Journal of the Reading Association of South Africa*, vol. 7, no. 1, pp.1-10.

Senter, R & Smith, EA 1967, 'Automated readability index', University of Cincinnati, Ohio.

Sibanda, L 2013, A case study of the readability of two Grade 4 Natural Sciences textbooks currently used in South African schools, masters thesis, Rhodes University, Grahamstown.

Spencer, M, Gilmour, AF, Miller, AC, Emerson, AM, Saha, NM & Cutting, LE 2019, 'Understanding the influence of text complexity and question type on reading outcomes', *Reading and writing,* vol. 32, no. 3, pp. 603-637.

Stahl, SA 2003, 'Vocabulary and readability: How knowing word meanings affects comprehension', *Topics in language disorders,* vol. 23, no. 3, pp. 241-247,

Stenner, AJ, Burdick, H, Sanford, EE & Burdick, DS 2006, 'How accurate are Lexile text measures?', *Journal of Applied Measurement,* vol. 7, no. 3, pp. 307-322.

Stevens, KT, Stevens, KC & Stevens, WP 1992, Measuring the readability of business writing: The cloze procedure versus readability formulas, *The Journal of Business Communication (1973),* vol. 29, no. 4, pp. 367-382.

Tennent, W 2014, *Understanding reading comprehension: Processes and practices,* Sage, London.

Tshesane, TMJ 2014, Evaluating the functionality of the translated Sepedi Home Language CAPS for Grade 10-12, masters research report, University of the Witwatersrand, Johannesburg.

Vajjala, S & Meurers, D 2014, 'Readability assessment for text simplification: From analysing documents to identifying sentential simplifications', *ITL-International Journal of Applied Linguistics,* vol. 165, no. 2 pp. 194-222.

Van Der Walt, C 2010, 'Of shoes-and ships-and sealing-wax: A dynamic systems approach to language curriculum orientation', *Southern African Linguistics and Applied Language Studies,* vol. 28, no. 4, pp. 323-337.

Van Der Walt, C 2018, 'The category Language Structures and Conventions in the CAPS for English First Additional Language: a critical analysis', *Journal for Language Teaching,* vol. 52, no. 1, pp. 170-200.

Wissing, GJ, Blignaut, AS & Van Den Berg, K 2016, 'Using readability, comprehensibility and lexical coverage to evaluate the suitability of an introductory accountancy textbook to its readership', *Stellenbosch Papers in Linguistics,* vol. 46, pp. 155-179.

Zhou, S, Jeong, H & Green, PA, 2017, 'How consistent are the best-known readability equations in estimating the readability of design standards?', *IEEE Transactions on Professional Communication*, vol. 60, no. 1, pp. 97-111.

NLAPOST2021 1st Shared Task on Part-of-Speech Tagging for Nguni Languages

*Pannach, Franziska**
Göttingen Centre for Digital Humanities
franziska.pannach@uni-goettingen.de

Meyer, Francois
University of Cape Town
francoisrmeyer@gmail.com

Jembere, Edgar
University of KwaZulu-Natal
Jemberee@ukzn.ac.za

Dlamini, Sibonelo Zamokuhle
University of KwaZulu-Natal
DlaminiS4@ukzn.ac.za

Abstract

Part-of-speech tagging (POS tagging) is a process of assigning labels to each word in text, to indicate its lexical category based on the context it appears in. The POS tagging problem is considered a mostly solved problem in languages with a lot of NLP resources such as English. However, this problem is still an open problem for languages with fewer NLP resources such as the Nguni languages. This is owing to unavailability of large amounts of labelled data to train POS tagging models. The rich morphological structure and the agglutinative nature of these languages make the POS tagging problem more challenging when compared to a language like English. With this in mind, we have organised a challenge for training POS tagging models on a limited amount of data for four Nguni languages: isiZulu, Siswati, isiNdebele, and isiXhosa.

Keywords: Shared Task, Competition, Part-of-Speech Tagging, Southern African Languages

1 Introduction

In this paper, we present the shared task and combined results of NLAPOST2021, Nguni Languages Part-of-Speech Tagging, hosted jointly by the Digital Humanities Association of Southern Africa Conference (DHASA) [1] and the Southern African Conference for Artificial Intelligence Research (SACAIR) [2].

The objective of the shared task was to invite researchers, students and other interested parties to provide systems that can reliably predict part-of-speech tags for isiNdebele, isiXhosa, isiZulu and Siswati. The motivation behind the organization of this shared task was three-fold: Firstly, we wanted to invite young researchers and students to participate in a Digital Humanities and Artificial Intelligence related conference, where they could showcase their expertise. Secondly, we wanted to utilize the newly published CTexT POS dataset [3]. Thirdly, we hoped our participants would achieve good results with diverse machine learning systems.

This paper is structured as follows: Section 2 contains a description of the data, Section 3 describes the shared task, and related work is presented in Section 4. The winning team's submission is introduced in Section 5, and the results are presented in Section 6. The paper is concluded by Section 7.

2 Data

The first shared task on Nguni Languages Part-of-Speech Tagging (NLAPOST) covered four different African languages: isiNdebele, isiXhosa, isiZulu and Siswati.

The data was provided to us by the Centre for Text Technology at the North-West University [4].

Each tab separated data file consists of text tokens, morphological analysis, lemma, treebank-specific part-of-speech (XPOS), and universal part-of-speech (UPOS). As of date, the publication describing the part-of-speech annotated data is yet to be published. Therefore, the authors have to refer

to internal annotation protocols kindly provided to us by CTexT (CTexT 2020, Pienaar 2021).

Table 1 shows the number of tokens per language file. An example of the data format provided to the participants is shown in Table 2.

The shared task data consisted of columns for token, morphological analysis, and universal part-of-speech. The original dataset was split into training set (90 %) and test set (10 %) per language file.

3 Shared Task

The participants were asked to predict part-of-speech tags (UPOS) for all languages. They were provided with morphological segmentation, but not the full morphological analysis. The NLAPOST2021 shared task was published on codalab.com and announced on various mailing lists, social media channels, and grassroots research networks, such as Masakhane [5]. Participants were asked to register, after which they received the necessary information from the organizers. After an initial phase, in which the registered participants only had access to a small development set (two sentences per language), the training data was released. In total, participants had eight weeks to use the training data to develop their systems. Three weeks before the end of the competition, the test data was published.

Participants were free to make us of the morphological segmentation, but not required to do so. Furthermore, unified systems (one system for all languages) or individual systems (one system per language) were accepted. Participating teams were asked to submit individual files per language, containing only the token and the predicted UPOS tag.

Table 1: Size of Shared Task Dataset

	Tokens
isiNdebele	51,120
isiXhosa	49,104
isiZulu	50,166
Siswati	50,528

Table 2: Example Annotation (Siswati)

TOKEN	MORPH SEG	UPOS
Ngetulu	nga-tulu	ADV
kwaloko	kwa-loko	POSS
,	,	PUNC
kuba	ku-b-a	V
khona	khona	CONJ
kuniketela	ku-niket-el-a	V
kwekwakhiwa	kwe-ku-akh-iw-a	POSS
kwemaKomidi	kwe-ma-komidi	POSS
emaWadi	e-ma-wadi	N

Out of a number of registered participants, despite extension of the deadline for submitting systems, only one team submitted results (the submission presented in sections 5 and 6). However, the participating team delivered encouraging results across all four languages. Therefore, we invited the team to collaborate on this paper [6].

4 Related Work

A number of POS taggers have been developed over the years for poorly resourced agglutinative languages. The first reported work on POS tagging for the four Bantu languages we use for our shared task was done as part of a resource construction project for ten of South Africa's official languages (Eiselen & Puttkammer 2014). The open-source HunPOS tagger (Halácsy et al. 2007) was used on data for isiZulu, isiXhosa, isiNdebele, Siswati and achieved an accuracy of 83.83 %, 84.18 %, 82.57 %, and 82.08 % respectively.

Recently, Igbo, an agglutinative native language of Nigeria has been the subject of an effort to develop an effective POS tagger for the language (Onyenwe et al. 2019). A tagset of 70 tags was used to tag a combined corpus of 303 816 words. A wide range of POS tagging methods are used as a baseline in this study: unigram, ME, HMM, transformation-based learning and similarity-based learning. A rule-based algorithm is developed, which takes advantage of relatively accurate morphological analysis. Given the complexity of the morphosyntax of Igbo, the au-

thors were able to produce specific rules which exploit this linguistic knowledge to produce results which are superior to the baseline models developed for a non-agglutinative language like English. This rule-based approach produces accuracies ranging from 82-100% on subsets of the aggregate corpus.

Bengali is the most spoken language in Bangladesh and the second most spoken language in India. Two models were developed for this agglutinative language, a hidden Markov model (HMM) and an Maximum Entropy (ME) tagger (Dandapat et al. 2007). A tag set of 40 tags was used to annotate 3625 sentences, which amounted to approximately 40 000 words. Both the HMM and ME models also integrated morphological information in their feature set, which significantly improved the accuracy of both models. In this study, it was the HMM supervised model which performed best, achieving an accuracy of 88.75%.

The HMM also performed well on another Indic language, Assamese (Saharia et al. 2009). This national language of India is spoken by approximately 30 million people. In this case a tagset of 172 tags was used to annotate a 10 000 word corpus for training and testing the corpus. Although no morphological information was used, the POS tagger was able to achieve an accuracy of 85.64%. This figure is difficult to put into context as the paper didn't report any results for other models on the same dataset.

As popular as the HMM and ME models are for POS tagging, other techniques have also been tried on the task. Conditional Random Fields (CRFs), Support Vector Machines (SVMs) and a rule-based approach were compared on Kokborok POS tagging (Patra et al. 2012). Kokborok is a language spoken by approximately 2.5 million native Indians based in the northern region of the country. A tagset of 26 tags was used to tag a corpus of 42 537 words. Morphological analysis was used to break the words down into morphemes so that features could be developed manually for the rule-based and machine learning approaches. Of the three approaches, the SVM model performed best, achieving an accuracy of 84.46%

Table 3: The different design choices we tried while developing our system. Our submission to the shared task is indicated in bold.

Component	Option	Label
Model	Bi-LSTM	lstm
	Bi-LSTM + CRF	**crf**
Features	Characters	char
	Character 2-grams	**2gram**
	Character 3-grams	3gram
Composition	**Sum**	**sum**
	Bi-LSTM	lstm

5 Methods

Our final submission to the shared task was a bidirectional LSTM (bi-LSTM) with a conditional random field (CRF) layer, using character 2-grams as input features. We chose this system as our final submission after experimenting with different design choices, as listed in table 3. In this section we present a baseline we initially developed to compare our subsequent systems to. We then go through the different components of our own system, and explain how we arrived at our final system.

5.1 Baseline

Our baseline system is a hidden Markov model (HMM) (Baum & Petrie 1966) using words as input features. A HMM is a statistical model which assumes that each observation in a sequence is produced by an unobserved *hidden* state. The hidden states follow a Markov process (the probability distribution of each hidden state depends only on the previous hidden state), and they in turn produce observations according *emission probabilities* that only depend on the current hidden state.

For our POS tagging baseline, we model the words in a sentence as the observed sequence $x_1, x_2, ..., x_n$ and their parts of speech as the hidden states $z_1, z_2, ..., z_n$. Training a HMM requires learning transition probabilities $p(z_t|z_{t-1})$ and emission

probabilities $p(x_t|z_t)$. In our case we have a labelled training set available, so we train our model by simply counting transitions and emissions, and normalising them to obtain probabilities. To predict the POS tags of an unlabelled test sentence, we run the Viterbi algorithm, which computes the most likely hidden sequence given an observed sequence. It is a dynamic programming algorithm that computes hidden state sequence probabilities in a forward pass, and traces the most likely hidden state sequence in a backward pass.

5.2 System components

In developing our system we were faced with a number of decisions, regarding which methods to use in the components of our POS tagging system. Here we discuss the options we tried for the neural model, the input features, and word composition.

Neural models

Long short-term memory networks (LSTMs) (Hochreiter & Schmidhuber 1997) are recurrent neural networks for sequence modelling. At each step in a sequence, they update an internal hidden state vector through a number of learned *gates* that act as filters on the hidden state. The gates are computed from the current input vector $\mathbf{x_t}$ and the previous hidden state $\mathbf{h_{t-1}}$ as

$$\mathbf{f_t} = \sigma(W_f\mathbf{x_t} + U_f\mathbf{h_{t-1}} + \mathbf{b_f})$$
$$\mathbf{i_t} = \sigma(W_i\mathbf{x_t} + U_i\mathbf{h_{t-1}} + \mathbf{b_i})$$
$$\mathbf{o_t} = \sigma(W_o\mathbf{x_t} + U_o\mathbf{h_{t-1}} + \mathbf{b_o})$$

where $\mathbf{f_t}, \mathbf{i_t}, \mathbf{o_t}$ are referred to as the forget, input, and output gates respectively, and W, U, $\mathbf{b}$ are learned parameters. Using these gates, the hidden state $\mathbf{h_t}$ is computed as

$$\tilde{\mathbf{c}}_\mathbf{t} = \tanh(W_c\mathbf{x_t} + U_c\mathbf{h_{t-1}} + \mathbf{b_c})$$
$$\mathbf{c_t} = \mathbf{f_t} \odot \mathbf{c_{t-1}} + \mathbf{i_t} \odot \tilde{\mathbf{c}}_\mathbf{t}$$
$$\mathbf{h_t} = \mathbf{o_t} \odot \tanh(\mathbf{c_t}),$$

where $\mathbf{c_t}$ is referred to as the current cell state. A LSTM processes data sequentially in a single direction. A bi-LSTM is essentially two LSTMs combined - one processing the sequence in the forward direction and another processing it in a backward direction. The two are combined by concatenating the hidden states of the two LSTMs at each time step, so the combined network encodes information from both directions of the sequence. The hidden states of a LSTM are usually passed to further neural layers, that produce task specific output (POS tag probabilities in our case).

Conditional random fields (CRFs) (Lafferty et al. 2001) are undirected probabilistic graphical models (PGMs) for sequence labelling. The main advantage they offer over LSTMs is that they explicitly model dependencies between predicted outputs (LSTM predictions are conditionally independent). Given an input sequence $\mathbf{x}$ and a label sequence $\mathbf{y}$, a CRF models the conditional distribution $p(\mathbf{y}|\mathbf{x})$. It does so by computing scores $f(\mathbf{x}, y_i, y_{i+1})$ for each position in a sequence, where f is a called the feature function. These scores are then normalised to obtain the probability $p(\mathbf{y}|\mathbf{x})$, using a dynamic programming algorithm for efficient computation. f is a often a parameterised combination of handcrafted functions that encode rules for the sequence labels (e.g. syntactic rules in the case of POS tags), but it is also possible to model f as a fully trainable function (e.g. a statistical model, or neural network). As in the case of HMMs, the most likely label sequence given an observed input sequence is computed with the Viterbi algorithm.

We experimented with two neural models for our POS tagging system - a bi-LSTM and a CRF with a bi-LSTM as feature function. Our bi-LSTM model produces concatenated hidden states for all the words in a sentence. These hidden states are then passed to a fully connected neural layer, which produces probabilities for all possible POS tags. For our CRF model, a bi-LSTM produces scores for each position in a sentence (i.e. we parameterise the feature function with a bi-LSTM). The scores are taken as input by a CRF, which computes probabilities for the tagged sentence. Both our models are trained by maximising the probabilities of tagged sentences in a training sets. For optimisation we use the Adam optimiser (Kingma & Ba 2015), a popu-

lar variant of stochastic gradient descent that adapts learning rates on a per-parameter basis.

Input features

One of our earliest findings while developing our system was that subword-based systems comfortably outperformed word-based models. This is expected, because all the task languages are agglutinative (words are formed by stringing together morphemes), so subword information is crucial. Furthermore, because the task datasets are relatively small (compared to those of high-resource languages), any held-out dataset will contain many previously unseen words. Incorporating subword features enables the system to handle new words, since it can use subword information to infer word-level properties.

We initially considered training a morphological segmenter on the morphologically analysed data, but decided against it. The task data contains canonically segmented words (words are divided into their standardised morphemes, as opposed to their surface forms) and canonical segmentation is a challenging task. Moeng et al. (2021) applied various models to the task of supervised canonical segmentation for the Nguni languages, and failed to exceed 0.75 F1 for any of the languages. Therefore we reasoned that the effort of developing a supervised canonical segmenter might not be worth the potential benefit to the model.

Instead, we incorporated subword information through two conceptually simple methods that were easy to implement and experiment with. Our first method simply segments words into their characters - the word "kuba" is represented as the character sequence "k-u-b-a". Our second method represents words as sequences of character n-grams. In our experiments we found that 2-grams worked well, and found no improvement in using higher order n-grams. Here the word "kuba" is represented as the 2-gram sequence "<k-ku-ub-ba-a>", where < and > are special symbols indicating the start and end of words.

Word composition

The final component of our system concerns how subword representations are composed to form word representations. Since POS tagging is a word-level task, we need some way to build word representations that can be processed as input by our neural networks. The method we settled on consists of simply summing the subword vector representations for a word. This was shown by Zhu et al. (2019) to be robust across different languages, compared to other composition functions. However, it discards sequential and positional information, modelling each word as a "bag-of-subwords". We also experimented with a bi-LSTM that processes a word as a sequence of subword units, and produces a vector representation for the word. Ling et al. (2015) showed that this improved performance on POS tagging, especially for morphologically rich languages. However, this significantly increased the training times of our models, and we observed no performance improvement over sum-based composition. Therefore we converged on sum-based composition early in our experiments.

5.3 Experimental setup

In addition to the components discussed above, we also employed various strategies that aid the training of deep learning models. We used a schedule for the learning rate, which determines the gradient descent step size in optimisation. We repeatedly decreased the learning rate by some factor according to a specified schedule. This ensures a high learning rate at the start of training and lower learning rates as training progresses (since smaller optimisation steps are required in the vicinity of maxima). We trained the model for a predefined number of iterations (epochs) of the training set and processed the data in batches made up of a predefined number of sentences.

Furthermore, we employed two regularisation strategies to combat overfitting - dropout and weight decay. We used dropout in our neural network layers, which randomly drops (zeroes

Table 4: The hyperparameter values we used for training our models.

Input embedding size	512
Hidden state size	512
LSTM layers	1
Initial learning rate	0.01
Adjustment schedule	every 3 epochs
Shrinkage factor	0.5
Epochs	15
Batch size	64
Dropout rate	0.2
Weight decay	1e-5
Gradient clipping	1.0

out) some proportion of units in the computed vector representations during training. Weight decay regularises the model by penalising large parameter values. We also applied gradient clipping during training, which scales gradient values if they exceed some specified threshold. This prevents excessively large gradients, which can be a problem for LSTMs.

During development, we used cross-validation to assess our systems and hyperparameters. This involved training our systems on 90% of the training set, and evaluating it on the remaining 10% (we did not have access to the test set at all during development). To ensure that our assessments were not overly dependent on a particular train/validation split, we performed multiple cross-validation experiments per system, each time assessing performance on a different 90%-10% split. To assess a system, we computed the macro-averaged F1 scores across all POS tags on the validation set (and averaged this over different validation splits). We then compared different systems according to this metric, since this is the evaluation metric used to evaluate submissions on the shared task. We repeated this tuning procedure across all the languages, but generally the same optimal hyperparameter values emerged. The hyperparameter values that we settled on through this process are listed in table 4.

6 Results

Here we present the results obtained by our systems on the shared task test datasets, and discuss our main findings. The results of the systems we experimented with are summarised in table 5, and we include a full breakdown of the performance of our final submission (crf, char + sum) in the appendix. Overall our systems performed well, consistently achieving accuracies and F1 scores above 0.85, often surpassing 0.9, and even reaching 0.95 in the case of isiXhosa.

All our systems comfortably outperform the baseline. The combination of subword features and sequential neural networks proves much more effective than the word-based HMM. Deep learning models sometimes perform poorly on small datasets, but the results confirm that the shared task datasets are large enough for neural networks to train effectively and learn generalisable rules.

Among our own systems, there are two that emerge as the best performing systems across the board. These are the CRF with 2-gram features and the bi-LSTM with character features. It is not obvious why these particular combinations of features and neural models work well, but what is clear is that the introduction of subword information proves highly effective. Performance levels vary across the languages, with all models (including the baseline) achieving their highest scores on isiXhosa and lowest scores on Siswati. This points to the existence of language-specific characteristics that may contribute to the relative difficulty of the task.

7 Conclusion

In this paper, we introduced the first shared task on Nguni Languages Part-of-Speech Tagging (NLAPOST). The dataset, which includes POS tags for four African languages (isiNdebele, isiXhosa, isiZulu and Siswati), is publicly available at `https://repo.sadilar.org/handle/20.500.12185/546`.

We also presented the results of the submitting team, which introduced an CRF classifier and a bi-LSTM system. Both systems performed well on

Table 5: The results obtained on the shared task test sets by our systems. We report the F1 scores macro-averaged over POS tags, and the test set accuracies.

		isiNdebele		isiXhosa		isiZulu		Siswati	
Model	Feature	acc	F1	acc	F1	acc	F1	acc	F1
hmm	word	0.75	0.58	0.77	0.60	0.76	0.58	0.72	0.54
crf	char + sum	0.86	0.85	0.90	0.90	0.87	0.85	0.85	0.81
	ngram + sum	0.90	**0.88**	**0.95**	**0.94**	0.91	0.86	**0.90**	**0.84**
lstm	char + sum	**0.91**	0.87	**0.95**	**0.94**	**0.92**	**0.87**	**0.90**	**0.84**
	ngram + sum	0.86	0.84	0.90	0.90	0.88	0.86	0.85	0.81

the test set, with the best results for isiXhosa (0.94 F1 score) and the lowest scores for Siswati (0.84 F1 score). These results are an encouraging contribution to natural language processing for the languages in the dataset.

Although the participation in the shared task was relatively low (despite a much higher number of registered teams), we consider the NLAPOST21 shared task a success in the sense that firstly, the winning team presented encouraging results. Secondly, the organization proved a successful collaboration between SACAIR and DHASA. Lastly, even though only one team submitted results, the dataset was introduced and made accessible to many early-carreer researchers and interested scholars, who will hopefully engage with it further.

Notes

[1] `https://dh2021.digitalhumanities.org.za/`

[2] `https://2021.sacair.org.za/`

[3] `https://repo.sadilar.org/handle/20.500.12185/546`

[4] `http://humanities.nwu.ac.za/CTexT`

[5] https://www.masakhane.io/

[6] FP, EJ and SD organized the shared task, FM delivered the system as the only participating party.

Acknowledgements

We thank CTexT, North-West University, particularly Martin Puttkammer, for supplying the dataset and documentation to us and allowing us to use it in the NLAPOST21 shared task. The work of Francois Meyer was financially supported by Hasso Plattner Institute for Digital Engineering, through the HPI Research School at UCT.

References

Baum, L. E. & Petrie, T. (1966), 'Statistical inference for probabilistic functions of finite state markov chains', *The annals of mathematical statistics* **37**(6), 1554–1563.

CTexT (2020), Annotation protocol: Part-of-speech tagging (isiZulu). Unpublished Protocol.

Dandapat, S., Sarkar, S. & Basu, A. (2007), Automatic part-of-speech tagging for bengali: An approach for morphologically rich languages in a poor resource scenario, *in* 'Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions', pp. 221–224.

Eiselen, R. & Puttkammer, M. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3698–3703.

URL: *http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_paper.pdf*

Halácsy, P., Kornai, A. & Oravecz, C. (2007), Hunpos: An open source trigram tagger, *in* 'Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions', ACL '07, Association for Computational Linguistics, USA, p. 209–212.

Hochreiter, S. & Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Kingma, D. P. & Ba, J. (2015), Adam: A method for stochastic optimization, *in* 'International Conference on Learning Representations (ICLR)'.

Lafferty, J. D., McCallum, A. & Pereira, F. C. N. (2001), Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *in* 'Proceedings of the Eighteenth International Conference on Machine Learning', ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 282–289.

Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L. & Luís, T. (2015), Finding function in form: Compositional character models for open vocabulary word representation, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Lisbon, Portugal, pp. 1520–1530.

Moeng, T., Reay, S., Daniels, A. & Buys, J. (2021), 'Canonical and surface morphological segmentation for nguni languages'.

Onyenwe, I. E., Hepple, M., Chinedu, U. & Ezeani, I. (2019), 'Toward an effective igbo part-of-speech tagger', *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **18**(4), 1–26.

Patra, B. G., Debbarma, K., Das, D. & Bandyopadhyay, S. (2012), Part of speech (pos) tagger for kokborok, *in* 'Proceedings of COLING 2012: Posters', pp. 923–932.

Pienaar, W. (2021), Annotation protocol: Morphological analysis (isiZulu). Unpublished Protocol.

Saharia, N., Das, D., Sharma, U. & Kalita, J. (2009), Part of speech tagger for assamese text, *in* 'Proceedings of the ACL-IJCNLP 2009 Conference Short Papers', pp. 33–36.

Zhu, Y., Vulić, I. & Korhonen, A. (2019), A systematic study of leveraging subword information for learning word representations, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 912–932.

A Full submission results

On the next page we present a detailed breakdown of the performance of our final system on the test data. The results are broken down by part-of-speech tag, and the right-most column ("support") indicates the number of test examples per tag type. Our final system was a CRF on top of a bi-LSTM, using 2-grams as input features.

isiNdebele

	precision	recall	F1 score	support
ABBR	0.86	0.92	0.89	13
ADJ	0.98	0.88	0.93	106
ADV	0.95	0.96	0.95	693
CDEM	0.87	0.80	0.84	76
CONJ	0.93	0.95	0.94	176
COP	0.76	0.53	0.62	36
FOR	1.00	0.75	0.86	4
IDEO	1.00	0.80	0.89	10
INT	0.79	0.73	0.76	15
N	0.90	0.89	0.89	1375
NUM	0.89	1.00	0.94	8
POSS	0.91	0.92	0.91	769
PRO	0.93	0.94	0.94	71
PUNC	1.00	1.00	1.00	583
REL	0.83	0.85	0.84	444
V	0.80	0.81	0.81	647
accuracy			0.90	5026
macro avg	0.90	0.86	0.88	5026
weighted avg	0.90	0.90	0.90	5026

isiZulu

	precision	recall	F1 score	support
ABBR	1.00	0.75	0.86	8
ADJ	0.91	0.85	0.88	82
ADV	0.92	0.95	0.94	643
CDEM	0.92	0.91	0.92	107
CONJ	0.95	0.93	0.94	228
COP	0.73	0.62	0.67	65
FOR	0.67	0.67	0.67	9
IDEO	0.50	0.75	0.60	4
INT	0.82	0.75	0.78	12
N	0.89	0.91	0.90	1075
NUM	1.00	1.00	1.00	10
POSS	0.92	0.94	0.93	712
PRO	1.00	0.98	0.99	55
PUNC	1.00	1.00	1.00	590
REL	0.87	0.89	0.88	511
V	0.90	0.84	0.87	844
accuracy			0.91	4955
macro avg	0.87	0.86	0.86	4955
weighted avg	0.91	0.91	0.91	4955

isiXhosa

	precision	recall	F1 score	support
ABBR	0.94	0.94	0.94	16
ADJ	0.87	0.93	0.90	82
ADV	0.94	0.98	0.96	613
CDEM	0.98	0.96	0.97	84
CONJ	0.99	0.98	0.99	195
COP	0.86	0.74	0.80	167
FOR	0.92	0.75	0.83	16
IDEO	1.00	0.89	0.94	9
INT	0.92	1.00	0.96	12
N	0.96	0.97	0.97	1097
NUM	1.00	1.00	1.00	11
POSS	0.96	0.95	0.95	756
PRO	0.94	0.98	0.96	48
PUNC	1.00	1.00	1.00	599
REL	0.93	0.91	0.92	430
V	0.95	0.95	0.95	775
accuracy			0.95	4910
macro avg	0.95	0.93	0.94	4910
weighted avg	0.95	0.95	0.95	4910

Siswati

	precision	recall	F1 score	support
ABBR	0.88	0.64	0.74	11
ADJ	0.87	0.92	0.89	64
ADV	0.88	0.91	0.89	591
CDEM	0.82	0.68	0.74	78
CONJ	0.87	0.84	0.85	245
COP	0.72	0.57	0.64	49
FOR	1.00	0.20	0.33	10
IDEO	1.00	1.00	1.00	1
INT	0.84	0.73	0.78	22
N	0.90	0.90	0.90	1013
NUM	1.00	1.00	1.00	17
POSS	0.89	0.88	0.89	751
PRO	1.00	0.92	0.96	37
PUNC	1.00	1.00	1.00	605
REL	0.88	0.88	0.88	413
V	0.88	0.91	0.90	789
accuracy			0.90	4696
macro avg	0.90	0.81	0.84	4696
weighted avg	0.90	0.90	0.90	4696

Invited Speaker

Decolonising African Cultural/Historical Memory Data: A Digital Humanities Approach

Prof. Tunde Ope-Davies, Chair of the Digital Humanities, University of Lagos

Abstract

With the rapid expansion and deployment of technology in every sphere of human activities, it is becoming increasingly apparent that the application of approaches and tools in Digital Humanities (DH) can inspire new developments and innovation in African studies especially within the fields of humanities and liberal sciences. This presentation provides a context for scholars in this field to discuss how the application of new technologies can provoke a re-interpretation and proper (digital) documentation and globalization of a large body of existing data on historical and cultural memories in Africa. It argues that the new technology-driven approaches being postulated in DH scholarship and research orientations across some academic and research communities in Africa would provide a genuine and authentic framework for the [re-]construction and [re-]presentation of information on Africa, its histories, cultures and epistemes.