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 written 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 following 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