Deriving lexical statistics for psycholinguistic research on isiXhosa

Berghoff, Robyn
Stellenbosch University
berghoff@sun.ac.za

Abstract

Psycholinguistic research on isiXhosa and related Bantu languages is scarce. For research on lexical processing in particular, a prerequisite is data on lexical properties that impact word recognition, such as word frequency and neighbourhood density. This paper describes the derivation of these and related lexical statistics from a newly created 4.8-million-word isiXhosa corpus. It then reviews the potential applications of such a lexical database for research on language acquisition, language development, and language processing. The paper closes with recommendations for further work in this domain.

Keywords: isiXhosa, psycholinguistics, lexical statistics, lexical processing, corpus

Introduction

The vast majority of language acquisition and processing research focuses on a small subset of the world's languages, specifically those from the Germanic and Romance branches of the Indo-European language family (Bylund, Khafif & Berghoff 2022; Norcliffe, Harris & Jaeger 2015). Within the neglected language groups, the Bantu languages are particularly understudied. This narrow focus in terms of language typology severely limits the generalizability of theories of language processing.

isiXhosa is an agglutinating language, meaning that its words typically consist of multiple morphemes that are concatenated in a relatively transparent manner. It has two particular features that distinguish it from more commonly studied agglutinating languages such as Turkish, Basque, and Hungarian (see van de Velde et al. 2019 for discussion). Firstly, alongside suffixation, it makes widespread use of prefixation to produce morphologically complex words. Secondly, it has a rich grammatical gender or noun class system, whereby nouns are divided into 15 groups, with noun class agreement being marked on several syntactic constituents (e.g., verbs, adjectives/relatives, determiners). As language-specific properties of affixation (e.g., Boudelaa & Marslen-Wilson 2011) and grammatical gender (e.g., Colé, Pynte & Andriamamonjy 2003) are known to affect language processing, psycholinguistic examinations of isiXhosa and related languages have much to contribute to theories of lexical and morphological processing. Further, in terms of practical applications, an understanding of how language and literacy development proceeds in languages of this sort is indispensable in designing teaching and intervention materials for young learners (Pretorius 2019).

A sine qua non of robust psycholinguistic research into lexical processing are statistics on certain lexical properties known to influence word recognition, such as word frequency and neighbourhood density. Databases of such statistics are increasingly being developed and made available to facilitate research on more commonly studied languages. Examples include GreekLex, for Greek (Ktori, van Heuven & Pitchford 2008); EsPal, for Spanish (Duchon et al. 2013); Aralex, for Modern Standard Arabic (Boudelaa & Marslen-Wilson 2010); StimulStat, for Russian (Alexeeva, Slioussar & Chernova 2018); the Chinese Lexical Database, for Mandarin (Sun et al. 2018); P-PAL, for Portuguese (Soares et al. 2018); and E-Hitz, for Basque (Perea et al. 2006); as well as CLEARPOND (Cross-Linguistic Easy Access Resource for Phonological and Orthographic Neighbourhood Densities; Marian et al. 2012), which provides lexical data for five widely examined European languages.

Resources on African indigenous languages are scarce. Unsurprisingly, then, the kinds of data needed for robust psycholinguistic research on

isiXhosa processing and acquisition are lacking. This paper describes the generation of lexical statistics for isiXhosa. It identifies and defines the types of lexical statistics needed for lexical processing research on isiXhosa and exemplifies their calculation based on a newly created 4.8-million-word isiXhosa corpus. The paper concludes by reviewing potential applications of such a database and outlining steps for future work.

Characteristics to be included in a database of lexical statistics for isiXhosa

This section reviews the characteristics that should, at a minimum, be included in a database of lexical statistics for isiXhosa. This overview focuses on characteristics that are relevant specifically to visual word recognition.

Frequency

Frequency is arguably the most important variable in studies of lexical processing (van Heuven et al. 2014), where more frequent words are processed more rapidly than less frequent words. This effect can be explained on the basis of lexical activation, whereby more frequently encountered words have higher resting activation levels and are thus accessed more quickly than their low-frequency counterparts. A word's frequency is calculated based on the number of its occurrences in a corpus. It can be expressed on the standardized Zipf frequency scale, where the lower half of the scale (1–3) represents low-frequency words and the upper half of the scale (4–6) represents high-frequency words (words with a Zipf frequency above 7 tend to be function words). Zipf frequency is calculated as log10 (frequency per million words) + 3 (van Heuven et al. 2014).

Word length

Word length is calculated simply as the number of letters in a given word. At least in English, it has been found to have non-linear effects on word recognition independently of other variables such as number of syllables (New et al. 2006).

Neighbourhood statistics

A neighbour of a given word is any word that can be created by substituting, adding, or deleting a single letter (for example, the isiXhosa *ubisi* 'milk' has as neighbours *ubusi* 'honey' and *usisi* 'sister', among others). The neighbourhood density of a word is equal to the number of its neighbours. Neighbourhood density effects on lexical processing typically manifest as processing slowdowns for words with more neighbours (Andrews 1997), which is attributed to the fact that when recognizing a word with many neighbours, numerous candidate lexical items become activated and must consequently be inhibited for the correct item to be selected. A related variable that is also of importance is neighbourhood frequency, which is the average frequency of a given word's neighbours. Here, a word with high-frequency neighbours takes longer to recognize than a word with low-frequency neighbours (e.g., Brysbaert, Mandera & Keuleers 2018).

Method

The corpus

The calculation of lexical statistics requires a sizeable corpus of contemporary language materials. The corpus used in this paper was created by combining the isiXhosa corpora provided in the Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012) with a new corpus created by the author from the online isiXhosa newssite *Isolezwe lesiXhosa*. The Leipzig isiXhosa corpora consist of texts randomly collected from the web and Wikipedia and therefore cover a multitude of topics (see Goldhahn, Eckart & Quasthoff 2012 for discussion). The *Isolezwe* corpus, on the other hand, contains reports on general news, sports, entertainment, opinion, and agriculture. This corpus was created via web-scraping using the rvest package (version 1.0.2; Wickham 2021) in the R environment for statistical computing (version 4.2.1; R Core Team 2022). The entire history of articles that was available from the site's inception (26 June 2015) up until 24 June 2022 was scraped.

To create the final corpus, each subcorpus was read into R and subjected to basic cleaning (e.g., removal of digits) using the stringr package (version 1.4; Wickham 2019). Tokenization of each subcorpus was then performed using the "unnest_tokens" function from the tidytext package (version 0.3.3; Silge & Robinson 2016).

Details of each component of the final corpus are provided in Table 1.

Table 1: Components of final corpus

Name	Tokens
Leipzig 2013 corpus	400,323
Leipzig 2015 corpus	153,661
Leipzig 2016 corpus	424,146
Leipzig 2017 corpus	343,517
Leipzig 2018 corpus	443,931
Leipzig 2020 corpus	436,772
Isolezwe corpus	2,656,625
Total	4,858,975

All the subcorpora were combined prior to further processing. The final corpus contained 466,957 distinct tokens. This size is comparable to that used in the calculation of lexical statistics for other languages (e.g., Basque; Perea et al. 2006).

Calculation of statistics

Frequency numbers were obtained using the tidytext package in R. These raw numbers were then converted to Zipf frequencies. The other lexical statistics were calculated using the LexiCAL program (Chee et al. 2021). This Windows application allows the user to input a corpus file, which specifies the tokens and their frequency in the corpus, for any alphabetic language. It then calculates the selected metrics and outputs the results to a separate file. For neighbourhood statistics, it also provides the neighbours of the words that are included in the corpus.

Excerpts from the database

This section presents excerpts from the database. To begin with, Table 2 lists the 20 most frequent words in the corpus with their raw and Zipf frequencies.

Table 2: Twenty most frequent words in the corpus

Word	Raw freq.	Zipf freq.
ukuba	73,396	7.18
le	22,958	6.67
xa	22,712	6.67
emva	21,155	6.64
kwaye	19,588	6.61
ke	19,169	6.60
okanye	18,274	6.58
lo	16,624	6.53
kodwa	16,387	6.53
kuba	15,903	6.51
uthi	14,762	6.48
nto	14,519	6.48
abantu	13,533	6.44
kunye	12,742	6.42
kakhulu	11,809	6.39
kule	11,120	6.36
afrika	11,014	6.36
ukuze	10,996	6.35
utshilo	10,806	6.35
into	9,845	6.31

Unsurprisingly, the majority of the 20 most frequent words are function words, with the exception of *uthi* 'you/he/she says', *(i)nto* 'thing', *abantu* ‘people’, *kakhulu* ‘very, a lot’, *afrika* ‘Africa’, and *utshilo* ‘you/he/she said’.

The crucial factor in designing psycholinguistic experiments, however, is not the raw frequency of items, but matching frequency and other lexical properties across items. Table 3 presents example database entries for ten randomly selected items from the corpus. In creating an experiment, the aim would be to select lexical items that are as closely matched on the numerical values in Table 3 as possible.

Applications

A database of lexical statistics such as that described in this paper has numerous applications for research on lexical processing, which requires careful control of word-level properties such as frequency and neighbourhood density. There is, for example, significant interest in whether recognition of morphologically complex words takes place at the whole-word level or whether it entails breaking a word down into its constituent morphemes. To the best of the author’s knowledge, only one study has investigated this question in relation to a Bantu language (Setswana; Ciaccio, Kgolo & Clahsen 2020). The extent to which morphosyntactic processing differs across first- and second-language speakers of a language is also theoretically important and critically understudied in relation to Bantu languages (Spinner 2011).

Another set of applications arises in the domain of language development. Lexical databases of the type described in this paper have been developed specifically for use in psycholinguistic studies of children’s language processing and for evaluating literacy materials aimed at developing readers (e.g., Corral, Ferrero & Goikoetxea 2009; Masterson et al. 2010; Schroeder et al. 2015; Terzopoulos et al. 2017). For such applications, corpora are typically compiled using a large selection of materials created specifically for children in order to increase the likelihood of children having been exposed to the language it contains.

Table 3: Example database entries

	Word	Raw	Zipf	Neighbourhood	Example neighbours	N. freq.	N. freq
	length	freq.	freq.	(N) size		(mean)	(SD)
umntu	5	7,530	6.19	16	mntu, kumntu	212.37	569.68
amanzi	6	1,672	5.54	18	yamanzi, abanzi	70.05	88.16
ububele	7	46	3.98	11	ubuyele, ubukele	27.09	42.35
isakhiwo	8	135	4.44	7	izakhiwo, esakhiwo	52.86	89.6
ukulala	7	108	4.35	23	ukudlala, ukuhlala	131.35	268.59
ukubhala	8	508	5.02	18	ukubala, ukubhalwa	30.33	50.63
ukucinga	8	199	4.61	18	ukujinga, akucinga	8.15	7.21
ukuthengis	11	152	4.49	5	ukuthengiswa, ukumthengisa	36.2	50.77

a

kaninzi	7	48	3.99	9	baninzi, maninzi	111.78	173.47
phakathi	8	4,730	5.99	10	ephakathi, iphakathi	110.5	130.46

Limitations and suggestions for further work

There are several additional steps that can be taken to improve on and expand the database presented in this paper. For one, after compiling the corpus and before processing it to derive lexical statistics, it would be desirable to cross-reference the corpus word list with a word list from an official isiXhosa dictionary. This cross-referencing process would enable misspellings to be filtered out from the corpus, thus reducing the number of spurious neighbours identified, and also facilitate the removal of non-isiXhosa words. At the time of writing, no such dictionary word list could be obtained in a digital format, and so this step has not yet been taken.

Another notable consideration is that all of the above calculations were based on word forms rather than roots or lemmas. This means that instead of, for example, *-lala* being treated as a lemma that surfaces in *ukulala*, *uyalala*, *siyalala*, and so forth, each of these word forms is treated as an individual item. This can also lead to inflation of neighbourhood density (however, the words affected by this issue – most notably, verbs – will tend to have their neighbourhood density inflated to the same extent). It remains an empirical question whether it is properties of the word form or the lemma that are better predictors of, for example, word recognition latency in languages such as isiXhosa. In order to address this question, lemmas could be obtained from corpus data using a morphological analyzer (e.g., du Toit & Puttkammer 2021).

Lastly, the work presented here could also be expanded by deriving phonological statistics for isiXhosa, such as syllable number and phonological neighbourhood density. Such statistics can be obtained via LexiCAL if each word entry is paired with a phonetic transcription and would enable research on spoken word processing in the language.

Conclusion

Psycholinguistic techniques that can capture language processing as it unfolds in real time have yet to be leveraged to examine the processing of isiXhosa and related languages. This paper has discussed one kind of resource – a database of lexical statistics on the language, compiled based on a large-scale corpus – that is necessary to address this research gap and realize the considerable theoretical and practical benefits of doing so. Future collaboration between (computational) linguists and language specialists will allow for the refinement of this resource and the creation of others.

Acknowledgements

This work was made by possible by an NRF Thuthuka grant (grant no. 138180) awarded to R. Berghoff.

References

Alexeeva, S, Slioussar, N & Chernova, D 2018, ‘StimulStat: A lexical database for Russian’, *Behavior Research Methods*, vol. 50, no. 6, pp. 2305–15.

Andrews, S 1997, ‘The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts’, *Psychonomic Bulletin & Review*, vol. 4, no. 4, pp. 439–61.

Boudelaa, S & Marslen-Wilson, WD 2010, ‘Aralex: A lexical database for Modern Standard Arabic’, *Behavior Research Methods*, vol. 42, no. 2, pp. 481–7.

Boudelaa, S & Marslen-Wilson, WD 2011, ‘Productivity and priming: Morphemic decomposition in Arabic', *Language and Cognitive Processes*, vol. 26, 4-6, pp. 624–52.

Brysbaert, M, Mandera, P & Keuleers, E 2018, 'The word frequency effect in word processing: An updated review', *Current Directions in Psychological Science*, vol. 27, no. 1, pp. 45–50.

Bylund, E, Khafif, Z & Berghoff, R 2022, 'Linguistic and geographic diversity (or lack thereof) in research on second language acquisition and multilingualism', *Applied Linguistics*. Manuscript submitted for publication.

Chee, QW, Chow, KJ, Goh, WD, Yap, MJ & Miwa, K 2021, 'LexiCAL: A calculator for lexical variables', *PloS one*, vol. 16, no. 4, pp. e0250891.

Ciaccio, LA, Kgolo, N & Clahsen, H 2020, 'Morphological decomposition in Bantu: a masked priming study on Setswana prefixation', *Language, Cognition and Neuroscience*, vol. 35, no. 10, pp. 1257–71.

Colé, P, Pynte, J & Andriamamonjy, P 2003, 'Effect of grammatical gender on visual word recognition: Evidence from lexical decision and eye movement experiments', *Perception & Psychophysics*, vol. 65, no. 3, pp. 407–19.

Corral, S, Ferrero, M & Goikoetxea, E 2009, 'LEXIN: A lexical database from Spanish kindergarten and first-grade readers', *Behavior research methods*, vol. 41, no. 4, pp. 1009–17.

du Toit, Jakobus S. & Puttkammer, MJ 2021, 'Developing core technologies for resource-scarce Nguni languages', *Information*, vol. 12, no. 12, p. 520.

Duchon, A, Perea, M, Sebastián-Gallés, N, Martí, A & Carreiras, M 2013, 'EsPal: One-stop shopping for Spanish word properties', *Behavior Research Methods*, vol. 45, no. 4, pp. 1246–58.

Goldhahn, D, Eckart, T & Quasthoff, U 2012, 'Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages', *Proceedings of the 8th International Language Resources and Evaluation*.

Ktori, M, van Heuven, W & Pitchford, NJ 2008, 'GreekLex: A lexical database of Modern Greek', *Behavior Research Methods*, vol. 40, no. 3, pp. 773–83.

Marian, V, Bartolotti, J, Chabal, S, Shook, A & White, SA 2012, 'CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities', *PloS one*, vol. 7, no. 8, pp. e43230.

Masterson, J, Stuart, M, Dixon, M & Lovejoy, S 2010, 'Children's printed word database: Continuities and changes over time in children's early reading vocabulary', *British Journal of Psychology*, vol. 101, no. 2, pp. 221–42.

New, B, Ferrand, L, Pallier, C & Brysbaert, M 2006, 'Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project', *Psychonomic Bulletin & Review*, vol. 13, no. 1, pp. 45–52.

Norcliffe, E, Harris, AC & Jaeger, TF 2015, 'Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances', *Language, Cognition and Neuroscience*, vol. 30, no. 9, pp. 1009–32.

Perea, M, Urkia, M, Davis, CJ, Agirre, A, Laseka, E & Carreiras, M 2006, 'E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque)', *Behavior Research Methods*, vol. 38, no. 4, pp. 610–5.

Pretorius, E 2019, 'Getting it right from the start', in N Spaull & J Comings (eds), *Improving Early Literacy Outcomes*, BRILL, pp. 63–80.

R Core Team 2022, *R: A language and environment for statistical computing*, <https://www.R-project.org/>.

Schroeder, S, Würzner, K, Heister, J, Geyken, A & Kliegl, R 2015, 'childLex: a lexical database of German read by children', *Behavior Research Methods*, vol. 47, no. 4, pp. 1085–94.

Silge, J & Robinson, D 2016, 'tidytext: Text Mining and Analysis Using Tidy Data Principles in R', *JOSS*, vol. 1, no. 3, <http://dx.doi.org/10.21105/joss.00037>.

Soares, AP, Iriarte, Á, de Almeida, José João, Simões, A, Costa, A, Machado, J, França, P, Comesaña, M, Rauber, A, Rato, A & Perea, M 2018, 'Procura-PALavras (P-PAL): A Web-based interface for a new European Portuguese lexical database', *Behavior Research Methods*, vol. 50, no. 4, pp. 1461–81.

Spinner, P 2011, ‘Review article: Second language acquisition of Bantu languages: A (mostly) untapped research opportunity’, *Second Language Research*, vol. 27, no. 3, pp. 418–30.

Sun, CC, Hendrix, P, Ma, J & Baayen, RH 2018, ‘Chinese lexical database (CLD)’, *Behavior Research Methods*, vol. 50, no. 6, pp. 2606–29.

Terzopoulos, AR, Duncan, LG, Wilson, Mark A. J., Niolaki, GZ & Masterson, J 2017, ‘HelexKids: A word frequency database for Greek and Cypriot primary school children’, *Behavior Research Methods*, vol. 49, no. 1, pp. 83–96.

van de Velde, M, Bostoen, K, Nurse, D & Philippson, G (eds) 2019, *The Bantu Languages*, Routledge, New York.

van Heuven, W, Mandera, P, Keuleers, E & Brysbaert, M 2014, ‘Subtlex-UK: A new and improved word frequency database for British English’, *Quarterly Journal of Experimental Psychology*, vol. 67, no. 6, pp. 1176–90.

Wickham, H 2019, *stringr: Simple, Consistent Wrappers for Common String Operations*, <https://CRAN.R-project.org/package=stringr>.

Wickham, H 2021, *rvest: Easily Harvest (Scrape) Web Pages*, <https://CRAN.R-project.org/package=rvest>.