Multilingual Data from the Agricultural Domain: Presenting the NWU-Pula/Imvula Corpora

Tanja Gaustad and Cindy McKellar and Martin J. Puttkammer
Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
{tanja.gaustad|cindy.mckellar|martin.puttkammer}@nwu.ac.za

Abstract

This paper presents new multilingual corpora from the agricultural domain for seven South African Languages, namely Afrikaans, English, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, and Setswana, based on the Pula/Imvula magazine. After pre-processing, the data has been automatically sentencized, tokenized, lemmatized and annotated with part-of-speech information using the services available at https://v-ctx-lnx7.nwu.ac.za/. The final resources comprising between 774k and 1,38M tokens per language are included on the Corpus Cooperative at North-West University (COCO@NWU) corpus platform at https://coco.nwu.ac.za/ as searchable corpora. In addition, the data can be made available as text files for research purposes upon request. To highlight the value of this agricultural domain-specific data collection in relation to more general data, we also include some corpus-based statistics and comparisons with previous research.

1 Introduction

As has already been pointed out by Hanks (2000), language usage is dependent on domain. For instance, lexicography (Drouin, 2004), language learning (Barrière, 2009) and machine translation (van Noord et al., 2022) benefit greatly from domain-specific corpora. Domain-specific datasets are collections of data customized to a particular field or application. These datasets focus on distinct types of content, information, formats, or use cases relevant to a domain, such as healthcare, finance, or—in our case—agriculture. For example, a financial dataset could contain transaction records flagged for fraud, or a question-answer dataset could contain data tailored to medical records and linked diagnoses. Unlike general-purpose datasets, domain-specific datasets address niche problems: Their value lies in capturing real-world patterns which are pertinent to a field and which will not be found in generic data, enabling models to perform tasks particular to that field with higher accuracy.

Currently, in the age of Large Language Models (LLMs), there is a reliance on large amounts of general heterogeneous data implicitly assuming that different domains will be covered due to sheer volume. For languages with less data, however, coverage of different domains will generally be (very) low. As most of the South African languages are considered under-resourced, domain-specific corpora can hopefully add value to develop machine learning and AI solutions, e.g. for fine-tuning LLMs, in areas of importance for everyday life.

One such area is agriculture: The livelihood of billions of people worldwide depends on agriculture and AI-enabled solutions show promise to contribute to solutions geared at reaching sustainable development goals (see https://sdgs.un.org/goals). However, often the necessary training or information is not available in the native language of the farmers and therefore does not reach the work force. With the multilingual corpora presented in this paper, we hope to contribute a new valuable resource for the development of tools aimed at agriculture.

The rest of the paper is structured as follows: After a description of some background on domain-specific research for low-resource settings in section 2, we discuss the source data used (section 3) as well as the pre-processing and automatic annotations applied (section 4). We then describe where the corpora can be accessed (section 5), followed by some corpus-based statistics and lexicon comparisons to give the reader a better idea of the value of this agricultural domain-specific data in relation to more general data (section 6). We end with conclusions (section 7) and limitations.

2 Background

One of the aims of UNESCO's initiative "Language Technologies for All" (LT4All) is to advance language technologies in order to (help) preserve linguistic heritage and promote multilingualism. An indispensable building block to achieving this aim are language corpora as they are essential for language research, language learning, and the development of language technologies.

Echoing some of the aims of LT4All, the recent white paper by Pava et al. (2025) focuses on the challenges and strategies for developing LLMs for low-resource languages. These languages face limitations due to data scarcity as well as data that is not representative of the socio-cultural context. One of the recommendations of the paper is to invest in research that increases the availability and quality of low-resource language data. We believe this includes domain-specific data.

Lately, investigations into using domain-specific data have become popular. For instance, Edwards et al. (2020) show that using unlabeled domain-specific data in supervised text classification delivers more robust results than models trained on more but only general data, such as BERT (Devlin et al., 2019), even when BERT is pre-trained on domain-relevant data.

Experiments conducted by Singh et al. (2022) also indicate that models trained with domain-specific implicit reasonings significantly outperform domain-general models in both automatic and human evaluations.

Tang and Yang (2024) investigate if the development of domain-specific embedding models is necessary and even useful, specifically for the financial domain. They come to the conclusion that general-purpose models struggle to capture domain-specific linguistic and semantic patterns whereas using domain-specific data delivers better performance.

The main advantages of domain-specific AI based on domain-specific data is that it is more relevant to the task, more reliable, and more easily scalable. The major downside is the limited adaptability of a domain-specific system.

3 Source Data and Languages Included

The data included in the corpora presented here issue from Pula/Imvula, a South African magazine focusing on the developing farmer and published by Grain SA.[1] The main aim of the magazine is to support developing farmers in becoming sustainable commercial farmers. The magazine is distributed on a monthly basis and currently only available in English. However, until September 2024 it was published in five languages (English, isiXhosa, isiZulu, Sesotho, and Setswana), and previously also included Afrikaans and Sesotho sa Leboa (discontinued due to lack of funding). The Centre for Text Technology (CTexT) has a long-standing working relationship with the editors of the magazine and has been receiving and archiving the original MS Word documents since 2007.

For the seven corpora, we have combined files from editions acquired between 2007 and 2024 for English, isiXhosa, isiZulu, Sesotho, and Setswana, and between 2007 and 2019 for Afrikaans and Sesotho sa Leboa (see Table 1).

4 Data Pre-Processing and Automatic Annotations

To ready the data orginally received from the editors for deployment as corpora, pre-processing steps as well as automatic sentencization, tokenization, lemmatization and annotation with part-of-speech (POS) tags were required. These steps will now be described in more detail.

4.1 Pre-Processing

In a first step, all data was extracted from the original MS Word documents to UTF-8 encoded text files. To ensure the accuracy and quality of our corpora, all text was then run through a language identifier (Hocking, 2014; Puttkammer et al., 2018) to ensure the correct language is contained in the respective corpora since South African documents sometimes contain mixed languages. Data was also checked for encoding problems caused by the incorrect use of diacritics as this is often a problem found in South African texts. Lastly, some English headers with instructions to the editors and translators of the magazine were stripped out.

Each file in the corpora contains metadata detailing the source of the data, the language it is in as well as the publication date. See Table 1 for the total amount of tokens contained in the final data for each language as well as Section 6 for more detailed information on the data.

[1] https://www.grainsa.co.za/farmer-development

Language	Years	# Tokens	# Types	TTR
Afrikaans	2007–2019	798,067	33,047	0.041
English	2007–2024	1,148,871	31,266	0.027
isiXhosa	2007–2024	825,364	113,613	0.138
isiZulu	2007–2024	774,124	120,335	0.155
Sesotho	2007–2024	1,376,801	30,672	0.022
Sesotho sa Leboa	2007–2019	1,014,905	28,357	0.028
Setswana	2007–2024	1,335,512	38,746	0.029

Table 1: Overview of final data included in the NWU-Pula/Imvula Corpora, including number of tokens, number of types and type-token ratio (TTR).

4.2 Automatic Annotations

All the non-English data has been automatically sentencized, tokenized, lemmatized and annotated with part-of-speech (POS) information using the CTexT NCHLT Web Services available at https://v-ctx-lnx7.nwu.ac.za/.[2] An in-depth explanation of the principal technologies used as well as evaluation results for isiXhosa and isiZulu can be found in (du Toit and Puttkammer, 2021).

The English data has been processed using the Stanza pipeline (Qi et al., 2020), also resulting in sentencized, tokenized, lemmatized and POS-tagged text.

5 Availability of the Corpora

The Corpus Cooperative at North-West University (COCO@NWU) is an initiative of the North-West University's (NWU) Faculty of Humanities. The broad aim of the initiative is to advance corpus-based research, in particular in the digital humanities. More specifically, it also aims to make corpora developed at the NWU available to all researchers and students at the NWU, their collaborators, or researchers outside the NWU.

The corpora available on the COCO@NWU platform have mostly been developed in collaboration with various corpus suppliers, such as publishing houses, news websites, (literary) blog sites, libraries, etc. Therefore, these corpora may be used for academic research purposes only.

In addition to the Afrikaans corpora already available on the platform, the multilingual Pula/Imvula data has been added as a searchable corpus for each language at https://coco.nwu.ac.za/ where researchers can carry out (corpus) linguistic and statistical analyses of the data. The platform includes the possibility for simple word-based searches as well as more advanced queries using information on words, POS, lemmas, etc. and allows for exports of the query results. Figure 1 contains a screenshot as an example of how the platform can be used for queries. The corpora can also be made available as UTF-8 encoded text files for research purposes upon request.

[2]The underlying python packages are available at https://pypi.org/project/ctextcore/.

6 Corpus-based and Lexicon Statistics

Table 1 contains an overview of relevant statistics for the presented corpora detailing the years of data included, the number of tokens, number of types and the associated type-token ratio for each language. As expected, the type-token ration for isiXhosa and isiZulu, two Nguni languages, is higher on account of their agglutinative nature and conjunctive writing style. The three Sotho languages, Sesotho, Sesotho sa Leboa and Setswana, all have lower type-token ratios as a result of being written disjunctively. The higher value for Afrikaans compared to English is due to its abundant use of compounding.

In 2002, Prinsloo and de Schryver (2002) presented an instrument to measure the degree of conjunctivism/disjunctivism of the South African languages. Building on this work, they later proposed multidimensional lexicographic rulers for all South African languages, which were "prediction instruments aimed at assisting the South African lexicographers with the compilation of their national dictionaries" (Prinsloo and de Schryver, 2005). These rulers drew on electronic corpora available at the time as well as on dictionary data, and their goal is to help predict what the distribution of dictionary entries per letter should be for a given language. Looking at dictionary entries for English, for example, it is immediately obvious that there are far less entries starting with X, Y an Z than there are entries starting with S. The distribution is, of course,

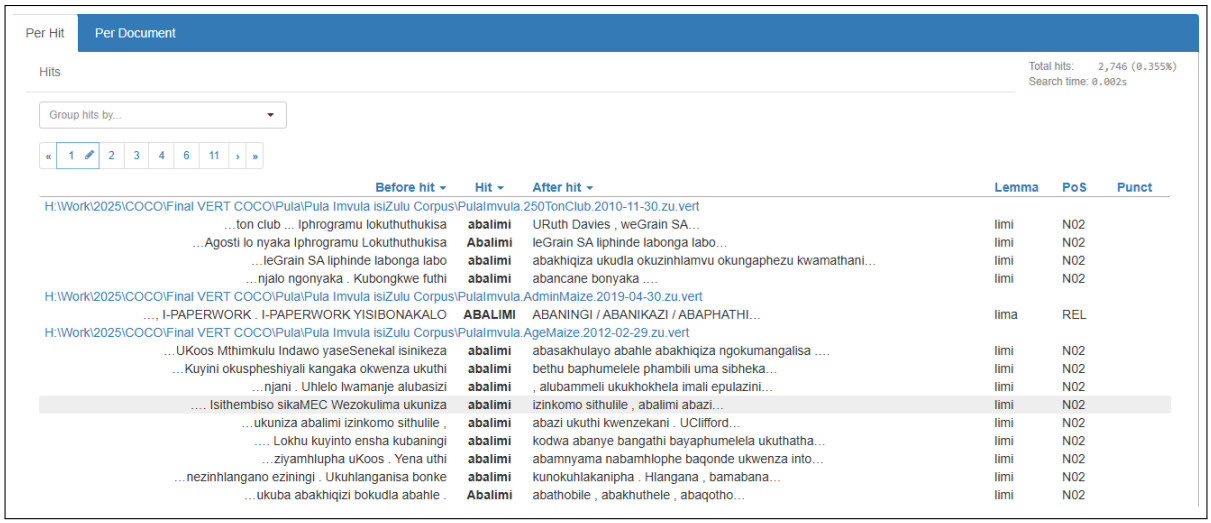

Figure 1: Screenshot of the COCO@NWU platform with a search result in the isiZulu Pula/Imvula corpus.

different for different languages. The authors were able to establish that so-called 'stretches', i.e. sections containing all lemmas starting with the letter A, then B, etc. could be modelled from corpora available for the language of interest.

To give the reader a better of idea of the valuable content of this agricultural domain-specific data in relation to more general data (albeit from 2005 and before), we compare the Pula/Imvula corpora to these rulers. This will help to identify any deviation from the expected distribution, possibly indicating new terms, loan words or spelling variations not present in the more general corpora used by (Prinsloo and de Schryver, 2005).

Figures 2 and 3 show the distribution of Afrikaans and English stretches compared to (Prinsloo and de Schryver, 2005). For Afrikaans, we see an increase of 7.55% in 'D' as well as an increase of 3.43% in 'I', both explained by the presence of agriculture-related terms such as *dier* ('animal'), *droog* ('dry'), *droogte* ('drought'), *insek* ('insect'), *inset* ('contribution') in the Pula/Imvula data, while e.g. 'S' is 5.31% lower compared to the data used by (Prinsloo and de Schryver, 2005). For English, only 'T' shows a large difference with an increase of 10.5%. Looking at tokens starting with 'T', we find several agricultural terms like *ton*, *tractor* or *tillage* that explain the increase.

For Figure 4, we calculated and plotted the difference between the ruler values in (Prinsloo and de Schryver, 2005) and in the lemmatised Pula/Imvula corpora for Sesotho (ST), Sesotho sa Leboa (NSO) and Setswana (TN). It is interesting to note that the variations are very similar for all three languages, with the exception of 'G' where the difference is only present for Sesotho sa Leboa and Setswana (8.88% and 10.29% respectively) and on 'H' where the difference is only present for Sesotho (10.44%).[3] The higher occurrence of 'G' in the Sesotho sa Leboa and Setswana data can be attributed to the higher than expected use of particles, for example *go* ('to/it') that is also the most frequent word in the Setswana corpus with 95,725 occurrences, *ga* ('I/it/he/she') and *gore* ('so that'). The Sesotho data shows similar observations with *ho* ('to/it') as the most frequent token occurring 71,868 times, followed by *ha* ('I/it/he/she') and *hore* ('so that'). Echoing the findings for 'G' and 'H', the increase for 'T' is mainly due to the high frequency of *tse/tše* ('this') and *tla* ('shall/will'). The higher use of particles can possibly be attributed to the instructional nature of the corpus.

The marked differences at 'K', 'L' and 'Y' can be explained by the use of specialised vocabulary such as *khemikale* ('chemical'), *kgwedi* ('month'), *kgwebo* ('business'), *laola* ('manage'), *laodi* (lemmatised form of *molaodi/balaodi* 'manager(s)') or agriculture-specific loanwords such as *yunite* ('unit') or *yield* ('yield'). The noticeable drop in lemmas starting with 'M' is most likely due to the deletion of class prefixes *mo-*, *ma-*, and *me-* (used in 5 different noun classes) during lemmatisation.

When examining the graphs for the isiXhosa (XH) and isiZulu (ZU) Pula/Imvula data in Figure 5, the differences compared to the rulers are quite similar for the two languages. The biggest

[3]This is due to a sound shift between 'g' and 'h' for these languages.

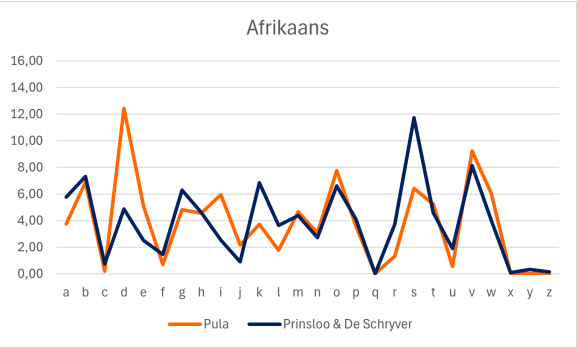

Figure 2: Graph showing the distribution of stretches for Afrikaans in Pula/Imvula compared to rulers in (Prinsloo and de Schryver, 2005).

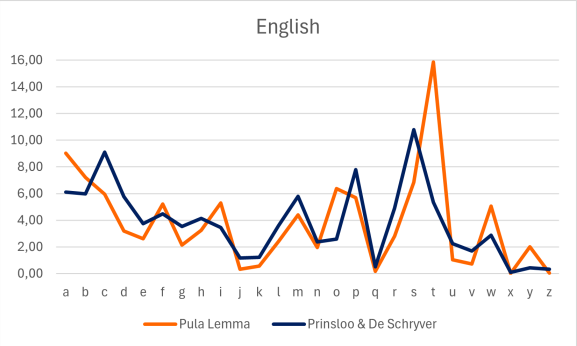

Figure 3: Graph showing the distribution of stretches for English in Pula/Imvula compared to rulers in (Prinsloo and de Schryver, 2005).

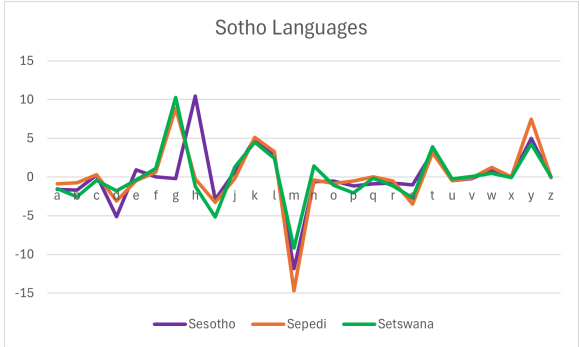

Figure 4: Graph showing the (calculated) differences in distribution for Sesotho, Sesotho sa Leboa and Setswana in the Pula/Imvula corpora.

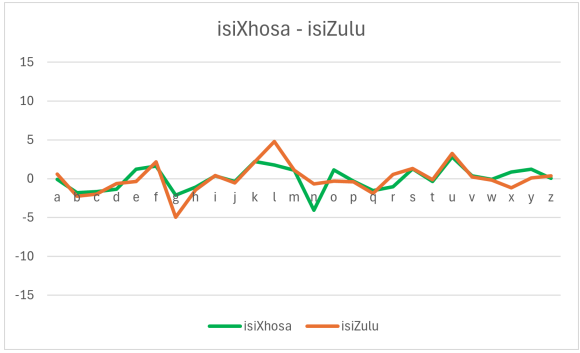

Figure 5: Graph showing the (calculated) differences in distribution for isiXhosa and isiZulu in the Pula/Imvula corpora.

dissimilarities are on 'F', 'K', 'L' and 'U'. Looking at examples from these four letters, we find that the large deviance is due to two reasons. On the one hand, words related to farming, learning or trade partners are prevalent, e.g. *fama* ('farm'), *fundo/a* ('study/learn'), *khemikali* ('chemical'), *khiqiza/o* ('produce/product'), *kg* ('kilogram'), *USA*, or *Ukraine*. On the other hand, like for the Sotho languages discussed above, the data contains a lot of particles used in instructions: *lokho/u* ('these'), *le/o* ('this'), *ukuthi* ('that'), or *ukuze* ('so that'). These particles together with lemmatised forms of 'farmer' and 'crop(s)' (*limi* and *limo*) also explain the bigger difference for isiZulu for 'L'.

In addition to the lexicographic rulers, we use AntConc (Anthony, 2025) to identify possible domain-specific terms from the Pula/Imvula corpora based on keyness calculated using log likelihood (Pojanapunya and Todd, 2018). The top 10 terms for each language are presented in Table 2.

As would be expected, the terms (based on full word forms, not on lemmas) are mostly of agricultural origin as can be seen from their translations. Namely the words *farmer* and *farmers* appear in the top ten selection for every language (Afrikaans: *boer/boere*; isiXhosa: *umlimi/abalimi*; isiZulu: *abalimi*; Sesotho/Sesotho sa Leboa: *molemi/balemi*; Setswana: *balemirui*). There are also many occurrences of other domain-specific terms across languages in just these top ten examples, such as *soil* (Afrikaans: *grond*; Sesotho: *mobu*; Sesotho sa Leboa: *mmu*) or *maize* (Afrikaans: *mielies*; isiZulu: *ummbila*; Sesotho: *poone*; Sesotho sa Leboa: *lehea*).

This further shows that the corpora are indeed a source that can be used to identify new terms specific to the agricultural domain and could be a valuable source for further linguistic research. It may also be possible, with a little help from a native speaker, to identify the terms along with their translations in other languages thereby creating translation lists of these domain-specific terms.

AF	Keyness	EN	Keyness	NSO	Keyness	ST	Keyness
graan (grain)	4046.24	crop	15112.39	balemi (farmers)	5416.89	poone (maize)	3850.75
boer (farmer)	3737.71	maize	14137.04	dibjalo (crops)	5410.29	mobu (soil)	3409.38
boere (farmer)	3566.30	grain	13348.70	ye (this one)	4434.32	dijothollo (cereal)	3258.05
grond (soil)	3238.18	soil	13049.94	lehea (maize)	4261.97	tlhahiso (production)	2585.10
mielies (maize)	3052.16	your	13047.17	tše (these)	3018.34	temo (farming)	2537.92
plant (plant)	2874.41	you	12350.11	puno (crop)	2891.40	balemi (farmers)	2514.88
sa (South Africa)	2361.17	farmers	11995.19	molemi (farmer)	2852.89	grain (grain)	2479.66
ha (hectare)	2290.34	farmer	9740.30	mme (madam)	2774.93	jala (sow)	2415.53
baie (much/very)	2077.75	production	8232.20	grain (grain)	2633.38	peo (seed)	2322.45
oes (harvest)	1970.48	can	8072.30	mmu (soil)	2283.31	molemi (farmer)	2217.17

TN	Keyness	XH	Keyness	ZU	Keyness
bokana (amount)	4539.75	abalimi (farmers)	3584.61	abalimi (farmers)	3825.05
go (to)	3949.11	sa (South Africa)	2298.85	sa (South Africa)	2718.37
balemirui (farmers)	3941.43	lokulima (of farming)	2150.72	isilimo (crop)	2333.58
jwala (sow)	3605.10	izityalo (plants)	1772.43	izilimo (crops)	2224.48
bolemirui (farming)	3445.27	ngehektare (hectare)	1750.17	ummbila (maize)	1997.58
tlhotlhwa (cost)	3216.86	zesoya (soya)	1667.59	isivuno (harvest)	1996.37
kumo (produce)	3190.06	isityalo (plant)	1547.00	ha (hectare)	1864.51
dijwalwa (seeds)	3129.57	ukuze (so that)	1542.46	ukhula (weed)	1725.21
kgono (skill)	3011.36	ngokunjalo (accordingly)	1532.64	isoya (soya)	1656.42
tlaa (shall/will)	2860.73	umlimi (farmer)	1435.55	kakhulu (very/mostly)	1486.99

Table 2: Top 10 words per language based on keyness for the NWU-Pula/Imvula Corpora (AF=Afrikaans, EN=English, NSO=Sesotho sa Leboa, ST= Sesotho, TN=Setswana, XH=isiXhosa, ZU=isiZulu). The included translations have been added for understanding and are not meant to be exhaustive.

7 Conclusion

This paper presented a collection of seven multilingual corpora in the agricultural domain containing data for Afrikaans, English, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, and Setswana. Next to a description of the data source, the processing applied and the availability of the final corpora, we included corpus-based statistics, i.e. type and token counts and type-token ratios. In addition, we provided the top 10 domain-specific terms in Pula/Imvula based on keyness as well as a comparison of lexical distributions between the Pula/Imvula data and more generic data compiled by (Prinsloo and de Schryver, 2005). These statistics aim to illustrate the added value of these corpora specific to the domain of agriculture.

Hopefully, these new resources are a useful addition to the already available corpora for under-resourced South African languages. Being accessible to researchers via a searchable online corpus platform will no doubt also increase the use of the presented corpora in the digital humanities and (corpus) linguistics community.

Limitations

As we have described in this article, the NWU-Pula/Imvula corpora are not comprehensive data sets on all aspects of agriculture, but contain explanatory articles geared towards the improvement of farming and agriculture in South Africa. The texts concentrate on crops cultivated in South Africa (mainly maize, and soy), planting/sowing, pest control, financial management, etc. and the subjects covered depend on the original content of the Pula/Imvula publications.

Furthermore, we cannot account for biases in the data arising from the magazine's editorial focus, the (possible) use of proprietory glossaries, spelling preferences or the allowance of including loan words.

Acknowledgments

We acknowledge the generous financial support of the North-West University's (NWU) Faculty of Humanities for subsidising the Corpus Cooperative at NWU (COCO@NWU). Through its aims to advance corpus-based research in the digital humanities, it directly supported and benefitted this research. Nonetheless, all searches, calculations, and interpretations on the data contained on the COCO@NWU platform are the researcher's only, and cannot be attributed to the NWU.

References

Laurence Anthony. 2025. Antconc (version 4.3.1) [computer software]. Tokyo, Japan: Waseda University.

Caroline Barrière. 2009. Finding domain specific collocations and concordances on the web. In *Proceedings*

of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning, pages 1–8, Borovets, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Drouin. 2004. Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jakobus S. du Toit and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce Nguni languages. *Information*, 12(520):1–12.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Hanks. 2000. Contributions of lexicography and corpus linguistics to a theory of language performance. In *Proceedings of the 9th EURALEX International Congress*, pages 3–13, Stuttgart, Germany. Institut für Maschinelle Sprachverarbeitung.

Justin Hocking. 2014. Language identification for South African languages. In *Proceedings of the Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. PRASA.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. Technical report, Stanford University.

Punjaporn Pojanapunya and Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.

Danie J. Prinsloo and Gilles-Maurice de Schryver. 2002. Towards an 11x11 array for the degree of conjunctivism / disjunctivism of the South African languages. *Nordic Journal of African Studies*, 11(2):249–265.

Danie J. Prinsloo and Gilles-Maurice de Schryver. 2005. Managing eleven parallel corpora and the extraction of data in all official South African languages. In Cobus Snyman Walter Daelemans, Theo du Plessis and Lut Teck, editors, *Multilingualism and Electronic Language Management*, pages 100–122. Van Schaik, Pretoria.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. NLP web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49, Melbourne, Australia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. 2022. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4674–4683, Marseille, France. European Language Resources Association (ELRA).

Yixuan Tang and Yi Yang. 2024. Do we need domain-specific embedding models? An empirical investigation. arXiv:2409.18511v3.

Rik van Noord, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere, and Antonio Toral. 2022. Building domain-specific corpora from the web: the case of European digital service infrastructures. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 23–32, Marseille, France. European Language Resources Association (ELRA).