Stop words in Khoekhoe

Menno van Zaanen
South African Centre for Digital Language Resources
North-West University
Potchefstroom, South Africa
menno.vanzaanen@nwu.ac.za

Alena Witzlack-Makarevich
University of Bremen
Bremen, Germany
witzlack@uni-bremen.de

Abstract

Stop word lists are useful resources that allow for the filtering of words in texts that typically do not carry (much) content. Filtering stop words can improve the efficiency and accuracy of data processing. Stop words are typically short and occur very frequently in texts. Stop word lists are language dependent and many low-resource languages currently do not have (accurate) stop word lists. In this article, we look at how we can create, based on word frequency, a stop word list for Khoekhoe, which is a low-resource language spoken in Southern Africa. Given that stop words do not carry much content, they can be expected to occur consistently across different texts. We compare lists of most frequent words between texts in different genres and which words feature in these lists consistently. We look at the overlap of frequent words in English texts and compare these to a known English stop word list as well, and compare the results with the overlap of frequent words in Khoekhoe texts. The results show that there is a high overlap between genres for English, but the overlap between the Khoekhoe genres is lower. This may be due to a different typological profile of Khoekhoe. This means that creating a stop word list for Khoekhoe is more complicated and most likely requires other techniques to produce a useful stop word list.

1 Introduction

The term *low-resource language* refers to languages that have limited or no digital resources. Low-resource languages present distinct challenges in the area of natural language processing due to both quality and quantity of both annotated and unannotated datasets and other linguistic resources needed for the development of NLP applications (see Pakray et al. 2025; Pava et al. 2025 for recent overviews). Even though some of these languages are often of limited or no commercial interest, this does not mean that they are less important or even that have fewer speakers speak them than high-resource languages (e.g., Persian with over 90 million speakers according to Eberhard et al. 2025 is among the low-resource languages). On the other hand, developing practical NLP applications for low-resource languages is central for preserving linguistic diversity, fostering inclusion within the digital world, and providing a robust foundation for expanding our understanding human linguistic capacity and linguistic diversity (Pakray et al., 2025).

Many of the languages spoken in Africa, in particular Southern Africa, are considered low-resource. One of these languages is Khoekhoe, which also lacks many resources. Khoekhoe does not have a strong presence online (e.g., on websites or social media), and (partially due to the lack of available computational linguistic training material) there is no support in many tools, such as in word processors, although an online crowd-sourced dictionary is currently being compiled[1] and a Universal Dependencies Treebank of 27,000 tokens has been recently released (Tulchynska et al., 2025b,a).

One of the basic NLP resources for a language is a list of stop words. Such a list provides the words that in general do not carry (much) meaning (see e.g., Luhn 1957; Lo et al. 2005; Manning et al. 2008). We discuss specifics of stop words in detail in Section 2.1. To tackle the scarcity of digital resources for Khoekhoe, in this article, we investigate the compilation of a Khoekhoe stop word list. To identify Khoekhoe stop words, we rely on the idea that stop words occur frequently in texts (Fox, 1989). If, however, we use texts on a particular topic, those texts will have a different distribution of word frequencies compared to texts on another topic (see e.g., Kilgarriff 1997). To identify stop words, we can compare the frequency lists of texts

[1] See https://lexikhoe.com/.

in different genres. Words that occur frequently even if the topics of the texts are different are likely to be stop words.

To test the idea of consistent high word frequencies of a set of words between texts of different genres, we investigate the consistency of the high frequency words in English texts across different genres first. We can compare these words against a known stop word list for English. Next, we apply the same approach to Khoekhoe texts in different genres.

This article is structured as follows. First we provide some background on stop words as well as the Khoekhoe language in Section 2. Next (Section 3) we present the methodology used, which includes a brief description of the corpora used and the practical approach taken. After applying the approach to the texts in the corpora, we present the results in Section 4 and discuss them in Section 5. Finally, in Section 6 we conclude.

2 Background

2.1 Stop words

Stop word lists are lists of words that do not carry (much) meaning and thus have low discrimination power, for instance, when indexing documents or processing queries (see e.g., Luhn 1957; Lo et al. 2005; Manning et al. 2008) and consequently may be of little help with information retrieval. These words typically occur very frequently (Fox, 1989) in all kinds of texts and are also typically very short, such as *the*, *of*, *and*, and *to*.

Stop word lists have traditionally been developed in the context of the field of information retrieval (van Rijsbergen, 1979; Baeza-Yates and Ribeiro-Neto, 1999). To search for documents in a large document collection based on a search query, one relies on indices that efficiently identify documents containing query terms. Including words in the index that do not describe content do not help with identifying relevant documents simply takes up memory and slows down the search. In particular, this holds for frequently occurring words that do not contribute to the identification of relevant documents. As such, removing words that do not typically describe content was expected to benefit the entire process (Huston and Croft, 2010).

Inspired by Luhn's (1957) work, several precompiled stop word lists have been generated. The earliest is the Van Rijsbergen stop word list (or "Van list", (van Rijsbergen, 1979)), which consists of 250 stop words based on word frequency. The Brown stop word list (Fox, 1989), which we use in this study, consists of 421 stop words and is also partially based on word frequency. The most frequent words include *the*, *and*, *a*, *that*, and *was*. A few of the frequent words were removed from the list because they were considered too important as potential index terms (hence, too useful for searching). They include *business*, *family*, and *house*. Furthermore, 26 words were manually added to the list as they were expected to be frequent in some types of genres. They include *sure*, *behind*, and *whether*. These lists are known as the *classic* or the *standard* stop word lists. These classical stop word lists have been criticized for being too generic and outdated, and various alternatives to the compilation and curation of stop word lists have been proposed including e.g., the ones based on TF-IDF (Term Frequency-Inverse Document Frequency) technique, see e.g., (Saif et al., 2014) for an overview and Chavan et al. (2024) for an example of its application in the low-resource language Marathi.

Several stop word lists exist for English. These may be relatively generic lists that focus on words that are not helpful in general for identifying relevant documents, but, additionally, depending on the task, some lists may also contain content carrying words are not useful in the identification of relevant documents in a particular subject area (Hvitfeldt and Silge, 2021).

A stop word list is an important digital resource which has an enabling function and as such is essential for many other digital resources. For example, stop words are useful in practical situations, such as information retrieval (van Rijsbergen, 1979; Baeza-Yates and Ribeiro-Neto, 1999), text classification (Uysal and Gunal, 2014), sentiment analysis (Saif et al., 2014), and even authorship attribution (Zhao and Zobel, 2005). Though in modern information retrieval systems, the use of stop lists is less common (Jurafsky and Martin, 2025), stop word removal remains useful in various NLP tasks.

2.2 The Khoekhoe language

Khoekhoe, also known as Nama-Damara (ISO 639-3: naq; Glottocode: nama1264), is a Khoe language of the Khoe-Kwadi family (Güldemann, 2014). Khoekhoe represents a dialect continuum (Haacke, 2018). It includes two northern varieties (Haiǁom and ǂĀkhoe) that differ considerably from the central and southern varieties, which are closer to what is considered as standardized Khoekhoe. This article is based on a corpus which contains texts written in the standardize Khoekhoe and includes spoken data only from the latter varieties.

Khoekhoe is spoken mainly in central and southern Namibia. With approximately 245,000 speakers (11.8% of the total population of Namibia), Khoekhoe is the second most spoken language in the country after Oshiwambo (Namibia Statistics Agency, 2011). Although Khoekhoe is well-documented and receives official recognition as a language of instruction in Namibia's educational system (Brenzinger, 2013), and despite being the largest non-Bantu click language in Africa, Haacke and Eiseb (2002) consider it to be an "endangered language". Khoekhoe has a limited literary tradition, historically confined to school readers and religious texts compiled by missionaries.

Other Khoisan languages, including other thirteen Khoe-Kwadi languages have even few digital resources than Khoekhoe. To our knowledge no accessible digital corpus of substantial size exists for these languages, not even a Bible translation, one of the most widely translated works for low-resource languages (Pava et al., 2025).

The phonemic inventory of Khoekhoe includes 20 click phonemes, featuring four primary click articulations (dental, alveolar, palatal, and lateral) and five secondary articulations (Haacke, 2013, pp. 54–55). As in other Khoisan languages (cf. Nakagawa et al. (2023)), Khoekhoe roots underlie the following phonotactic restrictions: The canonical root shape is $C_1V_1C_2V_2$ (e.g., *ǂkhoro* 'bottle'), where only four consonants can occur as $C2$: *w*, *r*, *m* and *n*. The other two possible root shapes are $C_1V_1V_2$ and C_1V_1N (e.g., *xai* 'kudu' and *xam* 'lion' respectively). These structures resulted historically from the elision of the intervocalic consonant C_2. N in the latter structure represents the nasal consonants *m* or *n*. When $V1$ and $V2$ are identical, they result in a long oral or nasal vowel. Long oral vowels are represented by the vowel character with a macron symbol, e.g., <*ā*>. Long nasal vowel are represented by the vowel character with a circumflex symbol, e.g., <*â*>. The corpus used in this study follows the standard orthography of Khoekhoe as outlined in Curriculum Committee for Khoekhoegowab (2003). Though Khoekhoe is a tone language, exhibiting four distinctive register tones (Haacke, 1999), tones are not marked in the standard orthography and are also absent in the Khoekhoe corpus presented here.

3 Methodology

In general, as mentioned in Section 2, stop words are highly frequent and are not content words. This gives us a means to identify them automatically. First, just like Fox (1989), we can take the most frequent words of the language (based on a corpus) as our stop word list. Second, based on the notion of context mentioned by Hvitfeldt and Silge (2021), we may improve on the frequency-only approach by taking word use in different text genres into account.

To identify potential stop words, we investigate which words are both highly frequent and, at the same time, genre-independent, by taking a data-driven approach. Starting from a corpus with texts in different genres, we can easily rank words based on their frequency per genre. Stop words are then those highly ranked that overlap between the different genres.

We apply this approach to a Khoekhoe corpus consisting of texts with corresponding genre information. We then compare the Khoekhoe results against those of an English corpus (with genre information). For English we can also compare the most frequently occurring genre-independent words against a known stop word list.

We first describe the corpora and the stop word list that we use (Section 3.1), followed by a description of the approach taken (Section 3.2).

3.1 Data collections

In the current study, we used two corpora. The first is a corpus of Khoekhoe texts, which is the main focus of the research. The second is the Brown corpus, which we use as an English reference corpus, which allows us to investigate whether the approach applied to Khoekhoe behaves in a similar way to that applied to English. Furthermore, we can compare the behavior of most frequent English words against a known stop word list. These resources are described here in more detail.

The Khoekhoe corpus used in this study was compiled from a range of sources. It includes texts of six genres as described in Table 3. The spoken genre (Conversation) includes primarily dyadic conversations between friends, relatives, and colleagues and was collected in Windhoek in 2023. The conversations were transcribed into Khoekhoe by students of the University of Namibia. The written genres include several books of fiction (Fiction), school books (Learned), newspaper articles of the

Khoekhoe edition of Namibian newspaper New Era (Press), a translation of the Bible into Khoekhoe (Religion), as well a few other text types collapsed to Misc (e.g., several procedural texts and translations of film subtitles). The corpus is processed in R and tokenized using `unnest_tokens()` from the `tidytext` package (Silge and Robinson, 2016).

The Brown corpus (Kučera and Francis, 1967) is a collection of text samples (500 samples of American English with each having just over 2000 tokens).[2] The text samples are divided into different genres: Belles-Lettres, Fiction (with subdivisions: general, mystery, science, adventure, romance), Humor, Learned, Miscellaneous: government & house organs (Misc), Popular lore, Press (with subdivisions: reportage, editorial, and reviews), Religion, and Skill and hobbies. An overview of the genres of the Brown corpus can be found in Table 4.

In the appendix (in Table 3 and 4) we provide an overview of the distributions of words for both corpora where we combine the subdivisions. For the analysis, however, in this article, we only focus on the five genres that can be found in both corpora. The distribution of words in these genres of both corpora can be found in Table 1. Note that both corpora are approximately of the same size and the sample used in the experiments is also of similar size, although the distribution of tokens is a bit different.

Table 1: Distribution of words per genre used in the approach from the Khoekhoe and Brown corpora.

Genre	Khoekhoe # tokens	Brown # tokens
Fiction	37,513	235,489
Learned	18,083	159,936
Misc	75,153	61,143
Press	76,328	177,177
Religion	656,808	34,308
Total	863,885	668,053

To be able to compare the most frequent words in the different English genres, we take the stop word list for English from the NLTK Stopwords Corpus (Bird et al., 2009).[3] This stop word list consists of 198 English words.

3.2 Method

To investigate whether we can use the top N words in Khoekhoe to identify stop words, we take the corpus as described in Section 3.1. We subdivide the corpus in the different genres and use the genres as described in Table 1. For each genre, we then rank the words based on frequency and take the top N words. Then we measure the overlap of the top N words pair-wise for all genres in the corpus. We also apply the same approach to the Brown corpus. This allows us to compare the behavior of the most frequent words across genres between Khoekhoe and English.

We expect stop words to occur in the top N words, but the ranking of these words (within the top N) is irrelevant. We are simply interested whether the same words occur at the top of the frequency lists. A metric that measures the overlap of elements between two sets is the Jaccard coefficient. Applying this metric to our frequency lists, computes the proportion of the number of words of the intersection of the two sets over the number of words in the union of the two sets. Jaccard coefficients range between 1 (complete overlap) and 0 (no overlap at all). When A and B are the sets of the top N most frequent words in two genres, the Jaccard coefficient can be calculated as follows.

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B}$$

Although we do not know how many words should be contained in a stop word list, the English stop word list from the NLTK gives us a rough estimate. Although different stop word lists may have different lengths, given that the NLTK English stop word lists contains approximately 200 words, we investigate the Jaccard coefficients by varying N from 1 to 200.

Note that for us to be able to compare the English stop word list against the English genres, we need to make sure that we have a frequency ranking for the words in the NLTK stop word list as well. Unfortunately, the NLTK stop word list is alphabetically ordered. We compute the ranking of these stop words based on the frequency of occurrence in the complete Brown corpus.

Summarizing, for each of the genres, we first rank all words based on their frequency. Next, we select the top N words (where N ranges from 1 to 200) and compute the Jaccard coefficients by

[2] The Brown corpus can be downloaded from https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/brown.zip.

[3] See https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/stopwords.zip.

comparing genres pairwise. To get an overview of these coefficients, they can be plotted. Note that in addition to the pair-wise comparisons between the genres, for English we also compute the Jaccard coefficients between the genres and the NLTK stop word list (Stop words) and the entire corpus (Total).

4 Results

For each pair of genres, we compare the overlap of the top N words using the Jaccard coefficients. We can now investigate the statistical significance of the effect of the genre pair under consideration and the source (either Khoekhoe or Brown) on the Jaccard coefficients. We see that both variables have statistically significant impact on the Jaccard coefficients ($p < .0001$). The independent variable genre pair accounted for 15% of the variance in the overlap of the top N words ($\eta_p^2 = .15$), whereas the corpora variable accounts for 26% ($\eta_p^2 = .26$) of the variation. Performing a TukeyHSD analysis on the statistical model indicates that there are statistically significant differences between the two corpora ($p < .0001$) and that there are significant differences ($p < .05$) between most genre pairs. We provide an overview of the comparisons between genre pairs that are not statistically significant (with $\alpha = .05$ in Table 2). Where the comparisons between the pairs are insignificant, this corresponds to those plots in Figure 1 whose the shapes are the same.

A more in-depth exploration can be done by visualizing the Jaccard coefficients for both the corpora and genre pairs, as in Figure 1. Each individual plot represents the Jaccard coefficients (represented on the y-axis) for the top N words (represented on the x-axis). All plots contain information on the Khoekhoe (dashed line) and Brown (solid line) corpora. High Jaccard coefficients indicate a large overlap and low coefficients a small overlap. Additionally, the bottom row represents the Jaccard coefficients for the different genres of the Brown corpus and the ranked English stop word list from the NLTK. A similar comparison is not possible for Khoekhoe as we do not have a Khoekhoe stop word list available.

Figure 1 suggests a number of observations. First, we see that with increasing N (i.e., the top most frequent words used for the comparison), the Jaccard coefficients go down, but stabilize at around 100 words for English and at less than 50 words for Khoekhoe. This indicates that when N is very small, many of the most frequent words between the genres are the same. Increasing the number of words under consideration leads to larger differences between the genres.

Second, overall, the Khoekhoe Jaccard coefficients are typically lower (except for the comparison between the Fiction and Learned genres and with low N for Fiction and Religion). This indicates that most of the compared genres do not share many words in the top N most frequent words.

Third, the Khoekhoe Jaccard coefficients typically do not show a consistent peak with low N. This is not what we expected. It means that even if we consider only the really most frequent words (taking N very small), the overlap between the words in the different genres is very small. In other words, Khoekhoe does not seem to have consistent frequent words or stop words like English does.

Fourth, looking at the Jaccard coefficients of the English stop word list and the different genres, we see that the coefficients (for most genres) go down after approximately 75 words, but overall remain relatively high (between 0.5 and 0.25). This indicates that there are relatively many stop words in the most frequent English words in the different genres. The lower Jaccard coefficients with higher N indicate that at that point frequent words are found that are not contained in the stop word list.

Fifth, we see that when we compare the stop word list against the entire Brown corpus, up to about $N = 75$ the Jaccard coefficients are extremely high. After that, the Jaccard coefficients go down, which indicates that non-stop words are found in the most frequent words of the corpus.

5 Discussion

The idea behind the study described here is that stop words are frequently occurring words that occur (frequently) in all types of text. As such, a comparison between the most frequently occurring words in different genres should allow one to identify these stop words. To investigate this, we compared the most frequent words between pairs of genres using the Jaccard coefficient and we did this for both Khoekhoe and English corpora.

The results show that the approach yields different results for the Khoekhoe and English corpora. Not only are the Jaccard coefficients for Khoekhoe mostly lower than those for English, the (small) peaks that can be seen in the English results are not present in the Khoekhoe results.

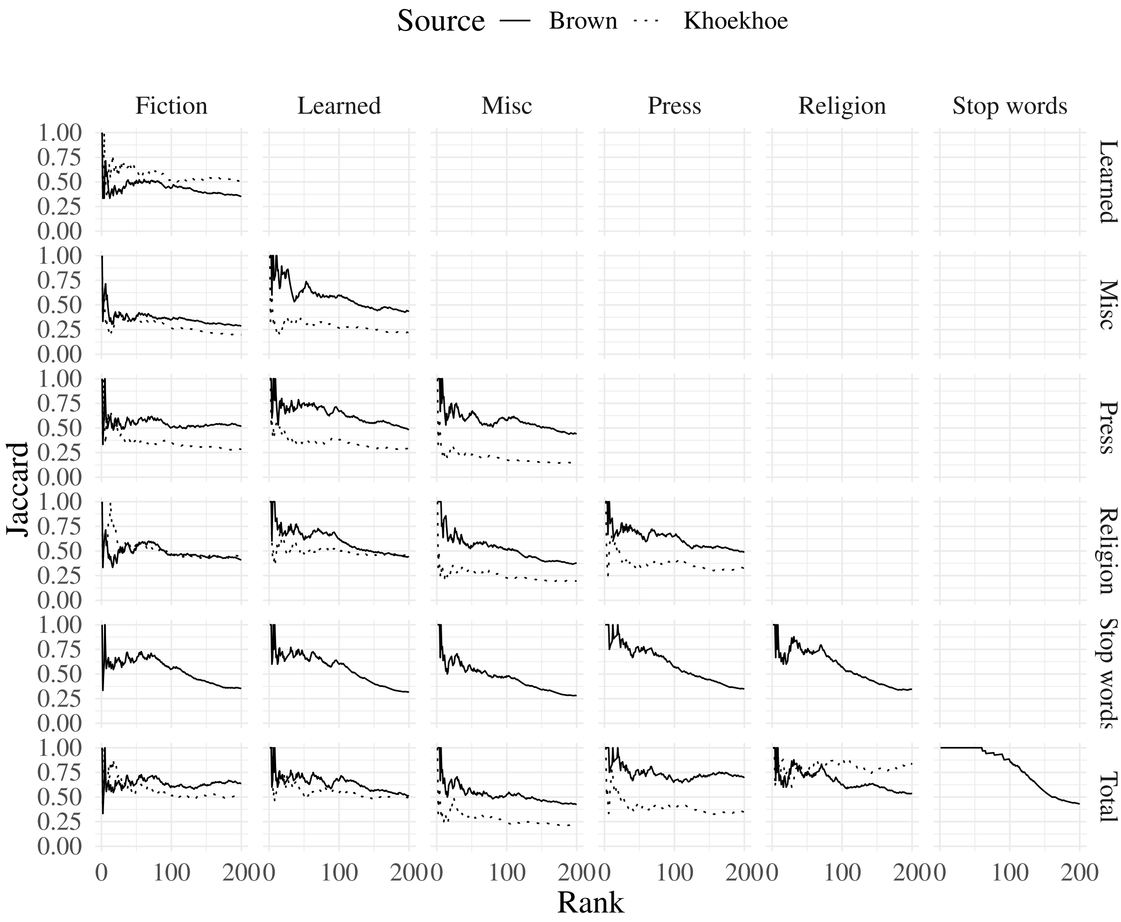

Figure 1: Overview of pairwise comparisons between genres in the Brown (solid line) and Khoekhoe (dotted line) corpora. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes). Note that the stop word data is only available for the Brown data.

Table 2: Overview of pair-wise comparisons that are not significantly different from other pair-wise comparisons.

Genre pair		Genre pair	p
Fiction and Religion	vs.	Fiction and Learned	1.000
Fiction and Stop words	vs.	Fiction and Learned	.145
Fiction and Stop words	vs.	Fiction and Religion	.054
Learned and Misc	vs.	Fiction and Press	.921
Learned and Press	vs.	Fiction and Learned	1.000
Learned and Press	vs.	Fiction and Religion	1.000
Learned and Press	vs.	Fiction and Stop words	.124
Learned and Religion	vs.	Fiction and Stop words	.998
Learned and Stop words	vs.	Fiction and Stop words	.999
Learned and Stop words	vs.	Learned and Religion	1.000
Learned and Total	vs.	Fiction and Total	1.000
Misc and Religion	vs.	Misc and Press	1.000
Misc and Stop words	vs.	Fiction and Learned	.166
Misc and Stop words	vs.	Fiction and Press	1.000
Misc and Stop words	vs.	Fiction and Religion	.355
Misc and Stop words	vs.	Learned and Misc	.315
Misc and Stop words	vs.	Learned and Press	.192
Misc and Total	vs.	Learned and Misc	.101
Misc and Total	vs.	Misc and Press	.979
Misc and Total	vs.	Misc and Religion	.986
Press and Religion	vs.	Fiction and Learned	1.000
Press and Religion	vs.	Fiction and Religion	1.000
Press and Religion	vs.	Fiction and Stop words	.114
Press and Religion	vs.	Learned and Press	1.000
Press and Religion	vs.	Misc and Stop words	.207
Press and Stop words	vs.	Fiction and Total	1.000
Press and Stop words	vs.	Learned and Stop words	.087
Press and Stop words	vs.	Learned and Total	1.000
Press and Total	vs.	Fiction and Stop words	.111
Press and Total	vs.	Fiction and Total	.271
Press and Total	vs.	Learned and Religion	.742
Press and Total	vs.	Learned and Stop words	.982
Press and Total	vs.	Learned and Total	.375
Press and Total	vs.	Press and Stop words	.819
Religion and Stop words	vs.	Fiction and Stop words	.305
Religion and Stop words	vs.	Fiction and Total	.707
Religion and Stop words	vs.	Learned and Religion	.936
Religion and Stop words	vs.	Learned and Stop words	.996
Religion and Stop words	vs.	Learned and Total	.795
Religion and Stop words	vs.	Press and Stop words	.956
Religion and Stop words	vs.	Press and Total	1.000

These results are somewhat unexpected; we expected the Khoekhoe stop words to behave like the English stop words. The linguistic properties of Khoekhoe, however, may provide an answer to this. Though Khoekhoe is comparable to English in terms of morphological complexity, its frequent grammatical words (e.g., particles and auxiliaries) only occasionally have more than one word form, whereas in English many frequent non-content words have frequent word forms, which jointly make up a substantial proportion of the top-200 list. For instance, there are only two tokens of the negation marking in Khoekhoe (the highly frequent *tama* in the present and past tenses and the less frequent *tide* in the future tense). In the NLTK stop word list for English, on the other hand, there are about 30 items with the negative marker including such orthographic forms as *aren*, *aren't*, *didn*, *didn't*, etc. (which are essentially related to variation in tokenization). Similar abundance of word forms is found with pronouns, e.g., in addition to *you*, the list includes *you'd*, *you'll*, *you're*, and *you've*, as well as *your* and *yours*. No comparable variation in frequent pronominal forms exists in Khoekhoe, though there are many more personal pronoun types including very infrequent ones, such as *sāse* for the first person inclusive plural feminine (i.e., "we" in the sense of the female speaker and female listeners including).

Essentially, whereas the range of high-frequency grammatical words is covered with some 50 words in Khoekhoe, after this various content words (nouns, verbs, adverb) start to show up in the top-200 list. They in turn vary more substantially between genres. On the other hand in English, there are many more high-frequency word forms of some of the most ubiquitous grammatical words, which are more uniformly distributed across genres.

If we look at the Jaccard coefficients between the English stop words from the stop word list and the different genres, we also see that the coefficients relatively quickly stabilize. Depending (a bit) on the genre under consideration, the coefficients are relatively high up to approximately $N = 75$. After that the Jaccard coefficients decrease. This indicates that around that point, content words are becoming more frequent. This corresponds to the fact that stop words are not always in the top most frequent words. From this we can learn that the development of a stop word list cannot only rely on the word frequency. An in-depth analysis of stop words is required.

6 Conclusion

Stop word lists are useful resources for a wide range of natural language processing tasks. For many low-resource languages, however, these lists are not available. In this article, we look at whether we can develop stop word lists based on word frequency information from corpora in different genres.

Previous work in identifying lists of stop words relied on the assumption that stop words are highly frequent and consistently highly frequent across documents. This means that by looking at the overlap of highly frequent words between documents (or documents from different genres), we should be able to identify stop words. Here, we apply this approach to an English (Brown) and Khoekhoe corpus that contains documents classified in various genres. We use the Jaccard coefficient to measure the overlap between the words in the different genres.

The results show that the English corpus has relatively high Jaccard coefficients for the comparison of different genres, which indicates that the stop words (taken from a known stop word list) are indeed often present in the top of frequency-ranked word lists. However, for Khoekhoe, the Jaccard coefficients are much lower, which means that there is more variation in the words that occur frequently in different genres.

Limitations

The research presented in this article has a number of limitations. First, the choice of genres and (amount of) texts may have an influence on the results. However, given that both corpora contain approximately 1 million tokens, the influence of this may be limited. On the other hand, it should be noted that overall fewer texts were included in the Khoekhoe corpus (e.g., only a couple of long texts in Fiction) and some genres (Religion) were massively overrepresented, this might have skewed both similarities and differences between some of the genres and between specific genres and the total frequencies.

Second, depending on the task, a slightly different stop word list may be required. Whereas for some tasks, for example, modal verbs or frequent adverbs are not relevant, for others they may be. This means that a generic stop word list is not always useful.

Finally, for practical reasons, we chose to use the existing stop word list for English as a baseline, however, as we discuss above, the English stop word list contains a large number of word forms and spelling variants (e.g., related to tokenization decisions) for many non-content items. A comparison of Khoekhoe high-frequency items with other languages might result in a more similar picture.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Addison-Wesley Publishing Company, Reading, MA, USA.

Steven Bird, Ewan Klein, and Edward Loper. 2009. NLTK: The natural language toolkit. In *Proceedings of the ACL 2009 Interactive Demonstrations*, pages 66–71. Association for Computational Linguistics.

Matthias Brenzinger. 2013. The twelve modern Khoisan languages. In Alena Witzlack-Makarevich and Martina Ernszt, editors, *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6-10, 2008, Riezlern/Kleinwalsertal*, Research in Khoisan Studies. Cologne: Rüdiger Köppe.

Rohan Chavan, Vishal Madle, Gaurav Patil, and Raviraj Joshi. 2024. Curating stopwords in Marathi - A TF-IDF approach. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, page 1–6. IEEE.

Curriculum Committee for Khoekhoegowab. 2003. *Khoekhoegowab: 3||î Xoaigaub*. Gamsberg Macmillan, Windhoek.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas, TX, USA.

Christopher Fox. 1989. A stop list for general text. *ACM SIGIR Forum*, 24(1-2):19–35.

Tom Güldemann. 2014. 'Khoisan' linguistic classification today. In Tom Güldemann and Anne-Maria Fehn, editors, *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, pages 1–41. John Benjamins, Amsterdam.

Wilfrid H. G. Haacke. 1999. *The tonology of Khoekhoe (Nama/Damara)*. Rüdiger Köppe, Cologne.

Wilfrid H. G. Haacke. 2018. Khoekhoegowab (Nama/Damara). In *The social and political history of Southern Africa's languages*, pages 133–158. Palgrave Macmillan, London.

Wilfrid H. G. Haacke and Eliphas Eiseb. 2002. *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Gamsberg Macmillan, Windhoek.

Wilfrid H.G. Haacke. 2013. Phonetics and phonology: Namibian Khoekhoe (Nama/Damara). In Rainer Vossen, editor, *The Khoesan languages*, pages 51–56. Routledge, London.

Samuel Huston and W. Bruce Croft. 2010. Evaluating verbose query processing techniques. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298, New York, NY, USA. Association for Computing Machinery.

Emil Hvitfeldt and Julia Silge. 2021. *Supervised Machine Learning for Text Analysis in R*. Chapman and Hall/CRC.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Rachel T.-W. Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. *Journal of Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR '05)*, 3(1):3–8.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press, Cambridge.

Hirosi Nakagawa, Alena Witzlack-Makarevich, Daniel Auer, Anne-Maria Fehn, Linda Ammann Gerlach, Tom Güldemann, Sylvanus Job, Florian Lionnet, Christfried Naumann, Hitomi Ono, and Lee J. Pratchett. 2023. Towards a phonological typology of the Kalahari Basin Area languages. *Linguistic Typology*, 27(2):509–535.

Namibia Statistics Agency. 2011. Namibia household income & expenditure survey 2009/2010. Technical report, Namibia Statistics Agency, Windhoek.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

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. White paper, Stanford Institute for Human-Centered AI (HAI) / The Asia Foundation / University of Pretoria. Available online.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817, Reykjavik, Iceland. European Language Resources Association (ELRA).

Julia Silge and David Robinson. 2016. Tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3):37.

Kira Tulchynska, Sylvanus Job, and Alena Witzlack-Makarevich. 2025a. Universal Dependencies treebank for Khoekhoe (KDT). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 119–128, Ljubljana, Slovenia. Association for Computational Linguistics.

Kira Tulchynska, Alena Witzlack-Makarevich, Sylvanus Job, and Michael Hahn. 2025b. Universal dependencies treebank for khoekhoe (KDT). In Daniel Zeman and et al., editors, *Universal Dependencies 2.16*. Universal Dependencies Consortium.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112.

C. J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. University of Glasgow, Glasgow, UK. Printout.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*, pages 174–189. Springer.

A Visualization of full data collections

In the article, we have selected only those genres from the Khoekhoe and Brown corpus that occur in both corpora. Here we provide similar information to Table 1, but now for the full Khoekhoe corpus in Table 3 and the full Brown corpus (after removing non-word tokens) in Table 4. Plots similar to that of Figure 1 can be found in Figure 2 for all Khoekhoe genres and in Figure 3 for all Brown genres.

Table 3: Distribution of words per genre in the complete Khoekhoe corpus.

Genre	# tokens
Conversion	155,292
Fiction	37,513
Learned	18,083
Misc	75,153
Press	76,328
Religion	656,808
Total	1,019,177

Table 4: Distribution of words per genre in the complete Brown corpus. The Press and Fiction genres are combined at that level. All non-word tokens are removed.

Genre	# tokens
Belles-Lettres	151,548
Fiction	235,489
Humor	18,265
Learned	159,936
Misc	61,143
Popular lore	96,691
Press	177,177
Religion	34,308
Skill and hobbies	71,531
Total	1,006,088

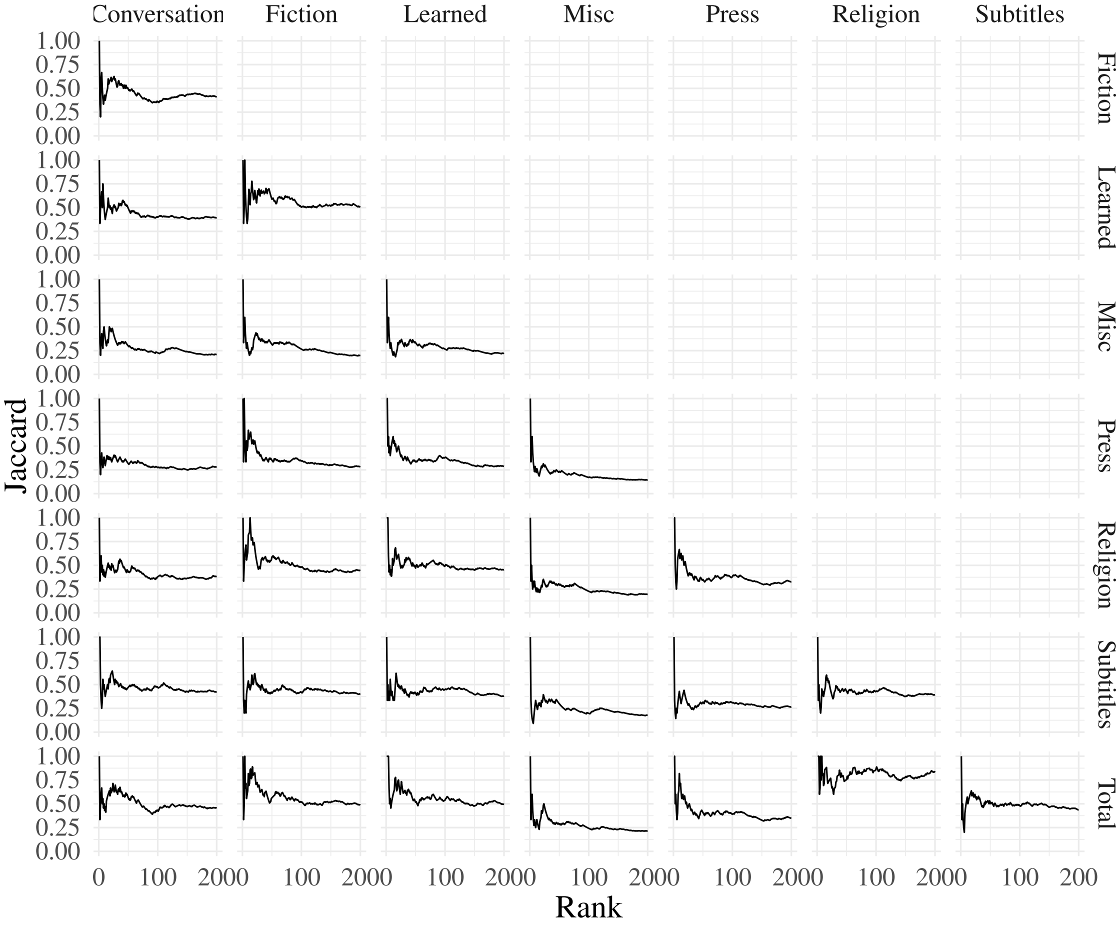

Figure 2: Overview of pairwise comparisons between genres in the complete Khoekhoe corpus. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes).

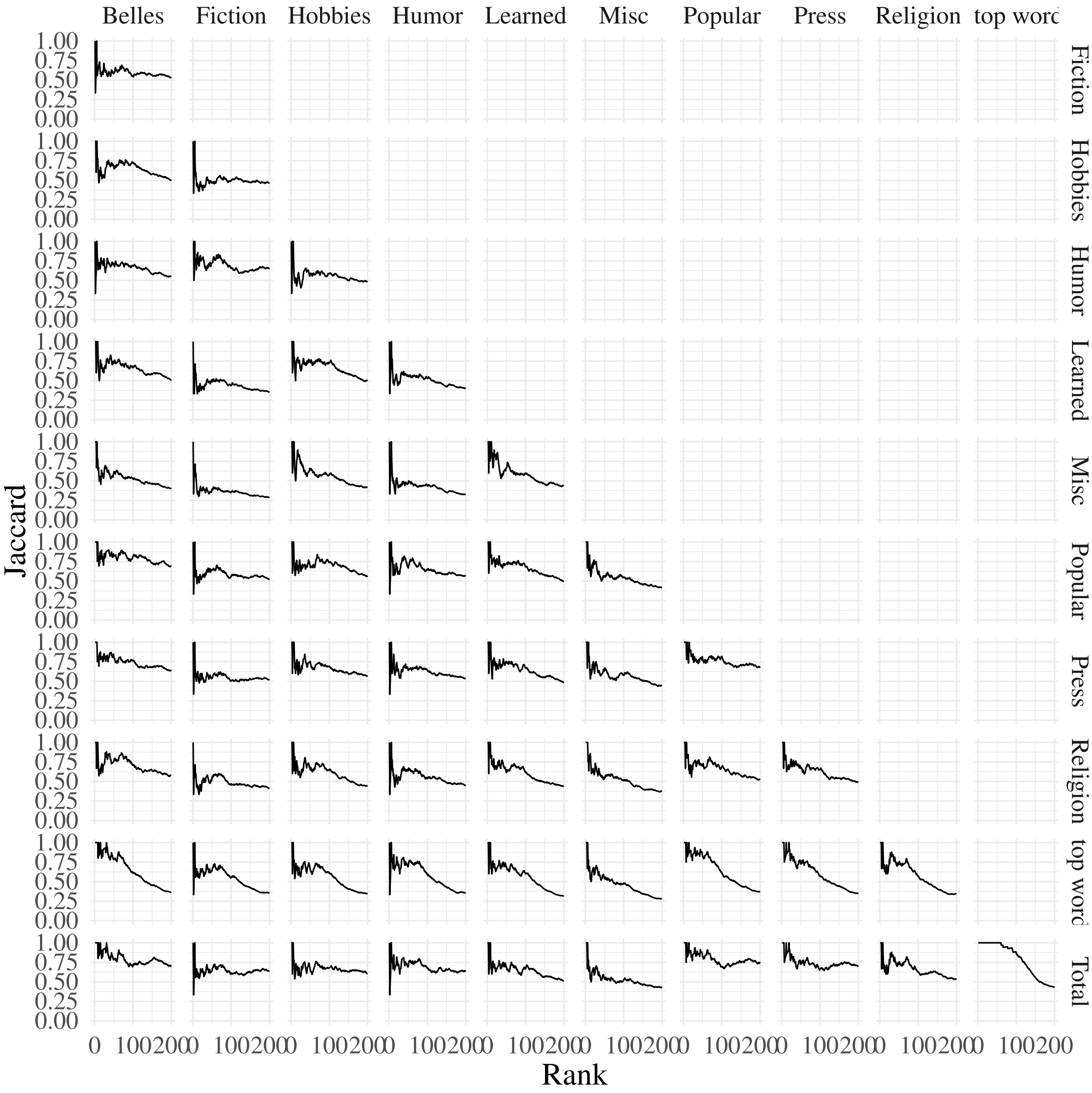

Figure 3: Overview of pairwise comparisons between genres in the complete Brown corpus. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes). Additionally the stop word data is provided.