# Extension of Linguistic Resources for South African Languages: Part-of-Speech Annotated Domain-Specific Data

**Tanja Gaustad, Roald Eiselen, Cindy McKellar**
Centre for Text Technology (CTexT)
North-West University, Potchefstroom, South Africa
{tanja.gaustad|roald.eiselen|cindy.mckellar}@nwu.ac.za

**Abstract**

In this paper, we present part-of-speech (POS) annotated domain-specific data for nine South African languages. The data has been sourced from five different domains (two academic domains, Caps and theses, two non-academic domains, news and magazines, and one fiction domain, novels), uniformly pre-processed, automatically POS-tagged and then corrected by linguistic experts. The widely used NCHLT government data sets (Eiselen and Puttkammer, 2014) have also been re-tagged with the current tag sets and manually corrected. Both the new domain-specific data sets and the re-tagged NCHL data sets have been uploaded into a public repository. To illustrate the characteristics of the domain data in comparison to government data, we include and discuss data statistics, namely type-token ration (TTR), tokens per sentence and out-of-vocabulary (OOV) rates, as well as POS tagging results with a baseline tagger trained on NCHLT data and applied to the different domains for all languages. Both the data statistics and the POS results clearly show that the domain data is significantly different to government data: For all domains and languages, the tagging accuracy decreases significantly compared to testing on in-domain government data. Also, POS results for the two domains with the highest OOV rates for all languages (Caps and novels) are much lower than for the other domains. These findings emphasise the need for more diverse data resources which in turn will aid in the development of more domain-independent language technologies.

**Keywords:** POS annotation, domain-specific data, under-resourced languages, South African languages

## 1. Introduction and Background

Data is key to the most recent developments in the field of Human Language Technology (HLT) and Machine Learning, e.g. Deep Learning. In addition, (diverse) language corpora benefit language research and language learning. However, the official languages of South Africa remain under-resourced[1] due to various reasons, such as historical inequality, economic disincentives or educational barriers to name but a few. Also, as a consequence of data scarcity, not a lot of variety in types of data can be found for South African languages.

In this paper, we describe new part-of-speech (POS) annotated data for nine South African languages sourced from several domains, namely academic texts, non-academic texts and fiction. Our aim is to extend the available POS-tagged data for the four official South African languages with a conjunctive orthography, i.e. isiNdebele (NR), isiXhosa (XH), isiZulu (ZU), and Siswati (SS), as well as for the five disjunctively written languages, i.e. Sesotho sa Leboa/Sepedi (NSO), Sesotho (ST), Setswana (TN), Tshivenḓa (VE), and Xitsonga (TS).[2]

With the data presented here, we hope to make a significant contribution to the availability of high quality linguistically annotated resources for the development of HLT technologies, but also for the further study of South African languages by (corpus) linguists, digital humanists and others, in a range of different domains.

Beyond the availability of the data, it is moreover well-established that Natural Language Processing (NLP) technologies trained on a particular domain generally perform substantially worse when applied to another domain (Van Asch and Daelemans, 2010; Derczynski et al., 2013; Schnabel and Schütze, 2013; Plank et al., 2014; Eiselen and Gaustad, 2026). These new data sets will also allow more robust taggers to be trained to improve automatic annotation across a broader variety of domains.

The remainder of the paper is organised as follows: Section 2 contains a description of the domain-specific data included in the newly released corpora, detailing the sources used, the amount of tokens included as well as a discussion of the preprocessing and data selection procedure followed to arrive at the final data sets. A presentation of various data statistics in Section 3, including e.g. Type-Token Ratio (TTR) and tokens per sentence, serves to demonstrate the difference in structure and content of the domain-specific data for the different languages. In Section 4, the POS annotation is described. Furthermore, we discuss the setup of (non-exhaustive) POS experiments carried out

[1] See the blog post "NLP with low-resourced languages: beyond bean counting artefacts" (https://keet.wordpress.com/2026/01/) and Keet and Khumalo (2026) for a discussion of language resourcedness.

[2] Work for Afrikaans is currently underway and that data will also be released shortly.



*Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026*, pages 7–19
12 May 2026. 

along with their results using a model trained on government data and applied to the different domain data sets. We conclude with a summary of our findings and remarks on future work in Section 5.

## 2. Domain-Specific Data

For the data resources described here, we have annotated data originating from different domains for the nine South African Bantu languages. This resulted in corpora of between 55,000 and 75,000 tokens per language containing equal portions of five different text types, namely two academic text types (National Senior Certificate examinations (Caps), MA/PhD theses), two non-academic text types (news, magazines) and fiction (novels). All data resources are publicly available (Gaustad, 2026a,b,c,m,n,o,p,q,r).[3] We also include a description of corpora from the government domain, the converted and corrected National Centre for Human Language Technology (NCHLT) data sets. See Table 1 for a full overview of the final data sets.

### 2.1. De Facto Standard: Government Domain

For the South African languages, the development of core technologies, like POS taggers or Named Entity Recognizers, has mostly been based on government domain corpora, making them the de facto standard data type. These corpora, originally released as NCHLT data sets in 2014 (Eiselen and Puttkammer, 2014)[4], contain data crawled from South African government websites as they are relatively easily accessible and free to use. The textual material is a combination of pamphlets, forms, legislative text, school materials, and informational content on a variety of subjects (health, local municipalities, tourism, etc.). As part of the work described here, the NCHLT corpora have been converted to the revised tag sets established as new standard and used in more recent research (du Toit and Puttkammer, 2021; Gaustad and Puttkammer, 2021; Puttkammer and Gaustad, 2021) and thoroughly quality checked. This resulted in uniformly linguistically annotated corpora for POS (and lemmas) for nine languages (Gaustad, 2026d,e,f,g,h,i,j,k,l), and also makes it possible to combine the NCHLT data sets with the government data sets for conjunctive languages released more recently (Gaustad and Puttkammer, 2022), increasing the available government data from roughly 50,000 tokens to 100,000 tokens for these four languages.

Even though government texts contain a variety of content and the developed core technologies perform adequately within the same domain, it has been shown that applying and evaluating these technologies on different domains results in a substantial loss in performance (Van Asch and Daelemans, 2010; Derczynski et al., 2013; Schnabel and Schütze, 2013; Plank et al., 2014; Eiselen and Gaustad, 2026). This performance loss could be due to differences in vocabulary, topics, and writing style between training and testing data (Manning, 2011), and can be felt most keenly by researchers in related fields, who want good results regardless of the original domain a model was trained on. Hence the initiative of annotating more diverse data.

### 2.2. Academic: National Senior Certificate Examinations (Caps)

A readily available source of academic data for all official languages of South Africa are National Senior Certificate examinations (commonly referred to as "matric exams"). Every high school student in South Africa is required to choose at least two of the twelve official South African languages[5] as subjects in order to qualify for a high school diploma. At least one of the chosen languages needs to be completed at "Home Language" level which refers to a language in which the learner has mastered reading, writing and interpersonal communication. The grade 12 test material of previous years is made available through the website of the Department of Basic Education[6] and typically contains reading comprehension questions as well as summary writing texts (see Sibeko and van Zaanen (2023) for a more in-depth discussion of the contents of these exams).

For the compilation of this corpus, we have downloaded the exams for all languages (except English) tested as Home Language from 2017 to 2024 and applied our pre-processing steps (see Section 2.7) for the final selection of data. Table 1 shows the number of tokens collected and annotated for the Caps domain per language.

### 2.3. Academic: MA and PhD Theses

Next to the Caps domain data which is aimed at grade 12 learners, the second example of aca-

[3] For isiZulu and Sepedi, the data is a combination of 20,000 and 30,000 previously annotated tokens respectively (Gaustad, 2024a,b) (used as a proof of concept to acquire funding) with extra data to make a complete corpus of at least 55,000 tokens per language.

[4] The original NCHLT data sets are available at Puttkammer et al. (2014a,b,c,d,e,f,g,h,i).

[5] Including South African Sign Language (SASL) which was recognised as an official language in 2023.

[6] https://www.education.gov.za/Curriculum/NationalSeniorCertificate(NSC)Examinations.aspx

| Language | | NCHLT train | Caps | Magazines | News | Novels | Theses | *Total tokens w/o NCHLT* |
|---|---|---|---|---|---|---|---|---|
| isiNdebele | NR | 38,427 | 14,659 | 0 | 17,838 | 16,440 | 12,157 | *61,094* |
| Siswati | SS | 39,486 | 13,750 | 0 | 20,227 | 14,736 | 12,122 | *60,835* |
| isiXhosa | XH | 42,049 | 12,005 | 11,437 | 11,336 | 12,480 | 12,060 | *59,318* |
| isiZulu | ZU | 41,580 | 11,895 | 11,587 | 14,244 | 15,657 | 14,492 | *67,875* |
| Sepedi | NSO | 65,920 | 15,510 | 13,320 | 16,475 | 14,991 | 14,721 | *75,017* |
| Sesotho | ST | 66,881 | 11,276 | 12,613 | 11,384 | 11,827 | 11,121 | *58,221* |
| Setswana | TN | 65,802 | 11,589 | 12,338 | 10,960 | 11,657 | 11,539 | *58,083* |
| Xitsonga | TS | 63,091 | 11,057 | 11,171 | 11,402 | 12,423 | 11,484 | *57,537* |
| Tshivenḓa | VE | 59,814 | 11,166 | 0 | 22,355 | 11,857 | 11,525 | *56,903* |

Table 1: Overview of token counts per domain and language for the final corpora.

demic writing we have included in our data collection are Masters and PhD theses. The writing in university theses represents highly academic and formal language and typically focuses on the critical analysis of a given subject. Most South African universities have electronic thesis and dissertation repositories, but the majority of the documents are in English, followed by Afrikaans. Other South African languages are less represented, the biggest hurdle proving to source theses in isiNdebele, Sesotho and Siswati.

Once the source data had been acquired, we applied our pre-processing steps to ensure uniform treatment of all different data types and to select the required number of tokens. See Table 1 for an overview of the token counts per language for academic theses.

## 2.4. Non-Academic: News

Curated news articles are informative, focused on delivering timely and objective news, and intended for a larger, more general audience, which requires the writing to be more accessible and less formal than most public administration or academic texts. Non-academic texts generally prioritise readability and engagement, employing less complex language that is crafted for emphasis, clarity, and in some cases sensationalism (Bhatia, 1993).

Even though news data is readily available for many high-resource languages, this is not the case for South African languages. This is partially due to copyright restrictions, but also because many news publishers revert to English rather than publishing multilingually (which requires more time, effort, and money).

The following sources were used to collect news data:

- Dizindaba newspaper[7], a small commercial newspaper of the Eastern and Western Cape provinces, for isiXhosa;
- Isolezwe (via Leipzig Corpora Collection[8], Goldhahn et al., 2012), a daily Durban-based newspaper, for isiZulu;
- Limpopo Mirror[9], a community-focused newspaper serving rural communities along the Limpopo River, for Tshivenḓa;
- KZN Namuhla[10], a community Newspaper with current and local news, for isiZulu;
- Seipone[11], a fortnightly local news publication, for Sepedi;
- Vuk'uzunzele[12], a government issued newsletter, for isiNdebele, Sepedi, Sesotho, Setswana, Siswati and Tshivenḓa.

Table 1 contains the final counts for newspaper data per language.

## 2.5. Non-Academic: Magazines

Similarly to news, magazines are considered popular literature as magazine articles are written for and read by the general public. They are sources of information and usually cover more in-depth, specialized or thematic content that is published less frequently than newspapers.

Unfortunately, not many magazines are (openly) available for the South African languages. Therefore, this data category is rather mixed and the contents vary significantly between languages. The following sources have been included:

- Bona magazine[13], a generic magazine, for Sesotho, isiXhosa and isiZulu;
- Pula/Imvula, a magazine aimed at educating emerging farmers, for Sesotho, Sesotho sa Leboa, Setswana, isiXhosa and isiZulu (see Gaustad et al. (2025) for more details);

[7] https://dizindaba.co.za/

[8] https://corpora.uni-leipzig.de/

[9] https://www.limpopomirror.co.za/

[10] https://kznnamuhlanews.co.za/

[11] https://seiponemadireng.co.za/

[12] https://www.vukuzenzele.gov.za/

[13] https://www.bona.co.za/

- VIV Mag[14], an online magazine, for Xitsonga.

For isiNdebele, Siswati and Tshivenḓa we could not source any magazines. In order to still reach a total of min. 50,000 tokens per language, we increased the data included for three other categories, namely news, novels and Caps, but as a consequence, only four types of domain corpora (instead of five) are available for these three languages. See Table 1 for a full overview on word counts for magazines.

## 2.6. Fiction: Novels

The last domain included in the released domain-specific corpora are novels, representing fictional narratives. Novels are generally written to convey a story, emotions, and experiences, and the language used varies from highly literary and descriptive to straightforward and conversational, depending on the style of writing and the type of novel. Unfortunately, we do not have detailed information on the content or target audience of all the novels included, as this information is often unavailable.

The novels in the described resources have been sourced as follows:

- Novels published by Oxford University Press as well as Shuter and Shooter from 2007 onwards and acquired by the South African Centre for Digital Language Resources (SADiLaR)[15], for isiNdebele, Siswati, Sepedi, Sesotho, Setswana, Tshivenḓa, isiXhosa and isiZulu;
- Children's books from African Story book[16] for Xitsonga;
- Children's books from Bookdash[17] for Xitsonga.

Table 1 shows the token counts for data from novels for each language.

## 2.7. Pre-Processing and Data Selection

As a first step, all collected documents were extracted to UTF-8 text files and combined by source type, including markup to keep the separate source files apart during processing. The files were sentence separated and all exact duplicates were removed. This pre-processing step resulted in five different domain-specific text files per language (or four where magazines were not available), each containing sentence separated, unique data. Due to the highly repetitive nature of the Caps data source, and complete or partial duplication of articles and stories in some of the news, novels and magazine sources, this deduplication step removed varying amounts of data for the different languages and domains: Caps data for all languages was reduced by 20%–25% whereas other domains experienced less dramatic cuts, often less than 10%, with a few outliers in isiNdebele novels (44%), Xitsonga magazines (63%) and isiXhosa and isiZulu news (42% and 89% respectively). The outcome of this first pre-processing step were sizeable text collections for each domain which ensured sufficient room to remove unwanted segments so that the final data was of the best possible quality. Word counts ranged between 500,000+ for the higher resource languages, like isiXhosa and Sepedi, and 30,000 per domain for smaller, lower resourced languages, like isiNdebele and Siswati.

Publications in the South African languages often contain many English or other South African language sentences or phrases mixed in with the main language of a document. To ensure the data used for POS tagging is mainly in the target language, all data was sorted using a proprietary language identifier. Sentences identified as belonging to a given language with a probability higher than the set threshold were kept, any sentences with lower probabilities were discarded. Given that some of the languages are considered resource-scarce, some types of domain-specific texts were very hard to locate, and some of the languages are very similar, making misidentification more likely, the probability barrier was set differently for each language and domain. This ensured the retained data was as good as possible without sacrificing too much and dropping below target amounts needed for the annotation project. All languages and domains were filtered for language identification on at minimum a 50% probability, with some of the more resource-rich languages and domains being filtered as high as 80%.

Since many of the data sources originated as PDF documents, they needed to be extracted to plain text. For more modern documents, this works very well, but older documents need to be OCRed to create their corresponding text formats. This can lead to the introduction of errors, such as non-existing characters (caused by broken diacritics), fragmented or conjoined words, and spelling errors. In order to minimize these problems, the text was first filtered through a clean-up script that removed any lines containing characters that were not common in that language in order to remove broken diacritics, and then spellchecked to remove spelling mistakes caused by incorrect OCR. During this cleaning phase, all sentences where spellchecked and filtered based on the percentage of correctly spelled words. As spelling checkers for the con-

[14] https://www.vivmag.co.za/
[15] https://sadilar.org/en/
[16] https://www.africanstorybook.org/
[17] https://bookdash.org/

junctive languages have a lower recognition rate and the different domains and languages had different amounts of extra data available, not all languages were filtered on the same percentage correctly spelled words. The lowest percentage used was 60%, only applied to the conjunctive languages due to the higher number of correctly spelled words that the spelling checkers cannot recognize. Other languages were filtered either on 70% or 80% correctly spelled words per sentence, depending on the amount of extra data available.

After spellchecking, the remaining data was furthermore filtered to remove fragmented sentences caused by PDF extraction, headings and list elements. Since POS tagging relies on sentence structure, the filtering was done in an attempt to only keep full sentences, i.e. sentences starting with capital letters and ending with either sentence termination punctuation or colons. At this stage, any sentences made up by more than half capital letters were also filtered out to remove instructions and partial headings present in the Caps data.

All these clean-up measures meant that the remaining data was no longer continuous running text as is found in ordinary documents. It also resulted in a second significant amount of data being removed. The data excluded at this point (not counting data already removed during deduplication) averaged out at 48% if counted across all languages and domains. The conjunctive languages did however experience a higher average data loss than the disjunctives (43% for disjunctives and 54% for conjunctives), mainly due to the previously mentioned problems with spellchecking highly agglutinative languages. Other variations in the amount of data discarded can be attributed to the quality of the original input with text originating in older PDFs being the most problematic. Given the set amount of data to be annotated per domain, the strict clean-up process ensured the final data would be of the highest possible quality and usability.

For some domains and languages much more data remained than required for annotation, so the remaining data was at this point further reduced to approximately the desired amount, with some leeway left for the final clean-up step. During this data selection step, random chunks of 10 sentences were chosen in such a way that data from each file from the original selection was included and larger files contributed more chunks than small files. In this way the data selection was evenly spread over all the original files.

The final pre-processing step applied was similarity checking. As POS annotation is expensive and the budget for this project was limited, only a set amount of data from each domain could be annotated. For this reason we tried to avoid (near)duplication of sentences in order to include more new and unique data. Sentences that were identical except for a single letter, number or word were especially prevalent in the Caps data. To remove these almost identical sentences, all the selected data (per domain) was put though a similarity filtering script. The script compared each new sentence to all the sentences that were kept before by measuring the Levenshtein distance between the sentences. Sentences with a similarity of 70% and higher were discarded while less similar sentences were added to the kept data to be compared to future sentences. This process is rather slow which is why it was only done after the data selection described in the previous paragraph. The similarity-based deduplication lead to very small amounts of data being discarded, with nearly every domain experiencing at most 1% loss. The exception to this was the Caps domain where data loss averaged at 6% due to the very repetitive nature of this domain's data.

## 3. Data Statistics

Table 2 shows a summary of relevant statistics for the language resources presented in this paper. We have aggregated the different counts for conjunctively and disjunctively written languages in the interest of readability of the table and report standard deviations.[18] This overview—together with the POS results discussed in Section 4—serves to showcase similarities and differences between language groups as well as domains, and hopefully illustrates the need for domain-specific data to diversify the content of available language resources.

The (normalised) type-token ratio (TTR)[19] highlights the impact of the two different orthographies used for South African Bantu languages: Conjunctively written languages (NR, SS, XH, ZU) have much higher TTR values than the disjunctively written languages (NSO, ST, TN, TS, VE). Additionally, sentences are typically shorter, i.e. contain less tokens, for the conjunctive languages when compared to the disjunctive ones (see Prinsloo and de Schryver (2002) for a more in-depth discussion of these measures and their significance for the South African context).

In order to gauge the lexical differences between the domains, out-of-vocabulary (OOV) rates are also included. These show the ratio of unknown words in relation to a generic corpus, in our case the NCHLT training corpus containing government data. The values reported in Table 2 illustrate well that all different domain corpora contain a significant amount of new vocabulary items not present in the

[18] The full data statistics per language are available in Table 3 in Appendix A.

[19] Type-token ratio is normalised per 1,000 tokens.

| Language | Domain | TTR/1000 | Tokens/Sent. | OOV |
|---|---|---|---|---|
| Conjunctive languages | NCHLT train | 0.62 (±0.01) | 15.69 (±0.72) | *na* |
| (NR, SS, XH, ZU) | NCHLT test | 0.68 (±0.01) | 12.69 (±0.75) | 0.33 (±0.01) |
| | Caps | 0.68 (±0.01) | 10.94 (±1.24) | 0.48 (±0.01) |
| | Magazines | 0.69 (±0.03) | 14.50 (±2.15) | 0.41 (±0.03) |
| | News | 0.71 (±0.00) | 16.86 (±2.61) | 0.42 (±0.00) |
| | Novels | 0.66 (±0.01) | 9.74 (±1.91) | 0.50 (±0.02) |
| | Theses | 0.64 (±0.06) | 14.85 (±4.31) | 0.45 (±0.04) |
| Disjunctive languages | NCHTL train | 0.33 (±0.01) | 25.12 (±1.26) | *na* |
| (NSO, ST, TN TS, VE) | NCHLT test | 0.37 (±0.01) | 20.83 (±0.86) | 0.09 (±0.01) |
| | Caps | 0.38 (±0.02) | 16.79 (±1.47) | 0.17 (±0.01) |
| | Magazines | 0.36 (±0.02) | 25.27 (±4.02) | 0.16 (±0.02) |
| | News | 0.38 (±0.01) | 27.63 (±3.63) | 0.12 (±0.01) |
| | Novels | 0.35 (±0.01) | 18.69 (±4.58) | 0.17 (±0.02) |
| | Theses | 0.33 (±0.02) | 24.92 (±2.07) | 0.14 (±0.02) |

Table 2: Summary of domain-specific data statistics with languages averaged per writing style (standard deviation in brackets).

NCHLT training data, and thereby contribute to the diversification of available lexical content.

When comparing the values for the different domains, we see that news (for conjunctively written languages) and Caps and news (for disjunctively written languages) have the highest TTR, whereas the NCHLT training corpus and theses (for both writing styles) have the lowest TTR. This points to more repetition in the two domains with lower TTR values, which aligns with more codified language use (theses and government data) as well as a more narrow focus/topic (theses). Similarly, news, NCHLT train and theses (for conjunctive languages) and news, magazines and NCHLT train (for disjunctive languages) all contain long sentences, while novels and Caps texts (for both orthographic styles) comprise shorter sentences. This can possibly be attributed to the expected level of language mastery of the intended audience, viz. grade 12 language learners for Caps and young readers for at least a proportion of novels. Interestingly, the observed OOV rates for novels and Caps (again for both writing styles) also show the highest amount of vocabulary not present in the NCHLT training data. As we will see in the discussion of the POS results in Section 4.3, this is indicative of the fact that these two domains are most dissimilar to government data.

## 4. POS Tagging

### 4.1. POS Annotation and Tag Sets

After finalising the data acquisition, pre-processing and final data selection, we proceeded to POS annotation. In a first step, all data was automatically annotated with POS information using the CTexT NCHLT Web Services available at https://hlt.nwu.ac.za/.[20] These automatic annotations were subsequently checked and if necessary corrected by linguistic experts.[21] During this stage, spelling mistakes, OCR errors and other issues due to imperfect digitisation and extraction processes were also manually identified and corrected.

In contrast to an English POS tag set, e.g. the Penn Treebank set with a total of 48 tags (Marcus et al., 1993; Taylor et al., 2003), or the Universal POS (UPOS) tag set (Petrov et al., 2012) with 17 tags, a comprehensive POS tag set for Bantu languages is typically substantially larger to accommodate the distinction of linguistic features such as noun classes (numbering between 12 and 20 different classes), various concords, as well as additional types of pronouns. Although there have been attempts for a few Bantu languages to use the language agnostic UPOS tag set (Dione et al., 2023; Gaustad et al., 2024), finding the corresponding UPOS tags for some Bantu-specific POS categories has proven challenging.

For the data presented here, the full POS tag sets range from 107 tags for the conjunctive languages to between 197 (NSO) and 243 (VE) tags for the disjunctive languages. The main reason for the larger POS tag sets for the disjunctively written languages lies in the orthography: tokens that are written agglutinatively for isiNdebele, isiXhosa, isiZulu and Siswati (and therefore not tagged separately), require their own POS tag in the disjunctively written languages, e.g. PART for particles or MORPH for tense, aspect and negative markers. The POS pro-

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

[21] For most of the languages included, we only had access to one qualified language expert (under-resourced also applies to linguistic experts), and hence we cannot report inter-annotator agreement.

tocols containing the relevant tag sets, examples and notes are distributed together with the data.

### 4.2. POS Experiments and Taggers

In an effort to get a first impression of the impact of different domains on POS tagging accuracy, we conducted a basic and non-exhaustive experiment using a model trained on government data for each language and then applied those baseline POS taggers to each different domain data set in turn.[22] The respective language-specific taggers were trained on the NCHLT POS annotated training data using the FLAIR biLSTM-CRF framework (Akbik et al., 2019) and FLAIR backward character embeddings (Eiselen, 2023a,b,c,d,e,f,g,h,i). The biLSTM consists of 512 hidden units trained for a maximum of 80 epochs with a batch-size of 128.

The embeddings were not fine-tuned during training. To ensure comparable results across languages and domains, the experiments used the same tagger architecture for all languages, even though there could be more optimal settings for individual languages, such as using different embedding models, different batch sizes, or different numbers of hidden units.

### 4.3. POS Results

The results of the taggers trained on the NCHLT data and applied to each domain and language are presented in Figures 1a (per language and domain) and 1b (per domain). These results clearly indicate the nature and extent of the degradation of the tagging accuracy when our data from different domains are automatically annotated.

Firstly, the in-domain NCHLT test evaluations in Figure 1a show the best performance across all languages, and in most cases substantially outperform the taggers in the other domains. This can also be observed in Figure 1b with isiNdebele the clear outlier the NCHLT test data. The degradation for other domains is most notable for the Caps and novel domains, with regressions of between ∼4% (ZU Caps) and ∼13% (ST and SS novels) in absolute accuracy.

These accuracy results also generally reflect the OOV rates of the various sets (see Table 2), with Caps and novels exhibiting the highest OOV rates, while magazines and news data have lower OOV rates, and generally outperform the Caps and novels domains, as illustrated in Figure 1b. This figure also nicely illustrated that novels have the widest spread of tagging accuracies, followed by Caps and theses, while news and magazines perform more uniformly across languages (with two notable outliers: NSO performing much better for magazines and NR performing much worse for news).

Moreover, there is (again) a clear distinction between the two orthographic styles: The disjunctive languages generally perform better than the conjunctive languages, although there are some language- and domain-specific outliers in the results. Siswati, for instance, performs on a similar level to the disjunctive languages on the NCHLT test and magazine/news sources, but performs worst on the theses domain. Sesotho also performs similarly to other disjunctive languages on the NCHLT test set, but performs comparably to the conjunctive languages for the Caps, novels and theses sets. Another noticeable outlier is magazines for Xitsonga, which is the worst performing domain for the language. This may in part be due to the fact that the magazine from which the data was sourced contains a relatively large number of serial short stories, which exhibit properties more similar to novels than typical magazine content.

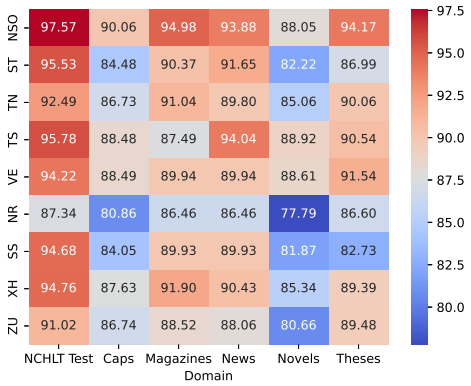
In conclusion, the Sesotho sa Leboa tagger performs the best across the different domains, while the isiNdebele tagging accuracy results are generally the worst for the different domains, with the exception of the theses domain where Siswati performs the worst. For all languages, novels is the most difficult domain to adapt to with the worst overall scores, followed by the Caps domain, while the magazines, news and theses data all perform similarly, with one or two exceptions.

## 5. Conclusion and Future Work

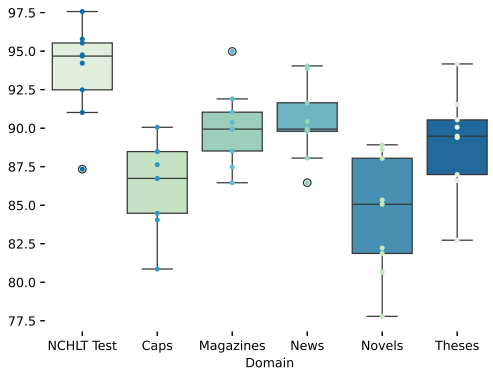
In this paper, we have presented new data from five different domains for the nine South African Bantu languages, as well as the uniformly POS annotated NCHLT data, with the aim to help diminish data sparseness and contribute to the diversification of language resources for these languages. Both the included data statistics and the POS accuracy results of a “generic” tagger trained on government data and applied to the different domains exemplify the differences in content and composition of data from non-government domains: In our data, we see that the non-academic domains of news and magazines are more similar to government data than academic theses, whereas Caps data (academic) and novels (fiction) are most dissimilar to government data.

These findings show the importance of including as diverse data sources as possible in resources for any given language and can potentially be used to guide future data acquisition and collection efforts. Moreover, domain-specific data is essential to test the generalisation power and performance of

[22] A more extensive analysis of domain adaptation is beyond the scope of this paper. However, results for POS domain adaptation on isiZulu and Sepedi data can be found in Eiselen and Gaustad (2026).




(a) Heatmap of POS tagging accuracy per language and domain. Note that results for magazines and news in NR, SS, and VE are kept the same (as described in Section 2.5).




(b) Boxplot of POS tagging accuracy per domain.

Figure 1: POS tagging accuracy results per language and domain for a model trained on NCHLT data.

core technologies, such as POS taggers, especially if these technologies are applied to new domains in Digital Humanities for instance. Adding various types of domain-specific data to training sets can also increase accuracy of POS tagging or Named Entity recognition, which can in turn improve downstream tasks, such as dependency parsing or machine translation.

Using the data presented here, future work will include building improved and more domain-independent core technologies. We plan to make these resources available as open-source python packages, as well as integrating them in the existing NCHLT web API[23] for easy access and use by researchers, also outside of HLT. Furthermore, we will continue exploring the repercussions of different domains and data sparsity on core technology performance.

## 6. Ethical Considerations and Limitations

The data described in this paper has been collected with the utmost care adhering to the highest ethical standards possible, and we hope to be fostering more equitable access to diverse data by making these resources openly available in the SADiLaR repository.[24]

The main limitation of the work presented originates in the available data: As detailed in Section 2, with the limited choice for South African Bantu languages, all possible options had to be pursued to find enough high-quality data for the endeavoured domain-specific corpora of min. 10,000 tokens each. This results in more heterogeneous data than would be the case for more highly resourced languages, in turn influencing the generalisability of the data and the results obtained on the data. However, we believe more diverse data resources are valuable regardless of this potential limitation.

[23] https://hlt.nwu.ac.za/

[24] https://repo.sadilar.org/home

## 7. Acknowledgements

This research was funded by the South African Centre for Digital Language Resources (SADiLaR) under projects “Linguistic corpus enrichment for South African languages” and “Update and extension of linguistic resources and core technologies for South African languages”. SADiLaR is a research infrastructure established by the Department of Science and Innovation (DSI) of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

## 8. Bibliographical References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Vijay K. Bhatia. 1993. *Analyzing Genre: Language use in professional settings*. Longman, New York.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria. INCOMA Ltd. Shoumen.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

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

Roald Eiselen and Tanja Gaustad. 2026. Domain adaptation in sequence labelling: A case study for two South African languages. *Northern European Journal of Language Technology*, 12(1).

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tanja Gaustad, Ansu Berg, Rigardt Pretorius, and Roald Eiselen. 2024. The first universal dependency treebank for Tswana: Tswana-Popapolelo. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 55–65, Torino, Italy. ELRA and ICCL.

Tanja Gaustad, Cindy A. McKellar, and Martin J. Puttkammer. 2025. Multilingual data from the agricultural domain: Presenting the NWU-Pula/Imvula Corpora. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 6(2).

Tanja Gaustad and Martin J. Puttkammer. 2021. Development of linguistically annotated parallel language resources for four South African languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA): Proceedings of the 2nd workshop on Resources for African Indigenous Language (RAIL) at DHASA 2021*, 3(3).

Tanja Gaustad and Martin J. Puttkammer. 2022. Linguistically annotated dataset for four official South African languages with a conjunctive orthography: isiNdebele, isiXhosa, isiZulu, and Siswati. *Data in Brief*, 41.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resource Association (ELRA).

C. Maria Keet and Langa Khumalo. 2026. Contextualising levels of language resourcedness for NLP tasks. Technical report, arXiv.

Chris Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing. CICLing 2011.*, volume 6608 of *Lecture Notes in Computer Science*, pages 171–189, Berlin, Heidelberg. Springer.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar. Association for Computational Linguistics.

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.

Tobias Schnabel and Hinrich Schütze. 2013. Towards robust cross-domain domain adaptation for part-of-speech tagging. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 198–206, Nagoya, Japan. Asian Federation of Natural Language Processing.

Johannes Sibeko and Menno van Zaanen. 2023. A data set of final year high school examination texts of South African Home and First Additional Language subjects. *Journal of Open Humanities Data*, 9(1).

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé, editor, *Treebanks: Building and using Parsed corpora*, volume 20 of *Text, Speech and Language Technology*, pages 5–22. Springer, Dordrecht.

Vincent Van Asch and Walter Daelemans. 2010. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 31–36, Uppsala, Sweden. Association for Computational Linguistics.

## 9. Language Resource References

Roald Eiselen. 2023a. *NCHLT isiNdebele FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/607.

Roald Eiselen. 2023b. *NCHLT isiXhosa FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/614.

Roald Eiselen. 2023c. *NCHLT isiZulu FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/615.

Roald Eiselen. 2023d. *NCHLT Sepedi FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/608.

Roald Eiselen. 2023e. *NCHLT Sesotho FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/610.

Roald Eiselen. 2023f. *NCHLT Setswana FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/611.

Roald Eiselen. 2023g. *NCHLT Siswati FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/609.

Roald Eiselen. 2023h. *NCHLT Tshivenda FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/613.

Roald Eiselen. 2023i. *NCHLT Xitsonga FLAIR-backward embeddings*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/612.

Tanja Gaustad. 2024a. *POS annotated corpus in 5 different genres for Sepedi*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/670.

Tanja Gaustad. 2024b. *POS annotated corpus with 5 different text types for isiZulu*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/671.

Tanja Gaustad. 2026a. *isiNdebele Domain corpus POS annotated (4 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/704.

Tanja Gaustad. 2026b. *isiXhosa Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/702.

Tanja Gaustad. 2026c. *isiZulu Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/701.

Tanja Gaustad. 2026d. *NCHLT isiNdebele POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/713.

Tanja Gaustad. 2026e. *NCHLT isiXhosa POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/711.

Tanja Gaustad. 2026f. *NCHLT isiZulu POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/712.

Tanja Gaustad. 2026g. *NCHLT Sepedi POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/708.

Tanja Gaustad. 2026h. *NCHLT Sesotho POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/709.

Tanja Gaustad. 2026i. *NCHLT Setswana POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/707.

Tanja Gaustad. 2026j. *NCHLT Siswati POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/710.

Tanja Gaustad. 2026k. *NCHLT Tshivenda POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/706.

Tanja Gaustad. 2026l. *NCHLT Xitsonga POS and Lemma annotated corpus*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/705.

Tanja Gaustad. 2026m. *Sepedi Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/700.

Tanja Gaustad. 2026n. *Sesotho Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/699.

Tanja Gaustad. 2026o. *Setswana Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/698.

Tanja Gaustad. 2026p. *Siswati Domain corpus POS annotated (4 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/703.

Tanja Gaustad. 2026q. *Tshivenda Domain corpus POS annotated (4 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/696.

Tanja Gaustad. 2026r. *Xitsonga Domain corpus POS annotated (5 domains)*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/697.

Martin Puttkammer and Tanja Gaustad. 2021. *Linguistically enriched corpora for conjunctively written South African languages*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/546.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014a. *NCHLT isiNdebele Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/302.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014b. *NCHLT isiXhosa Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/309.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014c. *NCHLT isiZulu Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/315.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014d. *NCHLT Sepedi Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR

Language Resource Repository. PID https://hdl.handle.net/20.500.12185/325.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014e. *NCHLT Sesotho Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/332.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014f. *NCHLT Setswana Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/337.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014g. *NCHLT Siswati Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/344.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014h. *NCHLT Tshivenda Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/353.

Martin Puttkammer and Martin Schlemmer and Ruan Bekker. 2014i. *NCHLT Xitsonga Annotated Text Corpora*. Centre for Text Technology (CTexT). SADiLaR Language Resource Repository. PID https://hdl.handle.net/20.500.12185/359.

## Appendix A: Data Statistics Table

| Lang. | Domain | TTR/ 1000 | Tokens/ Sent. | OOV | Lang. | Domain | TTR/ 1000 | Tokens/ Sent. | OOV |
|---|---|---|---|---|---|---|---|---|---|
| NSO | NCHLT train | 0.32 | 25.54 | *na* | NR | NCHLT train | 0.63 | 14.85 | *na* |
| | NCHLT test | 0.35 | 22.09 | 0.07 | | NCHLT test | 0.69 | 11.85 | 0.35 |
| | Caps | 0.37 | 14.19 | 0.16 | | Caps | 0.67 | 12.33 | 0.49 |
| | Magazines | 0.38 | 23.91 | 0.13 | | Magazines | 0 | 0 | 0 |
| | News | 0.37 | 25.50 | 0.11 | | NewsMags | 0.70 | 17.86 | 0.42 |
| | Novels | 0.35 | 18.46 | 0.19 | | Novels | 0.65 | 8.22 | 0.53 |
| | Theses | 0.33 | 24.74 | 0.13 | | Theses | 0.55 | 20.30 | 0.40 |
| ST | NCHLT train | 0.34 | 26.33 | *na* | SS | NCHLT train | 0.62 | 15.41 | *na* |
| | NCHLT test | 0.37 | 21.20 | 0.08 | | NCHLT test | 0.68 | 12.32 | 0.32 |
| | Caps | 0.38 | 17.62 | 0.17 | | Caps | 0.68 | 9.62 | 0.49 |
| | Magazines | 0.37 | 27.42 | 0.14 | | Magazines | 0 | 0 | 0 |
| | News | 0.37 | 26.54 | 0.10 | | NewsMags | 0.67 | 17.79 | 0.39 |
| | Novels | 0.34 | 25.22 | 0.17 | | Novels | 0.66 | 9.26 | 0.50 |
| | Theses | 0.31 | 22.70 | 0.17 | | Theses | 0.66 | 11.00 | 0.50 |
| TN | NCHLT train | 0.33 | 25.49 | *na* | XH | NCHLT train | 0.62 | 16.51 | *na* |
| | NCHLT test | 0.37 | 20.51 | 0.10 | | NCHLT test | 0.69 | 13.52 | 0.33 |
| | Caps | 0.38 | 17.72 | 0.17 | | Caps | 0.68 | 11.61 | 0.48 |
| | Magazines | 0.34 | 29.45 | 0.17 | | Magazines | 0.71 | 16.02 | 0.43 |
| | News | 0.37 | 33.01 | 0.14 | | News | 0.71 | 18.71 | 0.43 |
| | Novels | 0.35 | 18.27 | 0.18 | | Novels | 0.68 | 12.53 | 0.48 |
| | Theses | 0.31 | 28.28 | 0.15 | | Theses | 0.69 | 16.28 | 0.46 |
| TS | NCHLT train | 0.34 | 25.31 | *na* | ZU | NCHLT train | 0.60 | 16.00 | *na* |
| | NCHLT test | 0.37 | 19.86 | 0.08 | | NCHLT test | 0.66 | 13.08 | 0.33 |
| | Caps | 0.38 | 17.33 | 0.15 | | Caps | 0.69 | 10.21 | 0.47 |
| | Magazines | 0.37 | 20.31 | 0.18 | | Magazines | 0.66 | 12.98 | 0.39 |
| | News | 0.40 | 25.45 | 0.12 | | News | 0.71 | 15.01 | 0.42 |
| | Novels | 0.37 | 12.29 | 0.19 | | Novels | 0.66 | 8.95 | 0.50 |
| | Theses | 0.36 | 23.97 | 0.15 | | Theses | 0.66 | 11.81 | 0.43 |
| VE | NCHLT train | 0.34 | 22.98 | *na* | | | | | |
| | NCHLT test | 0.38 | 20.48 | 0.10 | | | | | |
| | Caps | 0.41 | 17.07 | 0.18 | | | | | |
| | Magazines | 0 | 0 | 0 | | | | | |
| | NewsMags | 0.40 | 27.56 | 0.14 | | | | | |
| | Novels | 0.36 | 19.19 | 0.15 | | | | | |
| | Theses | 0.35 | 24.89 | 0.12 | | | | | |

Table 3: Domain-specific data statistics per language.