Domain Adaptation in Sequence Labelling: A Case Study for Two South African Languages

Roald Eiselen, roald.eiselen@nwu.ac.za

Tanja Gaustad tanja.gaustad@nwu.ac.za

Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa

Abstract In this paper, we investigate domain adaptation for Part-of-speech (POS) tagging of two under-resourced South African languages, isiZulu and Sesotho sa Leboa, by studying its effect on the POS tagging results and how to possibly predict what quality can be expected when applying an existing POS tagger to a new domain. We carry out systematic experiments across six domains (governmental texts, exam texts for grade 12 South African learners, magazines, newspapers, novels, and PhD theses) to determine how POS tagger accuracy deteriorates when switching between domains. To mitigate this quality deterioration, three different domain adaptation strategies are tested to determine the most relevant approach in highly under-resourced scenarios. The results of these experiments show that adding even relatively small amounts of annotated data from a target domain delivers the highest accuracy on the target domain compared to other domain adaptation methods. To determine the underlying causes of the accuracy deterioration, a forward stepwise linear regression modelling experiment shows that a combination of lexical and syntactic divergence can account for a significant amount of the deterioration, and are good predictors of the expected deterioration when applying POS tagging to a new domain.

1 Introduction

Part-of-speech (POS) tagging has been a fundamental natural language processing (NLP) task for more than 40 years. Based on standard benchmarks, it is now considered nearly solved[1] in several well-resourced languages such as English, French, and German (Giesbrecht and Evert, 2009; Manning, 2011). High quality POS tagging is an important building block in more sophisticated NLP systems such as machine translation and sentiment analysis that depend on analysing the structure and meaning of a sentence. In under-resourced settings, building POS taggers is a relatively "cheap" and easy first step for many NLP pipelines used in both research and commercial systems as it allows the discovery of (crude) syntactic relationships and other patterns inherent in a text.

Beyond the utility of POS tagging in NLP, it is also used extensively in linguistic and digital humanities research, such as corpus linguistics, syntactic parsing, and computational literature investigations. However, when models are applied or evaluated on data from different domains, there is generally a substantial loss in performance (Derczynski et al., 2013; Schnabel and Schütze, 2013; Van Asch and Daelemans, 2010). This could be the result of differences in vocabulary, topics, and writing style between training and testing data (Manning, 2011). The accuracy loss can be especially jarring to researchers in related fields, who do not necessarily understand the impact that the domain change can have on the quality of the generated tags.

Although the degradation in quality has been well established in various high-resourced languages (Ferraro et al., 2013; Plank et al., 2014; Schnabel and Schütze, 2014), there has been comparatively little research on how much accuracy is lost when switching to different domains in settings where even the baseline taggers are trained on relatively little data.

One of the central challenges for under-resourced languages is limited data availability, often originating from a single domain only. With little data available to develop technologies such as POS taggers, it is crucial to understand what the quality of an existing POS tagger is when applied to data from a different domain. Furthermore, it is important to establish how much extra data from a different domain is necessary to improve POS tagging results in the new domain. This understanding will allow other researchers in different fields to better manage their expectations when using tech-

[1] See https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art).

https://doi.org/10.3384/nejlt.2000-1533.2026.6066

nologies such as POS taggers in these under-resourced settings.

With this in mind, the aim of this paper is to address the following questions related to POS tagging and domain change in a low-resource environment:

1. How much POS tagging accuracy is lost when changing domains?
2. To what degree does adding relatively small amounts of data from a specific domain improve tagging for that domain?
3. Does adding data from other domains improve tagging in a previously unseen domain?
4. What are the underlying causes of the degradation in accuracy when switching domain?

We address the questions above by training deep neural POS taggers under various experimental conditions for two under-resourced South African languages: isiZulu and Sesotho sa Leboa (also known as Sepedi or Northern Sotho). These agglutinative Niger-Congo-B languages are distinctive in their respective orthography, conjunctive and disjunctive, which could add additional insight into how domain change impacts different orthographies. Both languages have a relatively small amount of available POS-annotated data in the government domain from previous projects (Eiselen and Puttkammer, 2014; Gaustad and Puttkammer, 2022). As part of the current research, small sets of additional POS-annotated data were developed in five different domains, viz. exam texts for grade 12 South African learners, magazines, newspapers, novels, and PhD theses.

The following section provides a brief overview of work related to domain adaptation and POS tagging, both in the South African context and internationally. Section 3 gives a detailed description of the various data sets, POS taggers, and experiments that are used to address the research questions of this work. The results from the proposed experiments show substantial degradation in performance for both languages when applying these POS taggers in new domains. This degradation can be mitigated by adding relatively small amounts of data from the new domain, as discussed in Section 4. In Section 5 we use the forward stepwise linear regression procedure to identify the main contributors for losses in accuracy and show that a combination of vocabulary divergence and POS distribution account for most of the degradation in POS tagging accuracy.

2 Previous Work

2.1 Domain Adaptation

The problem with the loss of NLP model accuracy when changing domains is a well-established phenomenon. When applying a technology, such as POS tagging, that was trained on a particular domain to another domain, there is usually a divergence in the distributional properties of the language or target classes in the respective domains. There has been a substantial amount of research on different strategies to mitigate this underlying distributional change, not just for POS tagging, but also various other NLP technologies, such as machine translation (Chu et al., 2017), named entity recognition (Jia et al., 2019), and entity extraction (Grön et al., 2018). Although most of the research focuses on English domain adaptation, some work has been done for German (März et al., 2019), Dutch (Grön et al., 2018), and Chinese (Jiang et al., 2021).

For POS tagging, most of the previous research has focussed on adaptation of models trained on newswire data, such as the Wall Street Journal sections of the Penn Treebank (Taylor et al., 2003), to either the biomedical domain (Blitzer et al., 2006; Ferraro et al., 2013; Grön et al., 2018; Liu et al., 2007; Miller et al., 2006, 2007) or social media content (Andreevskaia and Bergler, 2008; März et al., 2019; Plank et al., 2014; Schnabel, 2013; Schnabel and Schütze, 2014). In both these instances, the taggers often show substantially degraded performance when applied to the new domain, generally between 5% and 10% absolute loss in accuracy, but up to 30% in some cases (Grön et al., 2018). There are two over-arching approaches in previous research to improving the tagging in the new domain, namely annotating data in the new domain or modelling distributional divergence.

Although annotating data in the new domain is the most obvious method to improve tagging in that domain, it is also prohibitively expensive, especially when considering the fact that new annotated data is required for every new domain to which the technology will be applied. To limit the amount of data requiring human annotation, several researchers have investigated strategies for selecting sentences in the target domain which are the most likely to improve the quality of the tagger.

The first methodology extends the training data by automatically adding sentences that have been tagged by multiple baseline taggers, and through different voting or ensemble strategies, selecting those sentences that show agreement, and therefore are likely to have been tagged correctly (Andreevskaia and Bergler, 2008; Jiang et al., 2021; Kübler and Baucom, 2011). Although the taggers trained on the extended data sets do show improvements on the target domain, the improvements

are relatively limited, typically only showing a 1-2% absolute improvement. Another shortcoming of this approach, which is especially relevant to isiZulu and Sesotho sa Leboa, is the fact that these methods typically require large amounts of unannotated data in the target domain, typically several hundred thousand tokens. Unfortunately, such data simply is not available for the languages and domains under consideration here, and therefore applying such automatic selection methods does not seem viable.

Unlike the automatic methods described above, both Liu et al. (2007) and Grön et al. (2018) explicitly show that adding even relatively small amounts of human annotated data to the baseline training data, significantly improves the quality of the classifier in the target domain. Liu et al. (2007) select sentences containing high frequency out-of-vocabulary (OOV) words, i.e. words that are not present in the classifier's original training data, and only annotate those sentences. They show that using this sentence selection strategy improves target domain accuracy from 79% to 92.7%, while a randomly selected set which is three times as large only improves the classifier accuracy to 93.9%. Grön et al. (2018) use a much larger random sample of clinical and biomedical texts, consisting of 27,590 tokens from five subdomains, split into train (67%) and test (33%) sets. By including this random sample of annotated text, they report an average absolute improvement on these specialised domains of approximately 19.90%[2] (from 66.48% to 86.39%).

In contrast to only adding annotated data, the other stream of research has focused on addressing the distributional differences between the source and target domains (Blitzer et al., 2006; Van Asch and Daelemans, 2010). As Van Asch and Daelemans (2010) show, there appears to be a linear relationship between the distributional lexical divergence of the source and target data, and the quality of classifiers trained on the source domain and applied to the target domain. Most of this work focussed on either extending the lexicon associated with the tagger or adding distributional features. Lexicon extension usually involves including OOV words, and more specifically including unambiguous specialist terminology, or providing ambiguity classes for OOV words (Ferraro et al., 2013; Kübler and Baucom, 2011; Miller et al., 2006, 2007). All of these approaches show improvements in tagging the target domain similar to annotating relatively large amounts of data (>50,000 words) in the target domain. This does however still require human annotators to review the lexicon entries, and assign either unambiguous tags, or adjust the most likely tag based on the target domain. Alternative strategies investigated adding additional distributional features to the target domain data (Blitzer et al., 2006; Schnabel, 2013; Schnabel and Schütze, 2013, 2014) by establishing features that behave similarly in both domains and show improvements over only annotating a small amount of data in the target domain. Since these methods are based on statistical distribution features, they require large amounts of unannotated data in both the source and target domain, but do not require any human intervention.

With the advent of deep learning models, much of the previous domain adaptation work has been superseded, since features have been replaced by various flavours of word embeddings. These embeddings do not rely on annotated data and usually have significantly larger lexicons also containing representations for words that do not occur in the classifier training data. The embeddings do still require large amounts of unlabelled text data, which remains a problem for most of the South African languages (de Wet et al., 2023).

2.2 POS Tagging for isiZulu and Sesotho sa Leboa

There have been several efforts to develop POS taggers and annotated data for isiZulu and Sesotho sa Leboa over the last two decades. Unfortunately, no common data sets or tag sets have emerged which makes for a slightly eclectic landscape. Important early discussions revolved around tag set creation as well as the choice of approach and sequencing of tasks best suited for Bantu languages with different orthographies without, however, reporting tagging results (Taljard and Bosch, 2006), or only reporting results on polysemous function words (Faaß et al., 2009). Other early work, describes the use of POS tagging in the context of lexicography and dictionary development (de Schryver and De Pauw, 2007) using a manually annotated corpus of 10,000 words for Sesotho sa Leboa, but without tag information. De Pauw et al. (2012) present a multi-lingual study on Swahili, isiZulu, Sesotho sa Leboa and Cilubà using a Memory-based Tagger reporting promising results. After the release of the original National Centre for Human Language Technology (NCHLT) data set (Eiselen and Puttkammer, 2014), a number of studies reported tagging results on this data with various approaches: using HunPoS, an opensource Hidden Markov Model tagger (Eiselen and Puttkammer, 2014), applying a deep neural network (DNN) (Loubser and Puttkammer, 2020) and a conditional random field (CRF) tagger (du Toit and Puttkammer, 2021). Most recently, Dione et al. (2023) investigated the use of a CRF tagger with pretrained Language Models on 20 languages, including isiZulu but not Sesotho sa Leboa, on the news domain.

Several limitations apply to the work that has been

[2] The authors report a 29.69% average improvement, but this figure is an average of the precision for 5 subdomains, rather than a weighted average dependent on the size of each subdomain.

done previously, especially when considering domain switching. Firstly, a number of the corpora developed in the early years of research on African language tagging are not available any more or are proprietary (De Pauw et al., 2012; Faaß et al., 2009; de Schryver and De Pauw, 2007; de Schryver and Prinsloo, 2000). This entails that it is not possible to build on this work and use the described data as part of current experiments, either within the same domain or in cross-domain experiments. Secondly, the definition of the tag sets differs between the respective studies, with some only considering the major parts-of-speech, as defined in the Universal POS tag set (Dione et al., 2023), others excluding class information from the tag set (De Pauw et al., 2012; de Schryver and De Pauw, 2007), while still others use detailed tag sets, with up to 190 tags (Eiselen and Puttkammer, 2014; Gaustad and Puttkammer, 2022; Loubser and Puttkammer, 2020; du Toit and Puttkammer, 2021). Lastly, the POS-annotated data sets that are available exclusively focus on a single domain, either data compiled from government publications (Eiselen and Puttkammer, 2014; Gaustad and Puttkammer, 2022), or data from Newspaper publications (Dione et al., 2023).

Most recently, data sets for isiZulu and Sesotho sa Leboa spanning multiple domains were released (Gaustad, 2024a,b). These data sets allow researchers to determine how different genres impact the quality of POS tagging, while also allowing multiple domain-specific taggers or a more general tagger (not limited to one domain) to be developed. We will now describe our experimental design to investigate the impact of domain on POS tagging accuracy as well as the different domain data sets and tag sets used.

3 Experimental Design

Although there has been a substantial amount of research in domain adaptation, and specifically for POS tagging, prior research focused on one of two domain adaptation strategies. Most previous research in this area has started with a very large source data set that originates from newspapers, e.g. the Wall Street Journal, and then adapted to either the biomedical domain or online social interactions, such as Twitter, blogs, and online comments. In contrast, for many under-resourced languages, the main source of data is typically not newspaper data, but rather content created by local administrations and national governments. Furthermore, the source data sets are also usually relatively limited in size. Both of these scenarios are relevant to the South African language contexts, where the majority of freely available language data originates from government publications, which typically have a different function than those found in various other domains, such as newspapers, novels, or scientific works. The current section therefore describes the various data sets developed and used in a group of experiments to determine how well (or poorly) POS tagging performs on different domains, and what the impact of domain-specific data is on the quality of the respective taggers. For this purpose, we consider a source annotated data set from the government domain, and experiment with annotated data in five different target domains, namely grade 12 high school exam texts, magazines, newspapers, novels, and PhD dissertations.

To make the results more insightful, and hopefully more broadly applicable, we consider two South African languages, Sesotho sa Leboa and isiZulu. Both languages are considered agglutinative Niger-Congo-B languages with (among other linguistic characteristics) a high number of noun classes (Doke, 1950; van der Velde et al., 2022). However, these two languages have very distinct orthographic properties: isiZulu is written conjunctively, where all morphemes of the linguistic word are combined in a single orthographic word (token), while Sesotho sa Leboa is written disjunctively, where the linguistic word (can) consist of multiple orthographic words (Louwrens and Poulos, 2006). See the example here below for an illustration (taken from Prinsloo and de Schryver (2002)).

Sesotho	*ke a mo rata*			
sa Leboa	ke	a	mo	rata
	I	[pres]	him/her	love
	'I love him/her'			
isiZulu	*ngiyamthanda*			
	ngi-	-ya-	-m-	-thanda
	I	[pres]	him/her	love
	'I love him/her'			

The implication of the different orthographies is that the vocabulary and type-token ratio (TTR) for isiZulu is substantially higher than for Sesotho sa Leboa. This also means that the number of unknown (or OOV) words is generally much higher for isiZulu than for Sesotho sa Leboa. The opposite is true for lexical ambiguity, which is low for isiZulu (due to the conjunctive writing style) and much higher for Sesotho sa Leboa (with many separately written highly ambiguous function words). Consequently, taggers trained for isiZulu in general perform substantially worse than those for Sesotho sa Leboa, regardless of domain (Eiselen and Gaustad, 2023; Eiselen and Puttkammer, 2014; Loubser and Puttkammer, 2020; du Toit and Puttkammer, 2021)

In the following subsections we provide details on the respective POS-annotated corpora and tag sets. The final subsection provides a description of the experimental design given the respective languages, data and tag sets, as well as POS architectures.

3.1 Data

3.1.1 Source: Government Domain

For Sesotho sa Leboa, the baseline corpus is the NCHLT data (Eiselen and Puttkammer, 2014) with approximately 65,000 tokens of training data and a test set of circa 7,100 tokens. The data was originally crawled from South African government websites and contains a combination of legislative text, pamphlets, forms, school learner material, and informational content.

For isiZulu, we used the Linguistically enriched corpora for conjunctively written South African languages data set (Gaustad and Puttkammer, 2022) as a baseline corpus. Its contents also originate from government websites with data types very similar to the NCHLT data described above. The isiZulu corpus contains approximately 44,000 tokens in the training set, and 5,000 tokens in the test set.

3.1.2 Target: Academic - National Senior Certificate Examinations (CAPS)

Our first example of an academic target corpus is a compilation of National Senior Certificate examinations (commonly referred to as "matric exams") from South Africa. All high school students in South Africa need to take at least two of the twelve official South African languages[3] as subjects, with at least one at "home language" level (defined as a language in which the learner has mastered reading, writing and interpersonal communication). The final test material of previous years is made available through the website of the Department of Basic Education[4] and these exams typically contain a portion of text for the students to answer reading comprehension questions as well as summary writing texts. For this corpus, we have downloaded the exams for isiZulu and Sesotho sa Leboa as Home Language from 2018-2019 and made a random selection of the content after processing and cleaning the data. The final corpora for this type of academic text amount to 4,000 tokens for isiZulu and 7,000 tokens for Sepedi.

3.1.3 Target: Non-Academic - Magazines

The choice of generic magazines available for isiZulu and Sesotho sa Leboa is unfortunately rather limited. The data in these target corpora issue from Pula/Imvula, a magazine focusing on farming. The aim of the magazine is to educate developing farmers and help them become sustainable commercial farmers. The magazine is distributed on a monthly basis and is currently only available in English. Until September 2024, however, it was published in five languages (English, isiXhosa, isiZulu, Sesotho, and Setswana), and until 2019 also used to include Afrikaans and Sesotho sa Leboa (Gaustad et al., 2025). From previously acquired editions, random files were selected and sentences extracted as a basis for each corpus. To ensure the quality of our corpora, both sources were run through a spell checker for the relevant language and only sentences which contain at least 90% of correctly spelled words were kept. In addition, all incomplete as well as duplicate sentences were removed. The final Magazine corpus contains 4,000 tokens for isiZulu and 6,000 tokens for Sepedi.

3.1.4 Target: Non-Academic - News

The next target corpus under consideration consists of curated news articles. The language used in these texts is informative by nature and intended for a larger, more general audience, requiring it to be more accessible and less formal than most government texts. Non-academic texts prioritise readability and engagement, giving way to less structured language that is crafted for emphasis, clarity, and even sensationalism (Bhatia, 1993). For isiZulu, we used a portion of the Leipzig corpus[5] (Goldhahn et al., 2012) originating from the newspaper *Isolezwe*, whereas the Sesotho sa Leboa data was acquired from a local newspaper[6]. Similar to the Magazine data, we applied a spellchecking control and removed duplicates and incomplete sentences. In a final step, the data was sorted randomly to ensure that the original articles and texts cannot be reproduced.[7] This resulted in an isiZulu News corpus of roughly 7,000 tokens and a Sepedi News corpus of about 9,500 tokens.

3.1.5 Target: Fiction - Novels

Another target corpus used in our experimental setup is a collection of fictional novels. These texts are predominately written to convey a story, emotions, and experiences. The language used varies widely, from highly literary and descriptive to straightforward and conversational, depending on the style of writing and the type of novel. All the data from novels have been acquired by the South African Centre for Digital Language Resources (SADiLaR)[8] from two publishers, Oxford University Press and Shuter and Shooter respectively, and comprise recently published novels (2007 onward). We selected three novels for each language, used the same cleaning and selection process as for the other target data, as well as the randomisation step. Unfortunately,

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

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

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

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

[7] This step was needed to comply with the usage rights of the data.

[8] https://sadilar.org/en/

Language	Domain	Token count	Type count	TTR/ 1000	Sentence count	Tokens/ sentence
ZU	Gov Train	44,142	13,355	0.70	2,713	16.27
	Gov Test	4,955	2,702	0.70	297	16.68
	CAPS	4,159	2,383	0.69	464	8.96
	Magazines	4,108	2,014	0.63	342	12.01
	News	6,860	3,626	0.69	598	11.47
	Novels	7,317	3,699	0.65	998	7.33
	PhDs	6,514	3,218	0.65	696	9.36
NSO	Gov Train	65,865	5,904	0.33	2,579	25.54
	Gov Test	7,153	1,485	0.36	324	22.08
	CAPS	7,265	1,552	0.36	523	13.89
	Magazines	6,012	1,201	0.37	305	19.71
	News	9,412	1,741	0.37	395	23.83
	Novels	7,655	1,538	0.35	498	15.37
	PhDs	7,960	1,257	0.32	348	22.87

Table 1: Summary of the respective data sources in isiZulu (ZU) and Sesotho sa Leboa (NSO)

we do not have further information on the content or target audience of these novels. The isiZulu novel corpus contains 7,500 tokens and the Sepedi novel corpus 7,700 tokens.

3.1.6 Target: Academic - PhD Theses

As the last target corpus and second example of academic writing, we have sourced PhD theses from the University of Pretoria's Electronic Thesis and Dissertations repository[9] for Sesotho sa Leboa and from the University of KwaZulu Natal's Research Space[10] for isiZulu. From the available theses in the relevant languages, a small number were selected and sentences compiled into a target corpus. Incidentally, all the theses have been submitted to the Faculties of Humanities in either language or literature studies. After applying the same processing and quality control steps as described above, we obtained a corpus of academic writing from PhD theses of 6,500 tokens for isiZulu and 8,000 tokens for Sepedi.

3.1.7 Overview and Comparison of Target Corpora

Table 1 provides a detailed summary of the various data sets used in subsequent experiments. Most notable in the data is the impact of the different orthographies used: The normalised type-token ratios (TTR)[11] for corpora of similar size show substantially different TTR values, with the conjunctively written isiZulu data having nearly double the TTR of the disjunctively written Sesotho sa Leboa. Furthermore, sentences are substantially shorter for isiZulu, averaging 12.78 tokens per sentence, while Sesotho sa Leboa averages 22.39 tokens per sentence.

When comparing the different data sets for Sesotho sa Leboa, we see that CAPS has the least tokens per sentence, followed by Novels and Magazines. The Government domain training data contains the longest sentences, pointing to more formal language use and possibly "legalese". The TTR for the Government domain training data is the second lowest, indicating a relatively large amount of repetition, especially compared to Magazines and Newspapers, where the content is more diverse. Surprisingly, the PhD theses have the lowest TTR which may be a consequence of the fact that a thesis is typically focussed on a single topic.

For isiZulu novels contain the least tokens per sentence, followed by CAPS and PhD theses. Again, as for Sesotho sa Leboa, the Government data has the longest sentences indicating very formal (and likely difficult) language using lots of relatives[12], etc. Unlike the case of Sesotho sa Leboa, the Government domain data has the highest normalised TTR, while Magazines have the lowest TTR. We will get back to some of these statistics during the discussion of the evaluation results.

3.2 POS Tag Sets

Typically, for English POS tagging, the Penn Treebank tag set with 36 POS tags and 12 other tags (for punctuation and currency symbols) is used (Marcus et al., 1993; Taylor et al., 2003). Recently, the Universal POS (UPOS) tag set (Petrov et al., 2012) also gained popularity due to its language agnostic nature. This tag set contains 17 tags: 6 for open classes (nouns, verbs, adjective, etc.), 8 for closed classes (e.g. pronouns, conjunctions) and 3 for miscellaneous items (such as punctua-

[9] https://repository.up.ac.za/
[10] https://researchspace.ukzn.ac.za/
[11] Type-token ratio is normalised per 1,000 tokens.
[12] Open class POS, similar to adjectives in English.

ZU Main POS categories		Counts	NSO Main POS categories		Counts
ABBR	Abbreviation	1	ABBR	Abbreviation	1
ADJ	Adjective	16	ADJ	Adjective	19
ADV	Adverb	2	ADV	Adverb	1
CONJ	Conjunction	1	CONJ	Conjunction	1
CDEM	Demonstrative class	14			
COP	Copulative	1			
			DEM	Demonstrative pronoun	15
			DEMCOP	Demonstrative copulative	17
FOR	Foreign	1	FOR	Foreign	1
IDEO	Ideophone	1	IDEO	Ideophone	1
INT	Interjection	1	INT	Interjection	1
INTER	Interrogative	1	INTER	Interrogative	1
			MORPH	Tense, aspect, negative marker	1
N	Noun	17	N	Noun	19
NPP	Place and brand name	1	NPP	Place and brand name	1
			NPRE	Nominal prefix class 15	1
NUM	Numerative	1	NUM	Numerative	1
			OC	Object concord	18
			PART	Particle	1
POSS	Possessive class	16	POSSC	Possessive concord	18
PROEMP	Emphatic pronoun	15	PROEMP	Emphatic pronoun	18
			PROPOSS	Possessive pronoun	18
PROQUANT	Quantitative pronoun	14	PROQUANT	Quantitative pronoun	17
PUNC	Punctuation	1	PUNC	Punctuation	1
REL	Relative	1			
			SC	Subject concord	20
V	Verb	1	V	Verb	1
VAUX	Auxiliary verb	1	VAUX	Auxiliary verb	1
			VCOP	Copulative verb	1

Table 2: Overview of main POS categories and total fine-grained tags for isiZulu (ZU) and Sesotho sa Leboa (NSO).

tion and symbols). In the context of Bantu languages, a comprehensive POS tag set accurately representing the linguistic features of the language in question warrants the use of more categories to account for phenomena such as concords and extra pronoun classes. In addition, the grammar of Bantu languages includes many noun classes (between 12 and 20) that need to be incorporated into the tag set (see Taljard and Bosch (2006) for a discussion on POS tagging for disjunctively vs. conjunctively written Bantu languages). For a few Bantu languages, attempts have been made to use the UPOS tag set (Dione et al., 2023; Gaustad et al., 2024), but it has proven quite a challenge to find the correct corresponding UPOS tags for some of the Bantu-specific categories.

Table 2 shows the main categories of POS tags in our tag set, as well as the number of different POS tags for each main category. For the main class of nouns N, for instance, there are 17 noun POS tags ranging from N00 (lexicalized nouns) to N15 for isiZulu and 18 noun POS tags including all possible noun classes (e.g. N01, NGA) for Sesotho sa Leboa. If all classes are collapsed, the tag set for isiZulu has 20 main tags whereas the one for Sesotho sa Leboa has 26. When including all classes, both tag sets are substantially bigger: a total of 107 tags for isiZulu and 195 tags total for Sesotho sa Leboa. One reason for the larger POS tag set for Sesotho sa Leboa is the disjunctive orthography: tokens that are written agglutinatively in isiZulu (and therefore not tagged separately) need their own POS tag in Sesotho sa Leboa, e.g. PART for particles or MORPH for tense, aspect and negative markers.

3.3 Experimental Setup

In order to determine the impact different domains have on the quality of the POS tagging task, we conduct 22 different experiments for each language, both to establish the effect of purely switching domain, but also how adding data from the target domain—or merely data from other domains—influences the tagging results in the target domain.

Given that we only want to determine the impact of data, rather than different algorithms, we use a single POS tagger architecture and constant parameters, to ensure that we are only measuring source and target data impact. The FLAIR framework (Akbik et al., 2018, 2019) is a Python library that facilitates sequence labelling, text classification, and language modelling,

based on the bidirectional long-short term memory (Bi-LSTM) algorithm. The framework initially achieved state-of-the-art results for various sequence labelling tasks, including POS tagging (Akbik et al., 2018, 2019). The main reasons for using this framework are that it is easily extensible to a variety of languages and tasks, while also supporting a large number of embedding and language modelling implementations.

For our POS tagging task, BiLSTM-CRF taggers are trained with a hidden size of 256, learning rate of 0.1, mini batch size of 32 for a maximum of 80 epochs. Based on previously reported results (Eiselen and Gaustad, 2023), the NCHLT Sesotho sa Leboa fastText-CBoW embeddings[13] and NCHLT isiZulu FLAIR backward embeddings[14] are used. Embeddings are not fine-tuned during training.

As a baseline for all the experiments, we train a POS tagger on each of the previously released POS-annotated data (Eiselen and Puttkammer, 2014; Gaustad and Puttkammer, 2022) consisting of circa 50,000 tokens per language, primarily from government website sources, and evaluate on a separate, distinct test set from the same source domain. The accuracy of this in-domain tagger is then compared to experimental taggers with different target domain data sets. Our experimental setup aims to provide information for answering the respective research questions (see Section1) by determining the impact of domain switching and addition of target domain data on POS tagging quality. The set of experiments performed for the two languages is as follows:

1. Apply the two baseline taggers to five different target domains, viz. CAPS, Magazines, Newspapers, Novels, and PhD theses, to determine the impact a domain switch has on POS quality (Question 1).

2. For each target domain select 10% of the annotated sentences with high OOV rates and combine them with the source domain training data to train an updated tagger. Test the tagger on the remaining 90% of the target domain data (Question 2).[15]

3. For each target domain perform 10-fold cross validation by combining the original government source domain data with 90% of target domain data and evaluating on the remaining 10% of the target domain data for each fold (Question 2).

4. Combine source domain training data with data from other domains, excluding the target domain, and evaluate on the target domain (Question 3).

The justification for each of these four experiments is as follows.

1. The first experiment estimates the impact that a change of domain has on the quality of the tagger when applied to five distinct target domains.

2. In the second experiment, the impact of a selection strategy targeting OOV words, in line with previous work (Liu et al., 2007) is done. Since embeddings, used as features for the tagger, include a much larger vocabulary than the source training data, this should in theory limit the impact of OOV words on tagger accuracy. This is true for both the embedding types used in these experiments, as they are able to provide representations for unseen words based on either character n-grams (fastText) or character sequences (FLAIR). There is however scant evidence of this in low-resource environments, and verifying this assumption in this context will also provide additional evidence for the underlying causes for the loss of tagger accuracy.

3. The third set of experiments provides an indication of how relatively small amounts of randomly selected data from the target domain impact the quality of the POS tagger for the target domain.

4. The last set of experiments aims to determine the impact of additional out-of-domain data on tagging accuracy for the target domain.

The results for each of these experiments for both isiZulu and Sesotho sa Leboa are provided in the next section, followed by a discussion of the possible underlying reasons for the degradation in quality.

4 Results and Discussion

For the first experiment, a baseline tagger is trained and evaluated on the source domain, viz. Government data, and then applied as-is to each of the respective target domains to determine the extent of tagger accuracy change in each domain. The results from this experiment are presented in the columns titled 'Exp. 1: Base' in Table 3. These results show that the accuracy of the Baseline taggers for isiZulu and Sesotho sa Leboa is 0.8864 and 0.9646 respectively. The difference in accuracy is primarily due to the distinct orthographies of the two languages, as explained in section 3.1. The performance of the Baseline tagger on all the target domains

[13] https://hdl.handle.net/20.500.12185/588
[14] https://hdl.handle.net/20.500.12185/615
[15] Selecting the 10% consists of finding the 20% of sentences with the highest OOV rates, and randomly selecting half of the sentences for training, while keeping the remaining half in the test set. This is done to prevent reducing the complexity of the test set by including all the sentences with the highest OOV rate in the training, and thereby "simplifying" the test set for the target domain.

	ZU				NSO			
	Exp. 1 Base	Exp. 2 10%	Exp. 3 90%	Exp. 4 All other	Exp. 1 Base	Exp. 2 10%	Exp. 3 90%	Exp. 4 All other
Gov	.8864			.8902	.9646			.9676
CAPS	.8454	.8474	.8698	.8675	.8959	.9152	.9439	.9319
Magazines	.8673	.8804	.9118	.8880	.9479	.9449	.9606	.9649
News	.8552	.8675	.8999	.8759	.9172	.9264	.9469	.9446
Novels	.7729	.8269	.8636	.8255	.8576	.9106	.9502	.9206
PhDs	.8416	.8717	.9160	.8781	.9303	.9431	.9629	.9603

Table 3: Accuracy results of Experiments 1-4 for isiZulu (ZU) and Sesotho sa Leboa (NSO).

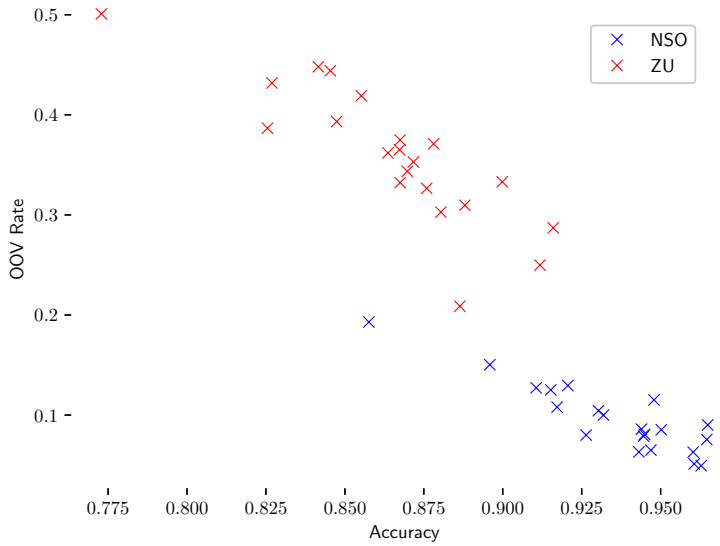

Figure 1: OOV rates and POS tagger accuracy of Experiments 1-4 for isiZulu (ZU) and Sesotho sa Leboa (NSO) over multiple domains.

in both languages is worse than on the Government test data: We observe a loss in accuracy of between 1.67% and 11.35%. Interestingly, for both languages, Novels appear to be by far the most difficult to accurately tag (11.35% and 10.7% worse), while Magazines see the lowest change in performance (a drop of only 1.91% and 1.67%).

For the second experiment, a small amount of target domain data with high OOV rates are added to the source domain training data, with results presented in column 'Exp. 2: 10%' in Table 3. The addition of a very small amount of target domain data containing a large number of OOV words does reduce the loss in accuracy of the taggers, e.g. halving the error rate in the case of Novels, but not consistently across all of the target domains. The CAPS domain, for instance, only shows very minor improvements for both languages. In general, this strategy appears to be more productive for isiZulu than for Sesotho sa Leboa. This is likely due to the fact that isiZulu has a much higher OOV rate to start with (see Figure 1), and that specifically targeting the OOV words therefore has a more substantial influence on the results. It is also clear from these results that the OOV rate, although correlated to accuracy, is not the only factor influencing the quality of the tagger. We discuss this in more detail when addressing the underlying causes for the differences in tagger accuracy in Section 5.

For the 10-fold cross validation experiments, where 90% of the target domain training data is added to the Source domain (presented in column 'Exp. 3: 90%' in Table 3), much more substantial improvements are observed for both languages and all domains. The loss in accuracy is reduced by more than 50% in all cases across both languages when compared to the baseline experiments. In some instances, the domain-specific taggers even outperform the baseline tagger, as is the case for isiZulu Magazines, News, and PhD target domains. In

these three cases, the inclusion of both additional instances of proper names, as well as field-specific terminology in the case of PhDs, seems to be the major contributing factor to the improvements.

For the fourth experiment, data from all domains except the target domain is added to the source data, with the aim of estimating the impact of just adding additional data to the original training set. In all cases, tagging accuracy is still improved on the target domain over the baseline tagger, as can be seen in the results in column 'Exp. 4: All other' in Table 3. However, the improvements are generally not as substantial as adding target domain data, even though the amount of training data is nearly double that of the baseline tagger. The implication of this is that tagging results on any target domain not covered by the source data will still show substantial reductions in accuracy, even though the general quality of the tagger is improved, including on the original source domain (cf. Gov results in Table 3).

Overall, we can say that for both isiZulu and Sesotho sa Leboa the domain target data containing Novels and CAPS data are the most difficult to tag correctly. Across all experiments, results on these two data sets are poorest.

The results of these experiments show that the addition of more substantial amounts of target domain data is essential for improving the tagger quality in the target domain. Adding small amounts of target domain data with high OOV rates or additional out-of-domain data show modest improvements to the quality of the tagger for a target domain. It is however still not obvious why some domains, e.g. Novels, show such large deterioration in quality, as it is clearly not only the OOV rates that cause these drops in accuracy. This makes it difficult for other users to determine what their expectations for tagger quality should be, and how much additional data is required to improve the target domain tagger quality. To address this, we consider four avenues of investigation in the next section to determine the underlying reasons for the poorer performance of the domain-specific taggers: vocabulary, vocabulary distribution, tag distributions, and sentence length.

5 Establishing Underlying Causes of Loss in Accuracy

When language technologies are applied to a new target domain, it is important to know in advance how well the technology will perform in the new domain. Most prior work in language technology evaluation in general typically considers a single factor when making projections on the expected quality of the technology in the new domain. For POS tagging specifically, most research only reports OOV rates (Plank et al., 2014; Schnabel and Schütze, 2014) or accuracy results on OOV words (Ferraro et al., 2013; Plank et al., 2016), while some investigations attempted to find more complex measures of the difference between the source and target, such as vocabulary distributional statistics (Van Asch and Daelemans, 2010). However, initial reviews of the results presented in the previous section indicate that the underlying cause for the loss in accuracy may not be accurately represented by a single measure. As an example, there is a correlation between OOV rates and tagger quality, but there are several cases where higher OOV rates do not align with worse tagger performance as can be seen in Figure 1. The first very obviously contradictory results are between languages: There are instances of the isiZulu taggers with OOV rates above 30% outperforming a Sesotho sa Leboa tagger with less than 20% OOV rate (see for example NSO-Novels vs. ZU-Magazines). To address this seeming discrepancy, we consider several possible factors as the source of loss in accuracy in the tagger results to establish which factors are correlated with the deterioration in quality, and also whether multiple factors combine to predict the deterioration.

The first predictor variable we consider is the Renyi vocabulary divergence, as proposed by Van Asch and Daelemans (2010). This measures the distribution divergence between the relative frequencies of the vocabulary in the source and target domains, where a higher Renyi divergence indicates a greater difference between the source and target vocabulary. Although this is related to the OOV rate (since higher OOV rates will also be reflected in the divergence score), calculating this divergence additionally takes into account the frequency of the vocabulary, which may be a better representation of the difference between the source and target domain.

The second set of additional variables attempts to account for the difference in POS distributions between the various domains. During the annotation process, some of the annotators noted that the assignment of parts-of-speech seemed to be very different from the initial annotation of the Government domain data (Prinsloo, 2024). To calculate this possible discrepancy, we applied the same procedure as Van Asch and Daelemans (2010) by calculating the Renyi divergence between the distributions of POS tags for individual tags, but also for tag combinations, i.e. tag n-grams with $n = \{2, 3\}$.

As noted earlier, there is a substantial difference in sentence length between the respective domains, and this is therefore also included as a possible predictor variable. While the difference in TTR between the source and target domains does not present such stark differences between the various texts, it is included as

a possible variable for the sake of completeness.

As a first step, we performed correlation tests to ensure that the variables under consideration are correlated with the change in accuracy. Using a two-tailed Pearson correlation coefficient, all of the variables except for normalised TTR show statistically significant negative correlations for both isiZulu and Sesotho sa Leboa. The tests are significant with $p < .001$ in all cases, except Sentence Length, where the p-values are $p = .002$ and $p = .008$ for isiZulu and Sesotho sa Leboa respectively. Since TTR is not statistically significantly correlated with the accuracy, it is not considered as predictor variable. Figure 2 and Figure 3 provide visual representations of the various regression models for each of the possible predictor variables for isiZulu and Sesotho sa Leboa respectively.

In order to determine which of the predictor variables contribute most to the difference in tagger results, we employed forward stepwise linear regression modelling to identify relevant predictor variables for the accuracy of the respective POS taggers. The forward stepwise linear regression modelling procedure fits a regression model by considering each predictor variable for addition to the model, and selecting the variable which gives the most statistically significant improvement to the fit of the model. Predictor variables are added until there is no significant improvement to the regression model. Although these models are typically used when there are a large number of predictor variables, it is also useful when identifying those variables that are the best predictor of the response variable, while excluding highly correlated predictor variables.

For Sesotho sa Leboa, the procedure yielded two models, one including only the Renyi vocabulary divergence score ($R^2 = .813$, $p < .001$) and the second including Renyi vocabulary divergence, as well as 2-gram POS divergence ($R^2 = .869$, $p < .001$). Similarly for isiZulu, the forward stepwise linear regression produces two statistically significant models. The first model only includes the 2-gram POS divergence ($R^2 = .793$, $p < .001$) and the second model also includes the OOV rate ($R^2 = .854$, $p < .001$). In both of these cases, approximately 85% of the variation in POS tagger accuracy can be accounted for by just two variables, a combination of a vocabulary divergence measure, and the 2-gram POS divergence.

Although the two languages indicate that different vocabulary divergence measures should be used as ideal models, the fact that Renyi vocabulary divergence and OOV rates are highly correlated (Pearson= .997, $p < .0001$), prompted us to consider an OOV model for Sesotho sa Leboa as well. Again, the stepwise procedure yielded two models, the first only including OOV rate ($R^2 = .807$, $p < .0001$) and the second combining OOV rate with the 2-gram POS divergence ($R^2 = .860$, $p < .001$). These models also account for a significant proportion of the variation in POS tagger quality deterioration, and are very similar to those using the Renyi divergence score. This is somewhat surprising, given the fact that the feature representations for these taggers are not limited to the vocabulary of the training set, but are based on embeddings trained on substantially larger data sets. This contradicts one of our initial expectations that using embeddings and LSTM deep neural networks would mitigate the impact of OOV words on tagger performance, as would be the case in instances where only the vocabulary of the training set is included as features.

Two conclusions can be drawn from the results of the forward stepwise linear regression modelling. Firstly, simple vocabulary divergence scores, either Renyi or OOV, can account for a significant amount of the deterioration in POS tagger accuracy. Secondly, it is important to note that there are other variables at play as well. In this case, the difference in distributions of the POS tags is also a crucial factor to take into account when attempting to predict how much worse a tagger will perform in a new target domain. In our isiZulu experiments for example, the tag divergence is a more important predictor than (only) vocabulary divergence.

Our investigations show that there are multiple variables to consider when applying taggers, and likely other language technologies, to new domains and making predictions about the expected quality of the results. However, a substantial, and in our case significant, amount of the divergence (usually loss) in accuracy can be accounted for by vocabulary divergence measures. Furthermore, the widely used, easily interpretable, and relatively simple OOV rate remains a good predictor of relative quality deterioration when a technology is applied in a different domain.

6 Conclusion and Future Work

In this paper, we investigated domain adaptation for POS tagging of two under-resourced South African languages, isiZulu and Sesotho sa Leboa, by studying its effect on the POS tagging results and how to possibly predict what quality can be expected when applying an existing POS tagger on a new domain. We carried out four systematic experiments applying an existing POS tagger developed on Government domain data to five new domains (exam texts for grade 12 South African learners, magazines, newspapers, novels, and PhD theses) varying the amount and type of additional data. The results of these experiments firstly show that there is typically a substantial loss in accuracy when switching domain, but that these changes in accuracy are related to underlying textual characteristics. Furthermore, the experiments show that adding even small

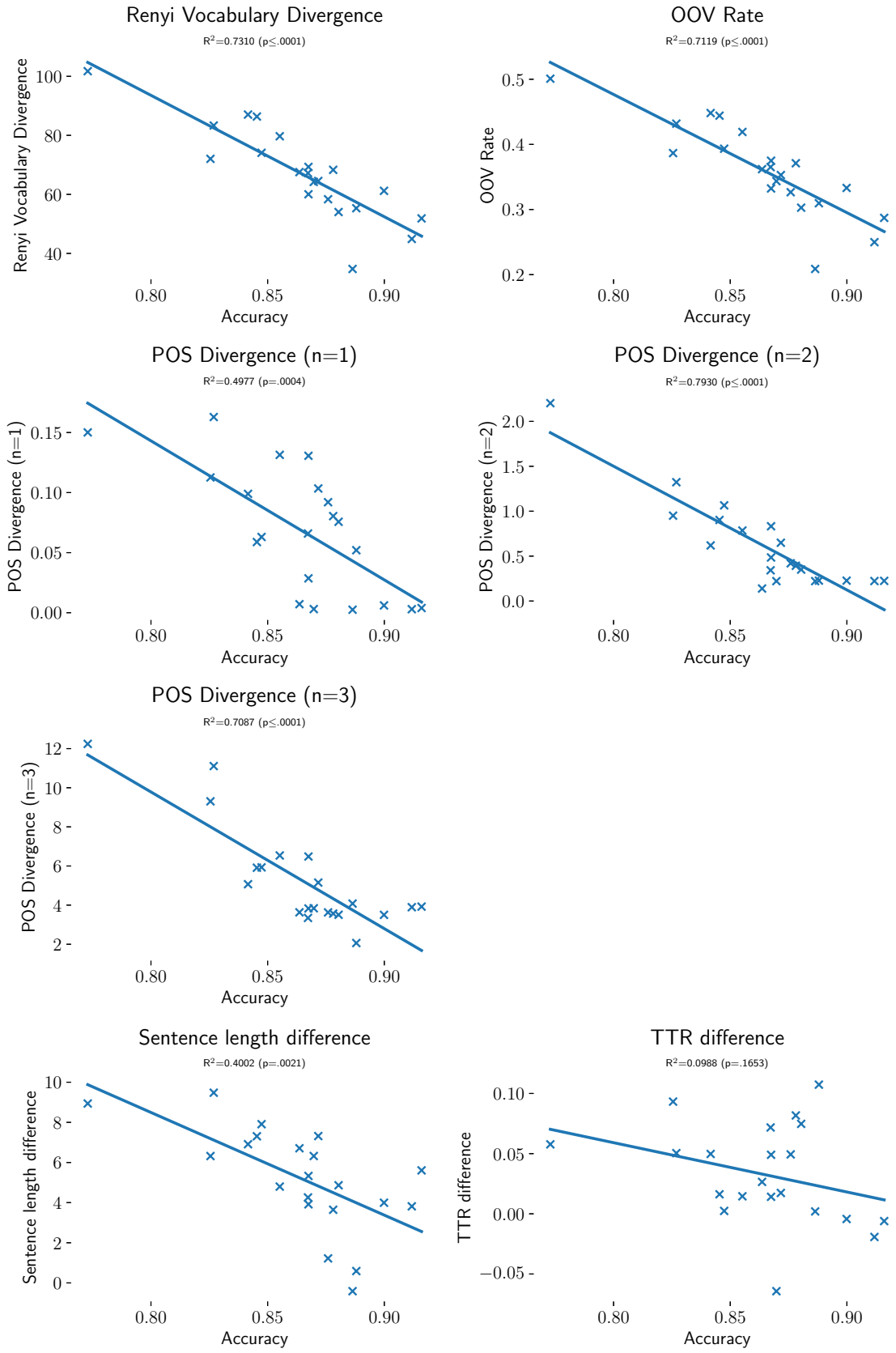

Figure 2: isiZulu regression models for individual predictor variables.

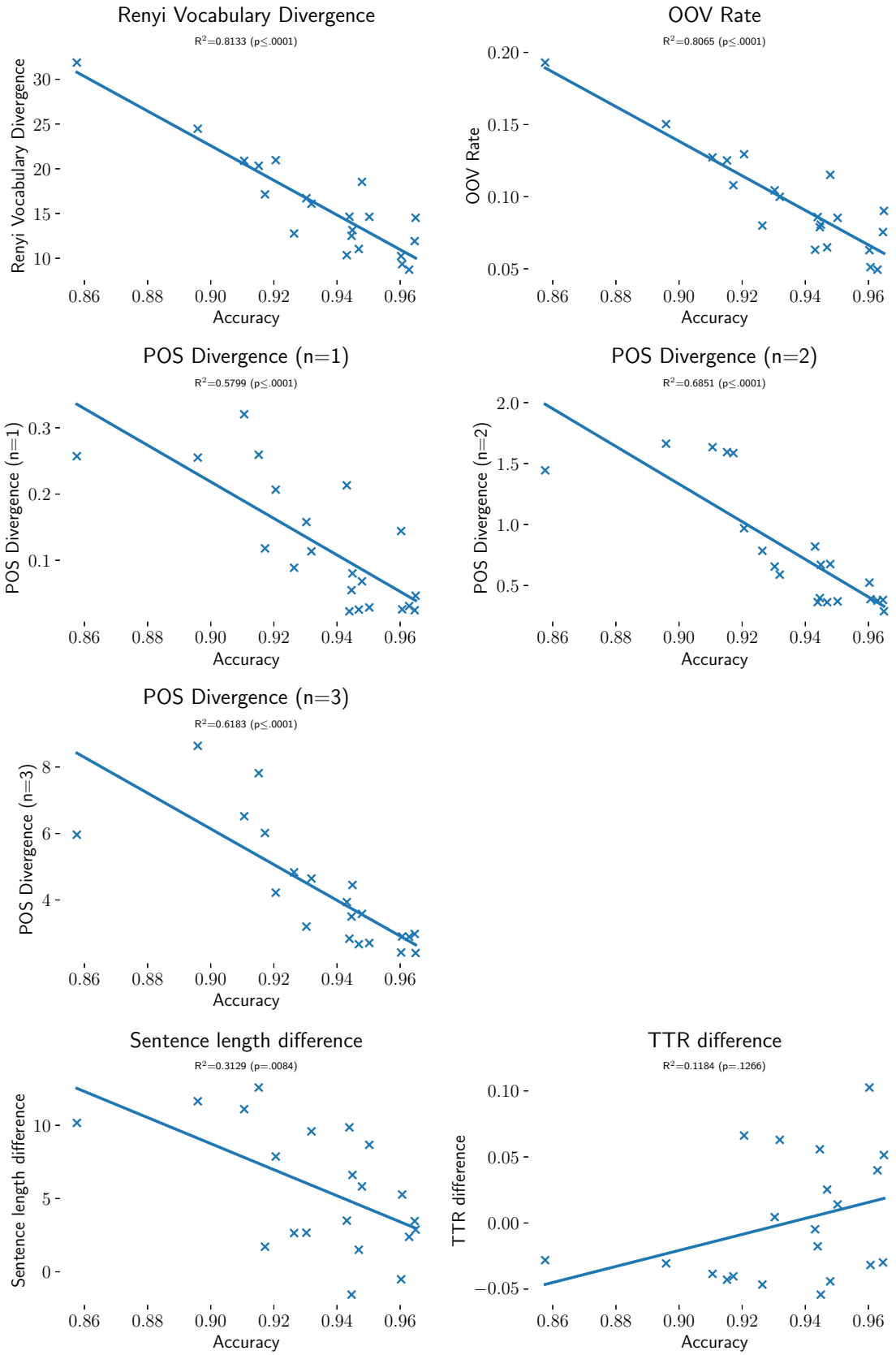

Figure 3: Sesotho sa Leboa regression models for individual predictor variables.

amounts of annotated data from the target domain delivers the largest accuracy improvement on the target domain, whereas adding data with a high OOV rate or data from other new domains but not the target domain ameliorate results to some degree, but not as much.

Comparing different predictor variables in a forward stepwise linear regression model, we find that a vocabulary divergence score such as Renyi or OOV rate accounts for a significant amount of loss in accuracy. However, the POS tag distribution also plays an important part when trying to predict the performance of a POS tagger on a new domain.

In brief: It is (most) effective to annotate even relatively small amounts of in-domain data and it is worthwhile to check the difference in OOV rate between source and target data to manage expectations on tagging results in a new domain.

Acknowledgements

We would like to thank Martin Puttkammer and Jaco du Toit for their input during initial discussions on the experiments described in this paper. We are also grateful to the anonymous reviewers for their feedback and for pointing out a flaw in our initial reporting and comparison of results. All remaining errors are our own.

References

Akbik, Alan, 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, page 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico. ICCL.

Andreevskaia, Alina and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *46th Annual Meeting of the Association for Computational Linguistics*, pages 290–298, Columbus, Ohio. Association for Computational Linguistics.

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

Blitzer, John, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Chu, Chenhui, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

De Pauw, Guy, Gilles-Maurice de Schryver, and Janneke van de Loo. 2012. Resource-light Bantu part-of-speech tagging. In *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL8 - AfLaT2012)*, pages 85–92, Istanbul, Turkey. CLiPS - Computational Linguistics Group University of Antwerp, Belgium, and School of Computing and Informatics, University of Nairobi.

Derczynski, Leon, 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.

Dione, Cheikh M. Bamba, 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.

Doke, Clement M. 1950. Bantu languages, inflexional with a tendency towards agglutination. *African Studies*, 9(1):1–19.

Eiselen, Roald and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Eiselen, Roald 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).

Faaß, Gertrud, Ulrich Heid, Elsabé Taljard, and Danie J. Prinsloo. 2009. Part-of-Speech tagging of Northern Sotho: Disambiguating polysemous function words. In *Proceedings of the EACL 2009 Workshop on Language Technology for African Languages - AfLaT 2009*, pages 38–45, Athens, Greece. Association for Computational Linguistics.

Ferraro, Jeffrey P., Hal Daumé III, Scott L. DuVall, Wendy W. Chapman, Henk Harkema, and Peter J. Haug. 2013. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, 20(5):931–939.

Gaustad, Tanja. 2024a. POS annotated corpus with 5 different genres for Sepedi. https://hdl.handle.net/20.500.12185/670.

Gaustad, Tanja. 2024b. POS annotated corpus with 5 different text types for isiZulu. https://hdl.handle.net/20.500.12185/671.

Gaustad, Tanja, 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.

Gaustad, Tanja, 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).

Gaustad, Tanja 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.

Giesbrecht, Eugenie and Stefan Evert. 2009. Is Part-of-Speech tagging a solved task? An evaluation of POS taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, pages 27–35, San Sebastian, Spain.

Goldhahn, Dirk, 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).

Grön, Leonie, Ann Bertels, and Kris Heylen. 2018. Is training worth the trouble? A PoS tagging experiment with Dutch clinical records. In *14th International Conference on Statistical Analysis of Textual Data*, pages 351–358, Rome, Italy. UniversItalia.

Jia, Chen, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Jiang, Peijie, Dingkun Long, Yueheng Sun, Meishan Zhang, Guangwei Xu, and Pengjun Xie. 2021. A fine-grained domain adaption model for joint word segmentation and POS tagging. In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 3587–3598, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kübler, Sandra and Eric Baucom. 2011. Fast domain adaptation for part of speech tagging for dialogues. In *International Conference Recent Advances in Natural Language Processing 2011*, pages 41–48, Hissar, Bulgaria. Association for Computational Linguistics.

Liu, Kaihong, Wendy Chapman, Rebecca Hwa, and Rebecca S. Crowley. 2007. Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *Journal of the American Medical Informatics Association*, 14(5):641–650.

Loubser, Melinda and Martin J. Puttkammer. 2020. Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11(1):41.

Louwrens, Louis J. and George Poulos. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389–401.

Manning, Chris. 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.

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

März, Luisa, Dietrich Trautmann, and Benjamin Roth. 2019. Domain adaptation for part-of-speech tagging of noisy user-generated text. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3415–3420, Minneapolis, MN, USA. Association for Computational Linguistics.

Miller, John, Michael Bloodgood, Manabu Torii, and K. Vijay-Shanker. 2006. Rapid adaptation of POS tagging for domain specific uses. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 118–119, New York, USA. Association for Computational Linguistics.

Miller, John, Manabu Torii, and K. Vijay-Shanker. 2007. Adaptation of POS tagging for multiple biomedical domains. In *Biological, translational, and clinical language processing*, pages 179–180, Prague, Czech Republic. Association for Computational Linguistics.

Petrov, Slav, 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).

Plank, Barbara, 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.

Plank, Barbara, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Prinsloo, Danie J. 2024. Personal communication.

Prinsloo, Danie J. 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.

Schnabel, Tobias. 2013. *Towards robust cross-domain domain adaptation for part-of-speech tagging*. PhD thesis, Institute for Natural Language Processing, University of Stuttgart.

Schnabel, Tobias 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.

Schnabel, Tobias and Hinrich Schütze. 2014. FLORS: Fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 2:15–26.

de Schryver, Gilles-Maurice and Guy De Pauw. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The case of Tshwanelex. *Lexikos*, 17:226–246.

de Schryver, Gilles-Maurice and Danie J. Prinsloo. 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies*, 18:87–104.

Taljard, Elsabé and Sonja E. Bosch. 2006. A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies*, 15(4):428–442.

Taylor, Ann, 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.

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

Van Asch, Vincent 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.

van der Velde, Mark, Koen Bostoen, Derek Nurse, and Gérard Philippson, editors. 2022. *The Bantu Languages*, 2nd edition. Routledge, London & New York.

de Wet, Febe, Roald Eiselen, Erwin Schillack, and Martin J. Puttkammer. 2023. Investigating the extent and usability of webtext available in South Africa's official languages. In *Southern African Conference for Artificial Intelligence Research*, Artificial Intelligence, pages 120–135, Muldersdrift, South Africa. Springer.