MORPHOLOGICAL PARSING USING DEEP LEARNING

Simbarashe Mawere
mwrsim003@myuct.ac.za
University of Cape Town
Cape Town, Western Cape, South Africa

ABSTRACT

This review focuses on literature on the natural language processing (NLP) task of morphological parsing (or labelling) to devise better ways of performing the task. The main approaches to this task will be training models from scratch and fine-tuning pre-trained language models (PLMs), both of which have not been significantly explored in practice. The task is to be performed in low-resourced Nguni languages. This review, therefore, also looks at models which deal with low-resource languages. Our goal in this review is highlight approaches that have worked for other tasks in the past. These approaches will be training models like Conditional Random Fields (CRFs) and fine-tuning PLMs.

KEYWORDS

machine learning, deep learning, morphological parsing, part-of-speech tagging, computational linguistics, natural language

1 INTRODUCTION AND MOTIVATION

1.1 Background

In Linguistics, there is the branch of Morphology concerned with the "forms of words", particularly how they are constructed from subword units[20]. From this, we get the concept of the morpheme - the smallest unit of linguistic meaning that a word can be split into. Each morpheme has a grammatical role, for instance, the word "largest" can be split into "large" (the stem) and "-est" (a suffix indication of superlative form/degree). This morphological information is useful for tasks like dependency parsing, translation and text filtering as proved by Klemen, Krsnik and Robnik-Sikonja [15].

However, for this morphological information to be obtained automatically there is a need to morphologically process plain text via tasks like segmentation [22], parsing [10] and paradigm learning [11]. Morphological parsing is the process of extracting morphological meaning through labelling individual morphemes in words. The goal of the task is to predict the syntactic role of morphemes which are described by a pre-specified set of tags.

This review will encompass literature surrounding the task of morphological parsing over recent years to evaluate the extent of the efforts to date. Further literature will be reviewed in the context of low-resource language exploration in NLP since the project will focus on very low-resourced Nguni languages. Lastly, along with the insights mentioned above, the intended approaches for study - training models from scratch and fine-tuning pre-trained models - will be analysed in the literature to devise effective strategies and find paths for further exploration.

Morphological parsing is the process of extracting this meaning, and there are tools for this available. However, for low-resource Nguni languages (isiNdebele, isiZulu, isiXhosa, and siSwati), there are none to a few such parsers. In these languages, morphological information is crucial because they are agglutinating, i.e., they carry a lot of sentence meaning in single words; for example, in isiZulu, the word "ngizokushada" is the sentence "I will marry you" in English as expanded into these morphemes: "ngi(I)-zo(will)-ku(you)-shada(marry).". Morphologically parsing these languages would help to extract meaning and use that meaning for a range of other tasks, including translation and text generation.

1.2 Motivations

Low-resource languages are a special part of Natural Language Processing because they represent an unexplored space. Magueresse, Carles and Heetderks [19] explore this issue in great depth in their review of the works surrounding low-resource languages. They identify that such languages are deemed low-resourced ever since NLP had a paradigm shift to take on more statistical-(or machine learning)-based approaches. Chiticariu, Li and Reiss [6] present evidence of this shift, attributing the increase in data collection as a proponent for using machine learning approaches since they are trainable and require less effort than rule-based approaches. To that end, there has been a surge to amass data for NLP and that has mostly happened for English and other major European languages. Tsvetkov [37] defines them as "languages lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications" which is relevant to our study because in the short history of NLP, African Languages [19] and by extension Nguni languages have had dismally fewer resources poured into them in proportion to their number of speakers/users. Exploring them has been shown to have great social benefits in language preservation and restoration.

Another major motivation for the research is the lack of deep learning approaches to the specific problem of Morphological Parsing and that is a gap that must be filled to determine their viability for the task. Deep learning approaches like fine-tuning pre-trained models [9] are fairly new approaches and they haven't yet been applied for Morphological Parsing.

1.3 Focus of Work

We are focused on the task of morphological parsing for Nguni languages. Given this task, there are many approaches to it like rule-based approaches and finite-state machines but we will focus our efforts on deep learning methods. As mentioned above, there is the option of fine-tuning PLMs on the task but there are also other approaches like training machine learning models for the task. The project has been split into these two separate approaches: I will focus on the former whereby the task will be handled through the application of PLMs to the task while my colleague, Cael Marquad will focus on methods to train models from scratch.

2 PRESENTATION OF EVIDENCE

2.1 Morphological Parsing

As mentioned earlier, morphological parsing is part of the NLP tasks involved with the sub-word structures known as morphemes, among other tasks like segmentation and paradigm learning. However, morphological segmentation is closer to parsing since the tasks are alike each other in the output they produce. For example, the Xhosa word "aliqela"; for segmentation, the output would be the individual morphemes in the word, "a-li-qela", while for parsing the output would be those same morphemes but with annotated tags, "a[RelConc6]-li[BPre5]-qela[NStem]". Therefore, there are some approaches to morphological parsing whereby the output from the segmenter can be used as input to the parser where the tags are added [24, 36].

Moeng et al. (2021) [22] tackle the task of segmentation using neural models like Bidirectional Long Short-Term Memory networks [14] (bi-LSTM), Conditional Random Fields [16] (CRFs) and Transformers with attention [38] and achieve relative success on the task for all four Nguni languages. This agrees with architectures devised by Tsarfaty et al. [36] for parsing morphologically rich languages (MRL) by segmenting them into their morphemes and then tagging the individual morphemes for their role in grammar. These architectures, however, are for other non-Nguni MRLs like Hungarian, Czech and Arabic. The paper rightfully identifies there has been a great success in morphologically parsing English on a word level [36] since grammar in English is usually determined by word order. This makes statistical inference models and handcrafted grammars (and lexical tree-adjoining grammars [3]) such an appealing option. It is harder to achieve the same success for MRLs because their grammars - syntax and semantics - are usually carried on a subword, morpheme, level.

The report by Tsarfaty et al. [36] explores other models to explore the task of parsing MRLs and concludes that there are distinct variations between the languages and there is no end-all model since they belong to different language classes, for example, German is Germanic while Hungarian is Finno-Ugric but both are MRLs. Moeng et al.[22], Pretorius and Bosch[26], and Dione et. al.[10] explored the similarities between the Nguni languages for segmentation, morphological analysis and part-of-speech (POS) tagging respectively and the results using different models indicate a solid relationship between the morphologies of the languages.

There have been efforts to increase the amount of data available for Nguni languages for training models on the tasks. Recently, in 2022, there has been a particular effort in creating data for our task of morphological parsing: a linguistically annotated dataset for the four Nguni languages by Gaustad and Puttkammer [12]. The dataset was acquired by scraping documents from South African government websites (*.gov.za) to acquire corpora in the languages along with parallel English data and from that annotating the data. The advantage of this dataset is that since South Africa has multiple official including these four Nguni languages, most documents will have copies in each language. This is important since for many applications the data needs to be acquired from "parallel texts" [18] for immediate and direct comparison of results. The dataset contains sentences which have been annotated in the following ways:

- Morphological Analysis: the focus of this research. The words are split into morphemes, each with a particular tag indicating the morpheme's role in the word.
- Lemmatization: extracting the roots and core parts of each word.
- POS tag: an annotation indicating the role of each word in the structure (usually sentence) of which it is a part.

Word	Morphological Analysis	Lemma	POS
aliqela	a[RelConc6]-li[BPre5]-qela[NStem]	qela	REL
izikhokelo	i[NPrePre8]-zi[BPre8]-khokelo[NStem]	khokelo	NO8

Table 1: Example lines from the annotation dataset.

Language	Total Tokens	Unique Tokens
isiNdebele	51,120	13,499
Siswati	48,816	14,142
isiXhosa	50,166	14,143
isiZulu	50,528	13,010
English	68,431	4926

Table 2: Distribution of words between the 4 Nguni languages in comparison to English.

Each line in the dataset is formatted with the word, its morphological analysis, lemmatization and POS tag as shown in Table 1. For the collation of the data, the data was pre-annotated and then presented to experts for more fine-grained annotation and correction. To ensure the validity of the data, there was immense quality control on all three facets verified using a rule-based generator. The entire distribution of the words/tokens in the dataset is given in Table 2 showing that there is a considerable amount of unique tokens in each Nguni language in comparison to English. The first two columns of Table 1 are the ones to be used in this project. On the final dataset, for each language, there is a 90% to 10% training-to-testing split. Spanning a total of 380 morphological tags, the dataset is very detailed and proves interesting for study.

Du Toit and Puttkammer [27] also attempt morphological parsing (referred to as morphological analysis in the paper) on an version of the above dataset along with canonical segmentation. They make use of two approaches namely the UDPipe [32] and MarMot [4] which are BERT- and CRF-based respectively. Their findings include the realisation of the difficulty in adapt the BERT-like UDPipe to Nguni morpheme tokens when it contextual embeddings were based on whole words. The MarMot model is trained to use the canonically segmented morphemes as input for parsing and this produces impressive results due to the morphological classes having context dependence and thereby being very similar to the POS word-level tagging that the model was made for. They achieved such results by treating each segmented morpheme as a word thereby making is capable of gaining context for each individual morpheme.

2.2 Related Tasks

While morphological parsing has not been greatly explored for Nguni languages, there are similar linguistic sequence labelling tasks that have been addressed for this language. One of the tasks is POS, which was mentioned earlier in the review. Morphological parsing (or tagging) is predicting the syntactic roles of *morphemes* while POS tagging is predicting the syntactic roles of *words* in sentences. The two tasks are closely related and some approaches for POS tagging can be transferred to morphological parsing. A more specific view on the task of POS on Nguni languages has been identified as a shared task that can be explored in NLP by the Digital Humanities Association of South Africa and the Southern African Conference for Artificial Intelligence Research in 2021 [24]. Morphological parsing is a task similar to but intrinsically distinct from POS tagging described in the above section. Pannach et. al [24] make use of a Bi-LSTM [14] with a layer of CRF [16]. They consider an approach to the task as described by Tsarfaty et al. [36] to add a segmenter before the parser but decided against it since the results from Moeng et. al [22] on segmenting the Nguni languages were not impressive enough to justify such use. The report concluded that they had relative success in the task achieving macro-averaged F1 scores between 84% and 95% and accuracy between 90% and 95%. The F1 score is a metric devised to harmonise the precision and recall of a model [8] as shown in Equation 1.

$$F1 = \frac{2 * (recall * precision)}{(recall + precision)} \tag{1}$$

As earlier stated, Dione et al.[10] perform the task in two Nguni languages: isiXhosa and isiZulu. Due to their attempts to generalise over many topologically diverse African languages, the training was not as successful with accuracy scores of 57.1% and 60.9% respectively. Their architectures were more complex as they had several baselines spanning from multilingual PLMs to CRFs; both having interesting architectures to be explored later in the review.

There are two datasets which have been created for the study of African languages are the MasakhaNER [1] and MasakhaPOS [10] been developed for their respectively named tasks: named entity recognition (NER) and POS tagging. MasakhaNER created high quality tagging data for ten West and East African languages with the bulk of them being Niger-Congo languages. MasakhaPOS alternatively, and more ambitiously, documented twenty typologically diverse African languages to ensure a broader representation of the continent's linguistic offerings. Among these twenty languages were the two most spoken Nguni languages, isiXhosa and isiZulu. The methods of collating the data were the same in both works: collecting from news corpora, manually tagging 100 lines, automating with PLMs then employing experts to check the automatically tagged data to fix errors. Similar approaches can be taken to augment the data on other low-resourced languages as explored earlier [19].

While in the context of NLP all African languages are low-resourced, there are still efforts going into creating more corpora and datasets for study and analysis. The problem is that there are few of these specific datasets and there might be a complete lack of data for the task. Creating these datasets is also expensive as they often require human input for cleaning and verification of correctness. One of the best ways to create new datasets is by having a working language tool like a morphological analyser which further motivates the need for this research.

Configuration	Hidden Size	Layers	Attention Heads	Parameters
BASE	768	12	12	110M
LARGE	1024	24	16	340M

Table 3: BERT Configurations from the initial paper.

2.3 Models

As discussed in the Subsection 1.3, this review focuses on PLMs to highlight their applicability in this project.

2.3.1 Pre-trained Language Models. Transfer learning [35] in machine learning is a way of applying a model trained on a dataset for one task to another task. The model is fine-tuned to work better on the second task to reduce the costs of training a model from scratch. With the fairly recent emergence of Transformer models [38] like the Generative Pre-trained Transformer (GPT) [28] by OpenAI and the Bidirectional Encoder Representations from Transformers (BERT) [9] by Google, PLMs have become some of the main approaches to language processing tasks because of their simplicity of fine-tuning.

2.3.2 BERT. Transformers [38] is a recently (2017) developed architecture with an encoder-decoder setup that processes sequence inputs and produces sequence outputs based on an attention mechanism. BERT is a transformer-based representation model which forms the basis of this research. Like the BiLSTMs [14] mentioned earlier BERT leverages context from both left and right directions to gain a better understanding of the context surrounding the word or token as compared to unidirectional models like GPT which only acquires context from the natural left to right direction. BERT is pre-trained with the primary task of "masked language modelling" (MLM); a task in random words in a text are obscured by a [MASK] token which the model has to unmask based on the context surrounding the token. The second task for pre-training is next sentence prediction in which the model is given a sentence and has to predict the next sentence in the sequence.

The architecture of BERT takes the encoder from the 2017 Transformer by Vaswani et al. [38] with two configurations: $BERT_{BASE}$ and $BERT_{LARGE}$ as shown in Table 3.

The details of the pre-training of the MLM follow very closely to the Cloze task by Taylor in 1956 [31, 33] in which words were masked at random and learners were asked to predict the missing words. 15% of tokens in the training text are selected for masking in the pre-training phase. Of this 15%, only 80% are replaced with the [MASK] token while 10% are changed to a random token and the last 10% are left unaltered. This is done to improve fine-tuning results since there is a mismatch between fine-tuning and pre-training: the

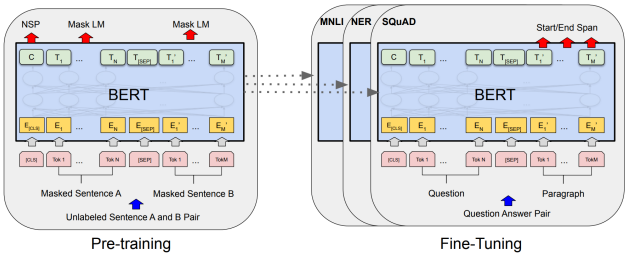

Figure 1: The operation of BERT, illustrating how the model is first pre-trained as an MLM and for next sentence prediction, then later fine-tuning it for downstream tasks like SQuAD and NER. The pre-training gets input as masked sentence pairs and in subsequent fine-tuning the input can be changed to question-answer pairs, e.t.c., depending on the downstream task [9].

fine-tuning data does not contain the [MASK] token. Masking is a process whereby contextual information is important because the masked word needs to be predicted from the surrounding context [30]. This context is heavily dependent on the semantics and the syntax of the language that the MLM is being trained on as that dictates the syntax.

After the pre-training on MLM and next sentence prediction, the original BERT was fine-tuned for tasks such as General Language Understanding Evaluation (GLUE) [39], the Stanford Question Answering Dataset (SQuAD) [29] and Named Entity Recognition (NER) [21, 34] with empirical results showing it outperforming its pre-trained contemporaries: OpenAI GPT [28], ELMo [25] and others. The overall operation of BERT is shown in Figure 1.

2.3.3 Fine-tuning PLMs for African language tasks. With the success of BERT, other BERT-inspired models like RoBERTa [13, 17] and AfriBERTa [23] came into existence improving results by changing pre-training procedures like removing NER or using different corpora.

Concerning African languages, there have been many works which focus on fine-tuning PLMs for African tasks including two extensive efforts on sequence labelling tasks. These two are MasakhaPOS by Dione et al. [10] and MasakhaNER by Adelani et al. [1].

In 2019, for the MasakhaNER 1.0 dataset as discussed earlier in the subsection 2.2, the team of researchers also trained the data for NER using various models including BERT-based architectures. Three years later they followed up on the project with MasakhaNER 2.0 by adding ten more languages, including isiXhosa and isiZulu. The Masakhane group later embarked on part-of-speech tagging data by developing MasakhaPOS [10] in 2023. For the first iteration of MasakhaNER, two variants of BERT were used as baselines:

- Multilingual BERT (mBERT) [9]: an MLM/NSP pre-trained model covering 104 languages including Swahili and Yorùbá from Wikipedia data.
- Cross-lingual RoBERTa (XLM-R) [7]: a RoBERTa [17] extension trained in 100 languages for tasks like cross-lingual classification from CommonCrawl data.

For each of the languages, they utilised language-adaptive fine-tuning as prescribed by Agić and Vulić [2] in which the hyper-parameters were adjusted on individual language corpus level. MasakhaPOS follows closely by using the above-mentioned models and, additionally, three Afro-centric PLMs:

- AfriBERTa [23]: a small low-resource PLM pre-trained on eleven West and East African languages.
- AfroXLMR: a multilingual adaptive fine-tuned [2] PLM pre-trained on 17 languages including isiXhosa and isiZulu.
- AfroLM: a dynamically learning PLM pre-trained on 23 languages including isiXhosa and isiZulu.

For both tasks, NER and POS, the PLMs performed better than competing approaches like convoluted neural networks with BiLSTMs and CRFs. There are conclusions reached from each of the experiments but the most prominent one is that AfriBERTa performed poorly for languages that weren't in the pre-training corpora. This information is especially relevant since the four Nguni languages are low resource [19] and there is little chance that there exists models pre-trained utilising them. Another issue encountered AfroLM having a very small pre-training dataset which leading to poor representations of the languages and inevitably subpar performance. The RoBERTa based PLMs, XLM-R and AfroXLMR, performed significantly better the PLMs which were pre-trained for named entity recognition showing that masked language modelling alone is enough to achieve effective results.

3 DISCUSSION

A significant amount of research has recently gone into NLP for low-resource languages and more particularly African languages after decades of most of the research going into European and American languages [19]. This has opened up many avenues for exploration of African Languages in morpheme-level NLP tasks including canonical segmentation, dependency parsing and morphological parsing [5] due to their agglutinating nature. [22]. There have been historically numerous approaches to the focused task morphological parsing and most of them have been finite state parsers for parsing the morphology on a word level. Neural models like Transformers, LSTMs and CRFs [14, 16, 22, 36] have been used for morphological tasks and mainly in English because English grammar is often on a word level with POS; which is not the case in Nguni languages where grammatical meaning is carried on a morpheme level.

Consequently, the linked tasks like POS for Nguni languages were identified [24], accepting novel approaches to the problem. Through the literature, an approach to the problem - fine-tuning pre-trained models - was reached and explored. It has had various applications including question answering [9], NER [1, 9], sentiment analysis and POS [13]. The initial pre-trained models like OpenAI GPT [28] and Google BERT [9] have proved the efficacy of transfer learning in natural language processing [35]. The benefits of this are apparent and useful in the context of this task since they include a reduction in training time, transfer of learned representations from other tasks and reduced need for copious amounts of data for effective training. A direct result of this is that we can apply these PLMs to low-resource languages for downstream tasks

like POS and NER as shown Dione et al. [10] and Adelani et al. [1] respectively through the performance of models like AfroXLMR and AfriBERTa.

It is noted that the performance would improved by injecting more data into the fine-tuning of the model [2, 10] but the base performance with limited data is enough to consider. Another limitation is the disparity between the base pre-training data and the Nguni fine-tuning data. It has already been shown that even in MLM the presence of the [MASK] token in the pre-training data is enough to skew the representations for fine-tuning in BERT, meaning a further difference with the segmented data on a morpheme level could prove a challenge for transferring the representations [9]. This poses a problem since even the multilingual version of BERT, mBERT, has little representation of African languages (only 2% of the 104 languages are African). In this regard, Dione et al. [10] for MasakhaPOS provide useful insight with their study of typologically diverse African languages because certain groupings of languages other than geographic origin can influence the quality of cross-lingual transfer learning; these groupings include morphological classes like agglutinating languages can have an impact on pre-training. Since mBERT was trained with some agglutinating languages like Hungarian and Finnish which could in turn have a positive effect on the agglutinating Nguni languages.

4 CONCLUSION

In conclusion, recent literature highlights an increasing interest in applying morphological tasks such as segmentation and parsing for Nguni languages. Particularly in South Africa, organisations like the Digital Humanities Association of South Africa have placed notable emphasis on part-of-speech tagging, shedding light on the potential for exploring morphological parsing. An observation from the review was the under-explored territory of utilising deep learning methods for this task, and by using methods like pre-training and fine-tuning, morphological parsing can feasibly be explored. The insights and information gained from this endeavor could prove invaluable for preserving these low-resource languages. Furthermore, the generated information stands to enrich existing datasets and contribute to broader efforts in linguistic and cultural preservation. Overall, the exploration of deep learning methods in morphological parsing holds significant promise for advancing and bettering our collective understanding and conservation of the Nguni languages.

REFERENCES

[1] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. arXiv:2103.11811 [cs.CL]

[2] Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3204–3210. https://doi.org/10.18653/v1/P19-1310

[3] Ali Basirat and Heshaam Faili. 2013. Bridge the gap between statistical and hand-crafted grammars. *Computer Speech Language* 27, 5 (2013), 1085–1104. https://doi.org/10.1016/j.csl.2013.02.001

[4] Anders Björkelund, Özlem Çetinoğlu, Richárd Farkas, Thomas Mueller, and Wolfgang Seeker. 2013. (Re) ranking meets morphosyntax: State-of-the-art results from the SPMRL 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. 135–145.

[5] Jan Buys and Jan A. Botha. 2016. Cross-Lingual Morphological Tagging for Low-Resource Languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1954–1964. https://doi.org/10.18653/v1/P16-1184

[6] Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! long live rule-based information extraction systems!. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 827–832.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[8] Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.* 4, 1, Article 3 (feb 2007), 34 pages. https://doi.org/10.1145/1187415.1187418

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[10] 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)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10883–10900. https://doi.org/10.18653/v1/2023.acl-long.609

[11] Greg Durrett and John DeNero. 2013. Supervised Learning of Complete Morphological Paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (Eds.). Association for Computational Linguistics, Atlanta, Georgia, 1185–1195. https://aclanthology.org/N13-1138

[12] 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 (2022), 107994. https://doi.org/10.1016/j.dib.2022.107994

[13] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open* 2 (2021), 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf

[15] Matej Klemen, Luka Krsnik, and Marko Robnik-Sikonja. 2020. Enhancing deep neural networks with morphological information. *CoRR* abs/2011.12432 (2020). arXiv:2011.12432 https://arxiv.org/abs/2011.12432

[16] John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:219683473

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[18] Friederike Lüpke. 2014. Data collection methods for field-based language documentation. *Language documentation and description* 6 (2014).

[19] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource Languages: A Review of Past Work and Future Challenges. arXiv:2006.07264 [cs.CL]

[20] P.H. Matthews. 1991. *Morphology*. Cambridge University Press. https://books.google.co.za/books?id=JGEAHLg1rmcC

[21] Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. 1–8.

[22] Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and Surface Morphological Segmentation for Nguni Languages. arXiv:2104.00767 [cs.CL]

[23] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 116–126. https://aclanthology.org/2021.mrl-1.11

[24] Franziska Pannach, Francois Meyer, Edgar Jembere, and Sibonelo Zamokuhle Dlamini. 2022. NLAPOST2021 1st Shared Task on Part-of-Speech Tagging for Nguni Languages. *Journal of the Digital Humanities Association of Southern Africa* 3, 01 (Feb. 2022). https://doi.org/10.55492/dhasa.v3i01.3865

[25] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365 [cs.CL]

[26] Laurette Pretorius and Sonja Bosch. 2010. Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues. In *Finite-State Methods and Natural Language Processing*, Anssi Yli-Jyrä, András Kornai, Jacques Sakarovitch, and Bruce Watson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 123–130.

[27] Martin Puttkammer and Jakobus Du Toit. 2021. Canonical Segmentation and Syntactic Morpheme Tagging of Four Resource- scarce Nguni Languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA)* 3 (01 2021). https://doi.org/10.55492/dhasa.v3i03.3818

[28] Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. https://api.semanticscholar.org/CorpusID:49313245

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL]

[30] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. arXiv:2104.06644 [cs.CL]

[31] Adam Skory and Maxine Eskenazi. 2010. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. 49–56.

[32] Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*. 197–207.

[33] Wilson L. Taylor. 1956. Recent Developments in the Use of "Cloze Procedure". *Journalism Quarterly* 33, 1 (1956), 42–99. https://doi.org/10.1177/107769905603300106 arXiv:https://doi.org/10.1177/107769905603300106

[34] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://aclanthology.org/W03-0419

[35] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 242–264.

[36] Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics* 39, 1 (03 2013), 15–22. https://doi.org/10.1162/COLI_a_00133 arXiv:https://direct.mit.edu/coli/article-pdf/39/1/15/1798976/coli_a_00133.pdf

[37] Yulia Tsvetkov. 2017. Opportunities and Challenges in Working with Low-Resource Languages.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[39] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).