Computational Modeling of Agglutinative Languages: The Challenge for Southern Bantu Languages

Farayi Kambarami[1], Scott McLachlan[4, 5], Bojan Bozic[3], Kudakwashe Dube[2] and Herbert Chimhundu[1]

[1] *Chinhoyi University of Technology, Zimbabwe*
[2] *Massey University, New Zealand*
[3] *Technological University Dublin, Ireland*
[4] *Queen Mary University of London, UK*
[5]*University of Birmingham, UK*

Abstract. In computational linguistics, language models are probabilistic models that predict the likelihood of words occurring within specific sentences. They are key components of many natural language processing systems. Traditional full word models do not work well for agglutinative languages. These are languages that have words built out of distinctly identifiable sub-parts that carry specific meanings and functions and can be combined in different ways to form new words. Sub-word language models have been considered to address this problem and have had success with some agglutinative languages. However the existing models do not appear to address the specific ways in which the sentences and words within the Southern Bantu languages, which are agglutinative, are formed. The adoption of sub-word models for these languages has also been low.

Keywords: computational linguistics, agglutinative languages, language models, word models, natural language processing

Languages: Southern Bantu

How to Cite this Article:

Kambarami, Farayi et al. 2021. "Computational Modeling of Agglutinative Languages: The Challenge for Southern Bantu Languages." *Arusha Working Papers in African Linguistics*, 3(1): 52-81.

1.0 Introduction

Natural Language Processing (NLP) technologies have seen extensive improvements in the last decade. Computers are now able to perform complex tasks like augmenting human authors in production of literature (Mulcahy and Wheeler, 2020) and unsupervised summarisation of collections of complex documents (Liu and Lapata, 2019). It is said that we are currently living through the golden age of NLP (Hedler, 2016). Sadly, these developments are not equitably shared across all of the world's languages and communities (Joshi, et al., 2020).

There is broad disparity in the availability of language processing resources between well-resourced and less resourced languages (King, 2015). This paper addresses the challenge of developing computational language models (CLM), which are mathematical and computational abstractions at the heart of many NLP tasks. CLM are receiving increasing research attention. However, most of this research has been directed towards massive pretrained language models that are used to perform a range of generic NLP tasks without having been explicitly trained for them (Petroni, et al., 2019). Two assumptions permeate the CLM domain. The first assumption is that all languages have access to sufficient resources. Also, given that the majority of this work focuses solely on the English language, the second implied assumption has been that what works for English may be easily transferred to other languages. These assumptions are known to be incorrect (Schwartz, et al., 2020; Nchabeleng and Byamugisha, 2020).

This paper considers the development of CLM for the Southern Bantu Languages (SBL), which are spoken in Southern Africa. SBL appear to violate both of the above assumptions that have been the basis of the state-of-the-art in language modeling. First, SBL are typologically different from English as they are largely agglutinative languages, while English is mostly fusional with only some minor agglutination. An agglutinative language is one whose words are made up by combining distinct meaning bearing units. These units are called morphs or morphemes. Each of them plays a distinct role and is clearly identifiable within the word. A language is fusional when the morphemes serve more than one morphosyntactic role. While these typological categories are known to have problems, reference is made to them in this paper because it has been shown by (Prinsloo and Schryver, 2004) that these typological differences have a significant impact on the size of the lexicon required to build a dictionary lookup based spell checker, for example. Second, and more importantly, SBL are largely under-resourced languages (Schwartz et al., 2020; Nchabeleng and

This research is supported by the Centre for Language and Communication Studies at the Chinhoyi University of Technology, the School of Fundamental Sciences in the Faculty of Engineering at Massey University, Queen Mary University as well as TU Dublin. SM acknowledges support from the EPSRC under project EP/P009964/1: PAMBAYESIAN: Patient Managed decision-support using Bayesian Networks.

Byamugisha, 2020). Whilst SBL have significant speaker communities and do not appear endangered from the perspective of daily use, their lack of CLM to enable sufficiently accurate electronic tools like spelling and grammar checkers threatens overall long-term digital vitality (Kornai, 2013).

There is recognition for the need to develop CLM and NLP tools for SBL. This has resulted in a number of initiatives in automatic speech recognition (ASR) and text processing. However, attention has generally been directed at only a subset of SBL languages (Faaß, et al., 2009; Prinsloo and Schryver, 2004; Prinsloo and Schryver, 2004; Prinsloo and Schryver, 2003; Mjaria and Keet, 2018; Langa Khumalo, et al., 2016) leaving other SBL unexplored. Still, progress has been relatively slow and the impact of these initiatives has been varied. Further, the outcomes of these initiatives have not always been easily transferable across all SBL. The main reason for this is that many of the initial attempts took a rules based approach – building models that are specific to the individual languages that they worked on. There have, however been some promising work on fast tracking the development of resources for new languages based on those for existing SBL (Bosch, et al., 2008; Pretorius and Bosch, 2009). Whilst the preceding are from an earlier phase in the development of NLP resources for SBL, more recently (Mahlaza and Keet, 2019) worked on a method to evaluate the similarity of languages with a view of using this information to inform the bootstrapping process for such rules based methods. Still, almost a decade after the initial phase in the development of NLP resources for the Nguni languages of South Africa, they point out that Zulu remains the most well-resourced, relative to the others, even though it can still be considered to be under-resourced in comparison to the other languages of the world.

We contend that availability of good quality CLM to support these languages would enable faster development of many NLP applications since language modeling is at the core of most NLP tasks (Petroni, et al., 2019; Jing and Xu, 2019). This paper presents a survey of the development of CLM for SBL, which the literature shows is largely lacking. This work provides researchers and those working on the development of NLP applications for these languages with a comprehensive understanding for the key challenges that have been encountered by others, as well as providing potential new directions for further investigation.

The remainder of this paper is organized as follows: Section 2 provides a brief linguistic and computational background to the problem, Section 3 introduces the problem, and the survey methodology for this research is found in Section 4. This is followed by a presentation of the results in Section 5, after which the results are discussed in Section 6. Recommendations and future work are discussed in Sections 7 and 8, respectively. Section 9 is a summary with some concluding remarks.

2.0 Background

This section presents a brief background to this research. The *linguistic background* situates this study and provides the context to the challenges that pertain to the modeling of SBL while the *computational background* gives a high level overview of CLM in general.

2.1 Linguistic Background

In this section we discuss the concept of *morphological typology* and place SBL within this framework. We then consider key characteristics of agglutinative languages, and discuss the NLP challenges they engender. Finally, we discuss the implications of all these factors on CLM of the SBL.

2.1.1 Morphological Typology of Languages

Contrary to the implicit language similarity assumption underlying development for most CLM, it is known that even in spite of numerous commonalities, known as *universals* (Croft, 2002), language complexity can differ significantly (Shosted, 2006). Linguists have a number of methods for classifying languages according to the existence of certain linguistic universals (Miti, 2006). One of the oldest ways of grouping languages according to these is known as *morphological typology*. Whilst the concept of words is problematic, and we will return to the reasons for this shortly, traditional morphological typology group languages on the basis of two features of "words". According to (Aikhenvald, 2007) the first classification is based on the *transparency of word-internal boundaries* and puts them into one of the following 3 groups: 1. ***isolating***, 2. ***agglutinative*** and 3. ***fusional*** languages according to how easy it is to identify the boundaries of the constituents of words in a given language. The same author gives a second grouping based on the *degree of internal complexity of words* and has two broad categories: one for ***analytic*** languages and a second one for ***synthetic*** languages. A third category of ***polysynthetic*** languages may also be included into this classification. We will now discuss these two classification systems in a bit more detail below.

Morphological Typology and the Transparency of Internal Word Boundaries

In this section we assume that the idea of words is well defined and is applicable in all languages. As we will see later, this is clearly not the case and the implications of this will be addressed in a subsequent sub-section below. For now, taking the above assumption, the languages of the world can be grouped according to how easy it is to identify and segment the morphemes that make up words in each of the languages. This yields the three classes mentioned above.

The first of these classes is the isolating languages. These arguably have the simplest structure from a morphological point of view. Words in these languages have only one morpheme. They do not use bound morphemes, which are morphemes that only occur attached to another morpheme in order to form the whole word. Languages in this category include Mandarin, Vietnamese and Cambodian (Aikhenvald, 2007).

A second category of languages based on the word-internal boundary criteria is that of the agglutinative languages, also termed agglutinating languages. These have at least three defining features. First, their words are composed of two or more morphemes - a *root*, which is sometimes referred to as a radical, and one or more *affixes*. Second, the affixes are all *bound morphemes* and each has a unique (singular) and well defined function within the word. Third the boundaries between each morpheme within the word are clear cut. For example, in the Shona word *vakaenda* (va-ka-end-a), the radical/stem/root of this word is [*end*] which means *to go*. The bound morpheme [*va*] indicates that the subject of this verb is either plural, or a respected individual. The morpheme [*ka*] indicates tense - in this case it is the remote past tense, and the last [*a*] is the final vowel. As we have previously stated, SBL, the group of languages that includes Shona, fall within the category of agglutinative languages.

The last category of languages in this typology is referred to as fusional languages. Whilst it is sometimes referred to as flectional languages, (Aikhenvald, 2007) argues that this is misleading. As with agglutinative languages, their words are also formed by bringing together several morphemes. However, unlike the former, the boundaries of the morphemes may not be easily distinguishable and some of the bound morphemes perform more than one morphosyntactic role within each word. An example of this is in the Spanish word habló . In this word, the morpheme [ó] indicates both the past tense and the third person singular subject.

Morphological Typology and the Degree of Internal Complexity of Words

A second classification of languages considers the internal complexity of words. (Sapir, 1921) identifies three different groupings into which languages could be placed based on this criteria. These three groups are the analytic, synthetic and the polysynthetic languages (Sapir, 1921). Analytic languages are defined as those that do not combine concepts into single words at all, or do so economically. In such languages the sentence is more important than the word. Synthetic languages, on the other hand incorporate more concepts into the word. Polysynthetic languages are a special case of synthetic languages in which the concepts that are incorporated into the typical word cross various morphological categories. These are languages which combine the features of the synthetic languages with even more complexity in the way that the morphemes exhibit allomorphy. They also have some verb forms where the morphemes refer to other entities other than the subject, a feature termed poly-personalisation. Most Amerindian languages are considered as polysynthetic.

These categories exist on a spectrum and no single language can be exclusively classified as belonging to one or the other of these categories. For example, whilst Mandarin is frequently cited as a typical analytical language, it has been shown to have some inflected words, see (Arcodia, 2012) for a detailed treatment of lexical derivation in the language.

Integrating the Two Views

Whilst the two systems of classifying languages could be considered independent of each other, (Aikhenvald, 2007) presents a possible way of merging the two concepts into one unified system as shown in **Error! Reference source not found.** below.

Figure 1: *Integration of Two Systems of Morphological Typology* (Adapted from Aikhenvald (2007)

Techniques of joining morphemes Meanings/Functions per morpheme	fusional	English	Russian	Chiricahua Apache
	agglutinating		Hungarian Shona	Yupic Eskimo
	isolating	Vietnamese Classical Chinese		
		analytic	synthetic	polysynthetic
		Number of morphemes per word		

The Problematic Word

The above typology is premised on the idea that words are clearly defined entities that can be easily identified across all languages. Unfortunately the reality is much more nuanced. To start with (Haspelmath, 2011) has shown that there is no single universally applicable and agreed upon definition of what constitutes a word. They state that there are four criteria that could potentially be used to define words: these being semantic, orthographic, phonological and morphosyntactic. In a separate study (Bejan, 2017) expand these four into five senses in which a word could be defined, none of which are guaranteed to yield the same list of words for a given language by breaking the morphosyntactic into a separate morphological and syntactic category. The full list of possible word definitions according to this new list is given below.

The first one is the prosodic or phonological level, which is based on how it sounds in spoken language (Hildebrandt, 2014). A second one is the orthographic or graphemic word, which is determined by how it is written down and is defined as a string of letters that are found between spaces or punctuation marks in writing or printing. The third definition is that of the morphological word, which considers how the words are formed and what part they play in speech. Yet another way of defining words is at the lexical or semantic level. This considers words to be the smallest units that carry meaning within a given language. Finally, words can also be defined at the syntactic level where each word is the smallest element of a sentence within a given language. Whilst there are overlaps and synergies between the above levels of word definition, there is no guarantee that each definition would yield exactly the same list of words in any language. In fact, they often do not.

This study is specifically interested in the orthographic or graphemic word as this is the one that we encounter in written text. The first challenge that we encounter with the orthographic word within the SBL is that there are different writing systems that are at play. Some of the SBL, like Shona, Zulu, and Xhosa utilise a conjunctive writing system whilst others such as Tswana and Sotho use a disjunctive writing system (Taljard and Bosch, 2006). As Table 1 below shows, language constituents that carry the same meaning and function are written very differently in these languages. The English phrase “I fear him/her” is rendered as one word in Shona – “ndinomutya”. This “word” is composed of the following morphemes: a Subject Concord, [ndi] – which is in first person singular, the tense marker [no], indicating the present continuous tense, an object marker [mu] – class 1, single human object and the root of the verb [ty] and the final vowel [a]. The same phrase is rendered as separate “words” in Sotho, each of which roughly correspond to the above morphemes. [kea] combines the subject concord [ke] and the tense marker [a]. [mo] has the same meaning as [m], the object concord in the Shona rendering of the phrase. [tsab] is the root of the verb and [a] is the final vowel.

One of the consequences of the above differences in orthography is that different approaches are required to perform natural language processing tasks on languages using either writing system. For example in a study of spell checkers for the South African languages, (Schryver and Prinsloo, 2004) found that the lexical recall for word list based spell checkers was higher for the disjunctively written languages using a significantly smaller dictionary of words than they were for the conjunctively written languages.

Table 1: *Illustration of Orthographic Differences and Word-Count Impact*

Original Shona	**Tswana**	**Translation**	**Analogous Shona Disjunctive Spelling**
ndinomutya	kea mo tsaba	I fear him/her	ndino mu tya
ndinomutya	kea mo rata	I love him/her	ndino mu da
ndinomuziva	kea mo tseba	I know him/her	ndino mu ziva
ndinomubatsira	kea mo thusa	I help him/her	ndino mu batsira
ndinovatya	kea ba tsaba	I fear them	ndino va tya
ndinovada	kea ba rata	I love them	ndino va da
ndinovaziva	kea ba tseba	I know them	ndino va ziva
ndinovabatsira	kea ba thusa	I help them	ndino va batsira
ndinokutya	kea u tsaba	I fear you	ndino ku tya
ndinokuda	kea u rata	I love you	ndino ku da
ndinokuziva	kea u tseba	I know you	ndino ku ziva
ndinokubatsira	kea u thusa	I (will) help you	ndino ku batsira
twelve distinct orthographic words	eight distinct orthographic words	eight/nine distinct orthographic words	eight distinct words
no clear relationships at the word level			

Southern Bantu Languages (SBL)

The morphology of Bantu languages has been the subject of various studies including those of (Miti, 2006) who has provided a comprehensive analysis of their morphology and phonology. The SBL are a significant sub-group of the Bantu languages which fall into Guthrie's zone S (Guthrie, 2017). Whilst (Janson, n.d.) excludes Shona from this grouping, we follow (Guthrie, 2017)'s original classification and include it in our analysis. Although future work will primarily refer to Shona when testing the concepts that will arise out of this current research, the aim of this and all related future work is to address SBL in their entirety.

There are four features of the Southern Bantu languages that are of particular interest to the present analysis These are 1) the nominal class system, 2) the concordial agreement system, 3) derivational morphology and 4) allomorphy. In the nominal class system (Maho, 2001), each noun can be assigned according to their type to one of 21 classes. Concordial agreement requires that words in a sentence must agree with the noun class of the sentence in which they appear. Changing the subject of the sentence usually requires only the replacement of prefixes and other concords of most words within the sentence in order to retain a semantically valid sentence. Words, especially verbs, can be extended in various ways by the addition of suffixes which leads to the formation of new verb forms, and in some cases, changes in part of speech. This is referred to as *derivational morphology*. *Allomorphy* has two implications. First, a particular spelling of a word may have more than one meaning in different contexts depending on the stress that specific syllables are given leading to ambiguity in written words. Second, the same word may be spelled differently by speakers of different dialects of the language, also leading to the increase in the number of words. All of these features lead to a comparatively larger vocabulary than English for example. For a more detailed introduction to the languages, the reader is referred to (Doke, 2017; Miti, 2006; Nurse and Philippson, 2006).

2.2 Computational Background

In this section we introduce the concept of language models, the noisy channel model of communication which underpins the utilisation of language models in several natural language processing applications, as well as the ways in which language models and the key applications that utilise them are evaluated.

2.2.1 Language Models

The term *language model* refers to a *mathematical function* that utilises statistical analysis to estimate the probability of a given word within a specific context. Such language models can perform one of two complementary tasks. The first task is that, given a complete sentence, there is a requirement to provide the answer to the question: *What is the likelihood that the last word (or any other word) in the sentence would occur given the preceding words?* The second task answers the question: *What is the word that is most likely to follow a given sequence of words?*

(Jurafsky and Martin, 2018) provide a comprehensive introduction to language modeling and the role that language models play within NLP. Most of the work that they report on is targeted toward well-resourced languages. Most agglutinative languages are under-resourced and the techniques that apply to most well-resourced languages do not work well for them.

Statistical language models are not the only way of representing and modeling the structures of human languages. Loosely conceptualised, "language models" can be built using approaches that exist on a continuum based on the amount of amount of human/expert knowledge required in their development. At the one extreme are true rules based models that are built using handwritten rules supplied by experts whilst on the other end are fully data driven models that learn the structure of a given language from data that they are supplied with. On the rules based end of the spectrum, Finite State Automata (FSA) can be used to fully describe the grammar of a given language and to accept either valid words or sentences of the language (Bakaev and Shafiev, 2020; Thottingal, 2019). Their development require a detailed understanding of the grammar of the specific language being modelled, however they have the advantage of generating models whose results are fully explainable. In general, developing such systems is more labour intensive and time consuming that building an equivalent statistical or machine learning model. As a result, most of the recent work on language modeling is on statistical or data driven language approaches, and in fact the state of the art are built with very little explicit knowledge about a given language encoded into them. The state of the art on the data driven end of the spectrum include natural language generating models such as GPT 3 which learn everything about given languages from the data that is provided to them with almost no human supervision (Brown, et al., 2020).

There is a place for each of the aforementioned approaches. Most of the initial work on SBL has favored the rules based approach mainly due to the paucity of data on which to build good quality models (Bosch, et al., 2008; Pretorius and Bosch, 2009; Byamugisha, et al., 2016; Bosch and Pretorius, 2017). However, as stated earlier, the challenge that this brings is that it slows down the development of NLP resources for the languages as it is heavily reliant on appropriate collaborations between linguists and computing experts. The present work considers the data driven language models as a way to address the slow progress in the development of NLP tools for SBL.

The most basic type of statistical language models (henceforth language models) are the *n-gram language models* (NGLM). Given a sequence of *n* words, these models estimate the probability of the *n*th word, given the preceding *n-1* words. They form part of a broad class of language models called counting models. This is because the probabilities that they generate are computed by counting sequences of words. In their original form, these models assume words as the modeling unit. Given that it is impossible to encounter a training document that contains all the words in a given language, they address the problem of incorrectly estimating the probabilities of unseen words using three key approaches. These are 1) *smoothing*, 2) *interpolation* and 3) *back-off*. *Smoothing* refers to a number of algorithms that aim to improve the estimates for unseen n-grams by assigning some of the probabilities of seen n-grams to them. In *interpolation* each n-gram probability always considers the probability of its shorter sub-sequences. *Back-off*, on the other hand, uses the probability of shorter n-gram sequences, only when the information about longer sequences are not available.

None of the methods referred to above address the challenges that arise from the morphology of conjunctively written agglutinative languages. In these conjunctively written agglutinative languages new words are encountered due to the problems mentioned above, but also due to the fact that new words can be easily formed as described in the linguistic background. As a result, researchers in these languages have considered using sub-word units to improve their performance (Arısoy, et al., 2008). Other language model types that are not necessarily developed for agglutinative languages have also been considered. These include *class based language models* (CBLM) (Brown, et al., 1992) which assign each word to a class, and then compute the probability of a word of that given class co-occurring with words in the classes that would have preceded it. *Factored language models* (FLM) (Bilmes and Kirchhoff, 2003) were initially developed for the agglutinative languages. They model each language as a sequence of related factors - where each factor could be a part of speech tag, or a component of each word.

N*eural network language models* (NNLM) were introduced by (Bengio, et al., 2003). Unlike all the previous models, these do not represent words as discrete elements. Instead, they project words into continuous space. (Goldberg, 2017) provides a comprehensive treatment of NNLM. These are currently the state-of-the-art in language modeling. Whilst they address a number of shortcomings inherent in the linear counting based language models, these NNLM do not solve the data sparsity challenge posed by agglutinative languages. As a result, similar attempts to utilise modeling units other than the morphological word have also been considered (Cai, et al., 2017; Kudo, 2018; Labeau and Allauzen, 2017).

2.2.2 The Noisy Channel Mode

Many NLP applications could be considered to be instances of the *noisy channel model* that was introduced by (Shannon, 1948). In this model, a sender transmits the intended message *e*, which is distorted by the noisy channel resulting in the receiver receiving the message *f*. The noisy channel is modelled as the process that generates the probability P(*f*|*e*) and the originally intended message can be assumed to be the result of a language model P(*e*). Given this scenario, Bayes' Theorem can be used to attempt to recover the originally intended message by computing the probability P(*e*|*f*) as P(*f*|*e*)P(*e*). This approach can then be applied to problems as diverse as automatic speech recognition and optical character recognition.

2.2.3 Natural Language Processing of Agglutinative Languages

Agglutinating languages present a paradox to the computing expert. While it is deceptively simple for a human reader to parse the words of these languages into various meaning-bearing morphemes, the task of achieving the same computationally is not as easy. An additional challenge which is brought by the highly productive nature of these languages is the existence of almost unlimited vocabularies which, due to the Zipfian distribution of words in human languages, worsens the data sparsity problem (Zipf, 1949). In their study that looked at the relationship between typology and the limits of multi-language NLP, (Gerz, et al., 2018) found that on average the agglutinative languages that they sampled had an type to token ratio of 0.16 versus 0.14 for fusional languages, 0.11 for fusional languages and 0.05 for isolating languages. More importantly they found that the perplexity scores for three different types of language models were consistently worse for the agglutinative languages than they were for all the other language types – mirroring the type to token ratios stated above.

3.0 Research Problem

SBL have morphology that results in very large vocabularies as described in the linguistics section. This increases the number of unknown words that a CLM could encounter after it has been trained, potentially reducing its performance. Whilst increasing the size of the training corpus could alleviate this problem, SBL also suffer from having limited resources in the form of good quality corpora (Khumalo, 2020). At the same time the traditional approaches used for NGLM and other CLM do not adequately address the challenges that arise due to this (Abulimiti and Schultz, 2020). Sub-word CLM (SWCLM) have been considered for other agglutinative languages (Arısoy, et al., 2008) but have received limited attention for SBL. This paper is part of a broader research project which aims to develop novel CLM for SBL and other agglutinative languages. The aim of this research is to draw attention to the gap in the research and deployment of SWCLM for SBL by conducting a survey of the usage of SWCLM across agglutinative languages. It also aims to identify the factors that influence the performance of SWCLM in order to inform how future SWCLM for SBL can be developed. In doing this, it also seeks to understand how the sub-word units used in SWCLM are determined.

To achieve the preceding, the research objectives are to:

1. review the level of research attention on SWCLM in SBL and other agglutinative languages.
2. determine model performance factors that influence performance of SWCLM used for agglutinative languages and MRL.
3. to investigate word decomposition by reviewing different approaches for splitting words in agglutinative languages into their underlying building blocks. The broader project will in future continue with work
4. to propose a modeling approach for agglutinative languages based on word and sentence (de)composition.

4.0 Methodology

This section presents the approach of this paper to conducting a literature survey on language modeling for agglutinative languages with a focus on the SBL. We first present our search methodology and then the survey framework used in this paper.

4.1 Literature Search and Selection

The search undertaken drew literature from Google Scholar, the ACL Web as well as the ACM websites was performed. This search sought papers that discuss *language modeling* for any of the agglutinative/agglutinating languages. Additionally, papers that discussed CLM for other morphologically rich languages that share the same challenges as agglutinating languages were also included. The following search terms were used:

[("Language Model" AND agglutinative) AND (decomposition OR "word splitting")]

After this initial search, the search terms were expanded to include “sub-word”, “subword” and “sub word”. This is because in our initial screening we determined that the majority of papers that had word decomposition or splitting referred to sub-word CLM. We needed to ensure complete coverage of all such CLM.

[("Language Model" AND agglutinative) AND (decomposition OR "word splitting OR Sub-word OR Subword")]

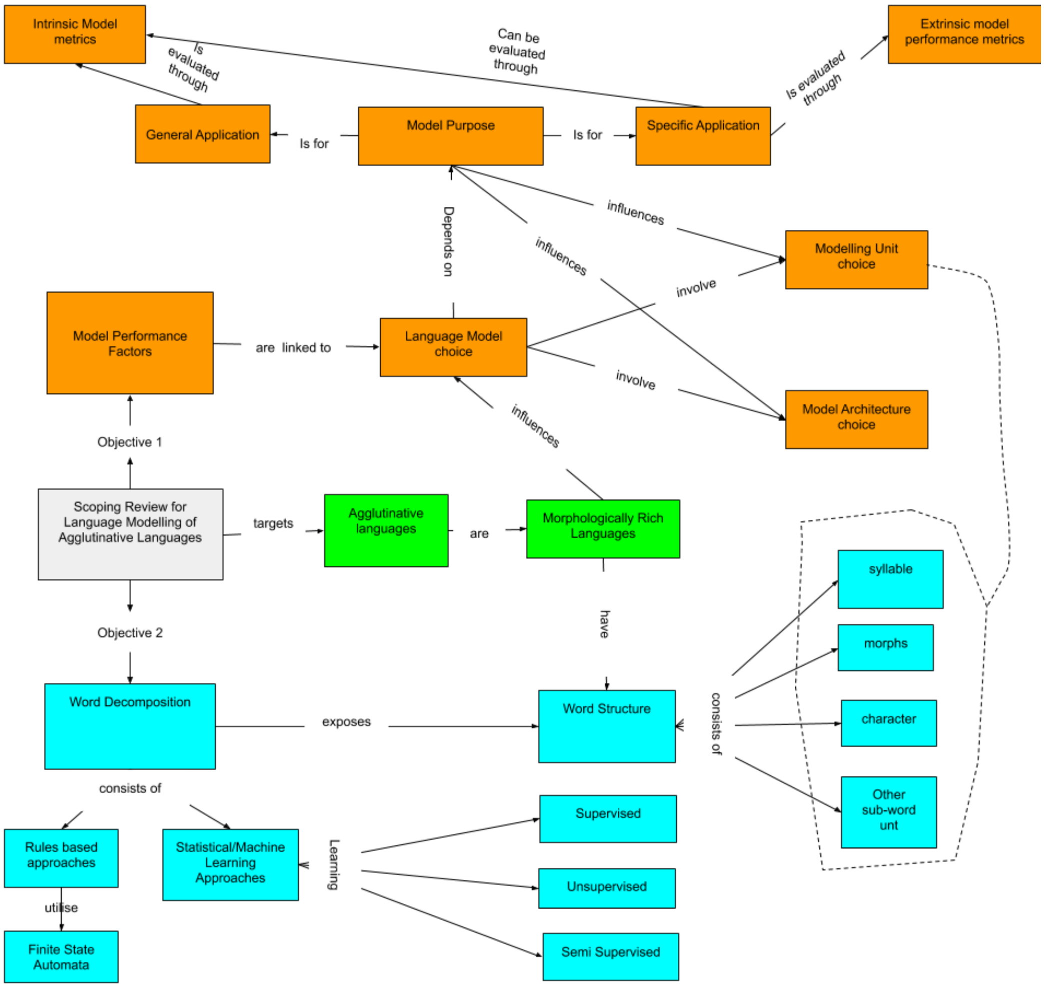
The initial collection of papers underwent screening on the basis of content, but not on chronology. The evaluation criteria for content considered:

1. Use of language models to solve a natural language processing problem in any domain.
2. Model use for at least one agglutinative language
3. Sub-word modeling

4.2 Survey Framework

Figure 1 presents the primary concepts collected and evaluated during review of the literature, as well as a visual representation of their relationships. The concept map defines the research questions. Concepts were deductively developed and inductively expanded during initial screening of the literature.

Figure 1: *Concept Map for the Literature Review of Language Models for Agglutinative Languages*



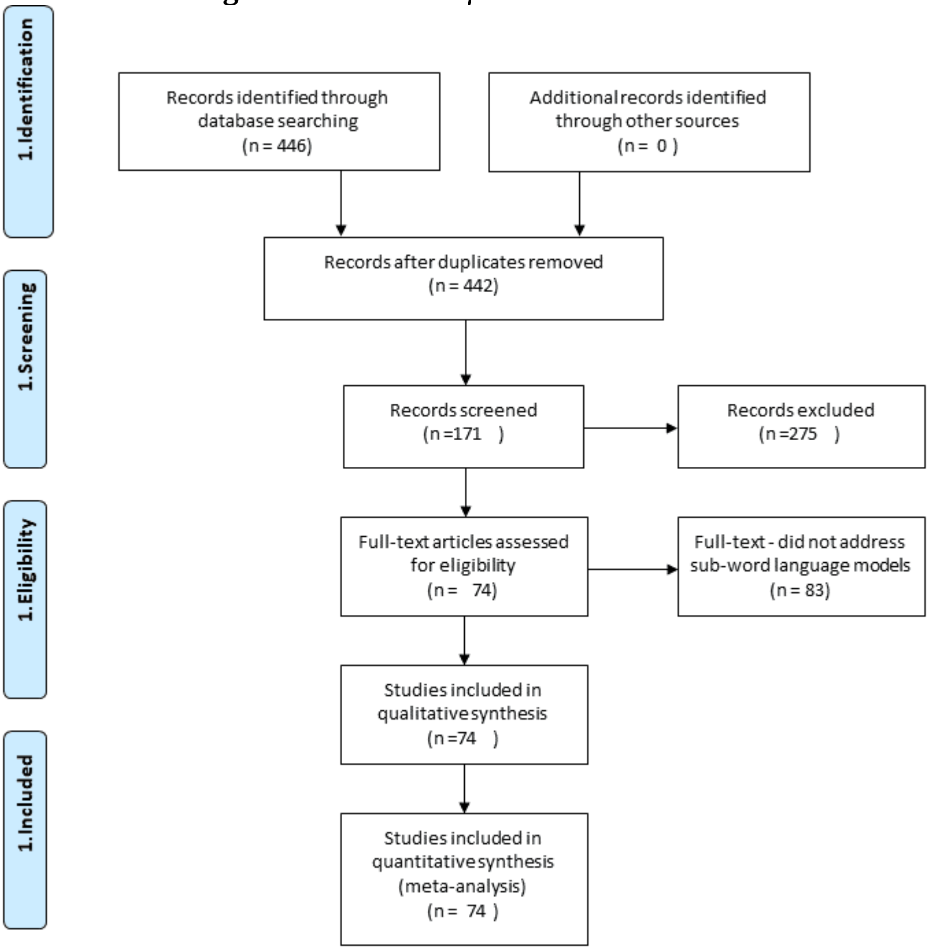
5.0 Results

This section presents the results of the survey, providing the outcome of the search process and detailing the findings for each of the elements of the concept map in figure 1.

5.1 Search and Collection Results

The results of this literature search and selection are summarized in Figure 2 below.

Figure 2: *Outcome of Assessment Process*


1. Identification
1. Screening
1. Eligibility
1. Included
Records identified through database searching (n = 446)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 442)
Records screened (n =171)
Records excluded (n =275)
Full-text articles assessed for eligibility (n = 74)
Full-text - did not address sub-word language models (n = 83)
Studies included in qualitative synthesis (n =74)
Studies included in quantitative synthesis (meta-analysis) (n = 74)

The literature search initially identified 442 papers for screening using the selection criteria listed in section 3.1, which resulted in a collection of 171 papers for full text assessment seeking papers that addressed sub-word language models. Upon completion of this process, seventy-four (74) papers remained for inclusion in this review.

5.2 Survey Results

This section presents the findings of the survey based on the conceptual framework in figure 1.

Lack of Research Attention on SBL CLM

Table 2 presents the number of publications for each agglutinative and morphologically rich language encountered in the literature. The majority of research on sub-word CLM (SWCLM) for agglutinative languages has been done on languages from outside Southern Africa. None of the papers addressed SBL models that operate at the *sub-word level.* For agglutinative languages, sub-word models are known to perform better than word level models.

Table 2: *Papers by Language and Language Type*

Language	**Papers Published**
Turkish	14
Other Agglutinative Languages	13
Other Morphologically Rich Languages	12
Arabic	11
Finnish	10
isiZulu	6
Hungarian	4
Estonian	3
chiShona	1
kiSwahili	1
Other Bantu	1

Model Performance Factors

Language model choice: The choice of language model is seen to be influenced by the purpose to which it is applied and involves the choice of modeling unit and modeling architecture. **Error! Reference source not found.** summarises the application domains in which language models were applied in the papers that we surveyed. Sub-word language models were observed to be widely utilised in the automatic speech recognition and statistical machine translation domains for agglutinative languages. By contrast, their use was observed to be very low in spell checking and optical character recognition domains.

Table 3: *Proportion of Papers with Applied Models to NLP Domains*

Application	**Papers Represented**
Automatic Speech Reognition (ASR)	68%
Statistical Machine Translation (SMT)	19%
None – General Model	7%
Spell Checking (Spell C)	2%
Keyword Search (KWS)	2%
Optical Character Recognition (OCR)	2%
Other	-

Two-thirds (66%) of papers reviewed discussed the application of language models in *automatic speech recognition* (ASR). A further 19% of papers presented *statistical machine translation* (SMT) for agglutinative and/or MRLs, including those covering the use of language models within SMT (Jayan, et al., 2015; Etchegoyhen, et al., 2018). Their work focused on less resourced languages such as the European language *Basque*.

Apart from specific applications, there were a number of papers that looked at fundamental research in language modeling. (Alexandrescu and Kirchhoff, 2006) introduced Factored Language models, partly to address the challenges encountered in the development of language models for agglutinative languages. Other work that reports on language models for agglutinative languages outside an application domain include that of (Labeau and Allauzen, 2017) who use a character based Neural Language model for Czech, and (Vania, 2020) who also looks at how character level models could be used to represent the morphology of agglutinative languages.

Model Architecture: In this paper, CLM architecture refers to the way in which the CLM is organized internally. As **Error! Reference source not found.** shows the majority of sub-word language models for agglutinative languages continue to utilise n-gram based architectures. These are followed by factored language models (FLM) and Class-based language models (CBLM), with neural network language models (NNLM) and factored neural network language models FNLM) having the least coverage.

Table 4: *Number of Papers by CLM Architecture*

Model Architecture	**Number of Papers**
N-Gram	51
Neural Network Language Models	11
Factored Language Models	6
Class Based Language Models	5
Other	4
Factored Neural Network Language Models	2

Modeling Unit: CLM can take different linguistic units as inputs. Despite the fact that we were explicitly looking for SWCLM, the majority of papers that we surveyed also covered some full word CLM. They were used as benchmarks against which all the SWCLM were compared. Apart from full words, the next popular modeling unit was the ambiguously defined morph. In some of the papers, this corresponded to actual morphemes like prefixes, stems and suffixes, but in others they corresponded to parts of words identified by specific machine learning algorithms as useful elements to use in modeling specific languages. The next set of popular sub-word elements were sub-words in general, followed by syllables, and characters as per **Error! Reference source not found.**.

Table 5: *Modeling Unit*

Modeling Unit	**Total Papers**
Word	48
Morph	40
Sub-word (/General)	36
Syllable	10
Character	8
Lemmas	1

Word Decomposition

Any sub-word language modeling assumes that the constituent components of words can be identified so that they can be given as input into the CLM. Our working assumption is that the quality of the word decomposition process has a significant influence on the performance of SWCLM. There are two main paths word decomposition can take. One method is a rules based approach, and a second one is a statistical machine learning approach.

Rules based approaches to word decomposition: This survey only came across one study that utilised a rules based approach to deduce sub-words *for use in a CLM*. The research by (Ablimit, et al., 2016) implemented a rules based stemmer as well as Statistical machine learning stemmers for Uyghur.

Statistical/Machine Learning approaches to word decomposition: Statistical/Machine Learning approaches appeared in the majority of papers that we surveyed. At a high level, there are three main approaches to the acquisition of the sub-words of agglutinative and other MLR. **Error! Reference source not found.** below shows the results of our survey on these high level methods used to decompose sub words in agglutinative languages.

Unsupervised approaches to word decomposition: The most preferred high level approach to splitting the words of agglutinative languages is through the use of *unsupervised learning approaches.* As **Error! Reference source not found.** below also shows, the majority of papers that used unsupervised methods, utilised the Morfessor tool to perform the morphological segmentation of words. Whilst Morfessor can be utilised in a semi-supervised and a supervised manner, the preference for many of the researchers was to utilise it in an unsupervised manner. Some of the papers did not report on how they acquired the sub-word units utilised in their models, and these constituted the second highest number of papers encountered. Figure 9 below provides more detailed information on the specific approaches that were utilised to perform the sub-word decomposition.

Table 6: *Papers by High-Level Word Decomposition Approaches*

Word Decomposition Approach	**Total Papers**
Unsupervised Learning	23
None	8
Semi-Supervised	6
Supervised Learning	5
Unspecified	4

Just under half of the papers surveyed (48%) papers surveyed utilised unsupervised methods to decompose the words in the agglutinative languages that they studied. The main method that was utilised to do this was using the Minimum Description Length principle which is implemented in the publicly available tool Morfessor as in the studies reported by (Botha and Blunsom, 2014; Creutz, et al., 2007), (Hirsimaki, et al., 2009), (Siivola, et al., 2003), (Kurimo, et al., 2006), (Mihajlik, et al., 2007), (Mihajlik, et al., 2010), (Sak, et al., 2012), (Tachbelie, et al., 2014), (Agenbag and Niesler, 2019), (Gupta and Boulianne, 2020) and (Vania, 2020). 76% of the papers utilised this method.

A few of the studies attempted to develop either their own stemmers or morphological analyzers as reported by (Demberg, 2007), (Mihajlik, et al., 2010) and (Ablimit, et al., 2014). More specifically (Mihajlik, et al., 2010) utilise a hybrid approach that takes both SMT induced as well as grammatical information to inform their approach for developing the word segments. On the other hand (Ablimit, et al., 2014) use Morfessor and a bespoke Morphological analyzer that utilises what they call “Discriminative learning”.

Supervised approaches to word decomposition: Figure 8 shows supervised learning methods are not as widely applied as unsupervised methods. However, supervised methods were used to segment words in a number of examples which include (Vergyri, et al., 2004), (Arisoy, et al., 2009), (Chahuneau, et al., 2013) and (Tachbelie, et al., 2014). One solution used a morphological analyzer that performed “shallow” morphological analysis of Arabic (Vergyri, et al., 2004; Darwish, 2002). Another utilised a finite state transducer to split the words of the languages that they reported on (Arisoy, et al., 2009), while (Chahuneau, et al., 2013) use a Finite state machine.

(Lajish, et al., 2015) developed their own sub-word modeling approach using the rules of Malayalam called Sandhi. The authors, using their knowledge of the language, drafted the rules to analyze the Malayalam words into their constituent components before passing these through to an n-gram language model (Lajish, et al., 2015).

Semi-supervised approaches to word decomposition: A small number of studies utilised semi-supervised machine learning methods to induce sub-words. For example, (Botha and Blunsom, 2014) utilised Morfessor in Semi-supervised mode to generate labelled morphemes which they then used to further perform word segmentation using another word segmentation model. (Mihajlik, et al., 2010) also used hybrid methods to achieve the word segmentation task. Both grammatical and statistical segmentation techniques were used to mitigate against the limitations that each type of segmentation method inherently possessed.

Table 7: *Papers Implementing Specific Word-Decomposition Approaches*

Word Decomposition Technique	**Number of Papers**
Morfessor	18
Minimum Description Length	16
None	8
Other (Rules-Based)	8
Bespoke Morphological Analyzer	7
Finite State Automata and Transducers	4
Bespoke Stemmers	3
Byte Pair Encoding	2
Sub-Word Regularization	1

Language Model Performance Metrics

In order to understand the factors that influence the performance of CLM we reviewed the ways in which CLM have been evaluated. CLM are evaluated either intrinsically, that is, by directly checking their performance outside of any specific application, or extrinsically by checking the performance of an application enabled by a specific CLM. We also reviewed the metrics in use for evaluating SWCLM to understand how different SWCLM performed comparatively. **Error! Reference source not found.** provides a view of the metrics and their usage.

Table 8: *Number of Citations Based on Specific Performance Metrics*

Measure	**Citations**
Perplexity	55
Word Error Rate	35
Out of Vocabulary Rate	33
Accuracy	10
Precision	9
BLEU	8
F1	8
Recall	8
Other Error Rate	7
N-Gram hits	6
Coverage Percentage	5
MTWV	4
Entropy	3
SLER	1

Perplexity, which is an intrinsic performance measure for CLM, was the most frequently utilised method in the papers surveyed. The next most cited performance measures were *word error rate* (WER) and *out of vocabulary rate* (OOV). The remaining metrics had relatively lower usages, with observed usage of less than 10 across the remaining literature. The choice of metrics used appeared highly dependent on the application domain. WER and OOV were the preferred metrics within the ASR community, hence their dominance in the literature.

Discussion of Findings

As we stated in section 3, the aims of this paper are to:

1. review the level of research attention on SWCLM in SBL and other agglutinative languages.

2. determine model performance factors that influence performance of SWCLM used for agglutinative languages and MRL.
3. investigate word decomposition by reviewing different approaches for splitting words in agglutinative languages into their underlying building blocks.

Review the level of research attention on SWCLM in SBL and other agglutinative languages

The results of the survey confirm the lack of research attention on SWCLM for SBL. It is not clear from the research results why this is the case.

Determine model performance factors that influence performance of SWCLM used for agglutinative languages and MRL

The results further show that the architecture and the choice of modeling units for SWCLM are dependent on the domain in which the model is deployed. In terms of the goal to determine the model performance factors, there were no consistent metrics for measuring CLM across a number of application domains in the literature. Whilst the textbooks recommend the use of perplexity to measure the intrinsic performance of a language model, this was found to not always be used. There is a strong link between the application domain and the evaluation method applied, as expected. However some of the evaluation methods cut across domains. In general it is not easy to compare the performance of the actual language models across domains. However the choice of model architecture and modeling units have an impact on the performance of CLM. This confirms the contention that developing language models that are attuned to the structure of the languages being modelled would yield better results.

There has also been limited attention paid to word decomposition techniques for SBL. The majority of work that has been done across MRL shows that researchers prefer unsupervised machine learning approaches to word decomposition - with Morfessor being the preferred tool with which to perform this. There are also some language specific approaches that have been developed with varying levels of supervision. However, rules based approaches are hardly used. Not all of the sub-words determined through these methods correspond to distinct linguistic phenomena. Despite this, they have proven to be useful in enhancing SWCLM. Thus in developing SWCLM for SBL, researchers need not strive to acquire exact representations of the underlying morphology.

A key limitation of our findings is that this study only considered studies that looked at SWCLM for agglutinative languages MRL. Research may be needed to expand the scope to include any SWCLM for any language typology as well as considering all other approaches to CLM in general. The study also limited its search to Google Scholar and the ACM digital archives. As a result, papers that are not visible from these two portals may have been missed. However, given the number of papers that were originally returned by the initial searches, there is a high likelihood that most of the papers that address this area have been reviewed.

Recommendations

We have established from this survey that there is a gap in the literature for SWCLM of SBL specifically and the Bantu languages in general. There is a need for researchers to establish the limits of these types of models for these languages.

The choice of SWCLM is largely dependent on the application of the model. Whilst the modeling unit plays a major role in determining how well the model performs, there is no consistent way of comparing the performance of models that are used to solve different NLP problems. Ultimately, the performance that matters for users of CLM is that of the complete applications that they develop. However, since CLM are usually one of at least two key model components in any application, it is also important to separate their performance from that of the complete application. This is to ensure that the right elements are improved upon if application performance is not optimal. Whilst perplexity as a measure is already established, its inconsistent usage needs to be further investigated. The metric is widely cited but not by all researchers as would be expected. It is also recommended that more work be done on establishing metrics and best practices for comparing CLM performance across domains.

SWCLM have been proven to be effective for other agglutinative languages. Their development requires the use of word decomposition techniques. Some of the techniques encountered have been previously tested for SBL. We recommend that further research be conducted to determine which ones work well for SBL and the conditions under which they cease to be effective needs to be conducted as a precursor to the development of SWCLM for SBL.

Future Work

This paper has established that whilst sub-word language models have been found to be very effective for other languages, they have not been investigated for SBL. It has also determined that the performance of these models is a function of the accuracy of the word segmentation models available for the specific languages. These two findings lead us to believe that there is merit in continuing our investigation into sub word language models for SBL. As a result, the project will be addressing the following broad research directions:

1. Evaluate the performance limits of SWCLM for the SBL family.

2. Develop robust criteria for evaluating the performance of CLM across application domains.
3. Investigate the conditions under which SWCLM perform best for SBL.

Summary and Conclusions

In this paper we carried out a survey of the literature on Language modeling of the agglutinative languages. We find that there is a lack of research attention on CLM for SBL. We further found out that there are inconsistencies in the utilisation of performance metrics, making it difficult to compare the performance of different languages utilised for different purposes across different domains. We propose that work be done on determining the limits of SWCLM for SWB. We also recommend the development of better and more harmonized metrics for evaluating Language modeling performance be considered.

References

Ablimit, M. et al. 2016. "Stem-Affix Based Uyghur Morphological Analyzer." *International Journal of Future Generation Communication and Networking*, 9(2): 59-72.

Ablimit, M. et al. 2014. "Lexicon Optimization Based on Discriminative Learning for Automatic Speech Recognition of Agglutinative Language." *Speech Communication*, 60: 78-87.

Abulimiti, A. and T. Schultz. 2020. "Building Language Models for Morphological Rich Low-Resource Languages Using Data from Related Donor Languages: the Case of Uyghur." *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL).* 271-276.

Agenbag, W. and N. Thomas. 2019. "Automatic Sub-Word Unit Discovery and Pronunciation Lexicon Induction for ASR with Application to Under-Resourced Languages." *Computer Speech & Language*, 57: 20-40.

Aikhenvald, Alexandra Y. 2007. "Typological Distinctions in Word-Formation." In *Language Typology and Syntactic Description: Grammatical Categories and the Lexicon, Second Edition - Volume III*, edited by Timothy Shopen, 1-65. Cambridge, UK: Cambridge University Press.

B.F.K. al A'amiri and A.F. Jameel. 2019. "Morphological Typology: A Comparative Study of Some Selected Languages." *Journal of College of Education / Wasit*, 1(37): 709-724.

Alexandrescu, A. and K. Kirchhoff. 2006. "Factored Neural Language Models." *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* 1-4.

Arcodia, G.F. 2012. *Lexical Derivation in Mandarin Chinese.* Crane.

Arisoy, E. et al. 2009. "Turkish Broadcast News Transcription and Retrieval." *IEEE Transactions On Audio Speech and Language Processing*, 17: 874-883.

Arısoy, E. et al. 2008. "Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages." *Speech Recognition: Technologies and Applications*, 10: 193-204.

Bakaev, I.I. and T.R. Shafiev. 2020. "Morphemic Analysis of Uzbek Nouns with Finite State Techniques." *Journal of Physics: Conference Series.* IOP Publishing.

Bejan, C. 2017. *English Words: Structure, Origin and Meaning.* Addleton Academic Publishers.

Bengio, Y. et al. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research*, 3: 1137-1155.

Bilmes, J.A., and K. Kirchhoff. 2003. "Factored Language Models and Generalized Parallel Backoff." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003—Short Papers - Volume 2.* Stroudsburg, PA: Association for Computational Linguistics. 4-6.

Bosch, S.E. and L. Pretorius. 2017. "A Computational Approach to Zulu Verb Morphology within the Context of Lexical Semantics." *Lexikos* (scieloza), 27: 152-182.

Bosch, S.E. et al. 2008. "Experimental Bootstrapping of Morphological Analysers for Nguni Languages." *Nordic Journal of African Studies,* 17: 66-88.

Botha, J. and P. Blunsom. 2014. "Compositional Morphology for Word Representations and Language Modeling." *International Conference on Machine Learning*, 1899-1907.

Brown, P.F. et al. 1992. "Class-Based N-gram Models of Natural Language." *Computational Linguistics* (MIT Press), 18: 467-479.

Brown, T.B. et al. 2020. "Language models are few-shot learners." *arXiv preprint* . arXiv preprint. doi:arXiv:2005.14165.

Byamugisha, J. et al. 2016. "Pluralising Nouns in isiZulu and Related Languages." *International Conference on Intelligent Text Processing and Computational Linguistics*, 9623: 271-283.

Cai, M. et al. 2017. "An Open Vocabulary OCR System with Hybrid Word-Subword language models." *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR).* IEEE. 519-524.

Chahuneau, V. et al. 2013. “Translating into Morphologically Rich Languages with Synthetic Phrases.” *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* 1677-1687.

Cotterell, R. et al. 2018. “Are All Languages Equally Hard to Language-Model?” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* 536-541

Creutz, M. et al. 2007. “Morph-Based Speech Recognition and Modeling of Out-of-Vocabulary Words across Languages.” *ACM Transactions on Speech and Language Processing.* (Association for Computing Machinery), 5: 1-29.

Croft, W. 2002. *Typology and Universals.* Cambridge University Press.

Darwish, K. 2002. “Building a Shallow Arabic Morphological Analyser in One Day.” *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages.*

de Schryver, G. and D.J. Prinsloo. 2004. “Spellcheckers for the South African Languages, Part 1: The Status Quo and Options for Improvement.” *South African Journal of African Languages*, 24: 57-82.

Demberg, V. 2007. “A Language-Independent Unsupervised Model for Morphological Segmentation.” *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* 920-927.

Doke, C.M. 2017 [1954]. *The Southern Bantu Languages: Handbook of African Languages.* Routledge.

Duvenhage, B. 2019. “Short Text Language Identification for Under Resourced Languages.” *33rd Conference on Neural Information Processing Systems (NeurIPS 2019).* Vancouver, Canada: arXiv.

Etchegoyhen, T. et al. 2018. “Neural Machine Translation of Basque.” (European Association for Machine Translation).

Faaß, G. et al. 2009. “Part-of-Speech Tagging of Northern Sotho: Disambiguating Polysemous Function Words.” *Proceedings of the First Workshop on Language Technologies for African Languages*, 38-45.

Gerz, D. et al. 2018. “On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling.” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 316-327.

Goldberg, Y. 2017. “Neural Network Methods for Natural Language Processing.” *Synthesis Lectures on Human Language Technologies*, 10(1): 1-309.

Gupta, V, and Boulianne, G. 2020. "Automatic Transcription Challenges for Inuktitut, a Low-Resource Polysynthetic Language." *Proceedings of The 12th Language Resources and Evaluation Conference,* 2521-2527.

Guthrie, M. 2017 [1948]. "Full Classified List of the Bantu Languages." In *The Classification of the Bantu Languages bound with Bantu Word Division*, 65-74.

Haspelmath, M. 2011. "The Indeterminacy of Word Segmentation and the Nature of Morphology and Syntax." *Folia Linguistica*, 45 (1): 31-80.

Hedler, F. 2016. "The Golden Age of NLP." *Research Live, MRS, London.* Available online at https://www. research-live. com/article/opinion/the-golden-age-of-nlp/id/5003342.

Hildebrandt, K.A. 2014. "The Prosodic Word." In *The Oxford Handbook of the Word,* edited by John R Taylor. Oxford. doi:10.1093/oxfordhb/9780199641604.013.035.

Hirsimaki, T. et al. 2009. "Importance of High-Order N-Gram Models in Morph-Based Speech Recognition." *IEEE Transactions On Audio Speech and Language Processing,* 17: 724-732.

Janson, T. 1991/2. "Southern Bantu and Makua." *Sprache und Geschichte in Africa,* 12/13: 63-106.

Jayan, V. et al. 2015. "Difficulties in Processing Malayalam Verbs for Statistical Machine Translation." 6: 13-24.

Jing, K, and Xu, J. 2019. *A Survey on Neural Network Language Models.* eprint arXiv:1906.03591.

Joshi, P. et al. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Jurafsky, D, and J.H. Martin. 2018. "Speech and Language Processing (draft)." *Chapter A: Hidden Markov Models,* 19: 2019.

Khumalo, L. 2020. "Corpora as Agency in the Intellectualisation of African Languages." In *The Transformative Power of Language: From Postcolonial to Knowledge Societies in Africa,* edited by R. Kaschula and H.E. Wolff, 247-258.

King, B P. 2015. "Practical Natural Language Processing for Low-Resource Languages." Ph.D. Dissertation.

Kornai, A. 2013. "Digital Language Death." 8. e77056.

Kudo, T. 2018. "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates." *arXiv preprint arXiv:1804.10959.*

Kurimo, M. et al. 2006. "Unlimited Vocabulary Speech Recognition for Agglutinative Languages." *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 487-494.

Labeau, M, and A. Allauzen. 2017. "Character and Subword-Based Word Representation for Neural Language Modeling Prediction." *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 1-13.

Lajish, V, L. et al. 2015. "A Morpheme Based Language Model for Malayalam Spoken Short Query Processing." Goa: Acoustical Society of India.

Liu, Y, and M. Lapata. 2019. "Text Summarization with Pretrained Encoders." *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP):* http://dx.doi.org/10.18653/v1/D19-1387

Mahlaza, Z. and C.M. Keet. 2019. "A Method for Measuring Verb Similarity for Two Closely Related Languages with Application to Zulu and Xhosa." *South African Computer Journal,* 31: 34-56. doi:http://dx.doi.org/10.18489/sacj.v31i2.698.

Maho, J.F. 2001. "A Comparative Study of Bantu Noun Classes." (elibrary.ru).

Mihajlik, P. et al. 2010. "Improved Recognition of Spontaneous Hungarian Speech—Morphological and Acoustic Modeling Techniques for a Less Resourced Task." *IEEE Transactions On Audio Speech and Language Processing*, 18: 1588-1600.

Mihajlik, P. et al. 2007. "A Morpho-Graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages-Like Hungarian." *Eighth Annual Conference of the International Speech Communication Association.*

Miti, L. 2006. *Comparative Bantu Phonology and Morphology: A Study of the Sound Systems and Word Structure of the Indigenous Languages of Southern Africa.* Centre for Advanced Studies of African Society.

Mjaria, F. and C.M. Keet. 2018. "A Statistical Approach to Error Correction for isiZulu Spellcheckers." *2018 IST-Africa Week Conference (IST-Africa).*

Mulcahy, L, and S. Wheeler. 2020. "'Couldn't You Have Got a Computer Program to Do That for You?' Reflections on the Impact that Machines Have on the Ways We Think About and Undertake Qualitative Research in the Socio-Legal Community." *Journal of Law and Society* (Wiley Online Library), 47: 149-163.

Nchabeleng, M, and J. Byamugisha. 2020. "Evaluating the Effectiveness of the Standard Insights Extraction Pipeline for Bantu Languages." *Advances in Information Retrieval,* 159-172.

Ndaba, et al. 2016. "The Effects of a Corpus on isiZulu Spellcheckers Based on N-grams." *2016 IST-Africa Week Conference.* IEEE.

Nurse, D. and G. Philippson. 2006. *The Bantu Languages.* Routledge.

Pereira, D,B Bastos, and Paraboni, I. 2007. “A Language Modeling Tool for Statistical NLP.” *Anais do V Workshop em Tecnologia da Informação e da Linguagem Humana.* Rio de Janeiro, RJ: TIL. 1679-1688.

Petroni, F. et al. 2019. “Language Models as Knowledge Bases?” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics. 2463-2473. doi:10.18653/v1/D19-1250.

Pretorius, L. and S. Bosch. 2009. “Exploiting Cross-Linguistic Similarities in Zulu and Xhosa Computational Morphology.” *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages–AfLaT 2009,* 96-103.

———. 2003. “Towards Second-Generation Spellcheckers for the South African Languages.” 6th International 'Terminology in Advanced Management Applications' Conference.

Prinsloo, D.J. and G. de Schryver. 2004. “Spellcheckers for the South African Languages, Part 2: The Utilisation of Clusters of Circumfixes.” *South African Journal of African Languages* 24: 83-94.

Sak, H. et al. 2012. “Morpholexical and Discriminative Language Models for Turkish Automatic Speech Recognition.” *IEEE Transactions On Audio Speech and Language Processing*, 20: 2341-2351.

Sapir, E. 1921. *An Introduction to the Study of Speech.* New York: Citeseer.

Schwartz, L. et al. 2020. “Neural Polysynthetic Language Modeling.” 2005.05477

Shannon, C.E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal*, 27: 379-423, 623-656.

Shosted, R.K. 2006. “Correlating Complexity: A Typological Approach.” *Linguistic Typology*, 10: 1-40.

Siivola, V. et al. 2003. “Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner.” *Eighth European Conference on Speech Communication and Technology.*

Tachbelie, M.Y. et al. 2014. “Using Different Acoustic, Lexical and Language Modeling Units for ASR of an Under-Resourced Language-Amharic.” *Speech Communication,* 56: 181-194.

Taljard, E. and S.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.

Thottingal, S. 2019. “Finite State Transducer Based Morphology Analysis for Malayalam Language.” *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages.* 1-5.

Vania, C. 2020. "On Understanding Character-Level Models for Representing Morphology." The University of Edinburgh. Ph.D. Dissertation.

Vergyri, D. et al. 2004. "Morphology-Based Language Modeling for Arabic Speech Recognition." *ICSLP-2004*, 2245-2248.

Zipf, G.K. 1949. "Human Behaviour and the Principle of Least Effort, Adisson." Wesley Press, Cambridge.