Background chapter

Victor Kabine, University of Cape Town

1. INTRODUCTION

This section explores the research that has been done in the field of spell checking using theory driven linguistic model. The theory driven linguistic model makes use of theoretical linguistics approaches. This means that we will focus on the nature and structure of the language. We will look focus on the syntax, basic grammar, basic notions and morphology of the language.

2. SPELL CHECKING AND MORPHOLOGY

A spell checker is a computer program that performs spell checking, in computing, spell checking can be seen as the process of detecting incorrectly spelled words and sometimes giving spelling suggestions on the incorrectly spelled words [Bassil and Alwani 2012]. spell checking is also a sub-field of natural language processing or as it is commonly referred to as computational linguistics [Bassil and Alwani 2012]. Why natural language processing (NLP) is so important is because the aim of natural language processing is to design software that will be able to analyse, generate and most importantly understand natural languages [Ritika Mishra 2013]. Since natural language processing research leads to better understanding of the language, it is clear to see that African languages are at a disadvantage because a significant amount of those languages are part of the less resourced languages of the world [Pretorius and Bosch 2003].

A lot of research has been done to combat this problem, one of the approaches developed has been the implementation of regular expressions for natural language languages [Karttunen et al. 1996]. Regular expressions can be compiled to create finite state transducers [Karttunen et al. 1996]. The finite state transducer is used to implement the morphological rules of the language, thus focusing on the nature and structure of the language [Karttunen et al. 1996]. regular expressions can be seen as a type of search string that can be used to search for a particular pattern of strings or numbers [Brüggemann-Klein 1993]. we can say that a text can be viewed as a string of characters and because of that we are able to process the language at a character level through the use of regular expressions. This is also called pattern matching [Bird and Klein 2006]. The use of regular expressions to analyse the morphological structure of the isiZulu language. The rich agglutinative morphological structure of the isiZulu language is based on two of the following principles which are the nominal classification system and the concordial agreement system [Bosch and Eiselen 2005].

The concordial agreement system is what governs the entire sentence structure of the isiZulu language. This system ascertains that the grammar of verbs, adjectives and other parts other than the noun are in correlation. The concordial agreement system is essentially brought about by the different noun classes in that the prefixes link the word to [Pretorius and Bosch 2009]. The concordial agreement system is far more complex than that of the nominal classification system. In the nominal classification system, the nouns are assigned prefixal morphemes and then assigned numbers [Pretorius and Bosch 2009]. The numbering of the classification helps to make it easier to distinguish between them as there are 23 noun prefixes in total [Bosch and Eiselen 2005]. Within the isiZulu language, there is also what is known as derivation. This is the combination of morphemes to create a new word, this word will also be found in a different category. Another term that can be

found is the infllectional morphology. In this structure, the morphemes that are added to a word do not change the category of the word [Bosch and Eiselen 2005]. Within the nominal classification system, there have been a number of classification methods that have been suggested in the past however the most widely accepted method of classifying nouns has been that of [Meinhof 1906]. This classification method identifies 17 different noun classes. The classification of these nouns into noun classes is based on the following three criteria: semantics, syntax as well as morphology [Twala 1992].

Class 1 nouns consists of two parts class 1 that contains a prefix of um(u) and class 1a that consists of the prefix u- and these nouns are used to identify a person. Class 2 can be seen as the plural form of class 1, it also consists of two parts class 2a which consists of the prefixes aba-, abe or ab. Class 3 consists of two parts as well, these nouns identify trees, plants and body parts. They have the same prefix as the first class. Class 4 is the plural version of the class 3 and has the following prefixes imi or im. Unlike the first class, the fourth class consists of only one part. Class 5 classifies nouns that identify fruits, body parts as well as words borrowed from another language and they contain the prefix i or ili. Class 6 contains plurals of class 5 and contain the prefix ama or ame. Class 7 contains nouns that describe objects or types of people, they have the following prefix isi or is. Class 8 contains plurals of class 7 and has the prefix izi or iz. Class 9 contains nouns that identify animals and have the prefix im, in or i. Class 10 contain plurals of class 9 and has the following prefixes izin or izim. Class 11 has prefixes ulu or u. Class 14 has the prefixes ubu or utsh. Class 15 has the prefixes uku or uk . Class 17 has the same prefixes as class 15 [Twala 1992].

3. RELATED WORKS

[Bosch and Eiselen 2005] makes use of regular expressions to create a finite state transducer that implements the morphological rules learned above to create a spell checker for the language isiZulu. [Pretorius and Bosch 2009] explores the cross linguistic similarities in terms of morphology between the nguni languages isiZulu and Xhosa. They make use of finite state tools to build the transducer. They tested the transducer using the ZulMorph corpus. [Brüggemann-Klein 1993] explores regular expressions within finite state automata and how regular expressions can be compiled to create a non deterministic finite state automata. [Karttunen et al. 1996] showcases the use of regular expressions in natural languages and the different ways they can be compiled or implemented. [Joubert et al. 2004]. introduces a framework with boot strapping approaches with the purpose of making the creation of linguistic resources quick and cost effective. [Kaur 2014] also introduces regular expressions in the field of natural language processing however this paper also includes a variety of tools that use regular expressions as well as a short description of each of the tools that are mentioned

REFERENCES

Youssef Bassil and Mohammad Alwani. 2012. Context-sensitive spelling correction using google web 1t 5-gram information. *arXiv preprint arXiv:1204.5852* (2012).

Steven Bird and Ewan Klein. 2006. Regular expressions for natural language processing. *University of Pennsylvania* (2006).

Sonja E Bosch and Roald Eiselen. 2005. The effectiveness of morphological rules for an isiZulu spelling checker. *South African Journal of African Languages* 25, 1 (2005), 25–36.

Anne Brüggemann-Klein. 1993. Regular expressions into finite automata. *Theoretical Computer Science* 120, 2 (1993), 197–213.

LJ Joubert, MH Davel, E Barnard, and V Zimu. 2004. A framework for bootstrapping morphological decomposition. (2004).

Lauri Karttunen, Jean-Pierre Chanod, Gregory Grefenstette, and A Schille. 1996. Regular expressions for language engineering. *Natural Language Engineering* 2, 04 (1996), 305–328.

Gaganpreet Kaur. 2014. Usage of Regular Expressions in NLP. (2014).

Carl Meinhof. 1906. *Grundzuge einer vergleichenden Grammatik der Bantusprachen*. Berlin: C. Reimer (E. Vohsen).

Laurette Pretorius and Sonja Bosch. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*. Association for Computational Linguistics, 96–103.

Laurette Pretorius and Sonja E Bosch. 2003. Finite-state computational morphology: An analyzer prototype for Zulu. *Machine Translation* 18, 3 (2003), 195–216.

Navjot Kaur Ritika Mishra. 2013. A Survey of Spelling Error Detection and Correction Techniques. *International Journal of Computer Trends and Technology (IJCTT)* 4, 3 (2013), 372–374.

Edith K. Twala. 1992. *The noun class system of isiZulu*. Master's thesis. University of Johannesburg, South Africa.