Surface realization for Nguni languages by focusing on weather summuary generation

Masters Proposal

Zola Mahlaza
Supervisor : Dr C. Maria Keet

ABSTRACT

There is a significant imbalance in the study of languages in South Africa within computer science. The South African government has been attempting to reform the imbalance through funding of academic work from its various departments. The efforts are a result of its constitutional responsibility of taking "practical and positive measures to elevate the status and advance the use of [South African] languages" [8] whose language family hails from Africa. These languages have seen significant neglect and suppression as a result of apartheid's exclusivism and prejudices. This work is a contribution to the reformation, and it is the investigation of methods of producing texts in Nguni languages from machine representations of knowledge. Specifically, we investigate the grammatical similarities between isiZulu and isiXhosa. These are languages which are part of the Nguni language group spoken chiefly in Southern Africa. The similarity will assist in determining whether a grammar can be reused between the languages when generating text thus saving development time.

1. INTRODUCTION

The history of South Africa and its general mood towards languages has resulted in human language technologies (HLTs) being seen as a priority by the government [13]. South Africa is a multilingual country with approximately 25 spoken languages, and 11 official languages. According to the country's 2011 census [1], three languages with the most first language speakers are isiZulu, isiXhosa and Afrikaans (in that order). English is only the forth on the list, yet it is the main language used in public and official discourse for most institutions. Furthermore, it is extensively studied in academia. This is because South African languages of foreign origin have seen significant investment than, and oftentimes at the expense of, other South African languages for various reasons. This is being corrected, according to Grover et al [13], through academic funding of HLTs received from the Department of Arts and Culture (DAC), Department of Science and Technology (DST), and National Research Foundation (NRF).

Natural language processing (NLP) tools and other HLTs that work with South African languages could have a very positive impact in the country. An area of NLP that is of interest to us is natural language generation (NLG). It is the study of techniques involved in the production of natural language texts from machine representations of knowledge. It has made it possible for computers to explain medical data to patients, summarise statistical data, etc [24, p2]. These are the benefits we would like to see for South African languages. However, we have not observed significant academic literature, if it exists, on NLG focusing on Nguni languages. This a group comprised of four languages which are spoken by a large population in Southern Africa.

Our work is an attempt to lay the groundwork for NLG in Nguni languages by focusing on the methods for auto-generation of weather reports for isiZulu and isiXhosa. This goal is made difficult the by the complex nature of these languages. NLP experts address the complexity of natural languages by first focusing on controlled languages. These are domain specific languages with a simplified grammar and restricted vocabulary. This is the reason our work focuses on weather generation. The use of this domain restricts the language constructs thus making the scope of the project manageable. In addition to using domain restriction, it is important to note that not all four languages can be dealt with at the same time. We will focus on isiZulu and isiXhosa. These are languages which have similarities, like the other Nguni languages. It is often assumed that the aforesaid similarities can be exploited when building Nguni language technologies. This assumption is the motivating factor to the work done by Pretorius and Bosch [23] on their isiZulu morphological analyser, ZulMorph. Our work will attempt to quantify the similarity between isiXhosa and isiZulu. This will help us in determining whether it is possible to use the same grammar rules for both languages when generating sentences in a specific domain. Finally, we will investigate the improvement that is brought on, if any, by the use of phonological conditioning rules in NL generation.

2. RELATED WORK

Natural language generation has uses in a variety of domains. The Bateman & Zock [34] list of NLG systems shows a variation of systems that exist in a number of fields ranging from systems which can produce flight information [2] to systems which can produce biographies [17]. The work done by Reiter and Dale [24] [25] details the principles of building NLG systems. They discuss not only the required material when building NLG systems, but they also go into detail explaining the tasks involved in the process of converting data into text. They present an architecture made up of three major steps. These are document planning, micro-planning and realisation. The pipeline encompasses a large number of steps which are not all compulsory for every NLG systems. These are content determination, discourse planning, sentence aggregation, lexicalisation, referring expression generation, syntax and morphology.

There has been variation in the approaches of developing NLG systems. Cimiano et al [7] point out that first generation systems used to rely on template-based approaches. They were followed by the use statistical architectures. Current systems, like those developed by Cimiano et al [7] and their contemporaries, use a combination of these two techniques. The work done by Cimiano et al [7] is the development and evaluation of a system capable of converting RDF (Resource Description Framework) data to a natural language. This work depends on the use of an ontology and an ontology lexicon. They follow a three step data-to-text pipeline which comprises of document planning, microplanning and surface realisation. Their hybrid development approach is such that only the last two stages make use of statistical techniques.

The work relies heavily on the ability to represent lexicon data through lemon (Lexicon model for ontologies) and the use of a lexicon database to retrieve inflectional variants. They also have access to a large domain corpus. Their approach cannot be easily be reproduced for Nguni languages because there is no large corpus of high quality, and at the moment there is no database to retrieve inflectional variants for South African Bantu languages. Lastly, Chavula and Keet [6] have shown that lemon, as-is, does not work with Bantu languages. They have proposed the use an additional ontology in order to scaffold lemon in order for it to be able to deal with the noun class systems which are a feature of Bantu languages.

The variety of systems, both in approach and function, means that there is a difference in the inputs expected by each system. For instance, Klein [19] developed a system capable of summarising essay paragraphs. The input to the aforementioned system is a paragraph of text. Davey [9] developed an NLG system whose input is a game of tic-tac-toe and produces English text describing the current state of the game. FOG [12], the weather NLG system which resides in the Forecast Production Assistant (FPA) system is comprised of three steps; data extraction, conceptual processing, and linguistic processing. All the steps in the work done by Cimiano et al [7] would make up the linguistic processing in FOG. The inputs to FOG are charts which are developed by a forecaster on the FPA. This wide variety shows us that NLG systems work with different types of data hence the forms of representation of said data should be dependent on the system. Common forms include conceptual graphs, RDBMSs, RDF/RDFS and OWL.

Bouayad-Agha et al [4] say that in spite of the numerous kinds of possible inputs to NLG systems, the 'natural' inputs are semantic/conceptual representations. This is because the said inputs result in more flexible NLG systems. Furthermore, the authors point out that the pipeline architecture presented by Dale and Reiter is not the only architecture in existence. There exists revision-based approaches, optimization approaches and monolithic approaches which map content into text [4, p3]. In all of these approaches, the complexity of the system is determined by the triplet inputs, context and output. There are other minor issues which contribute to the complexity, such as whether the system exists in isolation or within a larger system.

2.1 Foundations and Language similarities

A key aspect that cannot be forgotten is that NLG does not only require information pertaining to the application domain, Reiter and Dale [24] point out that it also requires knowledge about the language. It is for this reason that the work done by Twala [30] is relevant. In her dissertation, she discusses the evolution of the grouping of nouns in isiZulu by looking at the groupings presented by numerous authors over the years. The author also provides a comprehensible overview of the morphological, syntactic and semantic details for each noun class.

There are similarities between isiZulu and isiXhosa, but, be that as it may, the pond of academic literature attempting to quantify and/or document these differences has been moderately dry. An important notion that is brought forward by the likeness of Bantu languages is the generalization of techniques which currently apply to a specific language within this group to other Bantu languages. The work done by Pretorius & Bosch [23], which falls under natural language understanding (NLU), is evidence to that. The work in question attempted to document some differences pertaining to morphotatics and morphophonological alternations between the two languages. Their goal was to bootstrap the development of an isiXhosa morphological analyser by using their current prototype of an isiZulu morphological analyser, ZulMorph. This has implicit assumption the differences between the two languages are relatively small. The aspect of their work that is of interest to us is their enquiry into morphotatics. This refers to the ordering of morphemes when forming words. The authors point out that there are some differences in the workings of affixes between the two languages. For instance, isiXhosa unlike isiZulu has a temporal form for verbs and its role is to indicate when an action occurs [23, p98-p99]. The simple example given by Pretorius & Bosch to illustrate this point, however, is wrong as it violates the juxtaposed vowel prohibition (unless the vowels are the same) in isiZulu and isiXhosa. A better example is given in [20, p6] which shows future and past tense with the sentence "I have arrived in Grahamstown".

1. Ndi-fik-e e-Rhini
   (SC-arrive-PST LOC-Grahamstown)
2. Ndi-zaku-fik-a e-Rhini
   (SC-FUT-arrive-FV LOC-Grahamstown)

The sentences above are in isiXhosa and the abbreviations used in the above example are; SC the subject concord, PST is past tense, FUT is future tense, FV is final vowel and LOC is locative.

There are no methods for verbalising concepts from machine representations in isiXhosa, to our knowledge. This is why the work done by Keet and Khumalo [16] becomes a foundational aspect of our work. Keet and Khumalo [16] investigated the formation of methods which would allow one to be able to create an isiZulu controlled natural language (CNL). A CNL is a subset of a specific natural language. The difference is that the grammar and vocabulary is restricted. They have shown that a template based NLG system will not work with isiZulu, and correspondingly other Nguni languages, due to the complex noun class system. They developed verbalization patterns in isiZulu for logic constructs such as subsumption, negation, universal and existential quantification. Furthermore, they have also shown that a template based system that makes use of the developed verbalisation patterns will also not work - a "full-fledged grammar engine" [16, p23] is required. The transferability of these patterns to other Nguni languages such isiXhosa seems like a possibility. This is thanks to similarities between the Nguni languages.

There already has been work done to recycle the methods developed for isiZulu and reuse them for another Bantu language. There is the particular case were they were used as a scaffold in achieving the same goal for Runyankore. This is a language spoken in Uganda and other Central/East African countries such as Burundi and Kenya. Runyaronke shares similarities with isiZulu, however, the two languages also have some differences. An example of a difference is the observation that isiZulu has five distinct tenses whereas Runyankore has fourteen [5, p2]. The verbalization patterns faced similar issues and according to Byagumisha et al these are due to factors such as "the noun class of the name of the concept, the category of the concept, whether the concept is atomic or an expression, the quantifier use in the axiom, and the position of the concept in the axiom" [5, p7]. Nonetheless, their work provided more evidence that the bootstrapping approach when building human language technologies for Bantu languages significantly reduces development time and requires less resources.

A context free grammar (CFG) is a set of rules which determine how to form sentences/words in a formal language. Formal languages are not necessarily equivalent to natural languages. CFGs can be used, however, to model a controlled natural language. Controlled natural languages not only restrict vocabulary, they also restrict morphological forms, grammatical constructions, semantic interpretations and pragmatics [33]. The benefits of CNLs are that they can be realized through the use of grammar formalisms such as CFGs. There are also other forms of grammars which can be used towards this goal. Other examples include context sensitive grammars (CSG), probabilistic CFGs (PCFG), etc. CFGs and their variants, to our knowledge, see more use in NLG compared to CSGs. The scarcity of NLG systems that makes use of CSGs might be due to a number of issues. The main reason is the observation which was made by Simmons and Yu. They argued that context sensitive grammars were not attractive because they are conceptually and computationally difficult to deal with [28, p392]. It is for this reason that the class of grammars, mildly context-sensitive grammars, exists. This is a group of grammars which are more powerful than CFGs as they include the notion of context, however, do not face the same challenges as a CSG. Formalisms within this group, to name a few, include the tree-adjoining grammar (TAG), head grammar (HG) and the combinatory categorial grammar (CCG). The last of which was investigated by Karagol-Aya [15] who attempted using it to model the morphotactics and syntax of Turkish. CCGs are generally used when mapping natural language to a logic form in NLU. However, Karagol-Aya [15] uses an adapted version of the "semantic head-driven bottom-up generation" [15, p139] algorithm to generate a natural language.

There are authors who have taken alternative approaches in the treatment of language in the development of tools to assist in realization. An example is simpleNLG, a so-called realization engine. Its responsibility, unlike traditional surface realisers, is to create the lexicon and syntactical representations along with actions to allow a developer full control over the realization process. The major distinction is that its key feature is the ability to move the responsibility of "making appropriate linguistic choices given the semantic input" [11, p90] out of the realization tool and into the hands of the developer. A benefit of this is that the engine does not require a strict input formalism hence giving the developer flexibility in deciding on any suitable representation for the realization process. This tool is made exclusively for English. It has been adapted to other languages such as French [31], German [3] and Brazilian Portuguese [10].

2.2 Weather text generation

According to the Bateman and Zock [34] list, the automation of the production of weather summary text is the second most favourite application of NLG systems. It follows behind health-care/medicine. An example of such a system is the Forecast Generator (FOG), it is capable of creating forecast test summaries from weather maps. FOG exists within a bigger system called the Forecast Production Assistant (FPA). The goal of the FPA is the automation of routing aspects of weather reporting in order to allow forecasters to focus on "scientific questions"[12, pg45]. The key aspect about the FOG is that it is bilingual. FOG had two lexical challenges it attempted to solve. These are deciding which professional words to use when describing weather concepts and how to generate text in two languages from the same input. The authors dealt with the first challenge by using words which were decided upon by the forecasters. The second challenge was dealt with the introduction of an abstract interlingua that will capture the syntax irrespective of language. This interlingua will be used when generating a deep syntactic representation for each language. The deep syntactic representation is further used when determining the surface syntax. The introduction of different surface syntax representations is done because FOG will make use of independent language grammars to map surface syntax into text.

The authors point out that the deep syntactic representation is capable of modelling English and French because these languages have a similar communication/semantic style. This representation is not guaranteed to be passable for other languages. It is not clear whether such a technique is necessary for isiZulu and isiXhosa. It would only be necessary in the event that isiXhosa and isiZulu cannot be realized using the same grammar rules.

They made use of Meaning-Text Theory (MTT) because it is flexible thus can “accommodate a variety of forecast types, and regional needs and tastes”[12, pg50] and it allows better maintenance of the system. They also chose to use this model because they needed to support speech synthesis. The reason for this was that they needed the system to work with telephone-answering systems. This requirement is great as it leads to accessibility of weather services to people who are visually impaired. MTT is not a new theory/model. Kittredge et al [18] say that it was first proposed by Igor Mel'čuk in his 1970 and 80's work [21] [22]. A number of its qualities are not necessary when dealing with controlled natural languages. We will not be investigating it due to its complexity, our limited time frame and lack of literature pertaining to the use of MTT with Nguni languages. It might be, in the future, worth investigating the use of MTT for the generation of text in Nguni languages.

It is important that NLG systems account for possible imprecision in weather forecasts. The work done by Sripada et al [29] for the UK's national weather service generates predictions for numerous days. It accounted for the loss in accuracy in the text it generated to makes sure that users of the system are not mislead. Sripada et al [29] also faced the issue of a lack of a corpus when building the system. They dealt with this by obtaining text samples from experts and supplementing them by a domain language; weatherese. It is not clear whether or not their evaluation method verified that the respondents of their survey were not weather experts. The internal details and techniques used in the system are not given. However, we know that it is based on the Arria NLG engine [27]. The engine is a commercial product that is used in areas such as financial services, advertising and marketing, oil and gas, etc. The technical overview does claim that it is language-independent. The engine is based on the traditional techniques which have already been discussed here. It has other aspects/modules which are incorporated for commercial appeal. Unfortunately, there is no evidence to support the idea that it could be used with Bantu languages.

There also has been variation in the approaches in NLG systems built within the weather domain. The work done by Winkler et al [32] uses the idea of a catalogue-based system. They mention that the idea has been used in generating severe weather warnings before but has never been used with a complex sentence type and a bigger domain. They built a system for the Swiss avalanche bulletin capable of producing text in German, French, Italian, and English. The system uses a collection of sentence templates where each sentence is split into at most 10 segments. This approach, however, is a more advance form of templates. We can therefore deduce that it will suffer the same constraints as templates when it comes to Bantu languages

3. PROBLEM STATEMENT

In our examination of the current state and use of Nguni languages, we have observed that there is no fast and large scale producer, automated or otherwise, of weather summaries in said languages. This is due to several factors such as (1) There are multiple Nguni languages, each of which has numerous dialects and hiring human authors to interpret weather data and produce these summaries is expensive and inefficient, (2) There is no automated system to achieve the stated goal because of, among other things, the complexity of Nguni languages (which is due to their noun class systems and concordial agreement) and the shortage of computer scientists working with Nguni languages.

Furthermore, the grammatical similarity between isiZulu and isiXhosa has never been formally quantified. The qualification would us allow to determine whether the same grammatical constructs can be reused between the two languages.

4. AIM

The aim of this work is to develop techniques of generating weather reports in isiZulu and isiXhosa. We will attempt to eliminate the dependence on a single language for an NLG system in order to be able to generate texts in two closely related languages using the same grammar. The ability to generate text for both isiZulu and isiXhosa depends on our quantification of the similarities between isiXhosa and isiZulu. Lastly, we investigate the degree to which a set of phonological conditioning rules can improve text thus giving NLG developers for Nguni languages more information on how to prioritise phonological conditioning.

5. RESEARCH QUESTIONS

1. How similar are isiZulu and isiXhosa?
2. Can the same grammar rules be used for both languages to produce comprehensible sentences?
3. What is the degree of improvement in grammatical correctness can be brought on by the introduction of phonological conditioning rules?

6. METHODS

We will investigate technologies which are capable of sentence derivation using some context-aware grammar. We will then pick a suitable grammar formalism and tools for generating sentences. We will determine the weather concepts and the coverage of all syntactic forms/features with respect to the weather. This will be followed by the incremental development of a grammar for each language, restricted by the determined coverage. This will be done until we're capable of generating $>= 75\%$ correctness for the generated sentences. The posed first research question put forth the notion of similarity. This necessitates the use of a metric to determine this similarity. We will devise metrics which capture the effect of tense in the grammar, the order of certain sentence constituents and therefore semantic style. Furthermore, we will attempt to create a unified grammar should our metrics show that there is a similarity. The scale of deciding whether a similarity exists will depend on the developed metrics. The third research question shall be tested by implementing the three main vowel processes gliding, coalescence and deletion. We will then measure the correct values that are given by these processes.

The focus of our work is in surface realization, a step which does not exist in isolation and as such, we will assume the pipeline NLG architecture which has three major steps; document planning, micro-planning and realisation. We will not focus our attention on building a system capable of generating text for different audiences. It is for this reason and others that are not mentioned that will see document planning and micro-planning will not be given much attention. Nonetheless, the construction of entities and relations to be used is still required and it will done manually.

The most popular approach for the analysis of target text is corpus-based. We will make no attempt to use experts from organizations such as the South African Weather Service (SAWS). This is because the dependence on a second party to produce the text/terms is not guaranteed to be finished within a reasonable timeline. The provisional recourse is to make use of English sources of weather terms and language. We will study the work done by Reiter et al [26] which details the techniques they have used when choosing words to use for their generated weather summaries. We will manually translate the words by hand and supplement the translation with the use of literature [14] and, should it be necessary, make use of professional translation services.

We are aware that translations may sometimes distort meaning and fail to capture the essence of the original sentence. Language is not a fixed entity. The same language may sometimes vary across a group of people who have different social attributes such as ethnicity, religion, education, etc. This is a factor that will need to be considered when evaluating the correctness for the generated sentences. This is the reason why linguists at the School of African Languages and Literature at UCT will be used to determine correctness of the generated sentences.

7. ANTICIPATED OUTCOMES

We expect to have number of metrics highlighting the similarities between isiXhosa and isiZulu. The metrics which we will develop should not and will not be specific to the weather generation domain. If there are similarities between the two languages, we expect to be able to develop a restricted grammar capable of producing texts for both languages. If said grammar does not exist, we should be able to have sufficient reasons to indicate why the grammar cannot be developed. Additionally, we expect to reveal the importance and therefore the priority of phonological conditioning in generating Nguni sentences. The limitation of our work, however, is that we will not quantify the lexical similarities between isiZulu and isiXhosa. Therefore, there will be no coefficient highlighting that similarity. Furthermore, the metrics we produce may take different forms from lexical similarity coefficients and could possibly have different ranges. Their nature cannot be predetermined as they are dependent on the grammar formalism, its representation and other details. Finally, we should have two grammars capable of producing weather reports in isiXhosa and isiZulu with the correctness level mentioned in section 6.

8. SCHEDULE

The following dates are given in chronological order from 2016 to 2017.

- Thesis Proposal Presentation 21/22 July
- Background Chapter 15 September
- Language similarity results 30 November
- Experimental evaluation 22 March
- Results chapter 27 April
- Introductory chapter 26 May
- Discussion chapter 30 Junes
- Second iteration of dissertation 28 July
- Final submission August

9. REFERENCES

[1] Statistics South Africa. 2011 census : Census in brief. `http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf`. Accessed: 03 May 2016.

[2] Scott Axelrod. Natural language generation in the ibm flight information system. In *Proceedings of the ANLP-NAACL 2000 Workshop on Conversational Systems*, pages 21–26, Seattle, Washington, U.S.A, 2000. Association for Computational Linguistics.

[3] Marcel Bollmann. Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138. Association for Computational Linguistics, 2011.

[4] Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513, 2014.

[5] Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. Bootstrapping a runyankore cnl from an isizulu cnl. In *Proceedings of the 5th Workshop on Controlled Natural Language (CNL'16), Aberdeen, UK, July 25-26, 2016*. Springer LNAI, in press 2016.

[6] Catherine Chavula and C. Maria Keet. Is *lemon* sufficient for building multilingual ontologies for bantu languages? In *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014), Riva del Garda, Italy, October 17-18, 2014.*, pages 61–72, 2014.

[7] Philipp Cimiano, Janna Lüker, David Nagel, and Christina Unger. Exploiting ontology lexica for generating natural language texts from rdf data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 10–19, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[8] South African Constitution. Founding provisions, chapter 1, section 1-6. `http://www.justice.gov.za/legislation/constitution/chp01.html#sthash.DthAk1WL.dpuf`. Accessed: 19 April 2016.

[9] Anthony Davey. *The formalisation of discourse production.* PhD thesis, The University of Edinburgh, 1974.

[10] Rodrigo De Oliveira and Somayajulu Sripada. Adapting simplenlg for brazilian portuguese realisation. *INLG 2014*, page 93, 2014.

[11] Albert Gatt and Ehud Reiter. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 90–93, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[12] Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[13] Aditi Sharma Grover, Gerhard B Van Huyssteen, and Marthinus W Pretorius. The south african human language technology audit. *Language resources and evaluation*, 45(3):271–288, 2011.

[14] Language Inc. Weather terminology (multilingual south africa). `https://www.language-inc.org/en/resources/Resources/Weather\%20Terminology\%20(Multilingual\%20South\%20Africa).pdf`. Accessed: 14 June 2016.

[15] Burcu Karagol-Ayan. Morphosyntactic generation of turkish surface forms. In *ESSLLI Student Papers*, pages 137–144, Utrecht, The Netherlands, 1999.

[16] C. Maria Keet and Langa Khumalo. Toward a knowledge-to-text controlled natural language of isizulu. *Language Resources and Evaluation*, pages 1–27, 2016.

[17] Sanghee Kim, Harith Alani, Wendy Hall, Paul H. Lewis, David E. Millard, Nigel R. Shadbolt, and Mark J. Weal. Artequakt: Generating tailored biographies with automatically annotated fragments from the web. In *Proceedings of the ECAI 2002 Workshop on Semantic Authoring, Annotation & Knowledge Markup, Lyon, July 22-26, 2002*, 2002.

[18] Richard Kittredge, Lidija Iordanskaja, and Alain Polguère. Multilingual text generation and the meaning-text theory. *Proceedings of TMI-88, Pittsburgh, PA, U.S.A, June*, 1988.

[19] Sheldon Klein. Automatic paraphrasing in essay format. *Mechanical Translation and Computational Linguistics*, 8:68–83, 1965.

[20] Xiujie Ma. An analysis of temporal relations in languages : a comparative study of mandarin and isixhosa. Master's thesis, Rhodes University, Faculty of Humanities, English Language and Linguistics, 2013.

[21] Igor A. Mel'čuk. Meaning-text Models - a Recent Trend in Soviet Linguistics. *Annual Review of Anthropology*, 10:27–62, 1981.

[22] Igor A. Mel'čuk and Aleksandr K. Žolkovskij. Towards a functioning'Meaning-Text'model of language. *Linguistics : An Interdisciplinary Journal of the Language Sciences*, 8:10–47, 1970.

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

[24] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(01):57–87, 1997.

[25] Ehud Reiter, Robert Dale, and Zhiwei Feng. *Building natural language generation systems*, volume 33. MIT Press, 2000.

[26] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169, 2005.

[27] Ehud Reiter, Y Sripada, and R Dale. Technical overview: The arria nlg engine. Technical report, Arria NLG, 2015. Accessed: 13 June 2016.

[28] Robert F. Simmons and Yeong-Ho Yu. The acquisition and use of context-dependent grammars for english. *Comput. Linguist.*, 18(4):391–418, December 1992.

[29] Somayajulu Sripada, Neil Burnett, Ross Turner, John Mastin, and Dave Evans. *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, chapter A Case Study: NLG meeting Weather Industry Demand for Quality and Quantity of Textual Weather Forecasts, pages 1–5. Association for Computational Linguistics, Philadelphia, Pennsylvania, U.S.A., 2014.

[30] Edith Khanyisile Twala. The noun class system of isizulu. Master's thesis, University of Johannesburg, 1992.

[31] Pierre-Luc Vaudry and Guy Lapalme. Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, 2013.

[32] Kurt Winkler, Tobias Kuhn, and Martin Volk. Evaluating the fully automatic multi-language translation of the swiss avalanche bulletin. In *Proceedings of the 4th International Workshop, CNL 2014, Galway, Ireland, August 20-22, 2014*, pages 44–54, 2014.

[33] Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert Fuchs, Stefan Hoefler, Ken Jones, Kaarel Kaljurand, Tobias Kuhn, Martin Luts, Jonathan Pool, Mike Rosner, Rolf Schwitter, and John Sowa. On controlled natural languages: Properties and prospects. In Norbert E. Fuchs, editor, *Controlled Natural Language: Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8-10, 2009. Revised Papers*, pages 281–289. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[34] John Bateman & Michael Zock. Bateman/zock list of nlg systems. `http://www.nlg-wiki.org/systems/`. Accessed: 14 April 2016.