Developing Text Resources for Ten South African Languages

Roald Eiselen, Martin J. Puttkammer

North-West University
Centre for Text Technology (CTexT), Internal Box 395, Private Bag X6001, Potchefstroom, 2531, South Africa
roald.eiselen@nwu.ac.za; martin.puttkammer@nwu.ac.za

Abstract

The development of linguistic resources for use in natural language processing is of utmost importance for the continued growth of research and development in the field, especially for resource-scarce languages. In this paper we describe the process and challenges of simultaneously developing multiple linguistic resources for ten of the official languages of South Africa. The project focussed on establishing a set of foundational resources that can foster further development of both resources and technologies for the NLP industry in South Africa. The development efforts during the project included creating monolingual unannotated corpora, of which a subset of the corpora for each language was annotated on token, orthographic, morphological and morphosyntactic layers. The annotated subsets includes both development and test sets and were used in the creation of five core-technologies, viz. a tokeniser, sentenciser, lemmatiser, part of speech tagger and morphological decomposer for each language. We report on the quality of these tools for each language and provide some more context of the importance of the resources within the South African context.

Keywords: Language Resource Development, African Languages, Lemmatisers, Part of Speech Taggers, Morphological Analysers

1. Introduction

One of the central requirements for research, development and evaluation in natural language processing (NLP) is access to high quality linguistic resources, including annotated corpora and core-technologies (Krauwer, 2003; Streiter et al., 2006). The process of collecting, processing and developing these resources is a time-consuming and expensive endeavour, especially in South Africa, which has eleven official languages, ten of which have a scarcity of electronic linguistic resources.

As part of an effort to improve both research and development in the area of human language technology (HLT) in South Africa, the South African Department of Arts and Culture funded a project to develop open-source text resources for ten of the official languages of South Africa. This project concluded in 2013 with the release of domain specific monolingual corpora, parallel annotated corpora, and first versions of language specific tokenisers, sentencisers, lemmatisers, part of speech taggers and morphological decomposers. Even though English is an official language in South Africa, it was not included in this project since most of the required resources are readily available for English.

Nine of the official South African languages are Southern Bantu languages, and can broadly be categorised into two groups, the conjunctively written Nguni languages, isiZulu (ZU), isiXhosa (XH), isiNdebele (NR), and Siswati (SS); and the disjunctively written languages, which includes the Sotho languages Setswana (TN), Sesotho (ST), and Sepedi (NSO); the Tswa-Ronga Xitsonga (TS) and finally Tshivenda (TS) (Prinsloo & De Schryver, 2000). The other official language, Afrikaans (AF), does not fit in this categorisation as it is a compounding Germanic language which does not strictly adhere to the conjunctive versus disjunctive paradigm.

Prinsloo & De Schryver (2002) provide the following example to illustrate the difference between conjunctive and disjunctive languages. The phrase "I love him/her" is written as a single word, *ngiyamthanda*, in isiZulu, while it is written as four separate words in Sepedi, *ke a mo rata*.

Given the diverse set of languages that formed part of this project, focusing on all languages simultaneously made it possible to share experiences and knowledge during the development cycle. This also provided several challenges where differences between the languages required different methods and approaches.

In this paper we provide a brief description of the resources that were developed and the different approaches required to develop these resources efficiently, given time and budget constraints. We firstly focus on the data resources that were developed, after which we describe the technologies developed based on these data resources. We then show evaluation results for the developed core technologies. The conclusion looks at the importance of these resources for the South African language community and how these resources can be used in future development efforts.

2. Data Resources

2.1. Corpora

Developing text resources through the collection and annotation of corpora is an important part of enabling research in the field of HLT (Eskander et al., 2013; Mitkov et al., 1999). In this project, the first step in this process entailed collecting unannotated, monolingual corpora for ten languages. Due to the limited availability of electronic data for many of the languages, the aim of this part of the project was to collect one million tokens per language. For some languages, such as Afrikaans, Sepedi and Sesotho, this was a relatively straightforward target to achieve, since there are large sets of data available from various freely available sources. The biggest constraint in terms of attaining enough data was that all data collected during the project would be made available as open-source resources

and the apprehension of data providers of making data available that would be released without limitations. As a result, most of the corpus data was sourced from South African government websites and documents, with some smaller sets of news articles, scientific articles, magazine articles and prose. Although some parts of the data are parallel, not all of the data available from these sources are available for all languages with the result that some languages have less data than others. After collection of the data, the data was classified according to genre, including newspaper articles, informational texts, discussions and instructional documents.

Language	Corpus size (millions)	Type lexicon	NE lexicon
AF	3.27	19,499	3,413
NR	1.19	50,687	5,148
NSO	2.77	15,491	7,916
SS	1.15	61,850	8,399
ST	2.35	15,426	3,613
TN	1.91	12,407	3,885
TS	1.64	10,101	2,707
VE	1.26	8,850	2,932
XH	1.71	46,136	9,869
ZU	1.64	54,374	11,634

Table 1: Sizes for various collected text resources

Based on the collected corpora, additional processing was performed to produce a frequency-based type lexicon and named entity lexicon for each language. The frequency based lexica were verified with spelling checkers, after which language experts reviewed unrecognised words for correct spelling according to the standard written variant of the specific language.

The variation in size of the corpora and lexica, as seen in Table 1, can be attributed to two factors. Firstly, not all languages had the same amount of available electronic data that could be included in the corpora. Secondly, the variance in lexicon size provides a clear indication of the differences between the language families. The concatinatively written languages have much larger frequency-based lexica than the disjunctively written languages, even in cases where the corpus for the disjunctively written language is much larger. As an example, the Sepedi corpus has 2.77 million tokens, but only 15,491 unique types, while Siswati has only 1.15 million tokens in total, but more than 61,000 unique types. This variance is a feature of the morphological complexity of the two languages: the conjunctive written languages have far more inflectional complexity included in a single token, while the disjunctive languages tend to write different grammatical forms, such as locatives and verbs, as separate words. This phenomenon is also applicable to named entities where the conjunctive languages tend to add inflectional affixes to the stems of named entities, and thus increasing the number of uniquely named entity types. Since Afrikaans is a compounding language, but not conjunctively written to the same degree as the other conjunctively written languages, the number of tokens is closer to that of the disjunctively written languages, although the ratio of types to tokens is larger than that of disjunctively written languages.

2.2. Protocols

The second set of deliverables required for this project was the annotation of a subset of data on four layers, specifically the token, orthographic, morphological, and morphosyntactic layers. In order to attain a high level of annotation accuracy it was important to ensure consistency in the annotation process for each of these layers. With this in mind, sets of annotation protocols for each language and each level of annotation were developed and used as guidelines for the annotation of the data. For each language, four protocols with varying degrees of granularity and refinement were developed, guided by existing international standards, mainly the Expert Advisory Group for Language Engineering Standards (EAGLES, 1996).

On the token layer the corpora was segmented into paragraphs, sentences, multiword expressions, words and punctuation. In order to ensure that the corpora are as error free as possible, the following corrections were indicated on the orthographic layer: non-words (e.g. tabel -> table); confusables (e.g. eye (am) -> I (am)), run-ons (e.g. heruns -> he runs), and splits (e.g. fire man -> fireman). These errors were annotated and corrected, while still preserving the original text. The identification and correction of these errors were performed manually in conjunction with the manual verification of the tokenisation.

On the morphological and morphosyntactic layers, tokens were annotated with lemmatisation, part of speech and morphological analysis information. For each token the lemma and part of speech (Noun, Verb, Adjective, Adverb, etc.) of each word was indicated. A second level of annotation was added for generic token components, such as subject and object concords, roots and transitivity. On the lowest level, each token had full morphological analysis annotated with detailed labelling of each morpheme, for example:

baphindele:

ba[SC:2]	Subject Concord Class2
phind[VRoot]	Verb Root
el[Appl:Ext]	Applicative extension
e [VTerm:Subj]	Verb terminative vowel

In order to facilitate the development of the protocols for all ten languages, it was decided to take a tiered approach to the development process. One language from each of the language families was selected as prototype for the family, and the protocol for that language was developed first. Based on previous experience and expertise it was decided to start with Sepedi and isiZulu as prototypes for the other nine related languages. Because Afrikaans has a different origin and differs significantly from the other languages, both in terms of morphological construction and syntactic structure, it was decided to allow the Afrikaans group to develop their own separate protocol, although it is still

based on the same principles and standards as the other languages.

These initial protocols served as blueprints for the other languages both within the language family and the project as a whole. Most of the linguistic issues that are of interest to this project are addressed in these initial protocols, many of which are not specific to one language, but to broader linguistic features. The remaining languages in the language family could then relatively easily adapt and improve their respective protocols on each of the different layers based on the experience and adaptations in the first three protocols.

The design and implementation of the relevant tag sets for each of the layers of annotation proved to be a very challenging task. Several language experts from various research institutions in South Africa were involved in the process of identifying the appropriate tags and describing those tags in a way that would serve both annotators in the project and researchers using the resources in future. The protocols were mostly completed during the first year of the project, but adaptations and improvements to these initial protocols was ongoing throughout the rest of the project.

These protocols are significant both from the project's perspective as well the broader NLP community in South Africa. Given these protocols, other researchers and developers can extend and reuse the annotated data developed in this project, with full knowledge of the principles used in the annotation process.

2.3. Annotated Data

The next part of the project focused on the development of annotated development and test sets for each of the languages, based on the protocols for each of the different layers. The datasets that would be annotated consisted of a subset of the collected corpora with the aim of annotating approximately 50,000 tokens per language for the development sets and 5,000 tokens for the test sets.

As a measure to improve the worth of the data, it was decided to select aligned data that would allow for automatic annotation strategies, as well as direct comparisons between the results for the different languages. English was chosen as a baseline, with approximately 55,000 English tokens selected, along with the aligned sentences for each of the other South African languages. The result of this strategy was that the size of the annotated data varied, especially between the conjunctively and disjunctively written languages, as can be seen in Table 2.

Language	Annotated tokens (aligned)
AF	58,096
NR	41,014
NSO	65,299
SS	42,049
ST	65,338
TN	65,319
TS	65,483
VE	62,427
XH	44,609
ZU	44,324

Table 2: Token count for aligned annotated corpora.

Every token in each language was annotated on the four layers, as discussed in the protocols. In order to ensure faster and more accurate annotation, three different methods for improving annotation speed and accuracy were implemented during the annotation process.

Firstly, data annotation was performed as a two phase process starting with one language from each of the language families and leveraging the annotations from these initial annotation efforts to produce base-line annotations for the other languages in the same language family. The main advantages of performing the annotations in this way relates to building knowledge of the problematic cases in the annotation process, which was then used to update the existing protocols for all languages with specific examples. Furthermore, the experience gained by the language experts in the first round of annotation could then be shared with annotators for the other languages, improving both their understanding of the scope of the annotation process and their annotation rate.

The second approach and a consequence of the two phase approach is that the annotations and subsequent technologies in one language could be leveraged in some of the related languages in a so-called technology transfer approach, where existing technologies from one language (*L1*) were ported/transferred/re-engineered to another, closely-related language (*L2*). The basic hypothesis is that "[if] the languages *L1* and *L2* are similar enough, then it should be easier [and quicker] to recycle software applicable to *L1* than to rewrite it from scratch for *L2*", thereby taking care of "most of the drudgery before any human has to become involved" (Rayner et al., 1997: 65). One of the most successful applications of this approach was used in the annotation of part of speech for the disjunctive languages. As a first step, the Sepedi corpus was annotated for part of speech, after which a POS tagger for Sepedi was trained using the open-source HunPoS [1] application. The Sepedi POS tagger was then used to automatically annotate Setswana, Sesotho, Tshivenda and Xitsonga. Even though these are different languages, the similarities between the languages are such that the initial automatic annotation accuracy on these languages was

[1] http://code.google.com/p/hunpos/

above 75%. This meant that annotators for these languages only had to review the existing tags, and make changes to tags in less than 25% of the cases they reviewed. With such high initial accuracy levels it reduced annotation times for these languages significantly.

The third measure for ensuring accurate annotation was the development of an annotation environment, LARA II, which enforced limitations on the specific annotations, such as the tag set options, and provided annotation suggestions, such as possible morphological analyses, for particular tokens. This was especially important for the full morphological analysis. Prior to the morphological analysis, rule-based morphological analysers generated many of the possible analyses for particular tokens. Instead of having to select each morpheme and root or stem and assigning the grammatical function of the individual morphemes, the different analyses for the token was made available to the annotator, and they only had to select the correct analysis from the set of possible analyses. This was especially useful for the agglutinative languages where one token can have up to a dozen different analyses with as many as twelve morpheme split points, where each morpheme has to be identified and provided with the grammatically correct class. By providing a list of analyses where the various breakpoints have been included and classes have been assigned to the individual morphemes, both cuts down on the number of errors that annotators make and reduces the amount of time required to assign the relevant analysis.

Although the annotated corpora are still relatively small, they represent some of the most complete annotated data sets available for several of the South African languages, and can be used as starting points for further development of both extended annotated data sets and initial versions of core technologies.

3. Core Technologies

In addition to the corpora collected and annotated in this project, we were also tasked with developing initial versions of core technologies associated with the various annotation layers. These core technologies can be used to extend the annotated resources and can also be used as a base-line for the improvement of the technologies. Furthermore, the technologies can be used as features in other technologies such as chunk parsing, named entity recognition, wordnet development and machine translation systems. Based on the annotated data collected and annotated during the first part of the project, lemmatisers, part of speech taggers and morphological decomposers were developed for each of the ten South African languages. Although not specifically chartered to do so, the nature of these tasks required the development of language specific tokenisers and sentencisers. Tokenisers are especially important for identifying individual tokens and to handle punctuation such as hyphens and full stops in a manner that would allow the other technologies to handle abbreviations and contracted forms correctly. Similarly, part of speech tagging specifically requires correct sentencisation in order to correctly tag tokens within particular sentences. Using beginning and end of sentence markers as part of the feature set is critical. The output from these two modules was used both in the annotation stage and to generate the input for the other core technologies described below.

3.1. Lemmatisers

The first core technology developed for each language was lemmatisers, i.e. modules that find the linguistic normalised form of a word (Plisson et al., 2004). With the exception of Afrikaans, all languages followed a rule-based approach to lemmatisation according to language specific normalisation rules. The normalisation rules for the conjunctively written languages are based on previous research by Bosch et al. (2006) for morphological analysis. The lemmatisers used a similar approach to identify the root or stem of individual tokens and adding the relevant terminative vowel to the stem or root. The major drawback to this approach is that several different analyses are often possible for a single word, some of which overlap with forms that are not the desired stem or root of the word. In these cases incorrect or inappropriate stems are identified, and consequently an incorrect lemma is assigned.

The one exception to this rule-based approach is the implementation of the Afrikaans lemmatiser which used an existing module developed by Groenewald (2006). The reason for using a different approach for Afrikaans stems from the fact that a significant amount of research has already been done for Afrikaans, much of which was reusable within the context of this project. Afrikaans is also far less regular than the other African languages and this significantly impacts the level of accuracy that can be attained with a rule-based approach.

3.2. Part of Speech Taggers

The second set of core technologies that was developed for this part of the project was the part of speech taggers for each language. The POS taggers were trained on the annotated part of speech data using HunPoS, an open-source Hidden Markov Model tagger. The decision to use this tagger was made based on the fact that it could be easily trained and redistributed within the project without requiring intensive feature engineering within the limited scope of the project.

3.3. Morphological Decomposers

After completion of the part of speech taggers, morphological decomposers for each language were developed. The decomposers split tokens into their constituent morphemes, including all constituent affixes and roots for verbs and stems for other parts of speech. As an example, the isiZulu word *ukusebenzisa* (“use”) is split into its constituent morphemes as *u-ku-sebsenz-is-a*, where each affix boundary is marked in conjunction with the verb root. The decomposers are different from morphological analysers where the individual morphemes are identified and assigned tags based on their grammatical function.

The decomposers for the conjunctive languages are rule-based implementations, based on work previously done by Bosch et al. (2006). The basic approach to decomposition is to identify all affixes recursively until no additional

affixes can be found. The remaining constituent is then verified against a lexicon of roots and stems, and only in those instances where a valid combination of affixes along with a valid stem or root is found, will the decomposition be successful. The set of affixes consists of various grammatical classes, including relatives, negatives, verbal extension, concord classes, locatives and various derivational affixes. The disjunctive language decomposers are also rule-based, with rule deductions and implementations based on the 50,000 annotated tokens.

Afrikaans was approached in a different manner, since the structure of the language is significantly different from that of the other South African languages. The Afrikaans morphological decomposer is a combination of the implementations of the lemmatiser developed by Groenewald (2006) and the compound analyser developed by Van Huyssteen & Van Zaanen (2004). In addition to producing lemmas for Afrikaans, Groenewald's implementation also provides the morpheme boundaries for particular tokens. Since Afrikaans is a compounding language, where words from various parts of speech can be combined to form larger morphological units, of which certain items can also be inflected through various affixes, it is necessary to identify both the affix and compound boundaries. As an example the word *appelboompie* ("little apple tree") consists of two compound constituents, where the second constituent is inflected as a diminutive form. The decomposition for this word should be appel-boom-pie, rather than just removing the inflectional suffix *–pie*. This additional complexity is solved by including a compound analyser as part of the decomposition process. The compound analyser determines compound boundaries, and each of the compound constituents are then analysed for inflectional affixes.

4. Evaluations

As with all language technologies, the creation of these resources required evaluations that could establish initial benchmarks for the quality of the core technologies. In addition to the annotated development sets, an additional 5,000 tokens were annotated for each language on each of the four layers. These additional datasets were then used to calculate recall, precision and *F*-score metrics for the lemmatisation and morphological decomposition, as well as accuracy for POS tagging across all languages. At the beginning of the project, the expectation was to reach an *F*-score of 70% for lemmatisation and 80% for morphological decomposition evaluated on morpheme level. For POS tagging, an accuracy of 80% on the simplified tag set was expected. The results presented in Table 3 show how different languages and technologies perform based on these evaluation sets.

	Lemmatisers	POS taggers	Morphological decomposers
AF	88.55%	95.71%	81.90%
NR	80.32%	82.57%	82.26%
NSO	77.90%	96.00%	89.57%
SS	81.60%	82.08%	83.42%
ST	76.43%	92.36%	89.53%
TN	74.86%	96.02%	89.22%
TS	76.09%	89.83%	88.51%
VE	77.54%	88.25%	89.31%
XH	79.82%	84.18%	84.66%
ZU	81.56%	83.83%	85.19%

Table 3: Scores of core technologies for ten South African languages.

These results indicate that there is still a lot of room for improvement, especially for lemmatisation and morphological decomposition, but there are also encouraging results for some of the technologies for particular language families. The rule-based lemmatisation strategy looks to be less effective than the machine learning based approach used for Afrikaans, but the limited amount of training data available in this project did not yield comparable results and further investigation is required to possibly find alternative strategies. The morphological decomposers for the disjunctively written languages perform relatively well, while more research is required for the conjunctively written languages. Similarly, the POS taggers for Afrikaans and the disjunctively written languages perform very well, while the results on disjunctively written languages are still relatively low. These results do, however, provide a baseline that is applicable to future experiments and further improvements.

5. Conclusion

For the continued development of human language technology for resource scarce languages it is essential that standard and widely available resources are developed and distributed. We have described one such effort funded by the South African government to simultaneously develop text resources for ten South African languages. These resources provide a good initial step toward future development and increased reach of language technology for the people of South Africa. All of the resources are available as open-source modules and data that can be used by researchers and developers for the improvement of these resources and extended reach of language technology[2].

Furthermore, these resources can aid the development of other language technology resources, such as named entity recognition systems, chunkers, parsers and language identification systems, and in the development of applications such as machine translation systems, automatic speech recognition and text to speech systems for these languages. These resources will also be used in research to determine the best approaches to solving each

[2] All project resources are available from http://rma.nwu.ac.za/

of the core technologies presented here. Work on implementing different machine learning and rule-based techniques to improve the initial technologies will be critical to both improve the technologies and make them useful to a broader audience.

6. Acknowledgements

The NCHLT Text project reported on in this paper was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Arts and Culture.
The authors would also like to thank the contributors from various South African research institutions and linguistic experts that assisted us on the project.

7. References

Bosch, S., Jones, J., Pretorius, L. & Anderson, W. (2006). Resource development for South African Bantu languages: computational morphological analysers and machine-readable lexicons. In *Proceedings on the Workshop on Networking the Development of Language Resources for African Languages. 5th International Conference on Language Resources and Evaluation*. pp. 38-43.

EAGLES, 1996. Expert advisory group on language engineering standards: recommendations for the morphosyntactic annotation of corpora. EAGLES Document EAG-TCWG-MAC/R.

Eskander, R., Habash, N., Bies, A., Kulick, S., & Maamouri, M. (2013). Automatic correction and extension of morphological annotations. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. pp. 1-10.

Groenewald, H. J. (2006). Automatic lemmatisation for Afrikaans. Potchefstroom: North-West University. (Doctoral dissertation).

Krauwer, S. (2003). The basic language resource kit (BLARK) as the first mile-stone for the language resources roadmap. In *Proceedings of the International Workshop "Speech and Computer", SPECOM 2003*, Moscow, Russia.

Mitkov, R., Orasan, C., & Evans, R. (1999). The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting. In *Proceedings of Corpora and NLP: Reflecting on Methodology Workshop*. pp 1-10.

Plisson, J., Lavrac, N., & Mladenić, D. (2004). A rule based approach to word lemmatization. In *Proceedings of the 7th International Multi-Conference Information Society*. pp. 83-86.

Prinsloo, D. J., & De Schryver, G. M. (2002). Towards an 11 x 11 array for the degree of conjunctivism /disjunctivism of the South African languages. *Nordic Journal of African Studies*, 11(2), pp. 249-265.

Rayner et al. (1997). Recycling Lingware in a Multilingual MT System. In Burstein, J. & Leacock, C. (Eds.), *From Research to Commercial Applications: Making NLP Work in Practice*. Somerset, NY: Association for Computational Linguistics. pp. 65-70.

Streiter, O., Scannell, K. P., & Stuflesser, M. (2006). Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4), pp. 267-289.

Van Huyssteen, G.B., & van Zaanen, M.M. (2004). Learning compound boundaries for Afrikaans spelling checking. In *Pre-Proceedings of the Workshop on International Proofing Tools and Language Technologies*, pp. 101-108.