

PRINTED BY: gerhard.vanhuysteen@nwu.ac.za. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

19

TRANSLATION TECHNOLOGY IN SOUTH AFRICA

Gerhard B. van Huyssteen

NORTH-WEST UNIVERSITY, SOUTH AFRICA

Marissa Griesel

NORTH-WEST UNIVERSITY, SOUTH AFRICA

Introduction

South Africa has a rich and diverse multilingual culture with eleven official languages, namely two Germanic languages (English and Afrikaans), four Nguni languages (Ndebele (isiNdebele), Swati (Siswati), Xhosa (isiXhosa), and Zulu (isiZulu)), three Sotho languages (Northern Sotho (Sesotho sa Lebo or Sepedi), Southern Sotho (Sesotho), and Tswana (Setswana)), and two other Bantu languages (Tsonga (Xitsonga) and Venda (Tshivenda)). These languages are granted official status in chapter one of the Constitution of the Republic of South Africa (6 of 1996), stating that ‘the state must take practical and positive measures to elevate the status and advance the use of these languages’. To this effect, the Pan South African Language Board (PanSALB) was established in terms of the Pan South African Language Board Act (59 of 1995), with the goal to promote multilingualism in South Africa. Recently, the Use of Official Languages Act (12 of 2012) was promulgated, in which various conditions for the use of the official languages by government and other institutions are set in order to further a truly multilingual society. In addition to these acts, various other acts and industry regulations also contribute to create a progressive regulatory environment prescribing the use of multiple official languages. These include, inter alia, the Banking Association of South Africa (2004), and the National Consumer Protection Act (68 of 2008).

Despite the fact that English is only the sixth largest language in South Africa (with 9.6 per cent of speakers indicating English as their home language in the 2011 South African National Census; see [Table 19.1](#)), information in the business, health and government sectors is generally available only in English. Coupled with the fact that only a small portion of official South African government websites are available in all the South African languages (De Schryver and Prinsloo 2000: 89–106), it becomes clear that language practitioners and translators working with South African languages need all the help they can get to create texts in the South African languages as efficiently as possible.

Machine translation (MT) offers an attractive and viable option that is being explored only now on a more widely level in South Africa. However, as is well known, the quality of automated translation is not yet at a level, even internationally, to replace human translators for translation of documents; human involvement in post-process editing of generated translations is still of the utmost importance. This is even more true in the South African context where MT is only available for a few select languages, with quality still reflecting the early days of such MT systems. However, using machine-aided human translation (where a human is responsible for the translation, but uses different technologies to ease and assist with the process) is already very useful and attainable in the South African context.

Table 19.1 South African languages¹

<i>South African languages 2011</i>		
<i>Language</i>	<i>Number of speakers</i>	<i>% of total</i>
Afrikaans	6 855 082	13.5%
English	4 892 623	9.6%
isiNdebele	1 090 223	2.1%
isiXhosa	8 154 258	16%
isiZulu	11 587 374	22.7%
Sepedi	4 618 576	9.1%
Sesotho	3 849 563	7.6%
Setswana	4 067 248	8%
Sign language	234 655	0.5%
SiSwati	1 297 046	2.5%
Tshivenda	1 209 388	2.4%
Xitsonga	2 277 148	4.5%
Other	828 258	1.6%
TOTAL	50 961 443	100%

This article focuses on the history and state-of-the-art of MT research and development in South Africa for South African languages.² We will first provide an overview of the lead-up to MT development in South Africa, highlighting some related research, as well as the development of tools and data that could support MT in South Africa indirectly. Thereafter we give an overview of the first initiatives by the South African government to support the development of MT for South African languages. We then discuss individual research and development projects on MT for South African languages, before describing the Autshumato project, South Africa's first consolidated national MT project for South African languages, in more detail. We conclude with a look-ahead to post-Autshumato initiatives and possibilities for MT in South Africa.

Background: linguistics and language technology in South Africa

Linguistic research in all eleven South African languages has always been a rich field of study. Various aspects of the grammars of most of the languages have long since been described in various scholarly publications; for instance, as early as 1862, Bleek (1862) compared various aspects of the different Bantu languages. However, various political, socio-economic and socio-linguistic factors have slowed down processes of grammatical and lexical standardization, as well as the development of terminology in domains where higher functions are required (e.g. business, the judiciary, science and technology, mainstream media, etc.). Nonetheless, over the past twenty years more and more specialized dictionaries and terminology lists have

PRINTED BY: gerhard.vanhuysteen@nwu.ac.za. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

been developed through the establishment of government-supported national terminology units (not-for-profit companies) for each language. In addition, the national language bodies of PanSALB are responsible for language standardization and the development of orthographies for each of the eleven official languages.

The Bible Fellowship of South Africa has also contributed greatly, albeit unintentionally, to standardization of the South African languages. The Bible is available in all official languages, plus a few local variants like Fanagalo (a pidgin artificially created to support communication between English settlers and the local people, used extensively in the mines of South Africa, and incorporating words and structures from many different languages (Adendorff 2002: 179–198)). Professional language practitioners' forums, such as the South African Translators Institute or ProLingua, have become hubs of both knowledge in human translation practice, as well as sources for data such as personal wordlists and translation memories. These organizations have also become key partners in empowering freelance translators with tools to incorporate electronic resources such as translation memories, electronic dictionaries and term banks into translation practice.

In recent years, the human language technology (HLT) fraternity in South Africa has also become an important enabler, addressing some of the needs of translators and language practitioners. Since the South African HLT industry as a whole is still fairly young, only a few good quality core technologies exist for only some of the official languages. For example, automatic part-of-speech (POS) taggers utilizing different machine learning techniques have been developed for Afrikaans (Pilon 2005) and Northern Sotho (Heid *et al.* 2009: 1–19); lemmatisers for Afrikaans (Groenewald and van Huyssteen, 2008: 65–91) and Tswana (Brits *et al.* 2006: 37–47); morphological analysers for Zulu (Pretorius and Bosch 2003: 191–212; Spiegler *et al.* 2008: 9–12), etc. (see Sharma Grover *et al.* (2011: 271–288) for an overview of technologies and resources available for the South African languages). Most of these and other similar core technologies have yielded good results and can, for instance, be used in pre- and post-processing to improve machine translation output quality.

Spelling checkers, like those developed by the Centre for Text Technology (CTexT) at the North-West University (NWU) in South Africa, can also contribute greatly to the usefulness of an MT system by providing spelling variants, or checking the validity of generated constructions. Languages with conjunctive orthographies (like Afrikaans and Zulu) form new words (and even phrases) by combining words and morphemes; spelling and grammar checkers could play an important role in validating such combinations in the context of MT.

Another related development has been the creation and expansion of wordnets for five South African languages. A good quality wordnet could add valuable linguistic information to any MT system or for MT evaluation, as it includes various semantic relations, definitions and usage examples. The Afrikaans wordnet (Kotzé 2008: 163–184; Botha *et al.* 2013: 1–6) currently has more than 11,000 synsets, and is modelled to the standards set in the Princeton WordNet and the Balkanet project. A joint effort by the University of South Africa (UNISA) and the NWU, funded by the South African Department of Arts and Culture (DAC), also saw the development of wordnets for Northern Sotho, Tswana, Xhosa and Zulu, with more than 5,000 synsets in each of these wordnets. The project received renewed funding from UNISA to expand these wordnets even further, and to add other South African languages from 2012 to 2014.

With a view on automated speech translation in the future, the Meraka Institute at the Council for Scientific and Industrial Research (CSIR) and the speech research group at NWU have been the driving forces behind many projects to create core speech technologies and resources which could eventually be used for spoken MT. These include, inter alia, grapheme-to-phoneme conversion, speech recognition and speech synthesis, as well as a large-scale data collection efforts in various projects. However, no spoken MT system is foreseen for the immediate future.

The South African government and HLT

The establishment of HLT as a viable industry in South Africa has a history extending back to 1988,

with the publication of the LEXINET Report by the Human Sciences Research Council (Morris 1988). This report highlighted the importance of technological developments to foster communication in a multilingual society.

From the 1990s, South Africa was consumed with more pressing political matters, and the next government report to mention HLT explicitly only appeared in 1996. The final report by the Language Plan Task Force of the then Department of Arts, Culture, Science and Technology (DACST) included both short and long term action plans for language equality in South Africa (LANGTAG 1996). As a direct result of this report, a steering committee on translation and interpreting, as part of PanSALB, was established in 1998. A second steering committee, in collaboration with DACST, was formed in 1999, and was tasked to investigate and advise regarding HLTs in South Africa. The report by this joint steering committee was released in 2000, and a ministerial committee was established to develop a strategy for developing HLT in South Africa. The ministerial committee's report appeared in 2002, at which stage DACST split into two sections, viz. Department of Arts and Culture (DAC) and Department of Science and Technology (DST), with DAC retaining the primary responsibility for the development of HLT. (For an overview of the early history of HLT in South Africa, see Roux and du Plessis 2005: 24–38.)

Following the recommendations of a ministerial advisory panel on HLT in 2002, three major research and development projects were funded subsequently by DAC, including a speech project to foster information access via interactive voice response systems (the Lwazi project), a project to develop spelling checkers for the ten indigenous languages, and a project to develop MT systems for three language pairs (the Autshumato project); see the section on 'The Autshumato Project' below for details.

Based on a decision taken by the South African cabinet on 3 December 2008, the National Centre for Human Language Technology (NCHLT) was established in 2009. As one of its first large-scale projects, the NCHLT announced a call for proposals to create reusable text and speech resources that are to serve as the basis for HLT development, to stimulate national interest in the field of HLT, and to demonstrate its potential impact on the multilingual South African society. CText, in collaboration with the University of Pretoria (UP) and language experts across the country, was designated as the agency responsible for the development of various text resources. The following resources have been completed by 2013:

- corpora (one million words for each language);
- aligned corpora (fifty thousand words for each language, aligned on sentence level);
- wordlists for all eleven languages; and
- part-of-speech taggers, morphological analysers and lemmatizers for the ten indigenous languages.

Given all these projects funded by government, it soon became clear that a central repository should be established to manage multilingual digital text and speech resources for all official languages in a sustainable manner, in order to ensure the availability and reusability of this data for educational, research and development purposes. In 2011, CText was appointed to set up

PRINTED BY: gerhard.vanhuysteen@nwu.ac.za. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

the so-called Resource Management Agency (RMA³); the RMA works in close co-operation with the Dutch TST-Centrale. Data hosted by the RMA include broad categories such as text, speech, language related video and multimodal resources (including sign language), as well as pathological and forensic language data. It is now required that all past, current and future HLT projects funded by government have to deliver project data to the RMA, in order to prevent loss of data and to promote reusability of the data. The RMA also aims to position South Africa strategically through collaboration with other similar agencies worldwide, with the long-term vision of becoming the hub for language resource management in Africa.

Early MT projects in South Africa

Since the beginning of this century, when the South African government made it clear that it would be investing in and supporting initiatives for developing HLTs for South African languages, numerous research projects with smaller goals have begun exploring the possibilities that MT could hold for the South African community. One of the earliest projects (established in 2002 at the University of Stellenbosch) concerned the development of an experimental South African Sign Language MT system (van Zijl and Barker 2003). We could unfortunately not find any details on the performance of the system – from the latest publication from the project it seems as if it might still be under development (van Zijl and Olivrin 2008: 7–12).

As Afrikaans has the most available resources (data and core technologies) compared to the other indigenous languages (Sharma Grover *et al.* 2011: 271–288), most of the early developments in MT research for South African languages have had Afrikaans as either the source or target language. Ronald and Barnard (2006: 136–140) showed that, even with very limited amounts of data, a first MT system for translation from English to Afrikaans was indeed possible, using a statistical MT approach. They used a parallel corpus of only 40,000 sentences, and achieved a BLEU score (Papineni *et al.*, 2002: 311–318) of 0.3. Their study also included systems with even smaller datasets (3,800 sentences per language pair), translating from English to Tswana (BLEU = 0.32), Xhosa (BLEU = 0.23), and Zulu (BLEU = 0.29). This study set the scene for machine translation in South Africa, and made it very clear that data collection was a big part of the effort needed to improve the quality of translation output.

Another early project, established in 2003 at the University of the Free State (UFS), was the EtsaTrans project, which built on a rule-based legacy system, Lexica. The EtsaTrans system used example-based MT for domain-specific purposes (i.e. for meeting administration at UFS). Initially, it provided only for English, Afrikaans and Southern Sotho, but later developments also aimed to include Xhosa and Zulu (Snyman *et al.* 2007: 225–238).

Another independent study is that of Pilon *et al.* (2010: 219–224), which investigated the possibility of recycling (port/transfer/re-engineer) existing technologies for Dutch to the benefit of Afrikaans, a language closely related to Dutch. They convert (i.e. as a basic form of translation) Afrikaans text to Dutch, so that the Afrikaans text resembles Dutch more closely. After conversion, they use Dutch technologies (e.g. part-of-speech taggers) to annotate or process the converted text, resulting in the fast-tracking of resources for Afrikaans. Their conversion approach is similar to grapheme-to-phoneme conversion, in the sense that transformations are only applied on the graphemic level, and not, for instance, changing word order, etc. Similarities and differences between these two languages are captured as rules and wordlists, and require very few other resources (such as large datasets and probability estimations usually used in statistical MT methods). Although their recycling approach holds much promise for resource development for closely related languages, as an MT approach it is, of course, inefficient, since it does not deal with translation units larger than words. Pilon *et al.* (2010: 219–224) reported a BLEU score of 0.22 for converting Dutch to Afrikaans (compared to Google Translate's 0.40), and a BLEU score of 0.16 for Afrikaans to Dutch (compared to Google Translate's 0.44).

The Autshumato Project

As mentioned earlier, the Autshumato project was the first investment of the South African government to make MT a reality for South African languages. The aim of the project was to develop three MT systems (English to Afrikaans, English to Northern Sotho, and English to Zulu), an integrated translation environment (incorporating the MT systems in a user-friendly editing environment), and an online terminology management system. It was explicated that all resources and systems should be released in the open-source domain.⁴ The project was funded by DAC, and executed by CText, in collaboration with UP.

The biggest portion of the budget and time for the MT subproject was spent on a drive to gather high-quality parallel corpora for the three chosen language pairs. These efforts commenced in early 2008, and included web crawling (mostly the government domain (gov.za), as this was to be the primary application domain), as well as acquiring personal translation memories, glossaries and other parallel text data from freelance translators and translation companies. Data collection was an on-going effort for the entire duration of the project, and proved to be a more difficult task than anticipated. Web crawling was especially ineffective for Zulu and Northern Sotho, as there simply are not that many parallel texts in these languages available on the web. Translators were also sceptical about sharing their parallel corpora, because of privacy concerns related to their clients. Subsequently the project team at CText developed an anonymizer that replaces names of people, places, organizations, monetary amounts, percentages, etc., in order to ensure that confidential information is not included in the parallel corpora; this proved to be an effective measure to convince some translators and companies to make their data available to the project. As a last resort, the project team decided to commission translations and create a custom corpus. This method is by no means ideal and was costly, but delivered excellent quality data as it was translated professionally.

While data collection continued, development of the three MT systems commenced in 2009 with the English–Afrikaans system. Based on the research of Ronald and Barnard (2006), statistical MT has seemed to be a viable option, and it was decided that the Autshumato systems would be based on the Moses statistical MT toolkit.⁵ Data resources for all three systems include aligned units (sentences), wordlists and translation memories.

Since Zulu is a morphologically rich language with a conjunctive orthography, it poses many challenges for the development of HLTs in general. The English–Zulu system incorporated a very basic, rule-based morphologic analyser, but as it was still in early stages of development, it hindered development more than it helped. Although Northern Sotho is to some degree easier to process morphologically, performance of the English–Northern Sotho system was only slightly better than the English–Zulu system; compare [Table 19.2](#) for a comparison of the three systems.

Table 19.2 Comparison of three MT systems

<i>Language pair</i>	<i>No. of aligned units</i>	<i>BLEU score</i>
English–Afrikaans	470,000	0.66
English–Northern Sotho	250,000	0.29
English–Zulu	230,000	0.26

PRINTED BY: gerhard.vanhuysteen@nwu.ac.za. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

All three systems include a pre-processing module to improve performance (e.g. Griesel *et al.* 2010: 205–210). In later stages of the project, the pre-processing module was further successfully adapted to manipulate the syntactic structure of the English source sentences to be more similar to the Afrikaans target structure, thereby eliminating some of the translation divergences before automatic translation (Griesel 2011). Since data was such a precious commodity in this project, efforts by McKellar (2011) to manipulate available data and selecting the best possible candidate sentences for human translation, were invaluable.

To make these MT systems practically available to translators, an integrated translation environment (ITE) was developed in a second subproject, which commenced in 2010. This computer-assisted translation (CAT) tool supports the workflow by incorporating the MT systems, glossaries, translation memories, and spellcheckers in a single, easy-to-use editing application. The ITE is based on the OmegaT platform,⁶ an internationally recognized base for CAT tools. The ITE was designed and developed with continual inputs and evaluations by translators working for government, ensuring that the application would fit well in an everyday working environment. Since one of the functionalities of the ITE is to update translation memories and glossaries as you translate, these valuable resources are also currently being used to continually improve the MT systems.

The third subproject in the Autshumato project was the development of a terminology management system (TMS). One of the important functions of translators working for government is to keep a log of terminology that they come across while translating school books, government documents and pamphlets. This log serves as a way to standardize terms and encourage their use. The TMS is used for the development and management of a database of terms, including their various translations (in the eleven official languages), definitions, usage examples, images, sounds, mathematical equations, and additional notes by terminologists. It is searchable⁷ by anyone outside of government, but only DAC translators can add terms or edit information. The database is continually expanded, while quality checks are performed regularly to ensure that the term base remains of a high standard.

In the course of these three subprojects, needs also arose for the development of various other tools, either for use by developers or by translators. These include a pdf-to-text convertor, language identifiers for all eleven official languages, text anonymizers (described earlier), and a graphical user interface for alignment of parallel texts on sentence level. These tools were also released on the official project website (see [note 4](#)) under open-source licences.

In addition, the parallel corpora and evaluation sets are also available to download from the project website – also under open-data licences. It is the intention that the Autshumato website, plus the accompanying forum, should become the central hub for the development of MT systems and related tools for the South African languages.

The first phase of the Autshumato project was completed in 2011, and the lessons learned by the development team will serve future projects well. Except for the scientific and technology benefits of the project, one of the most important accomplishments of the project was the engagement of the translation community in the development and eventual uptake of this new technology as an essential part of their workflow. Further uptake is ensured through continual training workshops for government translators, as well as support and maintenance of the existing systems.

Conclusion: the future of MT in South Africa

A few independent research projects and the government-funded Autshumato project have marked the entry of South Africa in the global MT field. Since the conclusion of the first phase of the Autshumato project in 2011, research and development of MT systems and tools for other language pairs gained momentum. For example, Wilken *et al.* (2012) reported on the development of a baseline English–Tswana MT system, using the same syntactic pre-processing techniques described earlier. Griesel and McKellar (2012) continued work on the improvement of the English–Northern Sotho system by utilizing data from the closely related language, Southern Sotho. Sentences from a Southern Sotho corpus that were classified by the language identification tool as Northern Sotho, were added to the

training data. This method improved the translation output quality noticeably, and showed that closely related languages could indeed benefit from pooling available resources.

The fact that the tools available in the Autshumato ITE are available for free in the open-source domain, also led to the development of a community of language practitioners using more sophisticated computer-based translation aids. Training workshops played a vital role in this regard, and also served as a marketing mechanism to draw the attention of businesses and other government departments. Through these workshops it has also become apparent that one of the biggest needs is for customization of translation memories and glossaries.

Resource scarceness is certainly the most pressing drawback for HLT and specifically MT development for the South African languages. As HLT and MT hold the potential to facilitate human–human and human–machine interaction through natural language, the continued investment by government in this budding industry is of vital importance. The South African government’s commitment in this regard is illustrated through DAC’s funding of the development of an English–Tsonga (a minority language) MT system from 2013 onwards, and with the hope of including more language pairs in future. It is an important step by DAC to ensure the momentum created in the Autshumato project does not go to waste, and to further establish MT as an area of interest for researchers, developers, and end-users.

Notes

- 1 <http://www.southafrica.info/about/people/language.htm#Ugo-V5LTw6A>.
- 2 We do not give an overview of machine translation aids developed internationally for South African languages. In this regard, suffice to mention that Google Translate included Afrikaans as one of its first fifty languages, and that performance has increased significantly during the first few years. In September 2013 Zulu was released in Google Translate as a potential language, depending on community feedback.
- 3 <http://rma.nwu.ac.za>.
- 4 <http://autshumato.sourceforge.net>.
- 5 <http://www.statmt.org/moses>.
- 6 <http://omegatplus.sourceforge.net>.
- 7 <https://ctext-data1.puk.ac.za:8080/tms2>.

References

- Adendorff, Ralph (2002) ‘Fanakalo – A Pidgin in South Africa’, in Ralph Adendorff (ed.) *Language in South Africa*, Cambridge: Cambridge University Press, 179–198.
- De Schryver, Gilles-Maurice and D.J. Prinsloo (2000) ‘The Compilation of Electronic Corpora, with Special Reference to the African Languages’, *Southern African Linguistics and Applied Language Studies* 18(1–4): 89–106.
- Banking Association of South Africa (2004) *Code of Banking Practice*. Available at: www.banking.org.za.
- Bleek, Wilhelm, Heinrich Immanuel (1862) *A Comparative Grammar of South African Languages*, London: Trübner & Co.
- Botha, Zandré, Roald Eiselen, and Gerhard B. van Huyssteen (2013) ‘Automatic Compound Semantic Analysis Using Wordnets’, in *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa*, 3 December 2013, University of Johannesburg, Johannesburg, South Africa, 1–6.
- Brits, J.C., Rigardt Pretorius, and Gerhard B. van Huyssteen (2006) ‘Automatic Lemmatisation in Setswana: Towards a Prototype’, *South African Journal of African Languages* 25: –47.
- Griesel, Marissa (2011) ‘Sintaktiese herrangskikking as voorprosessering in die ontwikkeling van Engels na Afrikaanse statistiese masjiënvertaalsisteem’ (Syntactic Reordering as Pre-processing in the Development of English to Afrikaans Statistic Machine Translation), Unpublished MA dissertation, Potchefstroom: North-West University.
- Griesel, Marissa and Cindy McKellar (2012) ‘Sharing Corpora Effectively between Closely Related Languages: A Pilot Study for Improving the Quality of Sepedi-English Machine Translation Output’, Poster presentation at the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 29–30 November 2012, Pretoria, South Africa.
- Griesel, Marissa, Cindy McKellar, and Danie Prinsloo (2010) ‘Syntactic Reordering as Pre-processing Step in

- Statistical Machine Translation from English to Sesotho sa Leboa and Afrikaans', in Fred Nicolls (ed.) *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)–23 November 2010*, Stellenbosch, South Africa, 205–210.
- Groenewald, Handre J. and Gerhard B. van Huyssteen (2008) 'Outomatiese Lemma-identifisering vir Afrikaans' (Automatic Lemmatisation for Afrikaans), *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies: Special Issue on Human language technology for South African Languages* 29(1): 65–91.
- Heid, Ulrich, Danie J. Prinsloo, Gertrud Faasz, and Elsabé Taljard (2009) 'Designing a Noun Guesser for Part of Speech Tagging in Northern Sotho', *South African Journal of African Languages* 29(1): 1–19.
- <http://autshumato.sourceforge.net>.
- <http://omegatplus.sourceforge.net>.
- <http://rma.nwu.ac.za>.
- <http://www.southafrica.info/about/people/language.htm#Ugo-V5LTw6A>.
- <http://www.statmt.org/moses>.
- <https://ctext-data1.puk.ac.za:8080/tms2>.
- Kotzé, Gideon (2008) 'Development of an Afrikaans Wordnet: Methodology and integration', *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies: Special Issue on Human language technology for South African Languages* 29(1): 163–184.
- LANGTAG (1996) *Towards a National Language Plan for South Africa: Final Report of LANGTAG*, Pretoria: Department of Arts, Culture, Science and Technology.
- McKellar, Cindy (2011) 'Dataselektering en –manipulering vir statistiese Engels–Afrikaanse masjienvertaling' (Data Selection and Manipulation for Statistical English-Afrikaans Machine Translation), Unpublished MA dissertation. Potchefstroom: North-West University.
- Morris, Robin (1988) *LEXINET and the Computer Processing of Language: Main Report of the LEXINET Programme, LEXI-3*, Pretoria: Human Sciences Research Council.
- Papineni, Kishore A., Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'BLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002*, 7–12 July 2002, University of Pennsylvania, PA, 311–318.
- Pilon, Suléne (2005) 'Outomatiese Afrikaanse Woordsoortetikettering' (Automatic Afrikaans Part-of-speech Tagging), Unpublished MA dissertation, Potchefstroom: North-West University.
- Pilon, Suléne, Gerhard van Huyssteen, and Liesbeth Augustinus (2010) 'Converting Afrikaans to Dutch for Technology Recycling', in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 22–23 November 2010, Stellenbosch, South Africa, 219–224.
- Pretorius, Laurette and Sonja E. Bosch (2003) 'Finite-state Computational Morphology: An Analyzer Prototype For Zulu', *Machine Translation* 18: 191–212.
- Ronald, Kato and Etienne Barnard (2006) 'Statistical Translation with Scarce Resources: A South African Case Study', *SAIEE Africa Research Journal* 98(4): 136–140.
- Roux, Justus and Theo du Plessis (2005) 'The Development of Human Language Technology Policy in South Africa', in Walter Daelemans Theo du Plessis Cobus Snyman Lut Teck (eds) *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, 22–23 September 2003, Bloemfontein, South Africa, Pretoria: Van Schaik, 24–38.
- Sharma Grover, Aditi, Gerhard B. van Huyssteen, and Marthinus Pretorius (2011) 'The South African Human Language Technology Audit', *Language Resources and Evaluation* 45(3): 271–288.
- Snyman, Cobus, Leandra Ehlers, and Jacobus A. Naudé (2007) 'Development of the EtsaTrans Translation System Prototype and Its Integration into the Parnassus Meeting Administration System', *Southern African Linguistics and Applied Language Studies* 25(2): 225–238.
- Spiegler, Sebastian, Bruno Golenia, Ksenia Shalnova, Peter Flach, and Roger Tucker (2008) 'Learning the Morphology of Zulu with Different Degrees of Supervision', in Sinivas Bangalore (ed.) *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology (SLT 2008)*, 15–18 December 2008, Goa, India, 9–12.
- van Zijl, Lynette and Dean Barker (2003) 'A Machine Translation System for South African Sign Language', in *Proceedings of the 2nd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa, Afrigraph 2003*, 3–5 February 2003, Cape Town, South Africa, 49–52.
- van Zijl, Lynette and Guillaume Olivrin (2008) 'South African Sign Language Assistive Translation', in Ronald Merrell (ed.) *Proceedings of the 4th Annual IASTED International Conference on Telehealth / Assistive Technologies*, 16–19 April 2008, Baltimore, MD, 7–12.
- Wilken, Ilana, Marissa Griesel, and Cindy McKellar (2012) 'Developing and Improving a Statistical Machine Translation System for English to Setswana: A Linguistically-motivated Approach', in Alta de Waal (ed.)

Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 29–30 November 2012, Pretoria, South Africa.