Lang Resources & Evaluation (2011) 45:263–269
DOI 10.1007/s10579-011-9157-9

ORIGINAL PAPER

Introduction to the special issue on African Language Technology

Guy De Pauw · Gilles-Maurice de Schryver · Laurette Pretorius · Lori Levin

Published online: 6 July 2011

In today's digital multilingual world, language technology is crucial for providing access to information and opportunities for economic development. With approximately two thousand different languages, Africa is a multilingual continent *par excellence*, presenting acute challenges for those seeking to promote and use African languages in the areas of business development, education and relief aid. In recent times a number of researchers and institutions, both from Africa and elsewhere, have come forward to share the common goal of developing capabilities in language technology for African languages. In 2009 and 2010, the first two workshops on African Language Technology were organized (De Pauw et al. 2009, 2010a) as a forum to bring together a wide range of researchers working in this domain.

The first author is funded as a Postdoctoral Fellow of the Research Foundation—Flanders (FWO).

G. De Pauw (✉)
CLiPS (AfLaT), Department of Linguistics, University of Antwerp, Antwerp, Belgium
e-mail: guy.depauw@ua.ac.be

G.-M. de Schryver
Department of African Languages and Cultures, Ghent University, Ghent, Belgium
e-mail: gillesmaurice.deschryver@ugent.be

G.-M. de Schryver
Xhosa Department, University of the Western Cape, Bellville, South Africa

L. Pretorius
School of Computing, University of South Africa, Pretoria, South Africa
e-mail: pretol@unisa.ac.za

L. Levin
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: lsl@cs.cmu.edu

This *Special Issue on African Language Technology* presents a cross-section of the state-of-the-art in the field through selected highlights from the AfLaT workshops, as well as original submissions. In this introduction we would not only like to introduce the papers themselves, but also present an overview of the most prominent research efforts in the field of African Language Technology.

1 A brief overview of African Language Technology

We define the scope of *African Language Technology* as follows: research and development in the fields of computational linguistics, natural language processing and human language technology (including speech technology) for sub-Saharan African languages from the four major African language groups: Afroasiatic, Khoesan, Nilosaharan and Niger-Congo languages. This excludes African variants of European languages, but also Arabic, which is already well researched by other scientific communities.

Language technology research for African languages has seen a steady increase in recent years. Whereas the world's largest languages have been under investigation by computational linguists since the 1950s, published language technology research for African languages only seems to have found its way to the public domain from the 1990s onwards. In this section we will identify the major players in the field.

Most African Language Technology research originates from South Africa. The University of Pretoria (UP) pioneered research in corpus linguistics for Bantu languages (de Schryver and Gauton 2002) and computational linguistics, by building the first prototype finite-state transducer for Northern Sotho (de Schryver 2002), as well as the first extensive digital corpora for up to a dozen Bantu languages. Joining hands with the University of South Africa (UNISA), they then furthered the research in morphological analysis, part-of-speech tagging and the development of advanced digital lexical resources of Bantu languages, particularly focusing on official South African languages, such as Zulu, Xhosa, Swati, Northern Sotho and Tswana (Taljard and Bosch 2006; Pretorius and Bosch 2007, 2009; Bosch et al. 2008; Faaß et al. 2009; Pala et al. 2010).

Also in South Africa, the Centre for Text Technology (CTexT) at the North-West University is a relatively young research center, focusing on more applied research for South African languages. Re-using resources initially created at UP and UNISA, in addition to their own, CTexT develop commercial applications such as spell checkers and language instruction packages, but also perform basic research on machine translation, morphological analysis and speech technology (Brits et al. 2006; Groenewald 2009; Roux et al. 2010; Oosthuizen et al. 2010).

Another major player in promoting language technology in South Africa is the Human Language Technology research center at the Meraka Institute in Pretoria. Most of the research conducted at Meraka focuses on speech technology, such as text-to-speech systems, speech recognition in the context of information access and spoken language identification (Louw et al. 2005; Peche et al. 2009; Barnard et al. 2010).

South African language technology research has greatly benefited from the protected status of its eleven official languages and is fostered by a number of government-funded projects. Elsewhere on the continent, awareness of the benefits of human language technology research has not reached such structural levels yet. Most other countries usually recognize only a handful of languages, typically one of which an Indo-European language such as French or English. The enormous linguistic diversity in sub-Saharan Africa therefore does not find itself reflected in a wide range of research projects and publications.

A notable exception to this impasse is the *African Languages Technology Initiative* (Alt-i), who have unlocked the technological potential of Nigeria's recognized national languages: Hausa, Igbo and Yoruba. Apart from on-going localization efforts (Adegbola et al. 2011), and research on speech recognition and synthesis (Finkel and Odejobi 2009; Odejobi 2011), a wide range of other human language technology resources and applications for Nigerian languages are being developed at Alt-i (Adegbola 2009).

In East Africa, the School of Computing & Informatics (SCI) at the University of Nairobi (Kenya) is an important hub for language technology research. Not only relevant work on Kenya's official language, Swahili, is being done at this institute (Ng'ang'a 2005, 2011; Muchemi 2008), but as one of the very few research groups in African Language Technology, do the researchers at SCI also tackle non-official, local languages such as Gikuyu, Luo and Kikamba (Wagacha et al. 2006a,b; Chege et al. 2010; Kituku et al. 2011). SCI is also committed to the mobile phone as a (language) technology platform, a particularly relevant field of study in the African context.

AfLaT (African Language Technology) started out as research collaboration between the University of Nairobi, the University of Antwerp and Ghent University (Belgium). The focus of AfLaT-centered research lies on knowledge-light approaches to African Language Technology, relying on (annotated) corpora and statistical and machine learning techniques to bootstrap language technology (De Pauw et al. 2006; Wagacha et al. 2006b; de Schryver and De Pauw 2007; De Pauw et al. 2010b). Since then AfLaT has evolved into an organization committed to promoting language technology research for sub-Saharan African languages, through a user-driven portal site, on-line demos and the organization of a yearly AfLaT workshop (De Pauw et al. 2009, 2010a).

The Faculty of Informatics at Addis Ababa University (Ethiopia) has become very active in the field as well, working on, amongst others, Amharic morphological analysis, part-of-speech tagging, machine translation and document classification (Alemayehu and Willett 2002; Weldesellassie 2003; Amsalu and Gibbon 2005; Adafre 2005; Argaw and Asker 2007; Anberbir and Takara 2009; Tachbelie and Menzel 2010). An overview of Amharic language technology research efforts can be found at http://nlp.amharic.org. Some work is also done for Oromo at Haramaya University (Adugna and Eisele 2010).

Alongside these hubs, a large number of individual researchers is working on African Language Technology, not only in Africa (Abdillahi et al. 2007; Muhirwe 2007), but in the Western world as well (Hurskainen 1992; Gambäck et al. 2009; Gasser 2010; Dione et al. 2010; Shah et al. 2010). It falls beyond the scope of this

overview to introduce all of the individual researchers in the field, but we would like to refer to AfLaT.org for a fairly exhaustive bibliography on African Language Technology.

On the commercial side a few companies are involved in African Language Technology, the most notable of which is Google Africa, who released a Swahili version of their translation system in 2009. Translate.org.za is a non-profit organization that focuses on translation and localization work of South Africa's eleven official languages. TshwaneDJe HLT, with offices in Africa and Europe, is mainly involved in computational lexicography (Joffe and de Schryver 2004), but has also developed African-language spell checkers, text messaging resources, digital corpora and localization components.

2 Special issue on African Language Technology

This special issue starts off with a contribution by Aditi Sharma Grover, Gerhard van Huyssteen and Marthinus Pretorius, the result of a survey conducted in 2009 that indexes human language technology components and applications available for South African languages and rates them according to a weighted *maturity index*. This paper presents the most exhaustive survey of its kind for any African country and as such we are confident that this paper will serve as one of the most important contributions to the field of African Language Technology in years to come. The methodology and recommendations formulated in this paper, may serve as a cornerstone to other similar surveys, not only on the African continent, but for any other multilingual country as well.

Jacob Badenhorst, Etienne Barnard, Charl Van Heerden and Marelie Davel describe ongoing research in Project Lwazi. Their paper first outlines the collection of the Lwazi corpus, an expansive audio corpus for all eleven official South African languages and then proceeds with a quantitative description of phoneme variability within the multilingual corpus. The insights gained from this computational description of the domain is used to streamline experiments in speech recognition, using the Lwazi corpus as training material. The paper then zeroes in on the challenges of resource-scarceness and formulates relevant ways to work around these.

The next two papers deal with the largest African language, Swahili, spoken by more than fifty million people and often suggested as a candidate for the *lingua franca* of the African Union. Ralf Steinberger, Sylvia Ombuya, Mijail Kabadjov, Bruno Pouliquen, Leo Della Rocca, Jenya Belyaeva, Monica de Paola, Camelia Ignat and Erik van der Goot present exciting work on a Swahili extension for a media monitoring and information extraction tool, available on-line. Rather than investing time in developing and/or integrating deep linguistic analysis components for Swahili, Steinberger and colleagues have opted for the rapid development of Swahili NLP components such as named entity recognition and geo-tagging. This paper not only performs a quantitative evaluation of these components, but also discusses the very relevant issue of development time, establishing a nice baseline and benchmark for other African languages as well.

Guy De Pauw, Peter Wagacha and Gilles-Maurice de Schryver tackle the problem of English - Swahili - English machine translation. This paper describes the collection and annotation of the 2.5 million word parallel SAWA corpus. This data is then used to perform projection of annotation experiments, where part-of-speech tags are transferred from English onto Swahili, an often suggested approach to knowledge-light annotation of under-resourced languages. The paper concludes with a description of a statistical machine translation experiment with fairly good results, beating Google Translate on some levels.

Researchers in African Language Technology are often faced with challenges of encoding, non-standard orthographies and different transcription systems. Steven Moran describes the implementation of an Ontology for Accessing Transcription Systems (OATS), a hierarchical knowledge base containing an orthographic and phonemic inventory for more than two hundred African languages. OATS not only provides an invaluable interface for researchers in the field of language technology, also typologists interested in phonetics and phonology will greatly benefit from this tool, as it allows for intelligent search, error-checking and conversion across orthographies for African languages.

The development of annotated corpora for African languages is paramount to the field. Christian Chiarcos, Ines Fiedler, Mira Grubic, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes and Malte Zimmermann describe work on 25 sub-Saharan languages [Gur, Kwa and Chadic (Hausa)], outlining ongoing annotation work, using both elicitation and machine learning techniques. This paper then proceeds to introduce ANNIS, a web-based corpus interface that can effectively visualize and query multilevel corpora. From the case study of the web-mined Hausa corpus, it is clear that ANNIS will provide a useful tool for researchers working on multilevel and multilingual corpora.

Many people are now turning to the Internet for digital language data. Unfortunately, this presents a big challenge in terms of encoding, as many African languages use diacritics in their orthography to mark phonemic variants or tone, often not or inconsistently used. Kevin Scannell caps this issue with a truly pan-African paper on language-independent *unicodification*. The paper presents quantitative experimental results on more than one hundred different African languages. The approach can be used to (semi-)automatically normalize and unicodify web-mined corpora. The possible advances such an approach has for corpus-based approaches to African Language technology cannot be understated.

References

Abdillahi, N., Nocera, P., Bêchet, F., & Bonastre, J.-F. (2007). Information retrieval strategies for accessing African audio corpora. In H. Van hamme & R. van Son (Eds.), *Proceedings of the eighth annual conference of the international speech communication association*. Antwerp, Belgium: INTERSPEECH.

Adafre, S. F. (2005). Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL workshop on computational approaches to semitic languages* (pp. 47–54). Ann Arbor, Michigan: Association for Computational Linguistics.

Adegbola, T. (2009). *Building capacities in human language technology for African languages*. In De Pauw et al. (2009) (pp. 53–58).

Adegbola, T., Owolabi, K., & Odejobi, T. (2011). Localising for Yoruba: Experience, challenges and future direction. In *Proceedings of conference on human language technology for development* (pp. 7–10). Alexandria, Egypt: Bibliotheca Alexandrina.

Adugna, S., & Eisele, A. (2010). *English—Oromo machine translation: An experiment using a statistical approach*. In Calzolari et al. (2010).

Alemayehu, N., & Willett, P. (2002). Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing, 17*, 1–17.

Amsalu, S., & Gibbon, D. (2005). Finite state morphology of Amharic. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Recent advances in natural language processing* (pp. 47–51). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Anberbir, T., & Takara, T. (2009). *Development of an Amharic text-to-speech system using cepstral method*. In De Pauw et al. (2009) (pp. 46–52).

Argaw, A. A., & Asker, L. (2007). An Amharic stemmer: Reducing words to their citation forms. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources* (pp. 104–110). Prague, Czech Republic: Association for Computational Linguistics.

Barnard, E., Davel, M., & van Huyssteen, G. (2010). Speech technology for information access: A South African case study. In *Proceedings of the AAAI spring symposium on artificial intelligence for development (AI-D)* (pp. 8–13). Palo Alto, USA: Association for the Advancement of Artificial Intelligence.

Bosch, S. E., Pretorius, L., & Fleisch, A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies, 17*(2), 66–88.

Brits, J., Pretorius, R., & van Huyssteen, G. (2006). Automatic lemmatisation in Setswana: Towards a prototype. *South African Journal of African Languages, 25*, 37–47.

Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.) (2010). *Proceedings of the seventh conference on international language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Chege, K., Wagacha, P. W., De Pauw, G., Muchemi, L., & Ng'ang'a, W. (2010). *Developing an open source spell checker for Gĩkũyũ*. In De Pauw et al. (2010) (pp. 31–35).

De Pauw, G., de Schryver, G.-M., & Levin, L. (Eds.) (2009). *Proceedings of the EACL 2009 workshop on language technologies for African languages (AfLaT 2009)*. Athens, Greece: Association for Computational Linguistics.

De Pauw, G., de Schryver, G.-M., & Wagacha, P. W. (2006). Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček & K. Pala (Eds.), *Proceedings of text, speech and dialogue, ninth international conference,* (Vol. 4188/2006 of lecture notes in computer science, pp. 197–204). Berlin, Germany: Springer.

De Pauw, G., Groenewald, H., & de Schryver, G.-M. (Eds.) (2010a). *Proceedings of the second workshop on African Language Technology (AfLaT 2010)*. Valetta, Malta: European Language Resources Association (ELRA).

De Pauw, G., Maajabu, N., & Wagacha, P. W. (2010b). *A knowledge-light approach to Luo machine translation and part-of-speech tagging*. In De Pauw et al. (2010) (pp. 15–20).

de Schryver, G.-M. (2002). *First steps in the finite-state morphological analysis of Northern Sotho. In AFRILEX 2002, Culture and dictionaries, programme & abstracts* (pp. 22–23). Pretoria, South Africa: $(SF)^2$ Press.

de Schryver, G.-M., & De Pauw, G. (2007). Dictionary writing system (DWS) + corpus query package (CQP): The case of TshwaneLex. *Lexikos, 17*, 226–246.

de Schryver, G.-M., & Gauton, R. (2002). The Zulu locative prefix ku- revisited: A corpus-based approach. *Southern African Linguistics and Applied Language Studies, 20*(4), 201–220.

Dione, C. M. B., Kuhn, J., & Zarrieß, S. (2010). *Design and development of part-of-speech-tagging resources for Wolof*. In Calzolari et al. (2010).

Faaß, G., Heid, U., Taljard, E., & Prinsloo, D. (2009). *Part-of-speech tagging of Northern Sotho: Disambiguating polysemous function words*. In De Pauw et al. (2009) (pp. 37–42).

Finkel, R., & Odejobi, O. A. (2009).*A computational approach to Yoruba morphology*. In De Pauw et al. (2009) (pp. 25–31).

Gambäck, B., Olsson, F., Argaw, A. A., & Asker, L. (2009). *Methods for Amharic part-of-speech tagging*. In De Pauw et al. (2009) (pp. 104–111).

Gasser, M. (2010). *Expanding the lexicon for a resource-poor language using a morphological analyzer and a web crawler*. In Calzolari et al. (2010).

Groenewald, H. (2009). *Using technology transfer to advance automatic lemmatisation for Setswana*. In De Pauw et al. (2009) (pp. 32–37).

Hurskainen, A. (1992). A two-level computer formalism for the analysis of Bantu morphology. An application to Swahili. *Nordic Journal of African Studies, 1*(1), 87–122.

Joffe, D., & de Schryver, G.-M. (2004). TshwaneLex—a state-of-the-art dictionary compilation program. In G. Williams & S. Vessier (Eds.), *Proceedings of the eleventh EURALEX international congress* (pp. 99–104). Lorient, France: Université de Bretagne Sud.

Kituku, B., Wagacha, P., & De Pauw, G. (2011). A memory-based approach to Kĩkamba named entity recognition. In *Proceedings of conference on human language technology for development* (pp. 106–111). Alexandria, Egypt: Bibliotheca Alexandrina.

Louw, J., Davel, M., & Barnard, E. (2005). A general-purpose IsiZulu speech synthesiser. *South African Journal of African Languages, 25*(2), 92–100.

Muchemi, L. (2008). Towards full comprehension of Swahili natural language statements for database querying. In J. Aisbett, G. Gibbon, A. J. Rodrigues, J. K. Migga, R. Nath & G. R. Renardel (Eds.), *Strengthening the role of ICT in development, special topics in computing and ICT research* (pp. 50–58). Kampala, Uganda: Fountain Publishers.

Muhirwe, J. (2007). Computational analysis of Kinyarwanda morphology: The morphological alternations. *International Journal of Computing and ICT Research, 1*(1), 85–92.

Ng'ang'a, W. (2005). *Word sense disambiguation of Swahili* (PhD Thesis). Helsinki, Finland: University of Helsinki.

Ng'ang'a, W. (2011). Swahili inflectional morphology for the grammatical framework. In *Proceedings of conference on human language technology for development* (pp. 100–105). Alexandria, Egypt: Bibliotheca Alexandrina.

Odejobi, O. (2011). Design of a text markup system for Yorùbá text-to-speech synthesis applications. In *Proceedings of conference on human language technology for development* (pp. 74–80). Alexandria, Egypt: Bibliotheca Alexandrina.

Oosthuizen, N., Puttkammer, M., & Schlemmer, M. (2010). *Improving orthographic transcriptions of speech corpora*. In De Pauw et al. (2010) (pp. 55–58).

Pala, K., Fellbaum, C., & Bosch, S. E. (2010). *Lexical resources for noun compounds in Czech, English and Zulu*. In Calzolari et al. (2010).

Peche, M., Davel, M., & Barnard, E. (2009). Development of a spoken language identification system for South African languages. *SAIEE Africa Research Journal, 100*(4), 97–105.

Pretorius, L., & Bosch, S. E. (2007). Containing overgeneration in Zulu computational morphology. In Z. Vetulani (Ed.) *Human language technologies as a challenge for computer science and linguistics: Proceedings of 3rd language and technology conference* (pp. 54–58). Poznań: Wydawnictwo Poznańskie.

Pretorius, L., & Bosch, S. E. (2009). *Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology*. In De Pauw et al. (2009) (pp. 96–103).

Roux, J. C., Scholtz, P., Klop, D., Povlsen, C., Jongejan, B., & Magnusdottir, A. (2010). *Incorporating speech synthesis in the development of a mobile platform for e-learning*. In Calzolari et al. (2010).

Shah, R., Lin, B., Gershman, A., & Frederking, R. (2010). *SYNERGY: A named entity recognition system for resource-scarce languages such as Swahili using online machine translation*. In De Pauw et al. (2010) (pp. 21–26).

Tachbelie, M., & Menzel, W. (2010). *Capturing word-level dependencies in morpheme-based language modeling*. In De Pauw et al. (2010) (pp. 43–48).

Taljard, E., & Bosch, S. E. (2006). A comparison of approaches to word class tagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies, 15*(4), 428–442.

Wagacha, P. W., De Pauw, G., & Getao, K. (2006a). Development of a corpus for Gĩkũyũ using machine learning techniques. In J. C. Roux (Ed.) *Proceedings of LREC workshop - Networking the development of language resources for African languages*. Genoa, Italy: ELRA.

Wagacha, P. W., De Pauw, G., & Githinji, P. W. (2006b). A grapheme-based approach to accent restoration in Gĩkũyũ. In *Proceedings of the fifth international conference on language resources and evaluation* (pp. 1937–1940). Genoa, Italy: ELRA.

Weldesellassie, S. (2003). *Automatic categorization of Amharic news text: a machine learning approach* (MSc Thesis). Addis Ababa, Ethopia: Addis Ababa University.