Multilingualism and Electronic Language Management

Walter Daelemans, Theo du Plessis, Cobus Snyman, and Lut Teck (editors)
(University of Antwerp; University of the Free State; and Institute for Higher Education in the Sciences and Arts, Brussels)

Pretoria: Van Schaik (Studies in language policy in South Africa, volume 4), xv+184 pp; paperbound, ISBN 0-627-02601-X, SAR 175.95

Reviewed by
David Nathan
*School of Oriental and African Studies, University of London**

The volume aims to inform the implementation of human language technologies (HLTs) in the service of language policy in South Africa. Its 12 papers are organized into five sections whose themes, except for the first, "HLT resources and policy development," considerably overlap. The contributors, most with backgrounds in computer science, computational linguistics, and language engineering, are international, with several from The Netherlands, Belgium, and South Africa.

1. Overview of the Book

The introduction (Daelemans, du Plessis, Snyman, and Teck) argues that HLTs are needed to provide equity of access to information under South Africa's eleven-language policy. They see HLTs as a means of enabling "all South Africans to use and share information in the language of their choice."

Chapter 1 then draws on Dutch and Flemish research and experience, in particular the Dutch BLARK (Basic Language Resources Kit) pursued as part of a Dutch Language Union project. An array of charts lists the various components that such a kit might consist of (classified as applications, modules, or data), together with the relationships, priorities, and availabilities amongst them, based upon the work of a research committee and survey responses from academics and commercial developers. The result is a clear picture of the state of HLT for Dutch — found to be incomplete and of variable quality — and short priority lists and criteria for future development.

One of the more important chapters is the second (Roux and du Plessis), which completes the policy section. It describes the South African government's efforts to foster the development of language and speech technologies and resources, recently expressed as the "Human Language Technologies initiative" of 2004, exactly ten years since the eleven-language policy was established. The authors survey reports and developments over the decade, finding repeated themes such as an emphasis on government co-ordination and industry partnerships, promotion of multilingualism, and support of "historically marginalized languages" (page 29). The authors describe the general trend of government policy as treating HLT as primarily "a language affair," that is, as "an instrument of language development and language promotion"; they identify a tension between such aims and "more general expectations" of HLT as a means of enhancing access to information. Generally critical of the present government in

* I am grateful to Dr. Nhlanhla Thwala (SOAS/Witwatersrand) for advice about the sociolinguistic and language policy situation in South Africa.

its pursuit of "political" aims, the authors prefer to see HLT as an integral part of an "information society" that strives for universal access and consequent social and economic advantages. This polarity, and the wide gap it exposes between positions, represents a weakness of the book as a whole: the absence of language communities as living entities, and the lack of research (and evidence) about their linguistic situations and needs.

Chapters 3 to 6 cover automatic translation. Chapter 3 (Weber) describes a continuum of MT (machine translation) technologies, ranging from fully automated to those involving considerable human intervention at various stages. This is an important distinction, not only for the technologies themselves (for example, machine learning can crucially involve human supervision), but also for matching technologies to needs and the resources available, although there is little in the book about the latter. The chapter is a good summary of MT terms and concepts, and phases in its development; most interesting for language planners and policy-makers will be to learn that decades of MT research have been largely unsuccessful, with its first paradigm — rule-based translation — largely abandoned as fruitless, and newer statistical methods yet unproven (several later chapters corroborate this conclusion). We find here the first grounds for doubt about the book's ability to fulfill its aims: while the author provides three examples of translations, not only are they (as he explains) moderately unsuccessful, but they involve translation between major European languages; the language planner will wonder therefore what prospects there are for African languages, and may question why more relevant examples could not have been found. While the chapter offers a good, non-technical background to automated translation, hinting at the difficulties of translating meaning and context, unfortunately, as in most chapters, technology rarely gets considered in the context of its users. This may lead to problems; for example, tolerating "less-than-perfect results" may assist implementation but may be precarious in communities who do not have much experience with information technology (IT) or, perhaps as a result of that, expect perfection from it. The strategy of constraining translation to particular domains may not be relevant to community concerns or modes of information-seeking; here the weather-forecast domain is mentioned, in chapter 10 we find IT job-postings to a Usenet newsgroup.

Chapter 4 (Champollion) moves from technologies to specific tools and functions, and is another chapter that planners and policy makers should find useful. While (non-automated) translation is of interest to researchers, it remains the ugly duckling of HLT in terms of deliverables (useful tools). Paradoxically, it is widely approached with skepticism while being one of the few areas where HLT has brought productivity benefits and "become part of a language practitioner's daily life" (page 61). The chapter nicely describes several relatively humble but effective tools, such as a segmenter (assists mechanical tasks of aligning source and target sentences), translation memory (an interactive database of translation segments already made), and terminology banks. Such tools allow the human translator to focus on the language task and to leverage existing work. The reviewer is reminded of a dyslexic friend who many years ago found that even the most basic speech synthesis was a boon — he could hear and check the e-mail he was typing, opening up new possibilities for communication. Voice recognition has considerably improved, according to the author, but is strongly language-dependent; perhaps its adoption will depend on whether keypress or voice input technologies catch the imagination of a new generation.

Messerschmidt et al. (chapter 5) provide a digestible account of the mechanisms of a rule-based parser and grammar that learn under human supervision, showing its application to Sesotho. This will be useful for non-specialists. Nevertheless, their

argument that HLTs are needed for African languages is based on a belief that HLTs for colonial languages (English and Afrikaans) would be sufficient if only Africans had a better command of them. While colonial languages, especially English, continue to play significant, even increasing, roles, especially in business and service delivery in South Africa, what is missing here is the larger picture: evidence about language and technology preferences within particular activities and domains, and acknowledgment that the context for technology development should at least include "capacity building," including, for example, increasing IT training and employment for speakers of African languages. Short-term projects channeling the richer information currently available in English (e.g., about AIDS) into languages such as Sesotho may not provide a means of addressing *future* health threats.

Given that current MT methods have matured but remain ineffective, Champollion's second contribution (chapter 6) considers future possibilities for MT. Starting from a transcendental discussion on complexity (and the number pi), he moves between abstract discussion and concrete advice on corpus creation: Don't attempt to replace "relevance with bulk, or quality with quantity" (page 83) — it is futile to build larger and larger translation databases in the expectation of enabling successful translation. He sees today's methods as merely "pumps" that work on elements of "how humans think languages are organized" (page 83), and foresees a day when computers produce detailed fractal grammars in their own pattern-matching terms.

The third section is more optimistic, with two chapters based on the use of parallel corpora. Marcu shows that statistical machine translation software using such corpora is not only at least as effective as rule-based methods but is cheaper to develop and, more importantly, adaptable to a variety of languages. However, while the software is generic, preparing parallel corpora, especially for pairs of "low density" languages (page 95), is difficult (if not impossible), or may involve reducing content to the trivial or to domains that are usually dealt with in another language anyway. African languages appear for the first time as the main focus in chapter 8 (Prinsloo and De Schryver). Parallel corpora can contribute to research in phonetics, linguistics, language change, and lexicography, as well as in practical applications such as language teaching and spelling checkers. The authors used widely available corpus software (WordSmith and ParaConc) to build collections of translation equivalents and extract document keywords, with careful evaluation of the results; however, the techniques have not yet been extended to the production of useful tools such as dictionaries and spelling checkers.

Localization — the fourth theme — could have received more attention than a single chapter (van Rooy) classifying and evaluating student spelling errors in order to show that "linguistic resources should be localized for South African English." Some useful concepts, such as precision and recall, are explained, and a comprehensive typology of errors is presented. The author reaches the unstartling conclusion that a spelling checker would be improved if it was provided with a list of place-names, acronyms, and frequently occurring errors. A reader (especially a planner or policy-maker) surely wants to know about spelling in African languages, about localized change that English is undergoing as it is adapted across different African communities (rather than the increasingly outdated concept of standard South African English), and about localization issues for African software and operating systems.

The theme of text mining also receives a single chapter (Mooney and Nahm), which examines the discovery of relationships and patterns in both structured and unstructured documents. The authors developed a system that learns to perform

information extraction (e.g., summarization, template-filling), and then use further data-mining of the structured results to improve its information extraction accuracy. The domain chosen for presentation of results — IT job-postings to a Usenet newsgroup in English — is surely over-constrained and barely relevant to the purposes of the book.

"Multilingual text and speech interfaces" is the sixth and final section. Chapter 11 (Roux and Louw) revealingly presents a project to develop multilingual speech databases. Over 5,000 telephone calls from recruited participants were recorded and then transcribed phonetically and orthographically (with time alignment). Here we find not only consideration of *sound* for the first time, but also research linked to relevant sociolinguistic factors such as domains of usage (the importance of English in business) and variation (e.g., regional and community-based varieties of English; L1 vs. L2 speakers). Various stages in database design, speaker recruitment, automated prompting and recording, data management, transcription, choice of standards, and error correction are described. The chapter is replete with surprises that could have perhaps been factored into the methodology rather than discovered, but remain interesting nonetheless — such as that many people don't wear watches (so that responses about time were often not as expected), or that few speakers of African languages had the appropriate telephone equipment to take part in the experiments.

The final chapter (Bosch) warns again that languages without adequate HLT "run the risk of being marginalized in the global Information Society" (page 169). However, the author's main task is to summarize a system under development for the morphological analysis of Zulu, which presents many morphological and morphophonological complexities not found in English.

2. Conclusions

Despite the second chapter's wariness about treating HLTs as political "language affairs," several chapters do jump off from problems of disadvantage direct to HLT solutions without providing any evidence that the solutions fit the problem. Is it the planners' job to connect these? Certainly it is to some extent; however, in general the book leaves the space between problems and solutions rather too wide, and the planners' task no less than before.

Planners need to be skeptical about claims that languages without "adequate computer processing" (page 1) will simply disappear. Such dramaticizations, if true, would condemn almost all of the world's 6,000 languages to extinction. We know, from research in language endangerment, that factors causing language death are complex, widespread, and dangerous, and that they dwarf any contributions that HLT can (or cannot) provide.

Several chapters have provided evidence of HLT's limited capabilities, either directly or by indicating that the effectiveness and benefits of HLT fall in domains and functions that are not likely to be at the forefront of language planners' minds.

Will a planner or policy-maker be able to make use of this volume's spectrum of information from practical advice on data collection to the theoretical capabilities and limits of statistical and fractal grammars? Examples scattered throughout the book indicate possible directions: for example, IT training and the involvement of speakers of African languages in projects would assist not only in providing richer sociolinguistic input but also in the advancement of the technologies themselves (e.g., for supervised-learning grammars); or that the limited prospect for HLT support of multilingualism suggests prioritization of resources for multilingual education of the population.

While some HLTs can spark imagination and raise awareness of linguistic issues, across the globe it is only in limited contexts that they have actually become vehicles for communication, delivering services, or language learning and transmission. It may be true that HLTs are more apt for servicing "language affairs" than they are for delivering a multilingual information utopia.

David Nathan is Director of the Endangered Languages Archive at SOAS, University of London. He works with computing and educational aspects of applications for endangered languages, and has taught computing, linguistics, cognitive science, and multimedia. His publications include "Australian Indigenous Languages"; papers on lexicography, the Internet, and multimedia; and multimedia CD-ROMs. He was co-author (with Peter Austin) of the Web's first hypertext bilingual dictionary. Nathan's address is: SOAS, University of London, Thornhaugh Street, Russell Square, London WC1H 0XG, U.K.; e-mail: david@dnathan.com; URL: www.dnathan.com.