Peter Spyns and Jan Odijk (Editors), Essential Speech and Language Technology for Dutch: Results by the STEVIN programme
Springer, 2013, ISBN: 978-3-642-30909-0, xvii + 413 pp

Antal van den Bosch

the date of receipt and acceptance should be inserted later

The first word of the book’s title, Essential, reminds one of the bookstore’s popular science corner. This book, published in Springer’s Theory and Applications of Natural Language Processing series, is rather a colorful collection of scientific papers spanning topics in language and speech technology that were en vogue in the previous decade, tailored to the Dutch language. All of these projects were part of a Dutch-Belgian research programme, STEVIN, Spraak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands (2004-2009).\(^1\) The book’s title is a loose translation of the project’s title—hence the Essential. While I have great admiration for the work described in Parts II, III, and IV of the volume, my first reflection is critical and regards the framing of the book.

Section 1.3 of the Introduction by Peter Spyns is titled Mission accomplished. The recent unfortunate historical connotation of president Bush’ speech set aside, this section title sets a stage that is unfitting for a collection of academic work. Surely the mission is never accomplished; surely there cannot be something like an essential collection of language and speech technologies for a single language. Some history is needed to understand why the book kicks off with such a self-assured attitude. As I explain below, we learn how Basic would have been a better term than Essential in Part I of this book, How it all started.

Dutch is spoken as a first or second language by approximately 25 million people worldwide, most of which live in the Netherlands and the Dutch-speaking part of Belgium, Flanders. The Dutch Language Union is an transnational organization, formally established as an intergovernmental organization by Belgium and the Netherlands in 1980, with the aim of conducting a joint language policy, to deal with issues regarding the role and position of the Dutch language in science, education, culture, and society at large. In the late 1990s, in which it became clear that languages such as Dutch were not automatically receiving the same amount of attention in technology development as that which major languages such as English receive, the Dutch Language Union recognized there was...

a need for a BLARK, a Basic Language Resource Kit [Krauwer, 2003] for Dutch. It thus called the language and speech technology communities in both countries into action. The two countries had, and still have, a relatively large number of language and speech technology groups, many of which were seeded directly and indirectly through large machine translation projects of the 1980s, such as EU-ROTRA. Commissioned by the Dutch Language Union, a group of researchers led by Walter Daelemans and Helmer Strik set out to design a priority list of desirable technologies that should be consolidated or developed to create a first coherent BLARK for Dutch [Strik et al., 2002]. Eventually this priority list was the blueprint for the five-year, twelve-million-Euro Dutch-Belgian STEVIN programme. The book edited by Spyns (STEVIN programme co-ordinator) and Odijk (chair of the STEVIN programme committee) looks back at this period with pride and the aforementioned overt sense of accomplishment.

Meanwhile, language and speech technology research and development finds itself in the next decade. Already now it is obvious that much of the work of the previous decade is superseded by new insights, and other methods to create the same resources through different means (new corpora, machine learning methods, crowd sourcing). Obviously, many of the objects of research in the STEVIN programme were also studied elsewhere at the same time, and many of the developed technologies in the project are language-independent. To the outside reader of this book this is probably the key point of interest: to learn about a large concerted action geared at developing a BLARK for a language such as Dutch. This is the largest payoff of the STEVIN programme and the book, although the concept of a BLARK seems to have lost its ‘essential’ status and is barely used anymore. This is also what can be heard in the positive tones of the evaluation of STEVIN by an international panel. Chapter 2 of the book quotes the ten recommendations of the evaluators. They start by stating “The integrated approach of STEVIN was a good approach and should be replicated in a potential follow-up of STEVIN. The focus should then be shifted from the BLARK and strategic research to application-oriented research and demonstration projects.” (p. 34).

The book is a showcase of the breadth of Dutch and Flemish language and speech technology. The editors did not opt for the predictable choice to divide the natural language processing technologies and speech technologies in different parts. Rather, they adopted a goal-oriented categorization of projects to create three parts. Part II, the heart of the book, is intended to deal with resource and toolkit development; Part III contains a number of more technical and evaluative papers; and Part IV provides some examples of applications. Despite this seemingly logical structure, tools, applications, resources, and evaluation are described in all parts in nearly equal distributions. The differences in emphasis are negligible. An alphabetized list of project chapters would have worked as well.

In Part II, HLT resource-project related papers, we learn about resources and tools. Most of this work has been presented and published earlier, for instance in several volumes of LREC proceedings; chapters are typically reworked and integrated versions of earlier papers. Resources described include the Dutch Parallel Corpus, a balanced parallel corpus of Dutch–English and Dutch–French texts, SoNaR, a 500-million word reference corpus of contemporary written Dutch, and the JASMIN corpus of spontaneous speech recorded from children, non-natives, and elderly people. The latter two corpora complement another Dutch resource predating the STEVIN project, the Spoken Dutch Corpus [Oostdijk et al., 2002].
Besides these valuable corpora, several lexical resources were created: Cornetto, a Dutch Wordnet and ontology; the DuELME lexical database of multi-word expressions; the LASSY Treebank containing one million words of manually verified dependency parsing annotations and very large corpora (including SoNaR) of automatically parsed texts; the DAESO corpus of syntactically annotated sentential paraphrase pairs; and the Autonomata Spoken Names Corpus.

Part II also harbors descriptions of systems and tools developed in the STEVIN programme. In speech technology, the well-known PRAAT (‘talk’) software\(^2\) was extended with several modules such as a robust formant frequency analyzer. Analogously, SPRAAK (‘speech’), a new toolkit for speech recognition and automatic annotation was developed within the programme as a new open-source alternative to other speech recognition engines. At the language technology side, systems for coreference resolution (COREA) and automatic tree matching (again DAESO) were developed.

In Part III, HLT-technology related papers, the machine translation project ‘Parse and Corpus-based Machine Translation’ (PACO-MT) is described, a system that uses synchronous tree substitution grammar to perform tree-to-tree mapping in MT. PACO-MT uses a blend of dependency and constituency information, setting it apart from other similar approaches. Further in Part III we find chapters on three speech-related projects on proper name recognition, a benchmark for large-vocabulary Dutch speech recognition, and a system for handling noise in speech recognition through missing-data techniques.

Finally, Part IV, HLT application related papers, is a showcase of STEVIN projects that aimed at higher-level systems: integrating speech technology into courseware for language learning, the role of summarization in question answering, and bootstrapping lexicons for sentiment analysis.

With its 430 pages of content the book offers an impressive documentation of transnational joint research and development in language and speech technology from the past decade, that has been characterized by international reviewers as a front-running, exemplary effort. There are many Dutch-sized languages that are spoken across several countries, and the book makes a good case for why and how concerted transnational action can make the best use of expertise available in academic research groups to arrive at a BLARK. Although most of the work was already presented in other publications elsewhere, it is a must-read for prospective BLARK developers, HLT resource centers, and language unions everywhere, especially as the final Part V of the book (And Now) offers a close look at the lessons learnt at the level of intellectual property rights and the major efforts involved in the maintenance and distribution of all STEVIN outcomes through the Dutch HLT Agency.\(^3\)

Yet, as argued, I do not think the mission is accomplished; neither is the collection essential. Several obvious technologies, such as spelling correction, speech synthesis, speaker identification, discourse modeling and other text-level topic modeling, computational dialectometry and stylometry, word sense disambiguation, and — as the reader of this journal will have assumed by now — statistical machine translation are missing from the book, and so are many types of applications in text analytics and speech data mining. Fortunately most of these technologies are

\(^2\) http://www.praat.org
\(^3\) http://tst-centrale.org/
firmly in focus in present-day research and development in the Low Countries, and it remains to be stressed that much of this current work stands on the shoulders of the giants of the last decade.

References


Summarising a research programme that lasted for more than 6 years is a demanding task due to the wealth of deliverables, publications and final results of each of the projects concerned. In addition to the content-related topics, which interest scientists, research programmes also lead to new insights for policy makers and programme managers. The former want to discover advances in the state of the