Intergovernmental language policy for Dutch and the language and speech technology infrastructure

Gosse Bouma
Rijksuniversiteit Groningen
gosse@let.rug.nl

Ineke Schuurman
Katholieke Universiteit Leuven
ineke.schuurman@ccl.kuleuven.ac.be

Abstract

The European Union has been actively stimulating fundamental and applied research on language and speech technology for more than a decade. A substantial number of European projects aimed at the development of resources and infrastructure. While the development of general standards and multilingual tools and resources are a concern of the European Union, the development of basic resources for a given language needs to be stimulated by national initiatives as well. For Dutch – the language spoken in the Netherlands and Flanders – the 'Nederlandse Taalunie' (*Dutch Language Union*) is a unique intergovernmental institution, responsible for language policy in both the Netherlands and Belgium. In this paper we investigate the state of affairs with respect to resources and infrastructure for Dutch and we address the question how national and intergovernmental language policy can stimulate the development of resources and infrastructure.

1. Introduction

Progress in hardware and software performance in combination with integration of information technology and telecommunication services are leading to an information society in which language and speech technology will become increasingly important. Whereas many aspects of the information age are genuinely global in character, the development of language and speech technology has an important national component.

The creation of a solid infrastructure for language and speech technology for any language requires that language and speech resources are being developed, maintained, and distributed. This requires organizations which determine the needs of education, research, and industry, which set priorities and standards, and which make sure that resources are maintained, documented, and available to interested parties. Coordination and funding by national governments is of key importance, especially for smaller languages: commercial applications will only be developed if the level of basic, publicly available, resources has sufficient quality to make the development of these applications economically feasible.

With approximately 20 million speakers, Dutch is a relatively small language (at least from the perspective of companies active in the field of language and speech technology) which is spoken in Belgium (Flanders) and the Netherlands. There are many striking differences between the language as it is used in Flanders and in the Netherlands. Nevertheless, both governments recognize the importance of a joint language policy for Dutch. For this reason, the *Nederlandse Taalunie* (NTU) (*Dutch Language Union*) was created, an intergovernmental institution responsible for language policy in the Netherlands and Flanders. The NTU is responsible for maintaining the position of the Dutch language, and as such it stimulates educational programme's (both in the Netherlands and Flanders and abroad), it supports projects leading to dictionaries, grammars, and other language resources, it promotes Dutch literature, and it advises the Dutch and Belgian government on language policy issues. One of its recent activities was advising the two governments on a spelling reform, which was implemented in 1995. The NTU also enables the Netherlands and Flanders to speak with in single voice in the international arena, for example in order to ensure the position of Dutch as one of the working languages of the European Union.

Language and speech technology is becoming increasingly important for the NTU. For a long time, language resources consisted almost exclusively of paper dictionaries and grammars. More and more, however, the emphasis is shifting from paper resources to electronic resources. For instance, it is currently investigated how the *Algemene Nederlandse Spraakkunst* (Haesereyn et al., 1997), a descriptive grammar of Dutch produced with support from the NTU, can be made accessible electronically. Another example is the recent spelling reform (Renkema, 1996), which not only lead to the publication of new paper word lists and dictionaries, but also necessitated the release of new versions of spelling correction and hyphenation programs. With a shift in emphasis to electronic resources, the NTU is also confronted with the certain responsibilities with respect to maintenance and distribution of such resources. For instance, the NTU will be responsible for the distribution of a large (10 million words) corpus of spoken Dutch, which is currently being developed by a consortium of research institutions in the Netherlands and Belgium, as well as for the distribution of several (bilingual) dictionary databases. Finally, the most important reason for the NTU to be concerned with issues of language and speech technology is the fact that language policy for Dutch will increasingly be a matter of creating and stimulating an infrastructure in which the production of tools and applications implementing state of the art language and speech technology for Dutch is technically and economically feasible.

As a starting point for its new task, the NTU is currently investigating what exactly constitutes the language and speech technology infrastructure for Dutch, how it compares to the infrastructure for other languages, what the most important needs are for education, research, and industry, and what coordinating actions are required.

2. Resources and tools for Dutch

The availability of sufficient language and speech resources is of crucial importance for the development of language and speech technology for any given language. Resources include simple word lists (for spelling correction and hyphenation), advanced lexical databases with syntactic, semantic, and phonological information (for grammar development, disambiguation, text-to-speech, etc), statistical information extracted from corpora, such as word-level and character-level N-gram statistics (for optical character and speech recognition tasks), tagged corpora (for training and testing part-of-speech taggers and grammars), parallel corpora (for automatic translation), speech corpora of different sorts (individual words and connected speech, read and spontaneous utterances, optimal, office, and (mobile) telephone recordings), as well as various (language-specific) tools to be used as components of larger systems (such as stemmers for document indexing, and part of speech taggers and formal grammars for grammatical error correction, translation and dialogue systems).

We investigated whether sufficient resources are currently available for Dutch and concluded that, while a number of corpora and lexical databases are available, the overall situation needs to be improved drastically.

2.1. Text Corpora

The most important supplier of text corpora for Dutch is the *Instituut voor Nederlandse Lexicografie* (INL). This institute, whose prime task is the production of an exhaustive, corpus-based, dictionary of Dutch, has collected several large (from 5 to 50 million words) corpora of modern Dutch, including newspaper, literary, and technical texts. The corpora are mostly labelled with part of speech information. It should be noted, however, that the labelling is the result of an automatic process, and thus contains a certain percentage of errors. Only a small fragment of the data (the 5 million INL corpus distributed as part of the European Corpus Initiative Multilingual Corpus CD-ROM) is conveniently accessible for language technology research purposes. Access to the other corpora is restricted, because of copyright issues.

A widely used text corpus is the so-called *Eindhoven* or *Uit den Boogaart* corpus. This (750.000 words) corpus was collected in the sixties and early seventies, as part of a research project aiming at the collection of frequency data for written and spoken Dutch (Uit den Boogaart 75; de Jong 79) . While the statistical results are still available (at least in paper form), the corpus itself is no longer actively maintained, and therefore only available through informal channels. The fact that it is still in use must be be attributed to the fact that this is the only available medium-size corpus that was carefully tagged.

ELRA offers two multi-lingual text corpora including Dutch (the Polylingual Document Collection, ELRA-W0007, and the Multilingual Parallel Corpus, ELRA-W0008). More recently created corpora, such as the ANNO corpus (a 640.000 word corpus automatically labelled with part of speech information (partly corrected), created at the Centre for Computational Linguistics, Leuven University (Schuurman, 1997)) and the OVIS corpus (a 10.000 utterance corpus with syntactic and semantic annotations, created in the context of the Dutch NWO priority programme for language and speech technology (Bonnema, Bod, and Scha, 1997; Nederhof et al., 1997)), and the PAROLE corpus (a 3 million word Dutch corpus with 250.000 words annotated with part of speech, currently under development at the INL) are quite domain and application specific (OVIS), of medium size, and not (yet) publicly available.

Text corpora for Dutch that are easily available and accessible are quite rare. For research which requires mostly *raw* text, this problem can be relatively easily overcome, as raw text is available from a number of sources (such as newsgroup archives, web-sites, and CD-ROM's containing large quantities of text). It should be noted, however, that such corpora tend to be unbalanced (i.e. do not contain a representative sample of the relevant texts) and that the size of the corpora that can be constructed this way may still be too small for certain applications (and there may be copyright problems). More problematic is the fact that *annotated* corpora are not sufficiently available, and that the corpora that exist are annotated automatically (introducing errors) and hard to access. The creation of such corpora for English, such as the Penn Treebank (Marcus, Santorini, and Marcinkiewicz, 1993) has had an enormous effect on the field. The development of comparable resources for Dutch, be it corpora with labelled with part of speech, syntactic, or semantic information, is expensive, especially if manual correction of automatically labelled texts is included, and requires careful planning and coordination with potential users. Currently, we are not aware of any attempts to produce such text corpora on a larger scale.

2.2. Lexical Resources

The most important lexical resource for Dutch is the CELEX database (Baayen, Piepenbrock, and van Rijn, 1993), available on CD-ROM and also accessible on the internet. This database (which also includes English and German data) contains detailed morpho-phonological information for approximately 125.000 lexemes and 400.000 word forms. The database provides information concerning, among other things, pronunciation, stress, syllabification, part of speech, and frequency (based on the INL corpora). Recently, the database has been made conformant with the new spelling rules for Dutch.

For spelling correction, several word lists (with hyphenation patterns) are available, including some public domain lists (`http://www.iaf.nl/Users/Meridian/words.htm`). Eurodicautom (`http://www2.echo.lu/ edic/`) is a large multilingual dictionary and terminology database, compiled by the Translation service of the EC.

The CELEX database is the only resource which is used widely in research on Dutch language and speech technology. It is an invaluable source of lexical, especially morpho-phonological, information. It is very difficult to obtain lexical information on area's that are not covered in CELEX. Syntactic valence information for a substantial number of predicates is a crucial ingredient of wide-coverage syntactic analyzers. Such information, comparable to what is provided for English in the COMPLEX lexicon (Grishman, Macleod, and Meyers, 1994) is currently not available for Dutch. Similarly, semantic information, in the form of concept networks, thesauri, etc., is not available. This is an obstacle for disambiguation, automatic translation, and information retrieval tasks. The EuroWordNet project (`http://www.let.uva.nl/~ewn/`) has as its goal the construction of a multilingual semantic lexical database (with an estimated size of 50.000 entries per language), in the style of WordNet.

2.3. Spoken Language Resources

The development of speech recognition and synthesis is heavily dependent on the availability of corpora and (pronunciation) dictionaries. The earliest annotated corpora of spoken language are the *Eindhoven* and *Groningen* (available from ELRA) corpora. More recently, the POLYPHONE-NL corpus (telephone speech) was created (by SPEX) as part of a joint European effort, and the COGEN corpus was created at ELIS, (Ghent University), as part of the Flemish short term programme on speech and language technology. Pronunciation information can be found in CELEX, and in FONILEX, a Flemish counterpart of CELEX developed at Leuven University focussing on pronunciation only, and also developed as part of the *korte termijnprogramma*.

Recently, the Dutch and Flemish government started a joint project (*Corpus Gesproken Nederlands*), aiming at the creation of a corpus of spoken Dutch with a size of approximately 10 million words. The corpus will be supplemented with detailed phonological as well as morphosyntactic (part of speech) labels, whereas a smaller part will get a syntactic annotation, making the corpus a valuable resource for the development of spoken language applications.

2.4. Language Specific Tools and Standards

Resources for language and speech processing not only include lexical databases and corpora, but also standards for the annotation of corpora and coding of dictionaries, tools to explore these databases, and tools to perform (elementary) language and speech processing tasks, and evaluation standards. In as far as these tools and standards are language specific, they are also a part of the infrastructure for language and speech technology for a given language. The situation for Dutch in this respect is extremely weak.

Standards for the phonetical encoding of pronunciation dictionaries appear to be well investigated. The CELEX database provides an encoding (DISC) which can automatically be translated into several standard encodings (such as SAMPA), and which has been used (with minor modifications) in related projects such as FONILEX. For part of speech tagging, several tag-sets are in use, sometimes product specific, sometimes developed in the context of European projects (such as PAROLE), and sometimes derived from the categories distinguished in the ANS. No consensus exists, a situation which apears to be a serious obstacle for the development of larger annotated corpora.

Part of speech disambiguators for Dutch exist mainly as research tools (Daelemans et al., 1996; Berghmans, 1994)). Xerox has developed a tagger for Dutch which is avialable for commercial and research purposes (`www.rxrc.xerox.com/research/mltt/Tools/pos.html`).

Resources for development of wide-coverage grammars are lacking. There are no explicit, formal, grammars for Dutch, no test suites or tree-banks, no valence dictionaries, etc. Given this state of affairs, it is not surprising that there have been only a few attempts to develop wide-coverage, robust, grammars for Dutch. One such grammar is the AMAZON parser using CELEX as its dictionary (Oltmans, 1994) . Given the fact that CELEX does not provide valence information, it is not surprising to see that the resulting system can only be seen as a first attempt at a wide-coverage grammar.

A consequence of the lack of evaluation metrics, benchmarks, tree-banks, tagged corpora, etc., is that evaluation of existing systems is practically impossible. Within the current project, the evaluation of existing tools had to be restricted to the reactions of researchers who had experience in working with these tools, and no formal evaluation could be carried out. We believe that this state of affairs is not only a problem for determining the quality of existing systems, but also prevents the development of new and better tools.

2.5. Conclusion

The language and speech infrastructure for Dutch, as far as resources are concerned, needs to be improved drastically in at least three area's. The available lexical databases need to be extended with fine-grained syntactic and semantic information. Publicly available medium-size and large corpora with part of speech, syntactic, or semantic information are currently virtually non-existent. Yet, such corpora are crucial for the development of robust natural language systems. Finally, the development of tools and standards needs to be stimulated, as these are hardly available for Dutch, and yet form an essential part of the language and speech infrastructure. With the development of a large Dutch-Flemish corpus of spoken language, the state of affairs for spoken language resources is looking rather promising.

3. Coordinating Institutions

Creating an infrastructure for language and speech technology implies that certain coordinating institutions exist, which set priorities, ensure that resources are not only developed but also maintained and distributed, which perform evaluations, which coordinate between industry, university research, national governments, and the European Union, etc.

We interviewed approximately thirty researchers from universities and industry, as well as a number of representatives of reserach organizations. Practically everyone was of the opinion that coordination of research and development in the area of language and speech technology is currently practically absent in the Netherlands and Flanders.

This has a number of undesirable consequences, especially for the development and distribution of resources. First, as most of the development of resources takes place within reserach projects with funding for limited periods of time, maintenance of resources after completion of the project is practically impossible. For instance, it is currently unclear how the CELEX database will be maintained in the near future. Second, the development of resources is often intended for use within a specific project. The use and distribution of resources outside the project is often not taken into consideration properly, and thus several potentially interesting resources cannot be used. Third, collaboration on language and speech technology in which both the language as it is spoken and written the Netherlands and as it is spoken and written in Flanders is taken into consideration, is rare. Sometimes this cooperation can be achieved within a European context (as was the case for EUROTRA, with a group based in Utrecht (NL) and Leuven (B)). Another example of collaboration is the Dutch-Flemish development of a spoken language corpus.

Such collaboration is incidental, however, and opportunities for collaboration are clearly missed. One example is the Dutch NWO priority programme on language and speech technology and the Flemish short term programme for speech and language technology, which were initiated and carried out independently. Another example concerns a recent program for Dutch-Flemish cultural cooperation (initiated by the *Vlaams-Nederlandse Comité voor Nederlandse Taal en Cultuur* (VNC), which stimulates collaborative research on language and literature. Language and speech *technology* projects appear to be outside the scope of this program.

Finally, the CELEX project has been carried out as a national (Dutch) initiative. A seperate effort (FONILEX, see `bach.arts.kuleuven.ac.be/fonilex/`) was required to provide a pronunciation dictionary for Flemish. Other aspects of the Flemish vocabulary have not been accounted for, however.

The situation for language and speech technology contrasts with that in more traditional areas of linguistics. The INL (*Instituut voor Nederlandse Lexicografie*) is a leading institute for lexicography and is funded by the Dutch and Belgian government. The NTU has actively promoted the development of dictionaries (such as the bilingual dictionaries under development by the *Commissie Lexicale Vertaalvoozieningen*) and grammars (such as the ANS) by teams in which both Dutch and Flemish partners cooperate.

The NTU could play an important role as coordinating institution for language and speech technology in the Netherlands and Flanders. As an intergovernmental organization responsible for language policy, it is capable of coordinating between initiatives on both sides of the border. As an organization which has been actively involved in the development of dictionaries and grammars, it should also be capable of stimulating the development of electronic linguistic resources, as well as ensuring the maintenance and distribution of these resources over longer periods of time.

4. Education

To improve the position of Dutch in the field of language and speech technology one not only needs an infrastructure and sufficient resources, one also needs people to carry out research and development. Most universities in the Netherlands offer programs (under various names) aiming at either computational linguistics or speech technology. These programs are typically located within the Faculty of Arts (the computational linguistics specialization of Computer Science at the University of Twente being the only exception). In Flanders, the possibilities for studying computational linguistics are limited to individual courses within a more general linguistics programme and postgraduate courses, while speech technology programs are exclusively offered within computer science departments

Several of our interview partners indicated that their institute or company had trouble finding adequate staff. This may be a consequence of the fact that computational linguistics is far more technical than the other programs within the Arts faculties, and tends to attract only small numbers of students. As the field of language and speech technology matures, it puts a stronger emphasis on applied, data-oriented, and statistical research. This tendency appears to be an indication that education programs should be organized more as interdisciplinary studies, open not only to Arts students, but also to students with a background in Computer Science and related fields.

The possibilities for creating such interdisciplinary programs on language and speech technology are limited at the moment. However, there are several European Socrates projects aiming at the innovation of programs for language and speech technology (the ACOHUM network working group on computational linguistics and language engineering, the Phonetics and Speech communication thematic network, and the proposed "European Masters in Language and Speech") which might provide input to further development of these ideas.

5. Towards an infrastructure for Dutch language and speech technology

In the information age, a language for which not all the necessary resources, tools, and end products are available will be marginalized. The current infrastructure for language and speech technology for Dutch is seriously underdeveloped and, if no new initiatives are being developed, might lead to a situation where the language will only be used in informal settings.

The development of resources for Dutch language and speech technology has not been at the center of attention of a single, coordinating, institution. Consequently, resources have only been developed incidentally and on a small scale, with little or no attention for the needs of the research community as a whole.

The recent initiatives of the NTU might well be a first step towards improving this situation. As a public, intergovernmental, institution, the NTU can play a central role in the development of resources, the maintenance of such resources, and in monitoring the quality of available resources and language and speech products. Furthermore, it can coordinate research initiatives, advise funding bodies (research councils, ministries, the European Union, etc) on their policy with respect to resource development, and provide a platform where university researchers, industry, and funding bodies can meet and exchange ideas.

6. Acknowledgements

The research reported on here was carried out for the Nederlandse Taalunie as part of the project *De infrastructuur van het Nederlands in taal- en spraaktechnologie*. The views expressed in this paper are the sole responsibility of the authors, however. We would like to thank Jeannine Beeken, Walter Daelemans, Elisabeth d'Halleweyn, and Bruno Krekels for their support.

7. References

Baayen, R. H., R. Piepenbrock, and H. van Rijn. 1993. *The CELEX Lexical Database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Berghmans, J. T. 1994. WOTAN: een automatische grammaticale tagger voor het Nederlands. Masters Thesis, Katholieke Universiteit Nijmegen.

Bonnema, Remko, Rens Bod, and Remko Scha. 1997. A DOP model for semantic interpretation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the ACL*, pages 159–167, Madrid.

Daelemans, Walter, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27, Copenhagen.

de Jong, Eveline D. 1979. *Spreektaal : woordfrequenties in gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema. Werkgroep Frequentie-onderzoek van het Nederlands.

den Boogaart, P.C. Uit. 1975. *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema. Werkgroep Frequentie-onderzoek van het Nederlands.

Grishman, Ralph, Catherine Macleod, and Adam Meyers. 1994. Complex syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Kyoto.

Haesereyn, W., K. Romijn, G. Geerts, J. De Rooij, and M.C. Van den Toorn. 1997. *Algemene Nederlandse Spraakkunst*. Martinus Nijhoff Uitgevers Groningen / Wolters Plantyn Deurne. Tweede, geheel herziene, druk.

Marcus, Mitch, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nederhof, Mark-Jan, Gosse Bouma, Rob Koeling, and Gertjan van Noord. 1997. Grammatical analysis in the OVIS spoken-dialogue system. In Julia Hirschberg, Candace Kamm, and Marilyn Walker, editors, *Interactive Spoken Dialog Systems: Bringing Speek and NLP together in Real Applications*, pages 66–73, Somerset, NJ. Association for Computational Linguistics.

Oltmans, Erik. 1994. AMAZON in AGFL. Een contextvrije herschrijfgrammatica voor de structurele component van het AMAZON/CASUS-systeem, beschreven in het AGFL-formalisme. Doctoraal scriptie, Universiteit Nijmegen.

Renkema, Jan. 1996. *Woordenlijst Nederlandse Taal*. Den Haag: Sdu, Antwerpen: Standaard. (in opdracht van de Nederlandse Taalunie).

Schuurman, I. 1997. ANNO, a Multi-functional Flemish Text Corpus. In J. Landsbergen, J. Odijk, K. Van Deemter, and G. Veldhuijzen van Zanten, editors, *Computational Linguistics in the Netherlands 1996. Papers from the Seventh CLIN Meeting*, pages 161–176. Technische Universiteit Eindhoven.