The Benefits of a Product-Independent Lexical Database with Formal Word Features

Janneke Froon and Franciska de Jong

Abstract
Dictionaries can be used as a basis for lexicon development for NLP applications. However, it often takes a lot of pre-processing before they are usable. In the last 5 years a product-independent database of formal word features has been developed on the basis of the Van Dale dictionaries for Dutch. The database has proven to be useful in various NLP applications. This paper describes the history, some advantages and the constraints in the development of this database.

1. Introduction

Using traditional dictionaries as a starting point to construct lexicons for NLP applications is obvious. Dictionaries can be deployed in end-user applications such as spelling-correction tools and development tools, for instance phonological lexicons for automatic speech recognition.

Several attempts to apply machine-readable dictionaries have been reported in the literature, for instance Boguraev and Briscoe (1989), Binot and Jensen (1993), Braden-Harder (1993) and Wilks et al (1996). The focus of this work is mostly on the application of semantic information in the dictionaries. Semantic information is only one type of possible information in dictionaries. They can also be used to derive formal word features like hyphenation, pronunciation, word structure and inflection.

This paper describes the development of a lexical database on the basis of the Van Dale dictionaries for Dutch containing formal word features. Reusability of the data has been a major goal while developing this database. Reusability has always been an important concept in the development of lexical databases (cf. Calzolari 1990). For the Van Dale publishing house this concept is important since the information in the database is meant to be used as a source for various dictionaries and for the development of other applications such as text-to-speech systems.

Rather than adapting the dictionary resources to particular applications, a resource is created from which the content needed to realize new products and applications can be extracted. This paper describes the history of the database, the advantages that the database proved to have, and the limitations of the development.

2. History

Van Dale Lexicografie is a major lexicographic publisher in the Netherlands. The Van Dale dictionaries are commonly considered to be the most authoritative dictionaries in the Dutch-speaking community. The development of a database using the dictionary files of Van Dale has been a gradual process. In this section, the background of this process is sketched.

2.1 Twenty years ago

Until the early 1980s, paper dictionaries published by Van Dale each had their own author. The authors were responsible for the contents of the dictionary. The role of the editors was to check the data for textual correctness and ensure that the books got printed and sold. Since each author was responsible only for one dictionary, the contents of the dictionaries were not related, causing unintended differences. At best, the same implicit lexical standards were adhered to by all authors.

2.2 Fifteen years ago

The situation changed during the mid-80s. A new series of bilingual dictionaries from Dutch to three languages (English, French and German) was derived from the same dictionary file of common Dutch words. The potential advantages were immediately recognized by the NLP-community and attempts were made to use this high quality source of data for the development of NLP-systems, e.g. in the machine translation project Rosetta (Rosetta 1994).

However, the in-principle ideal situation didn't last long. The different dictionaries got different authors who were again responsible for their own dictionary only. Although editors checked the material that the authors produced, the dictionaries started to drift apart and therefore many defects were reintroduced, such as inconsistencies of products and the fact that products were not taking advantage of each other's corrections. As a result of the independent editing of the dictionaries the same work was sometimes done more than once and some work could never be done, because it was too expensive for a single dictionary and it was impossible to combine the various efforts.

2.3 Seven years ago

In 1995 the Dutch government changed the rules of spelling. For Van Dale this implied a spelling adaptation of Dutch words in about 50 books, containing 6 million words. This had to be done in a very short period, since users would have stopped buying if new editions were expected soon. Besides the time pressure, the job to be done caused a problem too. The editors of dictionaries may be experts in the lexicographic area, but not in the new spelling regulation that had to be adhered to and internalized.

These new circumstances forced the publisher to adopt a new working method: creating a product-independent spelling database. For this aim, it was still necessary to look at the spelling of about half a million different words. However, at least, this had to be done only once, and not over and over again for all books.

The use of such a product-independent database of spellings proved to have many advantages and was soon followed by other product-independent databases, such as one with hyphenation information and another with pronunciation information. These databases were integrated into a single database with formal word features.

The database had to overcome problems like those described in Quazza and Van den Heuvel (2000), as phonemic information in dictionaries has a limited usability because it is available only for exceptional words and base words, not for all related words.

2.4 Current applications

The Van Dale database with formal word features has been used in various applications. The Dutch text-to-speech system *Fluency*[1] uses the phonemic transcriptions in the database. The speech synthesis of *Fluency* has been used in several of Van Dale's electronic dictionaries and in the *Fluency e-mail-reader*, a tool which automatically announces and reads aloud e-mail messages[2] . Furthermore, the database is the basis of the *Van Dale Spellingcorrector* (VDS 2000), a spell-checker for Dutch. These two applications are examples of end-user applications.

Another type of use that illustrates the importance of product-independent databases for the NLP research community, is the application in development projects. For instance, the Druid project[3] and the ECHO project[4] use the pronunciation information of the database to build an acoustic model for a system for Dutch speech recognition. This speech recognition module is meant to play a role in the development of technology for spoken document retrieval, particularly in video retrieval (see Ordelman et al 1999).

A similar product-independent approach has been used in the development of the VLIS database[5] from an earlier database which contains semantic word features. This semantic database has been used in the Dutch version of EuroWordNet (see Vossen et al 1999), and is now available under

license for commercial use. In cross-language retrieval tools the semantic database has proven to be valuable, in particular in the development of the disambiguation method applied in the Twenty-One search engine, which has been evaluated at several TREC-conferences[6] (cf. Hiemstra and Kraaij 1999, Hiemstra and De Jong 1999).

3. Advantages

The database with formal word features proved to have many advantages, three of which are illustrated here. Firstly, the consistency of products is easily attainable. Secondly, the information in the database is richly encoded. Finally, the information is flexible. As the illustration below will underline, these aspects are beneficial for the production of book dictionaries, for the development of NLP-products targeting the end-user market, and for the level of support for NLP research teams.

3.1 Consistency of products

For Van Dale it is important that the products are consistent. As explained above, Van Dale is an authority in the field of lexicographic information. Therefore, its credibility and authority can be damaged if different products manifest different information. If, for instance, the 'Groot Woordenboek der Nederlandse Taal' (large dictionary of the Dutch language, Geerts and Den Boon, 1999) contradicts the dictionary 'Hedendaags Nederlands' (contemporary Dutch, Van Sterkenburg, 1996) in the pronunciation of a word, the user of these dictionaries cannot rely on the information anymore.

Word information like hyphenation and pronunciation is sometimes difficult to describe and causes differences. Even spelling, which seems to be strictly regulated and therefore unambiguous, has many uncertainties. For instance, the use of hyphens in words like on-line-verbinding (on line connection) is not unambiguously prescribed.

The consistency of dictionaries is guaranteed if the information is drawn from the same source every time a new dictionary file is assembled. The dictionary doesn't contain the information itself, but only a dynamic link to the information that is in the central database. When a new edition is prepared, information from the central database is imported into the dictionary. The imported information cannot be edited in the product file itself. If changes are needed, for instance because errors are found, they have to be stored in the central database. All dictionaries will profit from the corrections.

Not only do book dictionaries thereby become consistent, but so do all applications derived from the database. New insights are shared in every product. The overall quality of the products can reach a higher level.

3.2 Richness of data

The second advantage is the richness of the database. A product-independent database will tend to represent data on a more abstract level than when the data are assembled for a special product, thereby resulting in a richer resource. The most important reason is that while working on the database, it is often not clear at first which information will be needed in which product. It is not desirable to leave information out just because it is not needed at the moment when the database is constructed.

The next two examples illustrate the benefits of rich codes in phonemic representations and in hyphenation marks.

The first example is the representation of underlying phonemes while representing pronunciation. In the pronunciation of *bezettoon* (*busy signal*) only one *t* is heard. The *t* of *bezet* (*busy*) disappears because of degemination with the *t* in *toon* (*signal*). If the representation is needed for the phonemic transcription in a dictionary, one *t* will do. If the representation is used to synthesize the pronunciation, a single *t* will sound unnatural, and the presence of a second *t* has to be indicated. A

code indicating such a special *t* can cause the dictionary generator to delete it, while causing the speech synthesis tool to pronounce the *t*'s in a special way. The same code can be used in different products.

A second example originates from the hyphenation of words. An investigation of hyphenation for Dutch showed that it was better to indicate syllable boundaries instead of hyphenation, although syllable boundaries coincide often - but not always - with hyphenation positions. The reason is that there is a Dutch hyphenation rule that prohibits hyphenating on a position that would cause a syllable of one letter to be separate from the rest of the word on a new line. So *radi-o* (id.) and *a-demen* (*breathe*) are not allowed. The rule also applies when a single-letter-syllable is separated from the rest of a compounding part or derivational part. Therefore *radi-otoestel* (*radio+toestel: radio set*) and *bea-demen* (*be+ademen: insufflate*) are not allowed. Lexicographers who enter the hyphenation marks serve two aims: firstly, they have to indicate the syllable boundaries; secondly, they have to check whether a single-letter-syllable will be created. In the encoding, this distinction has to be kept. For *ademen* this causes the encoding a:de-men, where a colon indicates a syllable boundary that doesn't coincide with a hyphenation position.

In conclusion, every stage in the production of the word information should be represented. If every step of the thinking process is explicitly encoded, it is possible to correct the result without having to recall what was going on. Besides, a rich representation has an advantage in itself. The maintainability of the database greatly improves if rich representations are being exploited. Using rich codes, it is possible to infer which processes are responsible for the formation of, for instance, pronunciation or hyphenation. By checking the soundness of these processes, the quality of the data can be improved.

3.3 Flexibility of data

The third advantage of a database is the flexibility of the data. When needed for new products, the information in the databases is readily available, and because of the richness of the data there will be no obstacles in adapting it to a new product. Therefore the database may aspire to do things with the available data that are otherwise unattainable.

Gibbon (2000) points out that phonemic transcripts from machine-readable dictionaries require "extensive pre-processing" before they can be used in system lexicons. However, in the Van Dale database the phonemic information is simply there, readily applicable in a variety of products.

An example is the use of phonemic information in the *Van Dale Spellingcorrector*, a spell-checker for Dutch which benefits from this information in two different ways. The first is the use of phonemic information in the assembling of a list with predicted errors in the spell-checker. This list is used to detect quickly and properly correct the spelling errors in the list. A lot of predictions about errors can be made on the basis of problematic spelling patterns. For instance the *c* in Dutch is often confused with *k*, resulting in well-known errors like *kontakt* for *contact* and *aktie* for *actie*. A large group of errors is caused by writers staying too close to modern pronunciation, disregarding historical aspects of the spelling of certain words. For instance the *b* in *ambtenaar* is often incorrectly omitted, because it cannot be heard. Another example is the word *quitte* that has a spelling which is very different from its pronunciation /kit/, resulting in the erroneous spelling error *kiet*. These spelling errors can be predicted if phonemic representations are used. A whole class of plausible errors can be incorporated in the spell-checker, that are beyond reach if phonological information isn't available.

The second way pronunciation information is used in the *Van Dale Spellingcorrector* is in finding homophones. Homophones are words that are pronounced similarly, but have different spellings. Examples in Dutch are *biljart* (*game of billiards*) and *biljard* (*number, thousand billion*), *boxer* (*type of dog*) and *bokser* (*someone who boxes*), and in English *discrete* (*separate*) and *discreet* (*tactful*). These words cause problems, because writers tend to mix them up, writing for instance

biljart when the number is meant. A spell-checker can be improved if attention is paid to the difficulties with homophones, by using the pronunciation information in the databases.

Due to, among others, the use of pronunciation information in various ways, the *Van Dale Spellingcorrector* can compete with spell-checkers for Dutch that are provided with word processors. Without the information in the database, pronunciation information would have been out of reach because of the high costs. The database provides an affordable opportunity to incorporate into a spell-check this valuable information source.

4. Constraints

Although the development of a multi-purpose database has many advantages, it has a price in both time and money. If information has only one purpose and can be used in one product only, the cost effectiveness is not optimal. However, every time the information is reused, the return on investment potentially increases and for some applications the use of a product-independent database may be the only source of data that is affordable.

For a company, making profit is crucial, and it is tempting to choose making money in the short term. The development of a multi-purpose database undoubtedly has advantages, but especially in the long run. There is thus always the risk that developments are stopped for economic reasons, just before the end-goal is reached, because the remaining work isn't profitable enough. Collaboration with non-profit institutions, such as NLP-research groups with research capacity and/or knowledge, can then be an incentive for sustained resource development.

5. Conclusion

Building a multi-purpose database for formal word features in Dutch, or any other language, is a difficult and expensive task. However, the more the information is used in applications, the cheaper the information gets. The product-independence of the database pays itself back in the long run.

The advantages of such a product-independent database are indisputable. The information in the dictionaries is easily available, more consistent and rich, which benefits any application using the database. However, endurance is demanded of the developing companies to make these advantages commercially viable.

Notes

1 http://www.fluency.nl
2 http://www.emaillezer.nl
3 http://dis.tpd.tno.nl/druid
4 http://pc-erato2.iei.pi.cnr.it/echo
5 Vlis is the Van Dale lexicographic information system and semantic network, in existence since 1992
6 The Twenty-One search engine is distributed in the Netherlands by Irion Technologies http://www.irion.nl.

References

Binot J. and K. Jensen. 1993. 'A Semantic Expert Using an Online Standard Dictionary.' In *Natural Language Processing: The PLNLP Approach*, K. Jensen et al. (ed.). Dordrecht: Kluwer Academic Publishers, pp. 135-149.
Boguraev B. and E. Briscoe (ed.). 1989. *Computational Lexicography for Natural Language Processing*. Harlow: Longman Group UK.
Braden-Harder L. 1993. 'Sense Disambiguation Using Online Dictionaries.' In *Natural Language Processing: The PLNLP Approach*, K. Jensen et al. (ed.). Dordrecht: Kluwer Academic Publishers, pp. 247-263.
Calzolari N. 1990. 'Lexical Databases and Ttextual Corpora: Perspectives of Integration for a Lexical Knowledge-Base.' In *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*,

U. Zernik (ed.). Hillsdale, NJ: Laurence Erlbaum, pp. 191-208.
Geerts G. and Den Boon C. 1999. *Van Dale Groot Woordenboek der Nederlandse Taal*. Utrecht: Van Dale Lexicografie.
Gibbon D. 2000. 'Computational Lexicography.' In *Lexicon Development for Speech and Language Processing*, F. Van Eynde and D. Gibbon (ed.). Dordrecht: Kluwer Academic Publishers, pp. 1-42.
Hiemstra D. and De Jong F. 1999. 'Disambiguation Strategies for Cross-language Information Retrieval.' In *Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries: ECDL'99*. Heidelberg: Springer-Verlag, pp. 274-293.
Hiemstra D. and W. Kraaij. 1999. 'Twenty-One at TREC-7: Ad-hoc and Cross-Language Track.' In *Proceedings of the seventh Text Retrieval Conference TREC-7*. NIST Special Publication 500-242, pp. 227-238.
Ordelman R., A. Van Hessen and D. Van Leeuwen. 1999. 'Dealing with Phrase Level Coarticulation (PLC) in Speech Recognition: A First Approach.' In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*. Cambridge: Cambridge University Press, pp. 64-68.
Quazza S. and H. Van den Heuvel. 2000. 'The Use of Lexica in Text-to-Speech Systems.' In *Lexicon Development for Speech and Language Processing*, F. Van Eynde and D. Gibbon (ed.). Dordrecht: Kluwer Academic Publishers, pp 207-233.
Rosetta M.T. 1994. *Compositional Translation*. Dordrecht: Kluwer Academic Publishers.
Van Sterkenburg P. 1996. *Van Dale Groot Woordenboek van Hedendaags Nederlands*. Utrecht: Van Dale Lexicografie.
VDS 2000. *Van Dale Spellingcorrector voor MS-Word*. Utrecht: Van Dale Lexicografie.
Vossen P., L. Bloksma and P. Boersma. 1999. *The Dutch Wordnet*. Amsterdam: The University of Amsterdam.
Wilks Y., B. Slator and L. Guthrie. 1996. *Electric Words - Dictionaries, Computers, and Meanings*. Cambridge, MA: MIT Press.

About the authors

Janneke Froon is a language technology coordinator for Van Dale Data. In 1997 she graduated as a computational linguist at the University of Utrecht, and has since been working for Van Dale using language technology to enhance lexicographic information for dictionaries and language products. She is preparing her PhD thesis, researching the improvement of spell-checkers using lexicographic information, and is working on the development of a large lexicographical database that integrates formal and semantic features.
janneke@vandale.nl

Franciska M.G. de Jong teaches language technology at the Computer Science Department of the University of Twente, Enschede, and works for TNO-TPD in Delft as a consultant in the area of multimedia technology. Her background is in theoretical and computational linguistics, and she worked as an assistant researcher at the Faculty of Arts of the University of Utrecht (1980-1985) and as a senior researcher at Philips Research on the Rosetta machine translation project (1985-1992). She is frequently involved in international program committees, expert groups and review panels, and has initiated a number of EU projects. Professor de Jong is

currently coordinating several projects aimed at multimedia indexing and retrieval, and chairs the Advisory Board of Van Dale Lexicografie.
fdejong@cs.utwente.nl

About van Dale Data
For over 100 years Van Dale Lexicography has been recognized as the foremost dictionary publishing source in the Netherlands. Since 1989 it has been publishing electronic dictionary applications. Van Dale Data BV has been an independent enterprise of van Dale Lexicografie BV since 1999, focusing on the management and commercial operation of linguistic databases and their applications within language and speech technology. Van Dale is part of the Veen Bosch en Keuning publisihing group.
www.vandaledata.nl

Van Dale Lexicographical Information System (VLIS)
o semantic hierarchical network
o multilanguage information
o phraseology, idioms
o classification
o word attributes
Contents
o 170,000 Dutch word definitions, 1,070,000 translations
o 145,000 semantic relationships
o 225,000 examples, 525,000 translations
o 250 different thematic labels
Applications
o lexicographical products for different media
o multilingual dictionaries
o multiple text retrieval and analysis techniques
o automatic classification and summarizing of texts
o development of databases
o building of indexing tools

Word attributes database and language technology
o spelling and hyphenation
o expansions
o word class
o frequency
o context relationships
o pronunciation
o transcriptions
o morphology
Quantity
o 250,000 Dutch keywords
o 1,250,000 expansions
Quality
o checks using language rules
o relationships between words
o inheritance of attributes
Enrichments
o editorial expertise
o parameterization
o corpus
o frequency
o reverse engineering

o hyphenation, expansions

Applications

o electronic dictionaries
o language tools
o speech applications
o rhyme engine
o games

Speech technology

o pronunciation indication for all words
o lexicon of more than 180,000 word forms
o rules for unknown words
o rules for interpreting numbers, punctuation, etc.
o prosody generation
o rules for length of sounds in context
o rules for sentence melody

Diphone synthesis

o diphone
o diphone database
o MBROLA synthesizer

Applications

o talking dictionaries
o tools for handicapped persons
o games
o fluency e-mail reader
o telephony and Internet

K Dictionaries Ltd
10 Nahum Street, Tel Aviv 63503 Israel
tel: 972-3-5468102 • fax: 972-3-5468103
kd@kdictionaries.com