

Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ...

I. Schuurman*, W. Goedertier†, H. Hoekstra‡, N. Oostdijk◊, R. Piepenbrock◊, M. Schouppe*

*Center for Computational Linguistics, University of Leuven
Maria-Theresiastraat 21, 3000 Leuven, Belgium
ineke.schuurman,machteld.schouppe@ccl.kuleuven.ac.be

†Electronics and Information Systems (ELIS)
University of Ghent, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
Wim.Goedertier@elis.ugent.be

‡Utrecht Institute of Linguistics OTS
University of Utrecht, Trans 10, 3512 JK Utrecht, The Netherlands
heleen.hoekstra@let.uu.nl

◊Department of Language and Speech, University of Nijmegen
P.O.Box 9103, 6500 HD Nijmegen, The Netherlands
n.oostdijk,r.piepenbrock@let.kun.nl

Abstract

After the successful completion of the Spoken Dutch Corpus (1998 – 2003) the time is ripe to take some time to sit back and reflect on our achievements and the procedures underlying them in order to learn from our experiences. In this paper we will in particular pay attention to issues affecting the levels of linguistic annotation, but some more general issues deserve to be treated as well (bug reporting, consistency). We will try to come up with solutions, but sometimes we want to invite further discussion from other researchers.

Introduction

In 1998, when the Spoken Dutch Corpus (CGN) project started, we basically had to start from scratch, as there was no previous experience with the large-scale compilation and annotation of Dutch corpora, let alone a corpus of spoken Dutch. The aims set in this project were quite ambitious: to compile a corpus of Dutch as spoken in Flanders¹ and the Netherlands that would be comparable in size to the spoken part of the British National Corpus and to provide various annotations (Oostdijk et al., 2002). At the start of the project many issues remained to be resolved. This was not just limited to the design of the corpus, which had to be negotiated between different interest groups, it also applied to such questions as what annotation schemes to adopt, and what standards or what guidelines to adhere to.

Now, five years later, a corpus is available comprising some 9 million words. For the entire corpus an orthographic transcription is available, as well as linguistic annotations which take the form of part-of speech (POS) tagging and lemmatisation and the identification of multiword units. For substantial parts of the corpus additional transcriptions and annotations are available. These include (manually verified) phonetic transcription, syntactic analysis (SA), prosodic annotation and (manually verified) word alignment with the audio files. Also an automatic word alignment is provided for the whole corpus.

In section 1. we will reflect on the practical organisation of CGN, in section 2. on some more general issues and in section 3. on issues related to linguistic annotation: what did not work out the way we had expected? In the last section some recommendations will be formulated.

1. The setup of CGN: practical organisation

It will be clear that, in order to be able to complete such a large corpus of spoken language within the time allotted, many transcriptions and annotations had to be done in parallel rather than sequentially. A factor which also complicated matters was the fact that the project was a joint Flemish-Dutch undertaking. Because of funding stipulations, and to fully appreciate the linguistic differences that exist between the two communities, the Flemish part of the corpus was to be compiled by academic groups in Flanders, while the Dutch part was the responsibility of groups in the Netherlands.

1.1. Working in parallel

As stated above, we had to work in parallel at several layers of annotation due to time constraints. This could mean that, especially in Flanders, for example the phonetic transcription started, using the orthographic transcription files that were still subject to change as errors were reported from other annotation layers such as POS tagging. In general, we tried to complete one layer of annotation for a certain file before it was made available for another layer. But even in that case, it did not mean that the layer would be frozen as changes or corrections were always to be expected (cf. 2.2.).

The fact that the various stages of annotation took place in parallel gave us the opportunity to influence major decisions taken in a particular working group in case they affected another layer. Thus it was decided that, for example, the Flemish dialectal combinations of verbal forms with the cliticized personal pronoun *de* should be written as two words (*hebt de*d* etc. instead of *hebde*d*) at all levels. But although there has been a certain level of consultation between working groups, and although both project leaders

¹Flanders is the Dutch-speaking part of Belgium

were members of all project groups, it turns out that especially at the beginning more interaction would have been desirable (cf. 3.3.1.).

Yet if we had been working completely sequentially from the start, it is likely that the 'later' levels would have had no say at all on the decisions of the 'earlier' ones, so working in parallel certainly had its merits.

1.2. Working side by side

With Dutch being the official language in both the Netherlands and Flanders, it seems to go without saying that all kinds of resources should be developed together. In practice it is not that easy for all kinds of practical reasons, although in general there is an intention to do so. This joint approach has been greatly influenced by the Dutch Language Union (NTU), an intergovernmental organisation responsible for the language policy in the Netherlands and Flanders.

For CGN, two project leaders were appointed, one in each region, who conferred on a regular basis. The division worked out reasonably well, considering the oft-cited difference in mentality in both regions. The way in which people in both regions organised their work was not always similar, and it turns out that this is not necessary either. On the contrary, the fact that often two different approaches were used (for example for bug handling, cf. 2.2.) ensured that there was always a fall-back option available. One of the problems that may arise is that of (lack of) consistency, cf. 2.1.

2. Some general issues

During the process of creating an annotated corpus there are at least two issues that everyone is confronted with: firstly, maintaining consistency throughout the corpus (2.1.) and secondly, processing the reported bugs (2.2.). Moreover, a third issue is likely to occur: shortage of time (2.3.).

2.1. Consistency

The first months of the project were devoted to the creation of solid protocols both for the orthographic transcription and all stages of annotation. This has been done in binational working groups. Only after these protocols had been completed, a first trial run had been executed and the protocols updated, students were hired. After a training session (instruction plus the annotation of a couple of test files) they could start work.

CGN employed lots of students for tasks like creating transcriptions, checking POS-tags and performing semi-automatic syntactic analysis of the sentences, which made it complicated to keep all their decisions in line. A partial solution has been to give some files to several students every now and then, compare the results and go through the differences with them. Other approaches could be to encourage them to discuss difficult cases among themselves (keeping in mind that they cannot spend too much time on this, as they have to do a certain amount of work per hour!) or to ask them to mark cases where they are not confident. We really urged them to do so, and it may have worked out well, but the point remains that the trickiest constructions are the ones they do not realize to be problematic.

As far as consistency checks between the Netherlands and Flanders are concerned, there was no time to perform them. Of course, problematic constructions were discussed. The very fact that two sites were involved does not seem to be the big issue, but rather both the lack of time and the (initial) lack of awareness of problematic cases.

2.2. Bug reports

From the very beginning a bug reporting system was set up, stating for example that one is to report the bug found to the person responsible for that specific layer in either Flanders or the Netherlands and that this person has to take action within three working days and report back. And until then, one is not allowed to take action. As such, this procedure was a valid one, as it ensures that everybody is working with the same version of the files. However, in Flanders this procedure was soon considered to be too bureaucratic: there are many mistakes, especially orthographic ones, of which you can be 100% certain that you will be allowed to correct them. In such cases waiting for feedback was considered a waste of time. Gradually this more flexible approach was also used for more complex cases, although in case of doubt there would be consultation between the working groups involved. This is not to say that in the end various versions of a file would be delivered! Alignment of all levels, in order to match them all up, was executed for every internal release (see also 2.3.). It will be clear that alignment in Flanders was relatively time consuming, but this was considered worth the effort. Working this way, good tools are essential.

In the Netherlands the original bug reporting procedure was followed more strictly: groups were not allowed to change files themselves. As a consequence, due to various reasons, bug reports were piling up and therefore took a long time to be processed.

It might have been worthwhile to start out by investing more time in an at least partially automatic way of bug reporting and handling.

2.3. Time allocation

In various places in this paper it is made clear that if we had had more time, we would have done this, done that. In fact we underestimated the time we would need to get started (like writing protocols, testing tag sets), the time we would need to process the bug reports in a consistent way and the time we would need for intermediate internal releases. We had one of these almost every 6 months, and every one of them took us at least one month of preparation, putting a heavy burden on our time schedule. On the other hand, the internal releases were a good means to keep everybody alert. Even though some of these releases were requested by the funding bodies, we might have been able to do with a few less.

Another delay was caused by technical problems with the recording of telephone conversations. As telephone recordings were planned to constitute 30% of the corpus, this delay had a negative impact on all layers of the project. Also, copyright and privacy issues may cause local delays, as in our case it sometimes took a long time to ob-

tain written permission from publishers, networks and individual speakers to (re-)use their recordings and electronic resources, such as lexicons.

3. Linguistic issues

3.1. Spoken language

Especially with respect to syntactic analysis (SA) we encountered many constructions that are not covered by any of the major grammars of Dutch, (e.g. Haeseryn et al., 1997; Donaldson, 1997), although these constructions are considered to be fully acceptable in spoken language, cf. De Vries (2000). Since SA aimed to serve the needs of as wide an audience as possible, we were challenged to devise a theory-neutral annotation for such constructions. As work progressed and more and more constructions were identified, the protocol had to be extended and revised repeatedly. The same holds to a lesser extent for POS tagging. Here both the use of new words and the novel use of known words were problematic. This confirms earlier observations as that language change almost always shows up first in spoken language. As far as SA is concerned, we might have done more research into spoken language phenomena before starting the annotation, but note that for example De Vries (2000) only became available during the project.

In spoken language many sentences are fraught with hesitations, interruptions, repetitions, omissions and other speech features, which sometimes make it difficult to decide which label to use. Therefore it is recommended not to use too many categories, with too fine-grained distinctions. And of course the protocol should be very explicit, not leaving room for doubt. Sometimes underspecification may come in very handy. Note that the sheer number of tags is not that important, but the way they are defined: in POS we have been using 188 tags for pronouns (cf. Van Eynde, 2004) and that as such was not problematic, because many pronouns could only be associated with one tag and therefore few mistakes were made by the tagger.

3.2. What is standard Dutch?

Another kind of problem is related to the fact that both Southern Standard Dutch (as spoken in Flanders) and Northern Standard Dutch (as spoken in the Netherlands) had to be covered whereas these variants do not always behave in the same way. Especially in spontaneous conversations, the language used in the two regions differs to a large extent. In the past, grammarians and lexicographers working on the Dutch language have taken written Northern Standard Dutch as the norm, paying relatively little attention to the southern variant. This explains why quite a few words and constructions that are typically found to occur in Flanders so far remained undocumented. Within the project therefore one of the issues was to decide when exactly something was to be considered (northern or southern) standard or non-standard (spoken) Dutch.

This turned out to be rather complicated, especially for the southern variant. In fact there are two tendencies, the first being that also in Flanders one should use the northern variant, and the second one, that the southern and the northern variant are in fact equivalent in usage and function.

After some time the Flemish partners have adopted a very pragmatic approach. At the level of orthography those words that do not occur in Dutch and Flemish dictionaries (like Van Dale (1999), Bakema (2003)) and are not commonly used in Flemish newspapers, radio and TV newscasts and the like with the intended meaning are marked with “*d” attached to the word. But at the level of POS, those words were provided with one of the regular tags whenever the tagset designed for standard Dutch was applicable. Other words got simplified tags like *N(soort,dial)* (for dialectal common noun).

A similar approach of dictionary look-up to verify non-standard usage was taken in the Netherlands.

3.3. Consequences of decisions taken elsewhere

A third kind of problem has to do with the consequences of decisions taken with respect to other levels of annotation.

3.3.1. How to recognize a sentence

In an early stage of the project it was decided that in the orthographic transcription only the full stop, question mark and ellipsis points were to be used as punctuation marks. The reason behind this decision was that it is very hard for the transcribers to use all the other punctuation marks, especially the comma, in a consistent way. As a consequence, at many places where one would expect a comma, either no punctuation mark at all or a full stop occurs. Each of these has its own complications.

A second complication was the ordering of sentences uttered by different speakers at (almost) the same time. Since the utterances in the orthographic transcription are represented in *tiers*, speaker by speaker, there is no problem there. In POS and SA, however, everything is ordered chronologically, based on the time markers in the orthographic transcription. But these do not necessarily coincide with sentence boundaries, i.e. a *chunk* of speech, delimited by two time markers, may contain several sentences. That is why in dialogs annotated for POS or SA, the answer to a particular question may show up before that question! This would not occur if the orthographic transcription contained the *exact* starting time of *every* sentence, but adding these would aggravate the orthographic transcription task and complicate handling of bug reports (every new sentence boundary will also need a new time marker). The problem could be solved using the time information from the automatic word alignment, but this information was only available at the end of the project, i.e. too late.

Next to this problem with respect to the ordering of units/sentences, there is another problem as well. Although the protocol for orthographic transcription states clearly that the punctuation marks should be placed on a syntactic base, it turned out that the orthographic transcribers were regularly misled by (long) pauses in the speech signal. This is explicable because they are focused more on the speech signal and the words pronounced and less on the syntactic structures. In spontaneous speech it can be very difficult to distinguish the syntactic sentences. In Flanders POS and SA had permission to correct punctuation errors immediately, in the Netherlands these errors were reported to the orthographic level.

3.3.2. The consequences of the spelling

This may seem an odd issue, but it certainly was an issue for Dutch. There is a list with the official spelling of words (Groene Boekje), and there are official rules. The point is that there are many cases in which the meaning of the sentence prescribes how a certain word is to be written, or rather whether it is to be written as one word or as two words. This is the case with separable verbs, which are comparable in function and some of their behaviour to English phrasal verbs:

- (1) a) *ze zijn samen gekomen* (they arrived together)
b) *ze zijn samengekomen* (they had a meeting)
- (2) a) *dat ze er op is gevallen* (that she attracted attention there)
b) *dat ze erop is gevallen* (that she fell on it)

In some cases it is even more difficult to determine the correct spelling, as dictionaries and style guides apply the rules differently, or base their decisions on the rather ill-defined criterion of frequency of collocational usage.

Still, spelling issues cannot be ignored, for example for the identification of multiword units.

Another problem concerns our treatment of the cliticized word *da's* (short for *dat is* (litt. *That is*)). The people dealing with orthography wanted to treat it as one word (and that is indeed the way it is usually spelled), while POS and SA were in favour of spelling it as *da 's*. As a compromise, POS and SA were allowed to spell it as two words, whereas it remained one word at the other levels, bringing up the all too familiar adage of 'we can write a script to take care of that, no problem'.

This procedure certainly does not bear repeating, as we would now opt for spelling it as two words at all levels. However, as things stand, it is far from straightforward to do so, due to the layer of word alignment which is based on the speech signal. This automatically treats the speech signal *da 's* as one unit, making it necessary to manually verify the times for the boundary between *da* and *'s*.

4. Recommendations

- Start off by investing in writing robust, well-defined protocols for each layer of annotation. Discuss them extensively among working groups, feel free to modify them during the early stages of the project on the basis of various types of feedback, but remember to finalize them after a pilot project to ensure consistency within and between all levels.
- Invest in a serious pilot project (10,000 words) in order to detect all the teething troubles (protocols, formats, tools)
- Realize that all annotations will be based on the orthographic transcription. Its quality should therefore be very good. Do not try to save money by hiring people without all the necessary skills.
- Do not underestimate the time you need for preparing protocols and writing the necessary tools to check (consistency of) annotations.
- Do not underestimate the time needed for intermediate releases.

- Realize it may take a lot of time and effort to negotiate contracts with (non-)commercial suppliers of tools and resources.

- Keep in mind the delicate issues of copyright and privacy protection. Never make secret recordings of people and always ask their written permission to use the fragments.

- Try to set up a clear-cut (semi-)automatic system for bug reporting and/or checking the consistency of the layers and distribute the results among all relevant groups.

- Design a comprehensive, centralized database system for version and revision management of (stable) annotation files, documents and meta-data. Always log the access and the changes made to files.

- Be aware of the impact a decision at one level of annotation can have on other ones. Always consult with the relevant working groups before making decisions or implementing changes.

- Organize a group of external experts for consultation about specialist topics. Also, create incentives for a well-informed user group in order to generate valuable feedback on linguistic and presentational issues. Keep all relevant interest groups abreast of project developments by sending out regular newsletters and organizing workshops.

- Create a facility allowing for handling of bugs and the like after the lifetime of the project (cf Beeken and Van der Kamp, 2004). Make an inventory of all tools and resources employed and document them in such a way to make them suitable for re-use by follow-up and related projects.

5. Acknowledgements

This paper was supported by the "Spoken Dutch Corpus" (CGN) project which is funded by the Netherlands Organisation for Scientific Research (NWO) and the Flemish Government.

6. References

- Bakema, P. (2003). *Vlaams-Nederlands Woordenboek*. Antwerpen/Utrecht: Standaard Uitgeverij/Het Spectrum.
- Beeken, J.C.T. and P.H.J. van der Kamp (2004). The Centre for Dutch Speech and Language Technology (TST Centre). In *Proceedings of LREC 2004*, Lisbon
- De Vries, J. (2000). *Onze Nederlandse Spreektaal*. Den Haag: Sdu Uitgevers.
- Donaldson, B. (1997). *Dutch: a Comprehensive Grammar*. London: Routledge.
- Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij and M.C. van der Toorn (1997). *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Martinus Nijhoff Uitgevers/Wolters Plantyn.
- Oostdijk, N, W. Goedertier, F. Van Eynde, L. Boves, J-P. Martens, M. Moortgat and H. Baayen (2002). Experiences from the Spoken Dutch Corpus. In *Proceedings of LREC 2002*, vol. 1, pp 340-347.
- Van Dale. *Groot Woordenboek der Nederlandse Taal*. 13rd, revised ed. (1999)
- Van Eynde, F. (2004). *Part of speech tagging and lemmatising*. Technical report. Leuven: CCL