The Syntactic Atlas of the Dutch Dialects

A discussion of choices in the SAND-project

Sjef Barbiers and Hans Bennis
Meertens Institute, Amsterdam

Abstract:

This paper discusses some of the advantages and disadvantages of the various choices we had to make in order to realize the Syntactic Atlas of the Dutch Dialects (SAND) in a relatively short period. The idea is that by presenting the SAND in this way, we enable the ScanDiaSyn project and other new dialect syntax projects to profit from our experience in a similar enterprise. The presentation and explicitation of the choices we had to make, the problems we had to face and the mistakes we have made will not necessarily be the same choices, problems, and mistakes that will arise in the Scandinavian project, but it might give an indication of where problems may be expected and how mistakes may be prevented.

1. Introduction

June 2005: the first volume of the Syntactic Atlas of the Dutch Dialects (SAND) has appeared in a Dutch and an English version (Barbiers et al. 2005). The SAND is the result of a Flemish-Dutch project that started in January 2000. The object of the project was to develop a database, an electronic atlas on the internet and a printed atlas of the syntactic variation that is found in varieties of Dutch in the Netherlands, Belgium, and France. June 2006: the electronic database DynaSAND is launched (http://www.meertens.nl/sand/zoeken/). The second and last volume of the paper atlas will appear in 2007, thereby marking the end of the SAND-project.

Given the fact that the symposium in Leikanger (and the proceedings of this symposium) marks the beginning of the Scandinavian Dialect Syntax-project (ScanDiaSyn) we will in this paper discuss some of the advantages and disadvantages of the various choices we had to make in order to realize the SAND in a relatively short period. The idea is that by presenting the SAND in this way, we might enable the ScanDiaSyn project to profit from our experience in a similar enterprise. The presentation and explication of the choices we had to make, the problems we had to face and the mistakes we have made will not necessarily be the same choices, problems, and mistakes that will arise in the Scandinavian project, but it might give an indication of where problems may be expected and how mistakes may be prevented.

In this introduction we will list a number of problems that we encountered during our project. In later paragraphs we will discuss most of these issues in more detail. Problemetic issues turned out to be: (i) the

Nordlyd 34: 53-72, © Barbiers and Bennis 2007
Scandinavian Dialect Syntax 2005
Edited by Kristine Bentzen and Øystein Alexander Vangsnes
CASTL, Tromsø. http://www.ub.uit.no/munin/nordlyd

control of the project that was located in five places in the Netherlands and Belgium; (ii) the amount of money that was needed to realize our goals; (iii) the methodology to be used for the interview sessions; (iv) the theoretical bias of the various participating research groups; (v) the way and the detail of transcription and tagging; (vi) the relation between empirical work and theoretical analysis of the participating linguists, especially with respect to the participating graduate students; (vii) the relation between the dynamic, electronic atlas (DynaSAND) and the printed atlas.

2. External History of SAND

The SAND-project was intended to achieve the following two goals:
(i) to create an electronic atlas serving as a tool for linguistic research. This web-based Dynamic Atlas consists of a data corpus, a user-friendly search-engine and cartographic software for the on-line generation of maps. It contains data from 267 dialects collected through oral and telephone interviews and in a postal survey.
(ii) to produce a more traditional printed atlas that visualizes syntactic variation in the dialects of Dutch. Every map in the atlas will be provided with: a linguistic description of specific syntactic variables, a discussion of the attested geographical distribution (including a diachronic perspective when that is applicable) and a bibliography.

In order to achieve these goals, a group of linguists in the Netherlands and Belgium coordinated the grant applications to the Dutch and the Belgium science foundations (NWO and FWO, respectively). This group consisted of Hans Bennis (Meertens Institute, Amsterdam, Royal Academy of Arts and Sciences), Hans den Besten (Linguistics department, University of Amsterdam), Magda Devos (Dutch department, University of Gent), Johan Rooryck (French department, University of Leiden), and Johan van der Auwera (University of Antwerp). The Netherlands-Belgium consortium thus consisted of five locations. After having been successful in applying for a substantial amount of money in 1999, the actual team of investigators, both junior and senior, was selected.

Sjef Barbiers (Meertens Institute) was selected as the leader of this enterprise, in the first period assisted by a Flemish collaegue, Guido Vanden Wyngaerd. The research team consisted of six PhD-students: Jeroen van Craenenbroeck (Leiden), Marjo van Koppen (Leiden), Annemie Neuckermans (Antwerp), Gunther de Vogelaer (Gent), Henk Wolf (Fryske Akademy, Leeuwarden), and Hedde Zeijlstra (Amsterdam). By now, March 2007, four of them – Van Craenenbroeck, Van Koppen, De Vogelaer and Zeijlstra have written their dissertation and received their PhD (cf. references). These researchers were assisted by a team from the

Meertens Institute: Margreet van der Ham (coordinator of the team and general assistance), Irene Haslinger, Mathilde Jansen, Alies MacLean, and Vivien Waszink. In Belgium, Vicky van den Heede assisted the Flemish part of the work. Technical assistance was provided by Jan Pieter Kunst and Ilse van Gemert (both Meertens Institute) and methodological assistance by Leonie Cornips and Willy Jongenburger (both Meertens Institute).

From the location of the SAND-staff it may already be evident that over time the decentralized organization was changed into a project that was by and large controlled from the Meertens Institute. Although it was not very easy to shift control to a centralized organization, it turned out to be necessary to create more central control in order to keep the project from disintegrating into five more or less independent enterprises. A lot of energy was invested by Barbiers and Van der Ham to keep everything together.

It turned out quite soon that the project was much more comprehensive than we had envisaged in the project proposal. All parts of the project needed more time, energy, and money than was foreseen. In addition to the financial support of the VNC (Flemish-Dutch committee, which is a joint committee of NWO and FWO) substantial financial support was granted by the Meertens Institute, the Royal Netherlands Academy of Arts and Sciences, and NWO. In addition to that, the participating universities donated money to make the project successful.

3. Theoretical perspective

It is sometimes argued that dialectology is not a branch of linguistics proper. It only defines the empirical domain, but not the linguistic approach, both with respect to the nature of the relevant linguistic unit and with respect to the theory that serves as a model for the interpretation of otherwise meaningless data. The SAND-project was directed towards syntactic variation, which quite clearly defines the linguistic units that were relevant for the project, although traditional dialectology often claims that syntactic variation is exceptional and relatively unimportant. However, the actual choice of the theoretical framework to be used in this project was more complicated. In general, linguists in the Netherlands are much more inclined towards the framework of generative grammar, whereas linguists in Flanders have a more typological perspective on linguistics. The project started from the assumption that it would be possible to have generative linguists and typologists working together in the same project. This was a practical necessity in order to have sufficient support in both countries. The idea was that it would be possible to reach a descriptive level that is acceptable and relevant for the two theoretical frameworks. It also had a

more contentful side. The advantages of such a combination of theoretical frameworks are that both theories make predictions about the amount and the nature of syntactic variation and both theories determine topics that are particularly interesting to investigate. Another advantage is that such a combination of theories has forced us to provide a theory-neutral description, in as far as that is possible. This has led to a description which is accessible to a general linguistic audience and which might be relevant in the future because of a general lack of preoccupation with fashionable theoretical issues.

However, it turned out that it was far from easy to reach an agreement between the two theories. Differences in perspective sometimes led to fierce debates between the Dutch generative group and the Flemish typological group. There were different views on the importance of a particular phenomenon, on which aspects should have priority in the description of a phenomenon, on what counts as an adequate 'theory-neutral' description of the facts, and even discussions on the concepts that can/should be used in describing the facts, e.g. concepts such as 'copula' or 'anaphora' caused a lot of discussion. Finally, the collaboration of linguists from various frameworks forced the participating linguists to make their manuals, papers, and talks accessible to a general linguistic audience, passing by interesting issues which are only relevant in terms of a particular theoretical model.

It is clear from this exercise that data cannot be taken as theory-neutral objects that receive an interpretation within a particular theoretical framework. The data themselves are theoretical objects. A descriptive enterprise thus requires a certain level of theoretical agreement between participating researchers in order to get a consistent set of data. Within the SAND-project, this agreement was achieved by discussing relevant issues in order to establish a practical consensus on topics that might have divided the team otherwise. In other words, we have developed a linguistic 'polder model,' which was quite successful in keeping the team together.

4. Empirical domains

4.1 General issues

The determination of the empirical domains that should be covered by the SAND was one of the first problems to be solved. From the dialectological literature it was known that several constructions are extremely variable in Dutch syntax. Topics such as complementizer agreement, subject pronoun doubling, and verbal clustering were well-known as variable properties of Dutch dialects and theoretically fascinating. The SAND-project was not intended to cover all syntactic constructions in which geographic variation

can be attested. The selection of topics was determined by (a) the existing literature on language variation within the Dutch language area, (b) the variation that was known to dialectologists in the Netherlands and Flanders, and (c) the theoretical importance of specific phenomena. After a study of the literature on syntactic variation and a consultation of dialectologists, we decided to restrict the SAND to four domains: the left-periphery of the clause, the right periphery of the clause, negation and quantification, and pronominal reference.

- The *left periphery* includes topics such as complementizer selection (a.o. double complementizers), complementizer agreement, subject pronouns, subject doubling, relative clauses, and questions.
- The study of the *right periphery* is mainly devoted to the organization of the verbal cluster, with issues such as the order of verbs in the verbal cluster, the Infinitivus Pro Participio (IPP) effect, and the interruption of the verbal cluster with non-verbal material such as particles, stranded prepositions, bare nouns, etc.
- *Negation and quantification* are included to study the variation with respect to negative particles, negative concord, scope, negative polarity, and negative quantifiers.
- The topic of *pronominal reference* is directed towards the study of the variation in the use, the form, and the referential properties of personal pronouns (weak vs. strong), reflexive pronouns, and reciprocals.

The approach in which a set of topics was selected in advance had several advantages. First, the amount of syntactic variation is too much to make it possible to cover all issues. Traditionally, syntactic variation has been mostly neglected or considered relatively scarce. After an in-depth study of the available data, we conclude that there is much more syntactic variation than might be expected on the basis of the impression of most specialists. In fact, it was astonishing that a relatively little language area shows such a wealth of syntactic variation. On the other hand, this richness of different types of syntactic variation made a selection of topics imperative. Another advantage of the selection approach is that it allows a concentration on topics, which makes the data available for in-depth investigation. Given that the selection was based on existing knowledge – through specialized literature and the consultation of specialists – there was not much risk that important domains were overlooked, nor that the interviews concentrated on domains in which little or no variation would be found. Finally, it allowed us to concentrate on those topics that are particularly interesting from a theoretical point of view.

However, such a structured approach concentrating on a set of preselected topics has disadvantages too. The main disadvantage of this procedure is the risk that interesting variation is not included. Although this risk is relative small due to the study of the literature and the consultation of dialectologists, it might be true that some variable constructions – in particular those constructions that are infrequent and not salient – have not been part of the SAND-interviews. This is even more problematic since it appears to be the case that dialect variation is rapidly disappearing or at least changing over the past decades, mainly due to changes in the social structure in the Netherlands and Belgium. Phenomena that are not included in the SAND-recordings, might be impossible to attest in the near future.

4.2 Some examples

Particularly well-represented in the data collection resulting from the SAND-project are data on complementizer agreement (ex. 1; cf. van Koppen 2005); complementizer doubling (ex. 2); subject pronoun doubling (ex. 3; cf. de Vogelaer 2005); relative and *wh*-clauses (ex. 4; cf. van Craenenbroeck 2004); word order in verbal clusters (ex. 5, cf. Barbiers 2005); verbal morphosyntax (such as the *Infinitivus pro Participio* (IPP) effect (ex. 6), the *Imperativus pro Infinitivo* (IPI) effect (ex. 7), the *Participium pro Infinitivo* (PPI) effect (ex. 8), DO-support (ex. 9), negative concord and the negative particle (ex. 10, 11; cf. Neuckermans to appear; Zeijlstra 2004) and the form and distribution of reflexives and pronouns (ex. 12; cf. Barbiers and Bennis 2003).

(1) Complementizer agreement

a. [...] da Lisa zo schoon is of Anna West-Flemish
that Lisa as beautiful is as Anna
'that Lisa is as beautiful as Anna'

b. [...] da-**n** Bart en Peter sterker zijn West-Flemish
that.PLUR Bart and Peter stronger are
'that Bart and Peter are stronger'

(2) Complementizer doubling

Ik weet niet **of dat** Jan komt. colloquial Dutch
I know not if that John comes
'I don't know whether John will come.'

(3) Subject pronoun doubling

As **ze** **zulder** voor hun werk leven, [...] Flemish
if they.WEAK they.STRONG for their work live
'If they live for their work, [...]'

(4) Morphosyntactic variation in short and long subject and object relatives

a. de man **die** ik denk **die** het verhaal verteld heeft E.Flemish
the man who I think who the story told has
'the man who I think told the story'

b. de man **die** ik denk **dat** ze geroepen hebben E.Flemish
the man who I think that they called have
'the man who I think they called'

(5) Word order in verbal clusters

a. [...] dat iedereen **moet kunnen zwemmen** Standard Dutch
that everbody must can.INF swim.INF
'that everybody should be able to swim'

b. [...] dat iedereen **zwemmen kunnen moet** Frisian
that everyone swim.INF can.INF must
'that everybody should be able to swim'

c. [...] dat iedereen **moet zwemmen kunnen** Eastern Dutch
that everyone must swim.INF can.INF
'that everybody should be able to swim'

(6) Infinitivus pro Participio

a. Vertel maar niet wie zij had roepen **kend**. Groningen
tell just not who she had call.INF can.PTC
'Just don't tell her who she could have called.'

b. Vertel maar niet wie zij had **kunnen** roepen. Standard Dutch
tell just not who she had can.INF call.INF
'Just don't tell her who she could have called.'

(7) Imperativus pro Infinitivo

Hij ging naar de bakker en **koop** een broodje. Groningen
he went to the baker and buy.IMP a sandwich
'He went to the bakery to buy a sandwich.'

(8) Participium pro Infinitivo

a. Zou hij dat **gedaan** hebben gekund? Frisian
would he that done.PTC have.INF can.PTC
'Would he have been able to do it?'

b. Zou hij dat hebben kunnen **doen**? Standard Dutch
would he that have.INF can.INF do.INF
'Would he have been able to do it?'

(9) DO-support

Ik **doe** even de kopjes afwassen. Northern-Brabantish
I do just the cups wash
'I am just washing the dishes.'

(10) Negative concord

't Wil **niemand nie** dansen. West-Flemish
it wants no one not dance
'Nobody wants to dance.'

(11) Negative particle

Pas op da ge nie **en** valt! Brussels
look out that you not NEG-PART fall
'Don't fall!'

(12) Reflexives

a. Jan kent **zich-zelf** goed. Standard Dutch
John knows refl.pron-self well
'John knows himself well.'

b. Jan kent **zijn-eigen** goed. Central Dutch
John knows his-own well

c. Jan kent **hem-zelf** goed. Frisian
John knows him-self well

d. Jan kent **zijn-zelf** goed. West-Flemish
John knows his-self well

5. Collecting the data

The data collection stage of the SAND-project consisted of four substages: (i) Inventarization, (ii) Postal pilot study, (iii) Field work (oral interviews), (iv) Telephone interviews.

5.1 Taking stock

The literature consulted in the inventarization substage was entered into a database and enriched with key words. This dialect syntax bibliography now consists of more than 1300 titles and has been an important source of information throughout the project. It is part of DynaSAND and available on-line. Another valuable source of information in the inventarization stage were interviews with dialect-speaking linguists.

5.2 Postal survey

Next, we did a more elaborated pilot study to get a first impression of the geographic distribution of syntactic variables and to test various types of tasks. A written questionnaire consisting of 424 test sentences was sent to the informants network of the Meertens Institute. These informants were not controlled for social variables. The elicitation tasks used in this pilot, and also in the fieldwork, include translation tasks, indirect relative grammaticality judgement tasks, empty spot tasks, completion tasks, meaning questions and picture response tasks. An example of an indirect judgement task is given in (13). The complete data-set obtained from the postal survey comprises about 156,000 question - answer pairs (424 test sentences x 368 informants).

(13) (i) Below you find the same sentence with six different word orders. Could you indicate which of the orders of *moet - kunnen - werken* occur in your dialect? Please put a circle around 'yes' or 'no.'

(ii) If your answer is yes, could you indicate how common this order is in your dialect? 5 means highly common, 1 means extremely uncommon.

	Occurs	uncommon <> common
a. Ik weet dat Jan hard moet kunnen werken. *I know that John hard must can.INF work.INF*	ja / nee	1 - 2 - 3 - 4 - 5
b. Ik weet dat Jan hard moet werken kunnen.	ja / nee	1 - 2 - 3 - 4 - 5
c. Ik weet dat Jan hard kunnen moet werken.	ja / nee	1 - 2 - 3 - 4 - 5
d. Ik weet dat Jan hard kunnen werken moet.	ja / nee	1 - 2 - 3 - 4 - 5
e. Ik weet dat Jan hard werken kunnen moet.	ja / nee	1 - 2 - 3 - 4 - 5
f. Ik weet dat Jan hard werken moet kunnen.	ja / nee	1 - 2 - 3 - 4 - 5

(iii) If none of the above orders occur in your dialect, could you translate the sentence in your own dialect?

(iv) If your answer to (i) was 'yes,' could you translate the sentence in (ii) that you think is most common?

The postal survey enabled us to make the oral interviews more efficient. When the existing literature and the results from the postal survey clearly showed that a certain type of variation did not occur in a particular area, all questions pertaining to that variation could be safely left out from the interview in that area. For example, questions concerning the negative particle were only asked in the Dutch speaking part of Belgium and France, not in the Netherlands.

Written interviews have some advantages over oral interviews. The number of sentences that can be tested is higher than in oral interviews, because the informants can take a break when they need it. Also, written interviews allow for a higher complexity of the sentences to be tested; informants can reread a sentence until they understand it. Finally, written interviews are much less time, people, and money consuming than oral interviews, for which the fieldworkers have to travel.

However, there are also clear disadvantages. First, since the research team cannot possibly know all the dialects involved in the project, it is impossible to offer the test sentences in the local dialect. Consequently, the validity of the data is not always clear, since a sentence may be rejected on phonological or lexical grounds that are irrelevant for syntactic purposes. Secondly, for most dialects the informants had to invent their own orthography, with the risk that syntactically relevant sounds would be omitted. Thirdly, a written mode may trigger more formal, hence less dialectal behavior. Finally, in the postal interviews it is impossible to observe and respond to answers and reactions of the informants.

The geographic patterns found on the basis of the data from the written questionnaire are very similar to those of the oral interviews in most cases, but there are also some clear differences. In the near future, we hope to carry out a detailed methodological study to compare the validity of the data from these two sources.

5.3 Fieldwork

In the next stage oral interviews with an average length of 1.45 hours were held in 267 locations in the Netherlands, Belgium, and France. The first criterion employed for the choice of locations was an even distribution of locations over the language area. The second criterion was the amount of variation to be expected. In areas with a greater variation more locations were chosen. For this reason, the number of locations in Belgium and in the transitional zone along the eastern border of the language area is relatively large.

In each location, we worked with at least two informants. For the oral interviews we built up a new network of informants consisting of 607 informants. To guarantee that the relation between geographic distribution and syntactic variation was investigated, the informants had to meet the following requirements: (i) between 55 and 70 years old; (ii) born and raised in the location of the interview; same for their parents; (iii) not been away from the location for a period longer than 7 years; (iv) active dialect speaker in at least one public domain; (v) no higher education. It will be clear that this choice of criteria makes it impossible to do sociolinguistic or language change research on the basis of the SAND-data alone. Another

disadvantage of these strong selection criteria is that it can make it very hard to find enough informants, in particular in areas where the dialects have a weak position.

The methodology to be used in the oral interviews was subject to long discussions in the SAND-project (cf. Cornips and Poletto 2005). The main issue was the role of the fieldworker during the interviews. On the Dutch side, almost all fieldworkers were native speakers of Standard Dutch and not familiar with the dialects involved. Therefore, they would not be able to judge the dialect quality of the answers of the informants, and, more importantly, the risk of accommodation, i.e. the informant shifting towards Standard Dutch, would be high. To avoid these complications, it was proposed to work with two informants in each location and let one informant interview the other in the local dialect.

On the Belgian side, all the fieldworkers were native dialect speakers, and the Belgian linguists considered it to be sufficient if these fieldworkers would use a regionalized version of their dialects during the interviews. They trusted that the fieldworkers would be able to judge whether the informants were using their own local dialects, and thought that the risk of accommodation towards the regionalized variant of the fieldworker was low.

Since the two parties did not manage to convince each other, the decision was made that the Dutch would use the one methodology and the Belgians the other. To reduce the risk of accommodation in the Belgian interviews it was decided that they would work with two informants in each location as well.

Although it remains to be established which of the two methods yields the most reliable results, in those cases where it can be checked with the literature the validity of the patterns of variation established on the basis of the SAND-data is high, regardless of the methodology employed (cf. the discussion of the distribution of complementizer agreement in Barbiers et al. 2005).

The Dutch method is clearly more time consuming than the Belgian one, since it is necessary to provide the informant who will be the interviewer with instructions in a separate session. Another disadvantage is the reduction of the role of the fieldworker. Since direct interruption should be avoided, additional or clarifying questions cannot be asked until after the interview. A clear advantage of the Dutch methodology, in addition to reducing the risk of accommodation, is that rejection of sentences on lexical or phonological grounds does not occur, as test sentences are offered in the local dialect. Unfortunately, however, certain phenomena can

only be tested with translation tasks, making it impossible to really do the entire interview in the local dialect.

5.4 Telephone interviews

Given the number of interviews and test sentences, it will hardly be a surprise that the data resulting from the oral interviews were not always as complete as we wanted them to be. Sometimes, questions had not been asked for various reasons, and sometimes questions had yielded irrelevant answers. For certain phenomena, we wanted to ask a number of additional questions to make the picture as complete as possible. In the final stage of data collection we therefore decided to do a final round of telephone interviews in all of the locations involved in the fieldwork. This is a very efficient way of interviewing that provided us with a lot of useful data. However, no measures were taken in this round to avoid accommodation, so the data from this round should be handled with care.

6. Data processing, storage, tagging, retrieval, and visualization

6.1 Digitalization

In the Netherlands, the fieldwork and telephone interviews were recorded with DAT-recorders. Therefore, these tape recordings could be read into the computer directly without conversion. The recordings were read into the computer with the system Sadie DAW. The sample frequency was 44.1 kHz, 16 bits. This sampling rate was chosen to ensure a sound quality high enough to make phonetic research possible. There is an important choice here. A lower sample frequence is possible, as was the case in the Belgian part of the project, where minidisks were used. This yields a sound quality which is certainly good enough for syntactic research, but it is doubtful whether it is good enough for phonetic research.

6.2 Transcription

Transcription was carried out in PRAAT, a free software tool developed by the phoneticians Paul Boersma and David Weenink (University of Amsterdam).[2] Although PRAAT is primarily intended as a tool for phonetic research, it can conveniently be used for transcription purposes. PRAAT enables the transcriber to line up the speech signal with the transcription directly, so that searching through the transcriptions makes it possible to find the corresponding sound fragment. Another advantage of PRAAT is that it allows the user to divide the transcription into different

[2] See http://www.fon.hum.uva.nl/praat/.

tiers, which makes it possible to separate the contributions of the speech participants.

For the SAND-interviews, we used separate tiers for the speech of the first informant, the second informant, the field worker, the comments of the transcriber, and for clitic clusters. The comment tier also contains metalinguistic information, i.e. codes referring to the location of the interview, the two informants and the fieldworker. For reasons of privacy, the personal data of the informants and the fieldworker are kept in a separate database. In the future, the personal data of the informants and fieldworker will be anonymized and then added to the database.

A detailed protocol was deviced for the transcription of the spoken data (cf. Barbiers and Vanden Wyngaerd 2002). Every question (test sentence) and every answer received a special code, to make it possible to search for a particular question or answer. For reasons having to do with resource limitations, we chose to transcribe orthographically instead of phonetically, with distinct guidelines for lexical and functional morphemes. Lexical words were transcribed according to Standard Dutch orthography and abstracting away from sound differences. For example, when the word for *know* in a dialect was *kinne* it was transcribed as Standard Dutch *kenne*. Such normalization hardly ever leads to a loss of information that is relevant for syntactic analysis. In cases where it does, it is still possible to check the original data. Some advantages of a normalized transcription of lexical morphemes are: (i) spelling is uniform across dialects; (ii) automatic pre-tagging with a probabilistic tagger becomes possible; (iii) automatic lemmatization becomes possible; (iv) searches yield a more complete result, e.g. when *kenne* is used as a search term both *kenne* and *kinne* will be found.

Since functional morphemes including inflection are relevant for syntactic analysis, they were transcribed 'literally.' In this case, this means a one-to-one correspondence between sound and grapheme, where the rules of Standard Dutch determine which grapheme had to be chosen for a particular sound. For example, when the translation of the Dutch sentence in (14a) reads (14b), the missing /t/ of *wat* will not be added to the transcription, and the additional /n/ on *wie* will be transcribed as such.

(14) a. Wat denk je wie ik gezien heb? Dutch

b. **Wa** denk je **wien** ik gezien heb? dialect
what think you who I seen have
'Who do you think I saw?'

Clitic clusters received a special treatment in the transcription. The problem with clitic clusters is that their segmentation is not given *a priori* and can be obscured by phonological processes. To find the correct

segmentation often requires a significant amount of analysis which takes into account the full system of the relevant dialect. Since such analysis would slow down the transcription process considerably, it is to be avoided. However, transcribing clitic clusters as unsegmented wholes has important disadvantages too. In particular, it will be harder to directly access the separate morphemes in a search or in an automatic tagging task. In addition, field workers/transcribers working in a particular dialect area often have valuable intuitions about the proper segmentation of clitic clusters. We therefore decided that the transcriptions should contain both an unsegmented and a segmented rendering of the clitic cluster. Thus, a clitic cluster like (15a) uttered by informant 1 would be transcribed as such on the informant 1 tier, whereas on the cluster tier it would come out as in (15b).

(15) a. danzetzunder
that-they-it-they

b. dan ze ‘t zunder
that.PLUR they.NOM.WEAK it they.NOM.STRONG

Other cases in which the orthographic conventions of Standard Dutch had to be lifted include separable particle verbs *opeten* lit. up-eat ‘eat up,’ pronominal PPs such as *daarmee* lit. there-with ‘with that’ and preposition-complementizer collocations such as *voordat* lit. before that ‘before.’ In all of these cases Dutch orthography prescribes that the two morphemes be written as one word. However, we decided to split these words up since this is advantageous for automatic tagging and search, and can be easily defended on syntactic grounds.

6.3 Part of Speech Tagging

The accessibility of a large corpus containing data from a large number of different dialects is greatly enhanced if the data are enriched with part of speech (POS) information. POS-tagging of large corpora is a tedious task that should preferably be carried out by computers. Unfortunately, automatic probabilistic taggers require a sufficiently large training set to perform well on a certain language or dialect. As we did not have enough data for each of the 267 dialects to train an automatic tagger, fully automatic tagging was not an option. However, since most of the lexical morphemes were transcribed as Standard Dutch morphemes, it proved worthwhile to use the automatic memory-based tagger developed at the University of Tilburg as a pretagger (cf. Daelemans et al. 2002). The tag set we used is based on the Corpus of Spoken Dutch (so called CGN-tags; cf.

Van Eynde 2001) which, in turn, is based on the EAGLES standard.[3] We added a number of refinements to be able to deal with dialect material (cf. Barbiers and Vanden Wyngaerd 2003 for the full tag set). In addition, we decided to build in a number of attributes that strictly-speaking belong to the level of syntactic annotation rather than to POS-tagging. These attributes mainly involve information about syntactic function, word order, and hierarchy and will be useful until a full syntactic annotation has been provided. A tagging application was built which automatically translates the tags assigned by the pretagger to SAND-tags. Manual tagging then consists of checking and correcting the assigned tags and supplying missing attributes/values. The tagging application also suggests tags on the basis of already assigned tags. A considerable part of the corpus (6400 question - answer pairs) has already been provided with tags.

There seems to be a natural tendency among linguists to make a tag set as fine-grained as possible. In the framework of a large data collection project one should ask, however, whether the advantages of a fine-grained tag set outweigh the amount of work required to add all that linguistic information to the corpus. If we had used a much rougher tag set in the SAND-project, we would have been able to tag the entire corpus.

6.4 Database, search engine, and cartographic tool

The SAND-corpus is a relational database at its core. Our reasons for using a relational database is to have more flexibility for adding annotations to linguistic data, because it is possible to keep the data and the annotations separate from each other, in different database tables, which are linked with unique keys. When the data is stored in this way, it is possible to add many different kinds of annotation to a single data set without having to use multiple copies of the linguistic data (with all the versioning problems that this entails), and without the different annotations interfering with each other. Moreover, if tagged text files are ever needed (if some other program expects XML-input, for example) it is straightforward to combine tagging and data and output the results in whichever textual format is desired. The downside of our technique is, of course, that it takes programming work to view the annotated data: data and annotation must be combined from their separate tables. It is impossible to simply open an annotated text file in a text editor.

We use non-proprietary formats and open-source tools to store and handle the data: the particular database engine we use is MySQL (version

[3] Cf. the EAGLES Home Page: http://www.ilc.cnr.it/EAGLES96/home.html.

4.1.9 at the time of writing), with the InnoDB table type to enforce relational integrity.

The interface for the end-user is a web application written in PHP, and the graphics format for the cartographic component of the application is Scalable Vector Graphics (SVG), with an option to save maps as JPEG if needed. SVG has become a new standard for cartography on the web. Some important properties of the SVG format are scalability, direct availability of the original data on which the map is based and retrievability by search engines of the objects that constitute a map.

The sound of the interviews is also accessible from the web interface. The begin and end times for each interval within its sound file are stored in the database. We use QuickTime Streaming Server to host the sound files on a separate server. Because we have the begin and end times for the intervals we can then add links to the web interface which refer to that particular slice of the audio file of the relevant interview on the audio server.

The main advantages of hosting the data centrally and using a web application are: there is only one master copy of the database, so there are no problems with distributing updates to end users if the structure of the database changes; the user needs no special software to access the data, any modern web standards-compliant browser (with SVG support for the cartographical component) will suffice.

Disadvantages are that a centrally hosted database of course also means a single point of failure (if the Meertens web site is down, the SAND-database is also unreachable); the user needs a net connection to access the database; we need to work within the limits of a web browser using the HTTP protocol. The trade-off for being cross-platform, and not needing anything besides a web browser and an internet connection, is of course that the amount of interactivity and 'bells and whistles' that would be possible with a dedicated desktop application is also out of reach.

The corpus is not distributed to end-users as raw data, although it would be possible to reconstruct the original form of the textual data from the database, or to output the tagged data as XML files, if that were desirable. It is already possible to view every interview in its entirety via the web interface. For now, the user is expected to use the web interface to search the data. Search modes which are available include textual searches (with basic regular expression support), searches for part-of-speech tags, searches for lemmata, searches by the name or code of the municipality; searches by test sentence/sentence number and searches by keywords.[4]

[4] More detailed information on technical and various other issues of the SAND-project is provided in Barbiers, Cornips, and Kunst (2007).

7. Further organizational issues and final remarks

There are some organizational issues that arose during the project that are worth discussing here, since it is quite likely that they will arise in the ScanDiaSyn project as well.

A first issue is the role of the graduate students, in particular those working in a theoretical framework. For the latter there was a clear tension between their work for the larger project and their own PhD-research. Theoretical linguistic research often involves in depth investigation of a very specific topic and it was with this expectation that most of the graduate students entered the SAND-project, considering the work for the SAND-project as a small service. When it turned out that the amount of work that the SAND-project required was much larger than foreseen, they started to worry about their own projects, in particular because they could not immediately see how their work for the SAND-project would be advantageous for their own projects. This was quite understandable, since the work for the SAND-project can be characterized as broad and superficial as compared to the average PhD-project. Since the primary goal of collecting the data was to develop an atlas, it was not even clear to them that the SAND-data would be of any use for their PhD-projects.

Still, we believe that research environments such as the one provided by the SAND-project have a great added value for graduate students even if they have to spend a lot of time on the larger project, and that it should not be concluded from the above mentioned tension that data collection and processing be better assigned to research assistants. Although the SAND-data could not immediately be used for in depth analysis, they were a perfect starting point for further data collection and in depth investigation, as they included many syntactic phenomena that had not been analyzed before. In addition, the SAND-environment provided the perfect infrastructure for such research, with informants, specialized colleagues in the entire language area and a fully developed methodology readily available to provide the missing data. As a result, the theoretical dissertations that came out of the SAND-project have an unusually strong empirical basis. Despite initial appearances, the SAND-project gave the graduate students a flying start that reduced the amount of time that they needed for their PhD-projects. Conversely, it was very advantageous to have graduate students involved in the SAND-project, since their advanced linguistic knowledge proved important in all stages of the project.

Another source of tension was the printed version of the atlas. In the research proposal of the SAND-project this was set as the primary goal. Soon after the project had started it became clear that an electronic atlas would be highly desirable, since this would make it possible to give access

to the data behind the maps and hence would provide a much more dynamic environment, enabling researchers to generate their own maps and investigate potential correlations at will. This in turn raised the question as to whether a printed version was still necessary in addition to the electronic atlas.[5] Although the printed version could be presented electronically as well, it is clear that it provides a lot more than the dynamic atlas, which contains just the data, a search engine, and a cartographic tool. The printed version presents a selection of maps that show the most salient and characteristic types of syntactic variation. The maps are supplemented by a description of the linguistic properties of these syntactic variables, their historical development, their treatment in the literature, and a bibliography. We believe that the printed version can very well be used as an introduction to syntactic variation in the Dutch dialects, whereas the electronic atlas is a suitable tool for research at various levels. We also believe that it is important for a large scale dialect syntax project in which linguists of different persuasions work together to have a printed atlas (or its electronic version) as one of the major common goals of the enterprise. It may help to keep the research group together while at the same time leaving time and space for other types of syntactic research.

A final remark involves a problem that may arise when many people work together in one big project. It may be tempting to try to create a broad basis for decisions within the project by working in a bottom-up fashion, but it is our experience that this is not always the best approach. For example, at some stage of the SAND-project a task group was formed with representatives from all research institutes involved to set up a transcription protocol. This group spent a lot of time to discuss all kinds of details concerning transcription, and got stuck without producing a transcription protocol. In many cases a top-down approach worked much better. In the case of the transcription protocol, the group was dismissed from its task. Two people wrote a first version and presented this to the whole research group. The result was that most of the protocol was immediately accepted and that some useful discussion was necessary about a very limited number of details. The top-down approach ensures that the project retains its speed without passing over members of the research team that may have useful input.

In general, we can conclude that we had to make a number of important choices during the SAND-project. It concerns issues that were not expected to arise in the preparatory phase, although the actual choices made have had a large impact on the SAND-project, both in the followed

[5] One financial sponsor of the printed version initially turned down the proposal for exactly this reason.

procedures and in the outcome. In retrospect, we think that most of the choices we made were justified. Still, it would have been much better if we could have learned from the experience of a similar project on geographic language variation. We hope that our experience as described above will be helpful to the ScanDiaSyn project.

References

Barbiers, Sjef. 2005. 'Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics,' in Leonie Cornips and Karen Corrigan (eds.) *Syntax and Variation: Reconciling the Biological and the Social*. John Benjamins, Amsterdam, 233-264.

Barbiers, Sjef, Hans Bennis, Gunther De Vogelaer, Magda Devos, and Margreet van der Ham. 2005. *Syntactic Atlas of the Dutch Dialects*, Volume 1. Amsterdam University Press, Amsterdam.

Barbiers, Sjef, Leonie Cornips and Jan Pieter Kunst. 2007. 'The Syntactic Atlas of the Dutch Dialects (SAND): A corpus of elicited speech and text as an on-line Dynamic Atlas,' in Joan Beal, Karen Corrigan, and Hermann Moisl (eds.) *Creating and Digitizing Language Corpora: Vol. 1, Synchronic Databases*. Palgrave-Macmillan, Hampshire.

Barbiers, Sjef and Hans Bennis. 2003. 'Reflexives in Dialects of Dutch,' in Jan Koster and Henk van Riemsdijk (eds.) *Germania et alia. A Linguistic Webschrift for Hans den Besten*. Electronic publication, University of Groningen, Groningen. http://odur.let.rug.nl/~koster/DenBesten/contents.htm

Barbiers, Sjef and Guido Vanden Wyngaerd. 2002. 'Transcriptieprotocol [transcription protocol].' Technical report Meertens Institute, Amsterdam.

Barbiers, Sjef and Guido Vanden Wyngaerd. 2003. 'Woordsoort-etikettering voor het project Een Syntactische Atlas van de Nederlandse Dialecten. [POS-tagging and tag set].' Technical report Meertens Institute, Amsterdam.

Barbiers, Sjef and Guido Vanden Wyngaerd. 2001. 'Schriftelijke vragenlijst [written questionnaire],' Meertens Institute, Amsterdam.

Barbiers, Sjef and Guido Vanden Wyngaerd. 2001. 'Mondelinge vragenlijst [oral questionnaire].' Meertens Institute, Amsterdam.

CGN: Corpus of Spoken Dutch. http://lands.let.kun.nl/cgn/home.htm

Cornips, Leonie and Cecilia Poletto. 2005. 'On standardising syntactic elicitation techniques,' *Lingua* 115/7, 939-957.

Craenenbroeck, Jeroen van. 2004. *Ellipsis in Dutch dialects*. Ph.D. dissertation Leiden University. LOT Dissertations 96.

Daelemans, Walter et al. 2002. 'MBT: Memory Based Tagger, version 1.0, Reference Guide.' ILK Technical Report 02-09, 2002. ILK pub: ILK-0209.

EAGLES: Expert Advisory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES96/home.html

Eynde, Frank van. 2001. 'Part of speech tagging en lemmatisering.' Technical report, Centrum voor Computerlinguistiek, K.U.Leuven. http://lands.let.kun.nl/cgn/publicat.htm

Koppen, Marjo van. 2005. *One Probe, Two Goals: Agreement Phenomena in Dutch Dialects*. Ph.D. Dissertation Leiden University.

Neuckermans, Annemie. to appear. *Negatie in de Vlaamse dialecten*. Ph.D. Dissertation University of Ghent.
Vogelaer, Gunther de. 2005. *Persoonsmarkering in de dialecten in het Nederlandse taalgebied*. Ph.D. Dissertation University of Ghent.
Zeijlstra, Hedde. 2004. *Sentential Negation and Negative Concord*. Ph.D. Dissertation University of Amsterdam, LOT Dissertations 101.