Dialect Corpora Taken Further: The DynaSAND corpus and its application in newer tools

Jan Pieter Kunst[a] and Franca Wesseling[b]

[a]Meertens Institute, Royal Netherlands Academy of Arts and Sciences,
Joan Muyskenweg 25, 1090 GG Amsterdam, The Netherlands
janpieter.kunst@meertens.knaw.nl
[b]Meertens Institute, Royal Netherlands Academy of Arts and Sciences,
Joan Muyskenweg 25, 1090 GG Amsterdam, The Netherlands
franca.wesseling@meertens.knaw.nl

Abstract. In this paper we will expand on the creation and structure of the DynaSAND database as a case study of a corpus tool. Furthermore we will focus on its implementation in other search engines, thereby illustrating how the underlying data is decoupled from its original interface and used in new ways.

Keywords: Dutch dialects, corpus, user interfaces, multiple corpora search

1 Introduction

The Syntactic Atlas of the Dutch Dialects (SAND) corpus is one of the results of a large-scale dialect syntax project conducted between 2000 and 2003 in the Netherlands and the Dutch-speaking parts of Belgium and France (cf. Barbiers, Cornips & Kunst 2007, Barbiers & Bennis 2007). The goal of the SAND project was to gain insight into the (morpho)syntactic variation within and between Dutch dialects, and to make this variation visible in a clear manner. The information uncovered is of great interest to linguists, in particular syntacticians and dialect researchers, because it provides them with a huge amount of dialect data, structured in a well-organized manner. The output of the SAND project, i.e. the dialect data, has been made accessible in various ways.

First, there are two printed atlases that bring into view the dialectal variation concerning various morphosyntactic features such as the use of reflexive pronouns and the construction of relative clauses. Second, there is a freely accessible web-based dynamic atlas (DynaSand, available at www.meertens.knaw.nl/sand/), which displays the dialect data in an orderly fashion. The DynaSAND web application is meant to make the data that was collected in the SAND project available to linguists who do not have advanced computational skills. The linguistic data is stored in a relational database, a search engine is included, and the data is linked to the audio files of the original fieldwork recordings. Also, the geographic coordinates of the locations of the fieldwork are added to the database, which are used for drawing dynamic maps on the basis of search results.

A new development is the incorporation of individual language resources like DynaSAND in larger entities that are used to access and search different language resources simultaneously. This is accomplished by the addition of a web service interface to the DynaSAND corpus, i.e. a search interface that is not meant for humans but for other computer programs, so that the data from the corpus can be used in other applications besides DynaSAND itself. Two examples of projects that make use of this extra interface to the DynaSAND corpus are Edisyn (European Dialect Syntax) and MIMORE (Microcomparative Morphosyntax Research Tool).

2 The SAND corpus and DynaSAND web application

The SAND corpus contains data from 267 dialects collected in oral and telephone interviews and in a postal survey. The collection of the data took place in three phases. The first phase consisted of a written questionnaire consisting of 393 test sentences that served as a pilot study. This questionnaire was sent out to informants of the Meertens Institute at 321 locations in the Netherlands, Belgium and French Flanders, with mostly one informant per location. The informants had to judge whether the test sentence was attested in their dialect, or were asked to translate or complete it. The aim of this pilot study was to get a first impression of the syntactic variation in the dialects of Dutch and of its distribution across the language area. In the second phase of the project oral interviews (involving elicited, not spontaneous speech) were conducted at 267 locations spread across the Netherlands, Belgium and French Flanders. The interviews were based on the responses given in the written questionnaire, thus enabling to focus on a specific phenomenon that had come up in the written questionnaire. In total 456 sentences were asked in these interviews. The last phase of the data collection involved telephone interviews, which served to clear up any uncertainties in the data. In total, 331 sentences were tested in this round. These were either sentences that had been tested in the oral interviews but had not received a clear answer or new sentences that were required to get a more complete picture of some particular phenomenon.

2.1 Technical details

The DynaSAND (Dynamic Syntactic Atlas of the Dutch Dialects) web application is meant to make the data that was collected in the SAND project available to linguists who do not have advanced computational skills. This means that something more had to be done than just offering the raw data for download: a user interface had to be built (Barbiers, Cornips & Kunst 2007; Barbiers & Kunst, to appear). The original (raw) data are transcriptions of interviews with dialect speakers made with the freely available PRAAT transcription program (see www.praat.org). PRAAT stores its data in structured text files in which the transcriptions are linked to time stamps that refer to the audio recording that was transcribed. On the basis of this we had to build a user-friendly web application.

The first thing we did was storing the data in a way that was more flexible and more suited for use in a web application: namely, a relational (MySQL) database. We wrote a set of command line PHP scripts to parse the PRAAT files, split them into individual words (as the words had to be enriched with part-of-speech tags at a later stage) and insert the result in a database. (These scripts are open sourced under the GPL and can be downloaded from www.dialectsyntax.org.)

The resulting table structure is roughly as pictured in Figure 1:

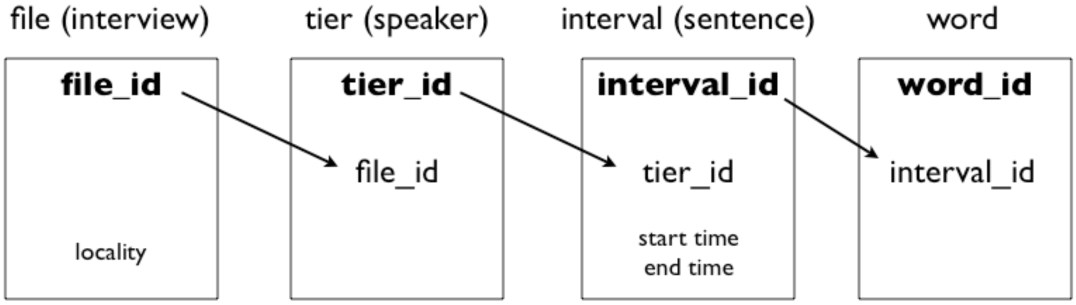


Figure 1: SAND table structure.

A table structure like this allows for unambiguously addressing individual words while preserving the context.

2.2 SAND user interface: DynaSAND

On top of this database we built a web application and search engine in PHP so that the data would be useable for linguists who do not have advanced computer skills. Some of the core functionality of the web application includes: the possibility to search for data on the basis of text strings, part-of-speech tags or sentence numbers (the interviews consisted of a fixed list of questions, so that comparing different dialects is possible). Searches can be limited on the basis of geographical information. It is also possible to view entire interviews.

The audio of the interviews is available in the web application. Notice that the 'interval (sentence)' table pictured above has fields for 'start time' and 'end time'. These are the boundaries of sentences which were assigned during the transcription phase in PRAAT, and which were preserved in the database. The original audio files where recorded with DAT (digital) recorders and were saved as AIFF, uncompressed digital audio. To be able to be used in the web application it was necessary that the audio of individual sentences could be retrieved. We accomplished this by converting the audio files to hinted QuickTime movies. If these files are served by QuickTime Streaming Server (or its open source variant, Darwin Streaming Server) it is possible to jump directly to an arbitrary point in the file. Since we have (from the PRAAT files) the begin and end points of the sentences in the database, we can simple instruct the audio server to play the fragments denoted by these timestamps by sending it a request containing the filename and the begin and end time.

Another notable part of the web application, and the reason it is called a 'Dynamic Atlas', is the cartographic component. The geographic coordinates of the locations where the interviews were held are stored in the database. This means that every found sentence in a result set has a point on the map associated with it, and this information is used to draw maps of search results. Since search results can be saved in the application, it is possible to combine different search results on a single map and hopefully show correlations between phenomena. Take for instance the phenomenon subject doubling: in some dialects of Dutch it is possible to double or triple a subject pronoun. This is exemplified in (1) and (2) below, where the first person plural subject pronoun (we) is doubled respectively tripled.

(1) Wij zijn-wij daar nog nooit geweest.
we(strong) are-we(strong) there yet never been
'We have never been there.' (Syntactic Atlas of the Dutch dialects 2005:53)

(2) We gaan me wij daar dikwijls naar toe.
we(weak) go we(weak) we(strong) there often to
'We often go there.' (Syntactic Atlas of the Dutch dialects 2005:53)

An example of a map generated by the DynaSAND application is shown in Figure 2, which shows the distribution of doubling and tripling of the second person singular.

Subject doubling vs. tripling

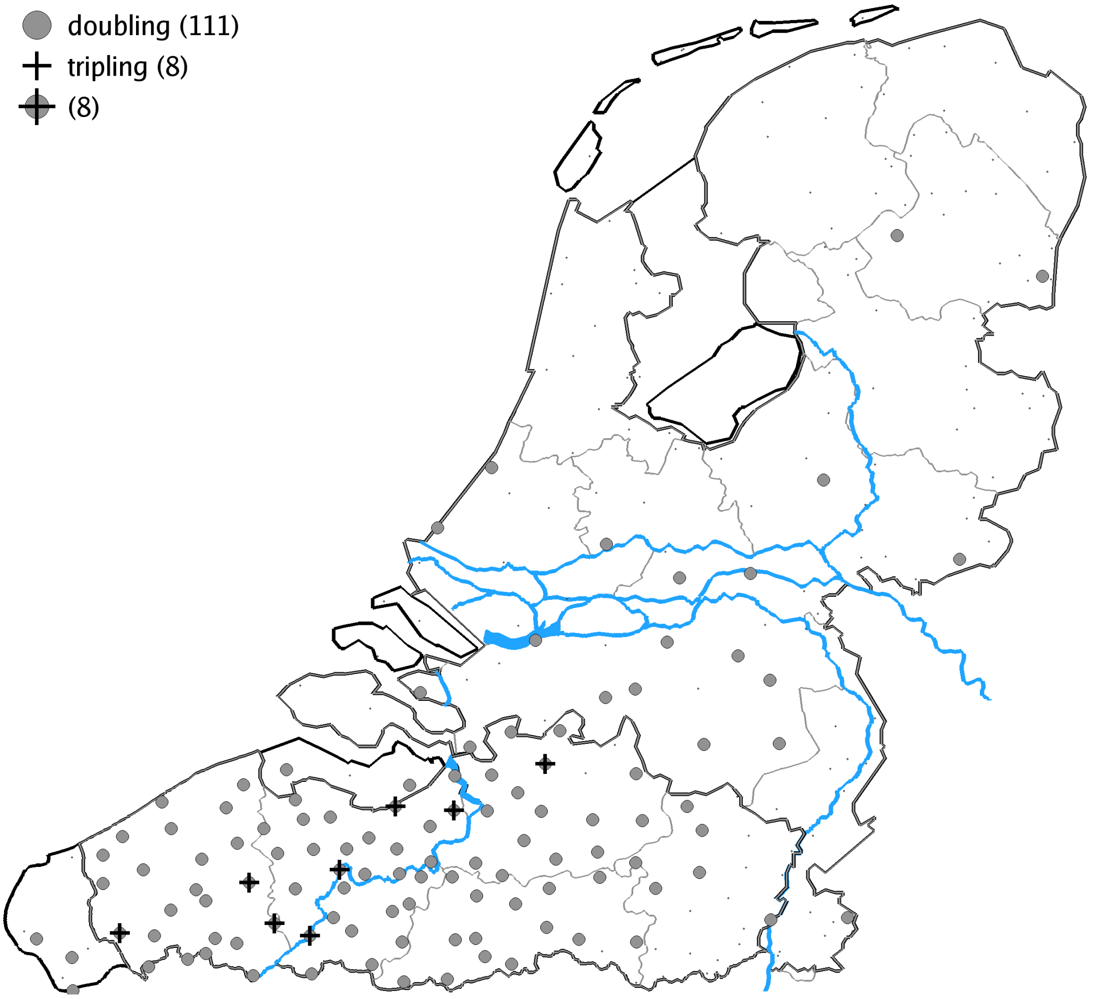


Figure 2: DynaSAND map.

The symbols used on the map can be customized by the user to make the map clearer, and large versions of maps can be exported for use in publications. It is also possible to show search results in Google Earth by exporting a map as a KML file.

The dialect data of the DynaSAND database is enriched with part of speech tags. These tags refer to common word categories such as verb, noun, pronoun, etc., in turn the categories are specified for number, gender, function, and other features. The tags are assigned to each individual word. In the first phase an automatic tagger was trained to tag the dialect data, in a later stage these tags have been manually checked and, if necessary, revised. It is possible to search on the basis of strings of part of speech tags, and on the basis of particular syntactic phenomena. Examples hereof include but are not limited to doubling patterns, complementizer agreement, and preposition stranding.

The DynaSAND has proven to be a useful tool for linguistic research, for example in the production of various dissertations, e.g. Subjectsmarkering in de Nederlandse en Friese dialecten (Subject marking in the Dutch and Frisian dialects), De Vogelaer (2005), Ellipsis in Dutch dialects, Van Craenenbroeck (2004), One Probe - Two Goals: Aspects of agreement in Dutch dialects, Van Koppen (2005), Sentential Negation and Negative Concord, Zeijlstra (2004) and in numerous articles and presentations.

3 The SAND corpus used in other search engines

3.1 Edisyn

Edisyn (www.dialectsyntax.org) is an ESF-funded project on dialect syntax. It runs at the Meertens Institute in Amsterdam from September 2005 until September 2010, with a partial extension until March 2012.[1] It aims at achieving two goals. The first is to establish a European network of (dialect)syntacticians that use similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography. The second goal is to use this network to compile an extensive list of so-called doubling phenomena from European languages/dialects and to study them as a coherent object. One of the deliverables of the Edisyn project is a web-based search engine to search different linguistic corpora, among those the DynaSAND corpus, simultaneously and show the combined search results.

The Edisyn search engine currently encompasses dialect data of Dutch (DynaSAND), Estonian (EMK corpus), Italian (ASIt), Scandinavian languages (Nordic Dialect Corpus) and Portuguese (Cordial-Sin). Data will be added in the near future on dialects of English (FRED) and Slovene (Slovene Dialect Database on Doubling). In addition, there are plans to incorporate databases of dialects of Basque (BasDiaSyn), French (Arbres), Spanish (COSER) and Finnish (Finnish syntax archive). By bringing together these various databases via one search engine, Edisyn enables dialect researchers to search and compare data of various dialects (and languages). The dataset of each database has its own specific enrichment and has been tagged with parts of speech tags that differ per database. Consequently these tags may vary from one database to the other. To facilitate a search which can be run through all these databases, a common tag set has been developed that can be mapped to the tags of any database. Note that the content of a database is not changed in any way; rather the tags of a particular database are 'translated' to the tag set of the Edisyn search engine, to be able to render a sound search result.

The Edisyn tag set consists of categories and features. The first are comparable to part of speech categories, these include verb, noun, pronoun, determiner, adposition, complementizer, adverb, adjective, conjunction, negation marker, clitic, particle. The second are specifications of the categories e.g. first person, singular, interrogative, nominative, comparative, etc. These features can be combined with a category to form a specific tag. More than one feature can be combined with a category, but a (composed) tag may consist of only one category. In the Edisyn search engine it is also possible to search on the basis of predefined tags such as V(fin,past,1sg) (first person singular form of a verb in the past tense). By leaving the set of categories unspecified many tags of various databases can be ‘translated’ to the Edisyn tag set. For instance, if a particular database does not specify the inflection of a verb, such as the tag mentioned above (V(fin,past,1sg), it can still be tagged by the Edisyn tag set because the category Verb need not be specified with the features past tense, first person and singular. In such as case, as in the Nordic Dialect Corpus, the corresponding tag in the Edisyn tag set will be simply V. In this manner it is almost always possible to relate a specific tag in a particular database to an equivalent tag in the Edisyn tag set. This is an imperative in the attempt to make different databases interoperable with one another. Another advantage of the unspecifiedness of the Edisyn tag set is that the tag set of each individual database can be left intact. The tags of each database are thus not changed, but are linked to the corresponding tag of the Edisyn search engine.

[1] ESF EURYI Grant 2004 to Sjef Barbiers for the project European Dialect Syntax.

The Edisyn tag set is also made compatible with the tags used in ISOcat (www.isocat.org). The ISOcat ISO 12620 provides a framework for defining data categories compliant with the ISO/IEC 11179 family of standards. The aim of ISOcat is to make a tag set available which can be used by all linguistic databases, thus making these databases more comparable. By translating the tags used in the Edisyn tag set to those that are part of ISOcat the databases that are made interoperable by the search engine are in turn also comparable to other databases that use the ISOcat tags. It should be noted that at the time of writing (August 2010) the ISOcat tag set is not yet complete and certain problems need to be discussed and overcome before the actual goal of ISOcat can be attained. In particular, the problem of information that is lost in the translation from one tag set to another and its consequences of search results needs to be addressed.

Being part of the Edisyn search engine, the structure of the DynaSAND database has not been changed in any way. It is simply added to the search engine, along with the other databases mentioned above. The enrichment of the database is still available, but now it is also possible to compare the Dutch dialect data with dialect data of other languages.

3.1.1. Edisyn technical details

From a technical point of view, the search engine consists of a central search module which translates the incoming search requests from its web interface to the native tag sets of the individual resources, sends it off, gathers the results and shows those on a result page. The coupling between the central search module and the individual resources is kept as loosely as possible: the individual resources have no knowledge about the central search module and its tag set.

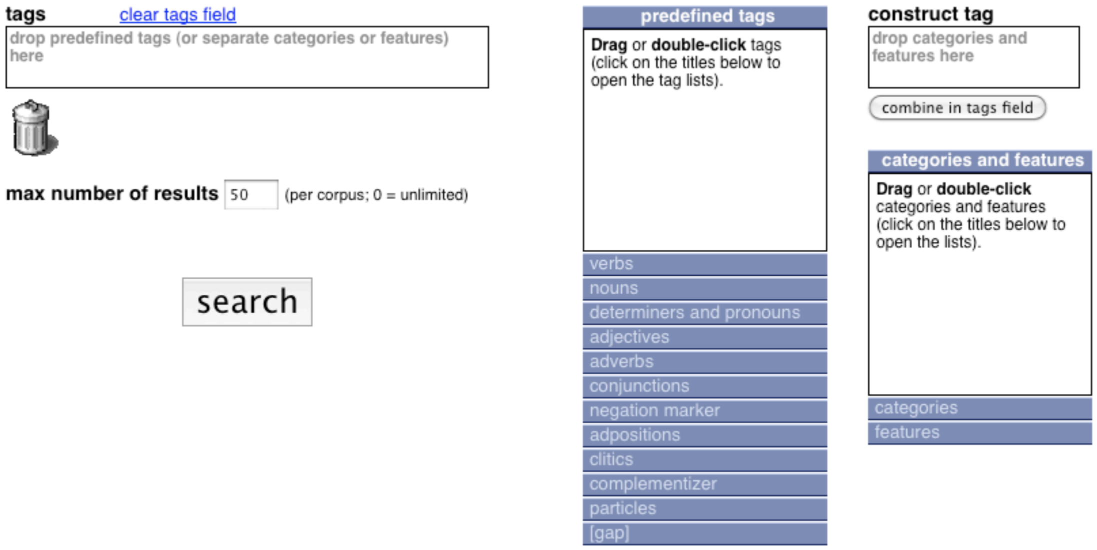
The ideal setup for such a constellation of resources would, in our view, be truly distributed: each group being responsible for hosting and maintaining its own resource, and opening up a web service interface which is used for queries by the central search component.

In practice, such a distributed setup is difficult to realize. Many research groups have created a corpus at some point in time, but do not have the resources to host it on the web, build search interfaces, etc. (with the exception of the Nordic Dialect Corpus). What we did in those cases was to host a copy of the resource locally on the Meertens Institute servers. We still use web services to communicate with the resources, though, so if the responsible groups are ever in the position to start hosting their corpus, it would be relatively easy to configure the central search component to deal with it.

3.1.2. Edisyn user interface

Since the tag set used in the central search component is very extensive (it consists more or less of the combination of all tags used in the individual resources) creating a useable interface for tag search was not trivial. We decided to turn to the JQuery Javascript library with its rich interface possibilities, specifically its ‘Accordion’ widget for quickly showing and hiding large lists of data (in our case: tags, categories and features) and its Drag and drop interaction behavior for constructing tags.

Below (Figure 3) is a screenshot of the latest incarnation (not yet stable at the moment of writing) of the Ediysn interface where one can search for tags.



drop predefined tags (or separate categories or features) here

max number of results 50 (per corpus; 0 = unlimited)

Drag or **double-click** tags (click on the titles below to open the tag lists).

drop categories and features here

Drag or **double-click** categories and features (click on the titles below to open the lists).

Figure 3: Edisyn search interface.

There are two ways of searching for tags: (1) choosing predefined tags from a list (the accordion to the left in the picture). This lets the user choose complete part of speech tags like V(fin,pres,1sg) or A(comp,sg) to be used as is in tag searches. The tags can be transferred to the tag search box on the extreme left by dragging and dropping or double-clicking; (2) constructing tags by combining categories and features, using the accordion on the right. This works by transferring a category and some features to the 'construct tag' box and then clicking the 'combine in tags field' button, which creates a tag in the tag search box. Tags can be removed from the tag search box by dragging them to the trash icon below it. Finally, clicking the large 'search' button performs the actual search in the resources chosen by the user.

All resources in the Edisyn Search Engine have geographical information associated with their data, and we plan to use Google Maps for drawing maps on the basis of search results. (The map-drawing component of the DynaSAND is Netherlands and Flanders only, so that cannot be used for data from other countries.) The implementation details of the map-drawing component will have to be worked out.

3.2 MIMORE

MIMORE (Microcomparative Morphosyntax Research Tool, http://www.clarin.nl/node/70#MIMORE) is a relatively small project that is funded by CLARIN (Common Language Resources and Technology Infrastructure, www.clarin.eu). It runs at the Meertens Institute from April 2010 to January 2011. Its goal is to create a common interface for three corpora containing data on Dutch dialects: (i) DynaSAND for syntactic variation at the clausal level, (ii) DiDDD, a corpus of elicited speech and text collected between 2005-2009 to chart the syntactic variation at the level of nominal groups in the same language area; (iii) MAND (www.meertens.knaw.nl/mand/database/) a corpus of elicited speech and text collected between 1980 and 1995 to chart morphological (word-level) variation. With a common search engine it will be possible to investigate potential correlations between variables at the three different linguistic levels (syntax, nominal groups, morphology).

3.2.1. MIMORE technical details

MIMORE is funded by CLARIN and therefore has to comply with various CLARIN guidelines and standards. The application and its resources must have harvestable metadata in the CMDI format, so that it can be found in catalogues that are built using this metadata. The linguistic categories must be compatible with the tags used in ISOcat. The application must have a web service interface, and the data it returns must be in a standardized format, so that it can also be used by other applications and not only by end users. In MIMORE we do all communication within the application with web services, i.e. the three constituent resources have a web service interface which is used by the central search component, and the central search component has a web service interface which is used by the end user interface. All these interfaces can of course also be used by external applications.

3.2.2. MIMORE user interface

At the moment of writing there is only a very bare bones test interface for using the MIMORE application. Details for the final user interface still have to be worked out, but it will in all likelihood look a lot like the Edisyn user interface, with the map-making component of the DynaSAND, which can be reused here because all data is from Dutch dialects.

4 Conclusion

The Edisyn search engine and MIMORE show that a database such as DynaSAND can be implemented in various ways. Being construed at first to display data of Dutch dialects, DynaSAND can now also be compared to data of other European dialects. In addition, DynaSAND can be implemented in a comparison of Dutch dialects at various linguistic levels, i.e. the morphological level, the level of the noun phrase and at the syntactic level.

References

Barbiers, S. & L. Cornips & J.P. Kunst. 2007. 'The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas.' In: J.C. Beal & K.C. Corrigan & H. Moisl [eds.] *Creating and digitizing language corpora. Volume 1: Synchronic databases*. Palgrave Macmillan, Hampshire, pp. 54-90.

Barbiers, S. & H. Bennis. 2007. 'The Syntactic Atlas of the Dutch Dialects. A discussion of choices in the SAND-project.' *Nordlyd* 34, 53-72.

Barbiers, S. & H. Bennis & G. De Vogelaer & M. Devos & M.H. van der Ham. 2005. Syntactic Atlas of the Dutch Dialects Volume I. Amsterdam University Press, Amsterdam,

Barbiers, S. & J. van der Auwera & H.J. Bennis & E. Boef & G. De Vogelaer & M.H. van der Ham. 2008. Syntactic Atlas of the Dutch Dialects Volume II. Amsterdam: Amsterdam University Press.

Barbiers, S. & J.P. Kunst, to appear. 'Generating maps on the internet'. In: Lamell, A. & R. Kehrein & S. Rabanus [eds.]. *Language Mapping. An international handbook*. Mouton de Gruyter, Berlin, pp. 401-415.

De Vogelaer, G. 2005. Subjectsmarkering in de Nederlandse en Friese dialecten. PhD Dissertation, Ghent University.

Van Craenenbroeck, J. 2004. Ellipsis in Dutch dialects. PhD Dissertation, Leiden University. Utrecht: LOT Publications.

Van Koppen, M. 2005. One Probe - Two Goals: Aspects of agreement in Dutch dialects. PhD Dissertation, Leiden University. Utrecht: LOT Publications.

Zeijlstra, H.H. 2004. Sentential Negation and Negative Concord. PhD Dissertation, University of Amsterdam. Utrecht: LOT Publications.

URLs

CLARIN Europe: www.clarin.eu
CLARIN Netherlands: www.clarin.nl
Darwin Streaming Server: dss.macosforge.org
DynaSAND: www.meertens.knaw.nl/sand
Edisyn: www.dialectsyntax.org
Edisyn search engine: www.meertens.knaw.nl/edisyn/searchengine
ISOcat: www.isocat.org
MIMORE: www.meertens.knaw.nl/mimore , http://www.clarin.nl/node/70#MIMORE
PRAAT: www.praat.org