

Phonological and Phonetic Databases at the Meertens Institute

Marc van Oostendorp

1. Introduction

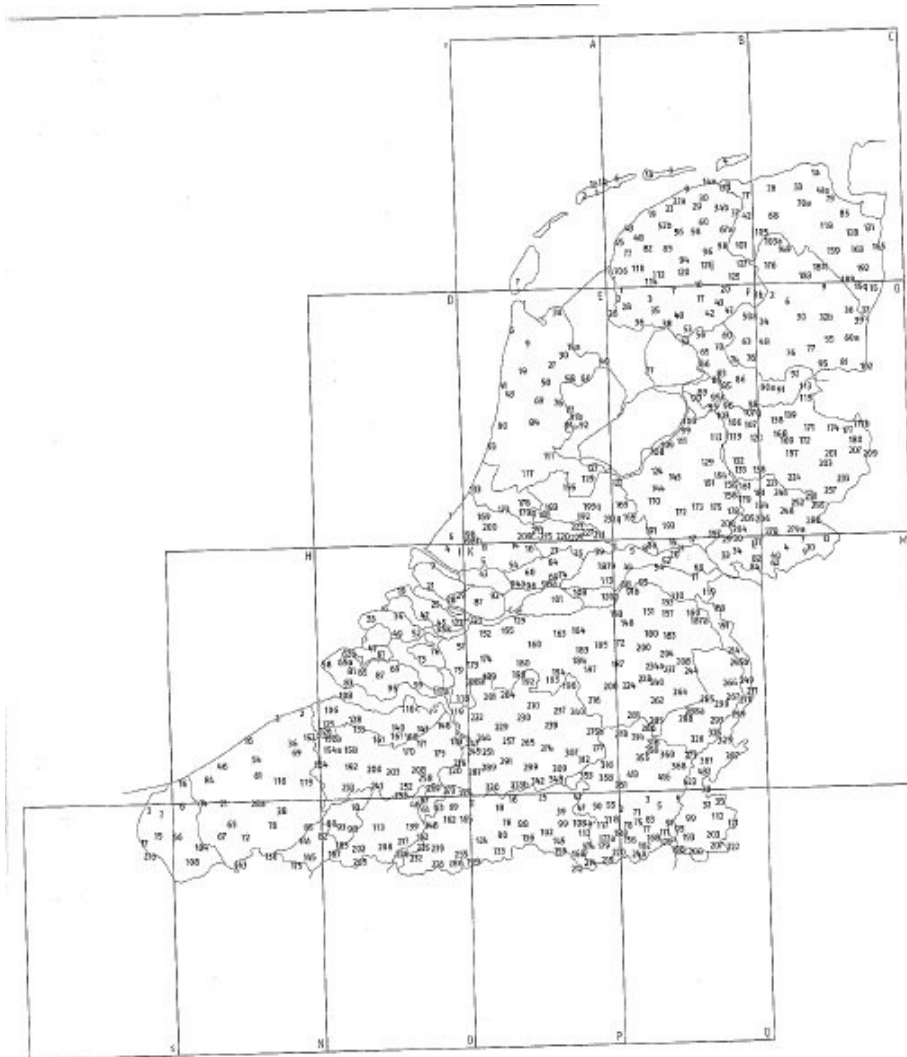
The Meertens Institute in Amsterdam was founded in 1930 under the name 'Dialect Bureau' (*Dialectenbureau*), and became an official institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) in 1952. In 1979, it was named after its first director, P.J. Meertens (1899-1985), a student of 17th century Dutch literature. Currently it comprises two departments, one of Dutch Ethnology and one of Variational Linguistics. Originally, the institute had as its primary goal the documentation of the traditional dialects (as well as folk culture) of the Netherlands. In the course of time, this focus has broadened in several ways; for instance the institute now also does its own research, and the linguistic department has widened its scope to other topics than the traditional dialects. At the same time, the documentation of dialects itself has progressed. Over the past 15 years, considerable effort has gone into digitizing material and putting it online.

This brief contribution seeks to describe the two most important databases on Dutch dialects which are available at the Meertens Institute: the *Goeman-Taeldeman-Van Reenen Database* and *Soundbites*; I will conclude with pointing out some future plans and desiderata.

2. The Goeman-Taeldeman-Van Reenen Database

At the core of phonological research at the Meertens Institute, we find the database of the so-called Goeman-Taeldeman-Van Reenen Project (GTRP; available at <http://www.meertens.knaw.nl/mand/database/>). It contains data about the phonology and

morphology of 613 dialects spoken in the Dutch language area of Europe - that is to say, in the Netherlands, in Flanders and in French Flanders. The locations are more or less evenly spread geographically on the basis of a hexagonal grid with one location per hexagon, as Figure 1 shows. In known transitional dialect areas the sampling of locations is denser



In most cases, there is one speaker per location; typically this is somebody close to the traditional NORM for classical dialectological fieldwork: a *non-mobile older rural male*, as the

following table shows (the reason for choosing such informants was that they are considered to be the more representative speakers of ‘traditional’ variants):

	Gender (% of females)	Mean Age	Age Range
Flanders	22%	65,2	37-91 (most speakers around 65)
The Netherlands	30%	61,7	25-84 (speakers more evenly spread)

Table 1: Informants for the GTRP Database

The fieldwork for this database was mostly done in the 1980s and 1990s. The core of the interview consisted of a questionnaire of 1876 items (mostly individual words, but also a few paradigms and sentences) in Standard Dutch which the informants were invited to translate into their own dialect. Since the emphasis of the project was on phonology and morphology from the outset, the design of the questionnaire and the actual interview were set up in such a way that translation into etymologically different items were avoided as much as possible. All interviews were recorded on audiotape and subsequently transcribed; the Dutch items at the Meertens Institute and the Flemish items at the university of Ghent. (To be precise, over 700 recordings were made, but in the end there was only money for the transcription of 613; these are the ones discussed here. The other recordings are digitized and may be added to the database in due course; the problem will be the transcription, which is specialized work.)

One of the problems with the database is that the work in the Netherlands and Flanders was done separately in the two different countries, by two different teams and at different times. This has as the effect that there are quite a few differences between the data from the two areas which cannot necessarily be understood in terms of dialect geography. For instance, the fieldwork in the Netherlands was mostly done in the 1980s while that in Flanders was performed almost a decade later. Also, as we can learn from Table 1 above, the selection of informants

was not carried out in exactly the same way in the two locations, with the Belgian and French informants adhering slightly closer to the NORM. Furthermore, the Dutch transcribers used a rather narrow phonetic transcription, which was notated into a computer legible ASCII format called KIPA ('keyboard IPA', an precursor to SAMPA which was developed specifically for this project). The Flemish transcriptions were more 'phonological' and furthermore notated directly in IPA; these files were later converted to KIPA, while the overall database has later been provided with an automatic translation into IPA.

The name of the GTRP project refers to its three founders: Ton Goeman from the Meertens Institute, who led the Dutch part of the fieldwork; Johan Taeldeman, from Ghent University, who did the same for Flanders and French Flanders; and Piet van Reenen, who has been mostly involved in setting up the database.

The resulting database has been the source of two linguistic atlases (on paper): the *Fonologische Atlas van de Nederlandse Dialecten (FAND; Phonological Atlas of Dutch Dialects; Goossens et al. 1998, 2000; De Wulf et al. 2005)* and the *Morfologische Atlas van de Nederlandse Dialecten (MAND; Morphological Atlas of Dutch Dialects; Goeman et al. 2005, 2008)*. The former comprises three volumes and was produced in Flanders around the turn of the century, whereas the latter consists of two volumes that have appeared in 2005 and 2008 respectively. The *FAND* is a monolingual publication in Dutch, whereas the *MAND* exists both in a Dutch and in an English version.

The database can also be consulted directly on the web in an interface built by Jan-Pieter Kunst of the Meertens Institute, in collaboration with the author of this article (<http://www.meertens.knaw.nl/mand/database/>). Unfortunately, this interface currently is available in Dutch only, although this should probably not be a problem when one uses some online translation tools. The database gives access to the original transcription files in a variety of ways (the files are also available on request from the Meertens Institute for those researchers who want to run their own scripts on the material; these files are simple UNIX textfiles in which

every record occurs on a separate line, and all the fields are separated by commas) in a variety of ways. During the last years, most of these data have been enriched by (highly trained) volunteers with the original sound files on which the transcriptions are based, and this work will continue in the next few years. Furthermore, the interface allows the user to save selections of the data, and to draw maps based on such selections of the language area. For example, one could select all the words which have a fronted pronunciation of the vowel in *aap* 'monkey', and see whether these form a coherent region.

One can search the transcriptions either in the KIPA form or in a simplified transcription, which is based on Standard Dutch orthography (although the data can be presented in an IPA form it is unfortunately not possible to search for them in that way). Thus, one can search for all occurrences of the high fronted rounded vowel [y] in all dialects. One can also search for items in the questionnaire, and thus get all translations of the item *brood* (bread). Another possibility is to search within word categories, and thus get all past tense forms of verbs (this obviously seems more useful for morphological searches), or to search for 'word endings', e.g. all words ending in a velar fricative. The reason to make a specific option for searching word *endings* rather than *beginnings* is that it is known from the historical phonology of Dutch that the initial segments of words have been more stable than the final ones (Goossens 1974); one can thus expect more variation towards the end of the word. Another search option inspired by historical phonology is that one can search for words which historically belong to the same 'vowel class' (Van Bree 1987)

Finally, one can also search for speaker properties, such as their town or province of origin, their age and their gender. It is also possible to combine various search dimensions and thus look for all instances of 'brood' containing an [y] and spoken by men (there are 21 such items, almost all of them in East Flanders).

The GTRP database is not finished yet. As has been noted above, currently, the audiofiles are added to the database. At the time of writing (2012), the institute is trying to integrate the

database with (at least) the database for the *Syntactic Atlas of Dutch Dialects (SAND)* and possibly with other dialectological databases. Another idea obviously is to have also the remaining approximately 100 recordings transcribed and added to the database; it turns out to be difficult to find money for this. So far, we have not been succesful in finding money for this project.

3. *Soundbites*

A second valuable tool for (phonetic) research into Dutch dialects is the so-called soundbites project, available at <http://www.meertens.knaw.nl/soundbites/>. (Again, the interface is in Dutch, but will be fairly transparent to the foreign user with some simple translation tools.) This website contains soundfiles with in total approximately 1,000 hours of spontaneous conversation in a variety of Dutch dialects (mostly from the Netherlands) spoken in the second half of the 20th century.

In the 1950s, researchers from the Meertens Institute started to more or less systematically record such spontaneous conversations from all over the language area on audiotape. (It should be noted that the definition of a spontaneous conversation was rather loose, and the dataset contains also monologues and interviews with the researcher.) They continued to record materials until some point in the 1980s, when it was decided that a representative coverage of the language area had been attained.

The recordings were (and are) stored in a climatized room at the Meertens Institute, but around the turn of the century the institute digitized all of its recordings, including those of this project (which never had an official name, as far as I know, but was commonly referred to within the institute as 'banden vrij gesprek' (tapes free conversation)). The digital copies were stored on cd's which were then also put in a climatized room. In principle, researchers from outside the institute could ask for copies of these recordings but this has seldom happened. As a matter of

fact, it was rather difficult to obtain insight into which recordings existed for which town even for researchers within the institute.

This changed in 2009, when money was obtained for putting the contents of all cd's on a single server, which was made this server accessible to the outside world through the internet. The result is a website which displays all the material with a simple interface. One can search either in an alphabetic list of names of provinces and place names, or visually on a two-dimensional map of the Netherlands, with red dots denoting all locations where dialect recordings can be found. Because of the latter interface, the website is also referred to as the 'Speaking Map'. The recordings themselves can be listened to on the website or downloaded, and are in many cases provided with metadata, although the quality and quantity of those metadata vary rather wildly from one example to the next. In any case, some of these metadata - age, town or origin, gender, year of recording - can also be searched for. Furthermore, the website shows a number of recordings from Flanders, from other regions in the world such as Brazil and Indonesia (these are either about Dutch emigrants or about colonial varieties) as well as a category of 'other' material, which contains e.g. recordings of radio programmes.

It goes without saying that the Soundbites material is potentially relevant not just for phonological and phonetic research. It could also be used to study other levels of linguistic structure, but also its contents are potentially interesting, e.g. for those studying oral history. At present, the site only presents raw material. The Meertens Institute has a set of transcriptions covering about 10% of the recordings; however, these transcriptions at present only exist in a paper version. In the ideal case, we would at least release all of these transcriptions also in a digital version, preferably in some way aligned with the sound material.

4. *Desiderata*

Next to the databases we already have put online, the Meertens Instituut still has a lot of material which it may try to publish on the Internet during the next years. First, more or less since the beginning until 2008, the Institute has sent out a yearly written questionnaire to a group of informants, asking them about all kinds of details in their dialects, including in many cases information about phonological topics. These questions were always formulated by researchers from the Meertens Institute. The answers to these questionnaires are currently being digitized and will be put online in the course of the next few years. (During the last years, the written questionnaire has been replaced by an Internet forum with the same function.)

Secondly, the institute hosts a large set of dialect grammars - books published both by professional and lay linguists describing fragments of the structure of individual dialects. There are plans for putting these online as well, possibly in a Wikipedia format, so that they can be edited and adjusted by readers.

Thirdly, the institute has started experimenting with the use of social network related types of data gathering. In particular, it has started a website Meldpunt Taal ('Reporting Language') together with a consortium of other Dutch language-related institutions, on which language users can report on any development in the Dutch language which they consider relevant. Data mining techniques will be applied to study the data from this project in more detail.

Another concrete step we are now taking is developing tools to make the *geographical* aspects of these data more directly relevant for research. Broadly speaking, the research focus at the Meertens Institute is on formal ('generative') grammar and on sociolinguistics. Although both disciplines are interested in linguistic (micro-)variation, neither of them have been particularly interested in the way in which linguistic phenomena are spread out on the map. Still, such geographic patterns may sometimes be taken as evidence, e.g. when a certain linguistic phenomenon A only occurs in areas which also have phenomenon B (which may show that the

two phenomena are not completely independent). In 2012, a research project started, involving several of the researchers to explore such issues in more detail.

References

Bree, Cor van, *Historische grammatica van het Nederlands*, Dordrecht: Foris.

Goeman, A.C.M. , G. de Schutter & B.L van den Berg en Thera de Jong (2005) *MAND Morphological Atlas of the Dutch Dialects Volume I*. Meertens Instituut/Koninklijke Academie voor Nederlandse Taal- en Letterkunde/Amsterdam University Press.

Goeman, A.C.M., M. van Oostendorp, P. van Reenen, O. Koornwinder, B.L. van den Berg & A. Van Reenen. (2008) *Morfologische atlas van de Nederlandse dialecten. Deel II/Morphological Atlas of the Dutch Dialects. Volume II*. Amsterdam : Amsterdam University Press.

Goossens, Jan. 1974. *Historische Phonologie des Niederlandischen*. Tübingen: Max Niemeyer.

Goossens, Jan, Johan Taeldeman en Geert Verleyen. 1998. *Fonologische atlas van de Nederlandse dialecten. Deel I: Het korte vocalisme*. Gent: KANTL.

Goossens, Jan, Johan Taeldeman en Geert Verleyen. 2000. *Fonologische atlas van de Nederlandse dialecten II - III. Deel II: De Westgermaanse korte vocalen in open syllaben. Deel III :De Westgermaanse lange vocalen en diftongen*. Gent: KANTL.

De Wulf, Chris, Jan Goossens en Johan Taeldeman. 2005. *Fonologische Atlas van de Nederlandse Dialecten. Deel IV (afl. 3) De consonanten*. Gent: KANTL.