Locating people with their language – An Applied Linguistics Course using linguistic microvariation databases and tools

Sjef Barbiers – Meertens Instituut and Utrecht University

Keywords: LADO, forensic linguistics, applied linguistics course, MIMORE, on line linguistic research tools and corpora, dialects, linguistic microvariation, sociolinguistics, asylum procedures, linguistic atlases, SAND, MAND, FAND

Abstract

Recent years have seen an increase in the on-line availability of dialect corpora, databases and search, analysis and visualisation tools (cf. www.dialectsyntax.org). Although primarily intended for linguistic research, this infrastructure provides rich resources for courses on sociolinguistics, dialectology, formal linguistics and linguistic methodology. This paper demonstrates the usefulness of the Dutch linguistic microvariation research tool MIMORE (www.meertens.knaw.nl/mimore) for a course in applied linguistics, more specifically on Language Analysis for the Determination of Origin (LADO). LADO is used in asylum procedures as a means to determine whether an asylum seeker originates from the country or area that s/he claims to originate from. It is a task of linguists to make clear if and how LADO can be a valid method, and what kind of linguistic expertise is needed. MIMORE contains three databases with (morpho-)syntactic and (morpho-)phonological data from a large number of locations in the Dutch language area. The paper describes how the MIMORE data and tools have been used in the course as training material, to introduce students to the different levels of language variation, to teach them how to recognize linguistic differences and make these explicit and to show them the complications involved in using linguistic properties to locate speakers.

1. Introduction

Recent years have seen a dramatic increase in the online availability of dialect corpora, databases and corresponding search, analysis and visualization tools for various dialect families such as Dutch, Scandinavian, Italian, English, Portuguese, Estonian and Slovenian (cf. www.dialectsyntax.org for an overview and examples). Although primarily intended for linguistic research, this digital infrastructure also provides a rich resource for introductory and advanced courses on sociolinguistics, dialectology and the methodology of linguistic research. This paper shows that the infrastructure is also very useful for courses in applied linguistics.

Three of the corpora in this online research infrastructure, DynaSAND (www.meertens.knaw.nl/sand), GTRP (www.meertens.knaw.nl/mand) and DIDDD, all available in the online tool MIMORE (www.meertens.knaw.nl/mimore),[1] have been used in a course on Language Analysis for the Determination of Origin (LADO). Various countries use LADO in asylum procedures as one of the means to determine whether an asylum seeker was socialized in the country or area that s/he claims to originate from (cf. the papers in Zwaan et al. 2010). The determination of origin is crucial in the asylum procedure because only asylum seekers who have a well-founded fear of being persecuted in their country of origin for reasons of race, religion, nationality, political opinion or membership of a particular social group may get a permit to stay. LADO is used in particular when other ways to establish the origin have failed and there is serious doubt about the origin of the asylum seeker.

It is the responsibility of linguists to make clear if and how LADO can be a valid method. As is well-known from sociolinguistic research, the relation between a speaker's language properties and geographic origin is often highly problematic (cf. Patrick 2010). Various complications are demonstrated and discussed during the course, on the basis of the data in MIMORE, including the lack of reliable data, speaker-internal variation partially dependent on language background, accommodation dependent on speech situation, register and participants, unclear or gradual as opposed to categorical geographic distribution of certain linguistic features, the difference between linguistic and political borders and the requirement to make intuitions about geographic origin explicit qualitatively and quantitatively

This paper describes the role of MIMORE and its databases in the course *Locating People with their Language*. From 2011, this course has been taught several times as part of the Utrecht University Master's Program *Language, People and Society*. To make the role of MIMORE in this course transparent and put it in context, I provide a description of the content of the course, the way it was taught and how it was received by the students. One clear finding is that students tend to overestimate their ability to derive

[1] MIMORE (Microcomparative Morphosyntactic Research tool) was developed by Jan Pieter Kunst, Matthijs Brouwer and Folkert de Vriend (Meertens Institute) in a CLARIN-project (www.clarin.nl/) led by Sjef Barbiers.

geographic origin from language properties. Therefore, the knowledge provided in this course is essential. The course provides insight in the kind of linguistic expertise that is required to qualify as an expert LADO analyst. The usefulness of this course is not restricted to students of linguistics. Interpreters and language specialists (i.e., native speakers) who play an important role in LADO procedures but usually do not have a background in linguistics would greatly benefit from this course as well, as a first step towards becoming an expert.

2. The MA-course *Locating people with their language*

2.1 Introduction to the course

The course starts with an introductory discussion of variation as an inherent property of natural language. The goal of this discussion is to replace commonplace ideas and myths about language variation with a scientific attitude that calls such ideas and myths into question. Fundamental questions are raised, such as:

(1) Why is not there just one human language instead of the plethora of languages and lects that we find?

(2) What could be the advantage of the human capacity to locate and identify people with their speech?

(3) On the basis of which linguistic properties do we locate people in everyday life?

(4) How reliable is this capacity?

(5) What do we mean when we say that a person speaks Dutch and does this make sense?

We start with the distinction between language as a cognitive and as a social phenomenon. The students get some basic insight in the relation between language variation and factors such as gender, age, social class, ethnicity, geographical origin. The effect of language variation is that a community (e.g., a country or a region) is divided into groups which are partly overlapping and dynamic, giving rise to highly complex patterns. To a certain extent the group defining capacity of language variation can be compared to that of clothing fashion.

The second question is whether such a division and the human capacity to recognize group membership has any advantage. To tentatively answer this question we look at language and humans from an evolutionary perspective. Human individuals live in groups and need groups to survive. It has been hypothesized that such groups have a maximal size of around 150 due to the limited social-cognitive capacities of human brains (cf. Dunbar 1998). This makes a division of communities into smaller groups necessary. Language variables function as shibboleths: it is important for humans to differentiate ingroups from outgroups.

The third question is meant to make the students aware that although we have clear intuitions about the group to which an individual belongs and hence about his or her geographic origin, it is very complicated to make explicit the linguistic properties on which such intuitions rely. It is clear that such explicitness should be required in LADO procedures, as it is in other legal procedures. Most intuitions rely on lexical and phonetic properties, which, however, are usually not categorical but gradual. There are also properties, in particular syntactic properties, that are below the level of consciousness and therefore hard to observe and report for non-linguists. The reliability of our locating capacity (question 4) is dependent on these factors. Later in the course the students get an assignment to test the reliability of their own judgements (see section 2.5).

As to question 5, the commonplace idea that the world can be divided into distinct dialects and languages is debunked. It is made clear that language varieties usually cannot be demarcated very sharply, neither in the individual nor in society and space, and that it makes more sense to speak of a continuum of language varieties. Therefore, it is often not possible to count languages and locate them in space in a sufficiently precise way, which is a serious complication for LADO procedures.

2.2 General language resources

In this part of the course the students are virtually placed in the position of a linguist who is responsible for a LADO procedure, as follows:[2]

[2] This is just one type of LADO procedure. A range of types exists, e.g. LADO procedures in which the interviewer is a linguist or uses no interpreter.

> *Suppose there is an asylum seeker who claims to come from area A where people speak language or dialect B.*[3] *If this claim is true, the asylum seeker will be granted a residence permit, because area A is dangerous (for him/her). The central part of the LADO procedure is an interview with the asylum seeker in his/her native language or in a second language that s/he speaks, carried out by an interviewer and an interpreter, both non-linguists. The recordings of the interview will then be analyzed by a native speaker, often a non-linguist, and you, the linguist who is responsible for the procedure and the final report. What kind of background information do you need to be able to carry out this procedure, which linguistic resources are available and how reliable are such resources?*

A number of questions must be addressed in carrying out this procedure. Which languages and dialects are spoken where in the relevant area when and by whom? What is the position of language B in this language situation? How (in-)stable is the language situation? What are the linguistic properties of language B? Which literature on this language is available? Are there linguistic atlases of the language areaa? Are there any linguistic specialists on this language? Are any of them native speakers of it?

The students are asked to explore general language resources (mainly on the web) that (may) play a role in LADO procedures. The resources used include, among others, Ethnologue (www.ethnologue.com/), UNESCO Atlas of the World's Languages in Danger (www.unesco.org/culture/languages-atlas/), World Atlas of Linguistic Structures (http://wals.info), Eurominority (www.eurominority.eu) and LinguistList (www.linguistlist.org). The students must write and present a report on which types of information these resources make available, and they also give an evaluation of the reliability of this information.

The latter is very important because quite a few students tend to take the reliability of such information for granted, especially if it comes from official sources such as SIL (originally known as the Summer Institute of Linguistics, the organization behind Ethnologue), UNESCO, Max Planck Institute (WALS) etc. They find out that the

[3] In the remainder of this paper I will use the term language to refer to both languages and dialects, except in cases where it is necessary to distiguish between the two.

number of linguistic atlases that can be useful in LADO procedures is very limited. They gain the fundamental insight that language situations are constantly changing and that therefore information on the number of languages in an area, number of speakers, location of the language, language properties etc. can never be complete and fully up to date.

If a resource nevertheless provides such information, checks need to be made on when the data on which this information is based were collected, how they were collected and by whom. For example, it makes a huge difference whether the data come from a recent census or reflect the intuitions of an individual linguist. It is also important to take into account the potential negative impact of self-reporting on the reliability of the data, especially self-reports by bilectal and bilingual speakers. As an example, we identify in the course, on the basis of the MIMORE data, a set of speakers in the western part of the Netherlands who consider themselves speakers of Standard Dutch but clearly are (also) speakers of an Hollandic dialect. Such speakers may deny the existence of a certain linguistic property in their dialect while at the same time using it in spontaneous speech.

Eventually, the students should be able to judge the extent to which a piece of information from such linguistic resources can be reliably used in the LADO report on the asylum seeker from area A who claims to speak language B. They should also understand the risks and consequences of the choice of a particular interpreter and language analyst in the framework of LADO. It is important to know whether these participants in the LADO procedure speak the same variety as the asylum seeker or a language variety related to it, and if the latter, how these varieties differ from each other.

2.3 Sociolinguistic variation and multilingualism

This part of the course addresses the problem of sociolinguistically determined variation and multilingualism, on the basis of Patrick (2010) and Muysken (2010), two papers that the students have to read. The goal is to understand language variability in its full complexity and to learn to relate this complexity to the possibility of determining the origin of a speaker on the basis of his/her language. I will give a summary of the insights presented in these two papers that play a central role in the course.

A key issue is whether the language background of a specific speaker allows us to determine his/her origin. A distinction needs to be made between geographic area and speech community as the origin of a speaker. There are various situations in which these two do not coincide. This is for example the case when a speaker is a member of an immigrant group that has retained their native language and not adopted a language of the receiving area. With LADO it may well be possible to determine the speech community from which the speaker originates, but impossible to determine his or her geographic origin. This is a serious complication, because it may very well be the situation in his/her geographic area, rather than the speech community, that brings the asylum seeker to the decision to fly. Another example of such a situation is when a speech community is spread over a geographical area that crosses a political border, with a safe and a dangerous side.

The language variety or varieties that a person speaks depend on a large array of factors, including the language(s) of the parents, the languages in the environment (especially of peers), age, sex, social class, ethnicity, education. This means that a speech community is never homogeneous. This raises the question of which variety has to be taken as the standard of comparison when it is tried to establish the speech community from which the speaker originates.

The mobility of the speaker and his/her family is another crucial factor. When a speaker has a history of moving from place to place, especially during childhood, s/he may speak a range of language varieties. These varieties may influence each other, with the result that in a concrete speech situation the speaker may switch between the language varieties and his/her proficiency may vary from variety to variety (Muysken 2010). The linguist responsible for a LADO-procedure therefore needs to investigate the language background of the speaker to establish whether s/he can be subjected to such a process.

Furthermore, the language use of a speaker is known to depend on the speech situation, a crucial factor for LADO-interviews which of course constitute highly artificial and formal situations (see the description of the interview situation in section 2.2). Situational factors determining language properties include the bureaucratic context, unequal power relations among the interview participants, necessity for language choice and interpretation, existence of ethnic, class, or racial conflicts which affect cross-cultural communication, pressures on minorities to assimilate linguistically to majorities, tendency to accommodate to the language that has a higher prestige, prevalence of language contact, code-switching and language mixing, prescriptive language attitudes and ideologies (Patrick 2010:79).

Another crucial insight that the students need to gain is that linguistic differences between language varieties are often not categorical but gradual and need to be measured both qualitatively and quantitatively. When students are asked to determine the geographic origin of a speech fragment, they typically motivate their answers with: the speaker is using word X which is typical for the south, or, the speaker does not pronounce sound Y, which is typical for the east. Students tend to take such properties as categorical, while they often are not. Furthermore, whether a particular sound is pronounced or not also depends on the linguistic context within a language variety, e.g. sound Y is pronounced in stressed syllables but not in unstressed ones. Also, two language varieties may differ in the number of times that a sound is pronounced, all things being equal. For example, a speaker of southern Dutch may pronounce the /t/ in /niet/ 'not' 20% of the time, while a speaker of central Dutch does that 80% of the time. Therefore, the students have to learn to analyze speech fragments and their transcriptions linguistically, both qualitatively and quantitatively.[4]

2.4 Microvariation databases, tools and atlases for language analysis

2.4.1 Introduction

In this part of the course the students get acquainted with the basics of linguistic analysis necessary for locating speech fragments, using a number of Dutch resources. There are several reasons to choose resources on varieties of Dutch, despite the fact that LADO in the Dutch situation usually involves language varieties spoken in parts of Africa and Asia.

First of all, the students in this course are Dutch themselves and hardly ever speak or understand one of the languages relevant for LADO in the Netherlands. It would therefore be very complicated to demonstrate the intricacies of linguistic intuitions and analysis with such foreign language varieties.

A second reason for choosing varieties of Dutch is that this part of the course has

[4] Note that such details of frequency are almost never available in LADO reports as they are unknown for the target languages.

as an additional goal to let students experience how difficult it is to make linguistically explicit an intuition about the geographic origin of a language variety related to their own. Varieties of Dutch serve this goal well. When a student who speaks Standard Dutch hears a speech fragment of a Dutch dialect variety for the first time s/he will have certain intuitions about the origin. S/he will then be asked to make these intuitions explicit by comparing the linguistic properties of this variety with those of Standard Dutch. Not only will the student experience how difficult this is, but also, the student is in this way in the same position as the (non-linguist) language specialist in a LADO procedure who is the speaker of a national language and has to judge a dialect of that language. Put differently, the student will get a first idea of whether a language specialist in such a situation will be able to come to valid conclusions and motivate them.

A third reason is that in this part of the course the students need to learn that certain linguistic properties are below their level of consciousness. They will not notice certain differences between their Standard Dutch variety and a Dutch dialect until the teacher has made them aware of these. This is particularly the case for (morpho-)syntactic differences. When asked to mention some (morpho-)syntactic differences between Standard Dutch and familiar Dutch dialects, students are generally not able to do so, while they are able to give examples of differences at the lexical or phonetic level. Similarly, when asked to give the linguistic differences between a particular speech fragment of a Dutch dialect and Standard Dutch, they typically overlook the (morpho-)syntactic differences.

For example, there are Dutch dialects in which, in addition to the finite verb, the complementizer agrees with a plural subject, with a /-n/ or /-ə/ suffix as the exponent. Such minimal (morpho-)syntactic differences usually go unnoticed, not only by the speakers of the relevant varieties themselves (cf. Pauwels 1958 and Barbiers 2015) but also by speakers of related varieties that do not have these properties. (Morpho-)syntactic properties therefore potentially are determining factors in a LADO-procedure, as a speaker trying to imitate a variety will typically omit (morpho-)syntactic properties that are below the level of consciousness.

To teach students the relevance of minimal, often subconscious linguistic differences and how to recognize and describe them, we introduce them to a number of

Dutch microvariation databases and corresponding software tools. The software tool MIMORE (www.meertens.knaw.nl/mimore) gives access to three databases: the GTRP database on phonological and morphological variation in the Dutch language area, the DynaSAND database on syntactic and morphosyntactic variation at the clausal level, and the DIDDD database on syntactic and morphosyntactic variation at the level of the nominal group. The students use these three databases for their assignments. There now follows a brief description of the content and functionality of GTRP, DynaSAND, DIDDD and MIMORE.

2.4.2 GTRP

The GTRP database includes the results of a data collection project carried out between 1979 and 2000 under the responsibility of the linguists Goeman, Taeldeman and van Reenen. Users who understand Dutch can access the data at www.meertens.knaw.nl/mand/database/. Other users should use the MIMORE tool (see section 2.4.5). For GTRP, data were collected in 611 locations across The Netherlands (including Frisia), Flanders (i.e. the Dutch speaking part of Belgium) and French Flanders, a small part of North-west France. The informants in these locations were asked to translate a list of 1876 items. These items were mainly individual words and phrases, and sometimes complete sentences. All informants had to meet the following requirements:

- The informant speaks the dialect of the community;
- The informant is born in the place of residence and has lived there preferably his/her whole life; the same goes for his/her parents;
- The informant is between 50 and 75 years old;
- The informant is preferably low-educated but with considerably high literacy skills.

Given the goal of the project, to chart phonological and morphological variation in the Dutch language area, these informant requirements were necessary to ensure that there was a relation between linguistic variable and geographic location and to reduce the potential influence of other sociolinguistic variables such as social class and age.

The atlases resulting from this project include the phonological atlas of the Dutch dialects FAND (three volumes; FAND I: Goossens et al. 1998; FAND II+III: Goossens et al. 2000; FAND IV: De Wulf et al. 2005) and the morphological atlas of the Dutch dialects MAND (two volumes: MAND I: de Schutter et al., 2005 and MAND II: Goeman et al. 2008). Together the three FAND volumes give a detailed overview of variation in the vowel and consonant systems of the Dutch dialects and the geographic distribution of this variation. The two MAND volumes give on overview of the variation in plural formation, diminutives, gender, comparatives and superlatives, possessive pronouns, subject and object pronouns, verbal inflection, participles and verb stem alternations.

2.4.3 DynaSAND

For an extensive description of DynaSAND and its background see Barbiers et al. (2007) and Barbiers and Bennis (2007). DynaSAND can be accessed at www.meertens.knaw.nl/sand and in MIMORE. The data in DynaSAND were collected between 2000 and 2003 in 267 locations in The Netherlands, Flanders and North-West France. The basis of the selection of locations was an even distribution across the language area, with higher density in areas where the dialects are still very strong and numerous, in transitional zones and in locations with special circumstances, e.g. (former) islands. The goal of the project was to chart the geographic distribution of (morpho-)syntactic variation at the clausal level. Therefore the informants in the fieldwork stage had to meet more or less similar requirements to the GTRP informants.

The methodology of data collection for DynaSAND was different from GTRP, though. There were three stages: a postal pilot study, oral interviews and telephone interviews. The atlases (SAND I, Barbiers et al 2005; SAND II, Barbiers et al 2008) are based on the oral and the telephone interviews. In the oral interviews in The Netherlands, there were two informants in each location and they did the interview together in the local dialect without interventions by the fieldworker. This was to reduce accommodation as much as possible and to avoid judgements based on phonetic and lexical differences. As opposed to the Dutch interviews, the Flemish interviews were carried out by linguists who spoke the dialect or regiolect of the area and there were two informants in each location.

Around 150 different syntactic properties in 424 test sentences were investigated. We mainly used translation tasks and concealed judgement tasks. The latter did not ask for the grammaticality but for the commonality of a construction in a particular dialect, to avoid influence of normativity on the judgements. Often translation and judgement tasks were combined, also to check whether translation and judgement were consistent. Other types of tasks that we used include cloze tests, completion tasks and picture response tasks.

The data available in DynaSAND and the two SAND volumes include the left periphery (complementizer system, complementizer agreement, Wh questions, relative clauses, other fronting constructions), subject pronouns, subject pronoun doubling and cliticization, reflexive and reciprocal pronouns, morphosyntax of verbal clusters and auxiliaries, verb cluster interruption, negation and quantification.

With the DynaSAND software tool it is possible to search the database with text strings, strings of Parts-of-Speech tags, test sentences, syntactic phenomena, locations, and areas. The results of these searches are lists of geo-referenced sentences with tagging, English glosses and translations and the corresponding sound fragments. This makes it possible to check the validity of the data and to select the locations that have a particular syntactic phenomenon. This selection can then be fed into a cartographic tool that depicts the geographic distribution of the phenomenon and in the case of multiple phenomena, the correlations between them. Most of the maps of the printed volumes SAND I and II are also available in DynaSAND by searching with syntactic phenomenon.

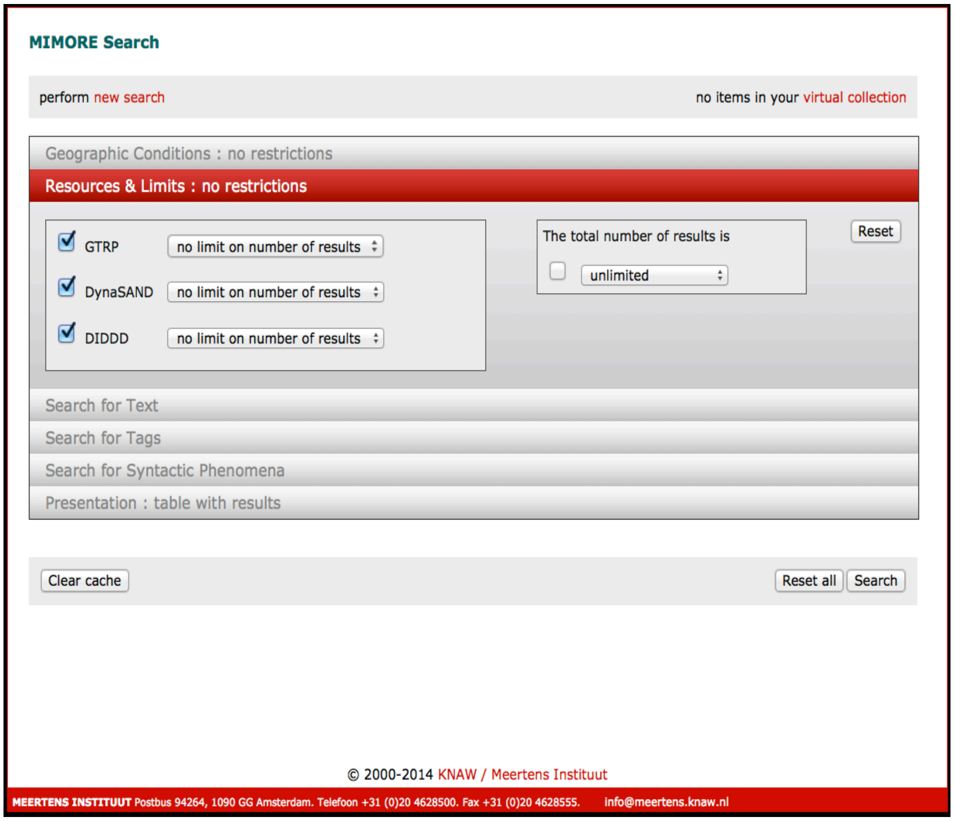
2.4.4 DIDDD

The data for the Diversity in Dutch DP Design database were collected between 2005 and 2009 in about 200 locations in the Netherlands, often the same locations as in DynaSAND and with a methodology comparable to that of DynaSAND. For a more extensive description see Corver et al. (2007). The DIDDD data can be accessed through MIMORE.

The goal of DIDDD was to describe the variation in nominal groups in the Dutch dialects. Phenomena investigated include noun phrase internal pronouns, substantivized pronouns, combinations of definite articles, demonstratives and possessive pronouns/phrases, number, negation, and quantification. For an overview of attested variation cf. Corver et al 2013.

2.4.5 MIMORE

MIMORE (www.meertens.knaw.nl/mimore) was developed to enable the researcher to search DynaSAND, GTRP and DIDDD at once and in a uniform way. In this way morphological properties, (morpho-)syntactic properties at the level of the nominal group and (morpho-)syntactic properties at the clausal level can be related to each other. It is possible to search with text strings, strings of Parts-of-Speech tags and syntactic phenomena.



MIMORE Search

perform new search

no items in your virtual collection

Geographic Conditions : no restrictions

Resources & Limits : no restrictions

☑ GTRP — no limit on number of results

☑ DynaSAND — no limit on number of results

☑ DIDDD — no limit on number of results

The total number of results is

□ unlimited

Search for Text

Search for Tags

Search for Syntactic Phenomena

Presentation : table with results

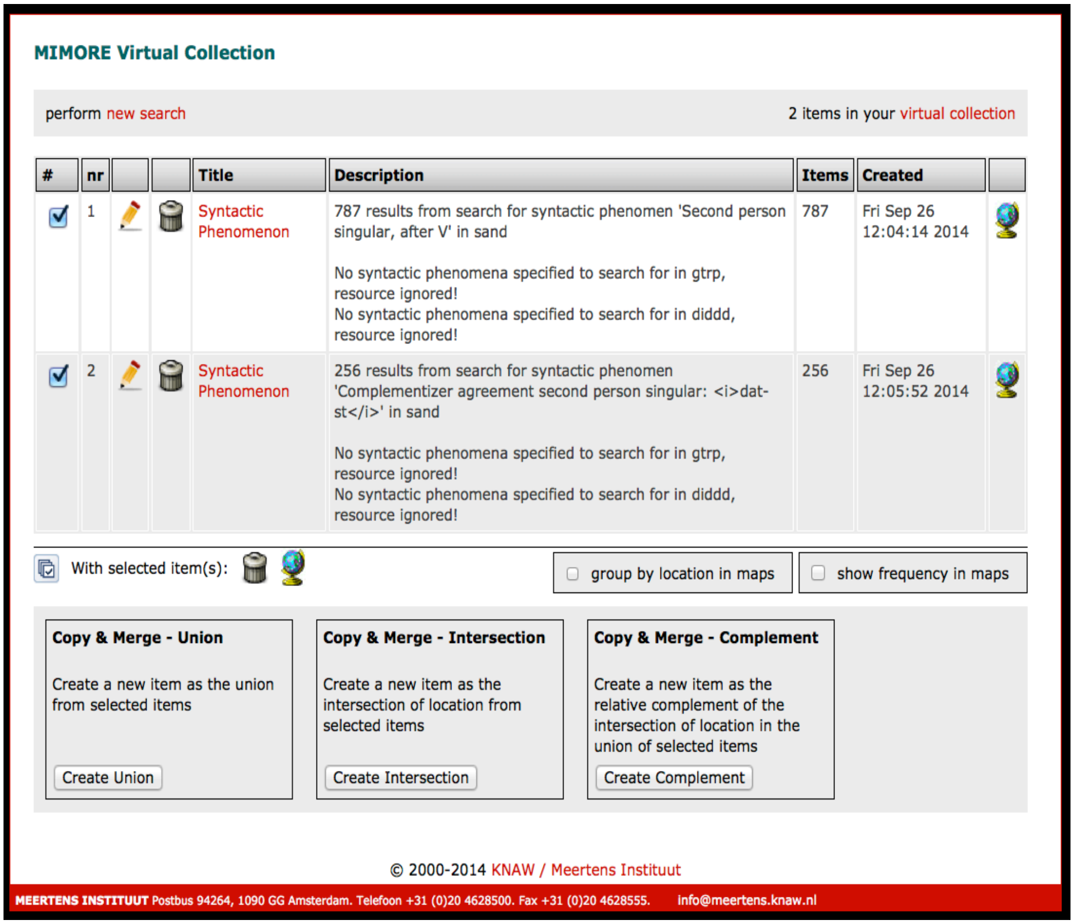
© 2000-2014 KNAW / Meertens Instituut

MEERTENS INSTITUUT Postbus 94264, 1090 GG Amsterdam. Telefoon +31 (0)20 4628500. Fax +31 (0)20 4628555. info@meertens.knaw.nl

Figure 1: The MIMORE search tool

The POS-tags to be included in the search can be constructed from a list of primitive categories and a list of primitive features, or one can use a list of predefined complex tags.

As in the case of DynaSAND, the result of a search is a list of geo-referenced sentences or sentence fragments with their POS-taggings, glosses and translations and the corresponding sound fragments (if available). In the case of search results from GTR, phonetic representations are given as well, which is relevant for students who need to compare pronunciation in detail. Selections of search results can be exported. One way of exporting is to a so-called virtual collection which makes further operations possible. It is possible to derive the intersection, union and complement set of sets of selected search results (i.e., of their locations). This way, potential correlations between two or more phenomena can be investigated. Sets of search results can also be presented on geographic maps.



MIMORE Virtual Collection

perform new search

2 items in your virtual collection

#	nr			Title	Description	Items	Created	
☑	1	✎	🗑	Syntactic Phenomenon	787 results from search for syntactic phenomen 'Second person singular, after V' in sand No syntactic phenomena specified to search for in gtrp, resource ignored! No syntactic phenomena specified to search for in diddd, resource ignored!	787	Fri Sep 26 12:04:14 2014	🌐
☑	2	✎	🗑	Syntactic Phenomenon	256 results from search for syntactic phenomen 'Complementizer agreement second person singular: <i>dat-st</i>' in sand No syntactic phenomena specified to search for in gtrp, resource ignored! No syntactic phenomena specified to search for in diddd, resource ignored!	256	Fri Sep 26 12:05:52 2014	🌐

With selected item(s): 🗑 🌐

□ group by location in maps

□ show frequency in maps

Copy & Merge - Union

Create a new item as the union from selected items

Create Union

Copy & Merge - Intersection

Create a new item as the intersection of location from selected items

Create Intersection

Copy & Merge - Complement

Create a new item as the relative complement of the intersection of location in the union of selected items

Create Complement

© 2000-2014 KNAW / Meertens Instituut

MEERTENS INSTITUUT Postbus 94264, 1090 GG Amsterdam. Telefoon +31 (0)20 4628500. Fax +31 (0)20 4628555. info@meertens.knaw.nl

Figure 2: MIMORE Virtual Collection

2.5 Use of microvariation databases and tools in the course

After having read some introductions to the databases, tools and atlases involved (Barbiers 2006, Goeman 2006, Taeldeman 2006, Corver et al. 2007), the students get an assignment developed by Wilbert Heeringa (Groningen University). It involves speech fragments with identical content from eight different locations/dialects in the Netherlands and Flanders. Six of these locations are close to the Dutch-Belgian national border, the fragment of the two other locations serve as control items. The main question of this assignment is: is it possible to determine whether a certain fragment is from the Belgian part of the language area or from the Netherlandic part? This mimics the situation in a LADO procedure, in which an informant must be located in the right area, where 'right area' is politically defined. In this assignment, we are dealing with a political division of one language area (the Dutch one) and the dialects on both sides of the border are closely related, i.e. we have three cross-border minimal pairs of Limburgian, Brabantish and Flemish dialects.

The assignment consists of the following steps. First, the students listen to the eight fragments. Then they are asked to transcribe (some of) the fragments orthographically and partly in IPA. Using these transcriptions, they give a detailed description of the lexical, phonetic, morphological and syntactic differences between each fragment and Standard Dutch[5]. Finally, they have to determine on the basis of these descriptions whether a speech fragment belongs to the Netherlandic or the Belgian part of the language area and where in the area the speech fragment is located. To be able to make these final steps they need to compare their descriptions of the sound fragments with the transcriptions and sound fragments that are available in the three databases in MIMORE and with the maps in the various printed atlases.

There are various ways in which the students can use the databases and tools. A first option is to search for location. This is useful in cases where the student has an intuition about where a speech fragment could be located but does not yet know how to support this intuition with linguistic properties. In such a case s/he can search, e.g., in DynaSAND with one or more location codes or names or with a geographic region or province. The results of such a search are the complete interviews for one or more

[5] Comparing dialect and standard language is not a typical part of LADO tasks but it is necessary to train the students to identify and describe linguistic features.

locations. The student can then start to compare the transcriptions and sound recordings of these interviews with the material to be located.

A second way of using the databases and tools obtains when the student has no idea where a speech fragment should be located, but s/he has found a number of linguistic properties that are distinct from Standard Dutch. In this situation, s/he can start searching with these linguistic properties, with text strings, POS tag strings, test sentences and syntactic phenomena, to find out in which locations or areas these properties occur.

Obviously, these two ways can and should be combined. For example, when a student uses the second method and has found a number of locations, the next step should be to go to the full interviews of these locations and make a detailed comparison between these interviews and the fragment to be located in order to find a list of common properties. Again, it is very important in this stage of the course that the student learns that it is not enough to say: I found this word/sound/construction, therefore the fragment must from location/area X. Rather, the outcome of this assignment should be that the student understands that only a list of (preferably quantified) properties from various linguistic domains will be convincing evidence for a particular location.[6]

The results of this assignment in the past four years show that students tend to stick to the lexical and phonetic level when describing the properties of a dialect. This is a striking result given the enormous amount of syntactic and morphological data available, and the fact that syntactic and morphological properties are often more distinctive and decisive than lexical and phonetic properties. As was suggested in section 2.4.1, they possibly neglect syntactic and morphological properties because such properties are often subconscious. The next question is of course why that would be the case. Needless to say, as soon as the students have been made aware of particular syntactic and morphological properties that are typical for dialects their perfomance on this decision task improves. Another finding is that there seems to be a discrepancy between the ability to decide where in the language area a fragment should be located and the ability to make such a decision linguistically explicit. That is, the average student

[6] Note that LADO analysis normally begins with a preexisting set of features that distinguish the targeted dialects, rather than generating a list of features that strike the analyst as salient while listening, as described here.

performs not flawlessly but reasonably well on the location task, but there is quite some variation in the quality of linguistic argumentation that is used to reach such a decision.

2.6 A location experiment

Following up on the course described here and inspired by Foulkes and Wilson (2011), Smidt van Gelder designed an experiment to find an answer to the question: Can analysts correctly locate speakers whose dialect is not the same as the analyst's, but a related one, e.g. can speakers of British English correctly locate speakers of American English dialects? The experiment and its results are reported in Smidt van Gelder (2012). The question she asked is directly relevant for the LADO-procedure. Recall that it is not always possible to find a native speaker of exactly the same dialect for the roles of language analyst and interpreter and that in such cases speakers of related language varieties will be chosen. It is important to know how this influences the validity of the conclusion.

Smidt van Gelder presented a number of speech fragments of Limburgian dialects close to the Dutch-Belgian border to groups of (non-linguist) subjects at increasing distances from the geographic origin of the speech fragment. Unsurprisingly, she found that the further the distance, the more mistakes in locating the speech fragment. However, she also found that even speakers of exactly the same local dialect make mistakes and are not always able to determine whether a speaker is from his/her own village or from a neighbouring village close by across the border.

Smidt van Gelder concludes that the speakers in her experiment were not very good at locating a speaker of a related dialect, that their performance got worse with increasing geographic/linguistic distances, that they were very bad in motivating their decisions, that these motivations were usually superficial and based on feelings rather than facts and that despite all of this the speakers were very confident that their locating judgements were right.

The question should therefore be raised whether native speakers who do not speak exactly the same language variety as the asylum seeker should play a role in LADO procedures. This research also shows that the people involved in LADO procedures should be trained in linguistics to be able to use and analyze linguistic resources and to provide explicit and factual motivation when they try to locate a person.

2.7 Remainder of the course

For the sake of completeness, I describe the final part of the course, in which MIMORE and its databases play a more limited role. Up to this point, the course has introduced the students to the availability and use of linguistic resources, the basics of linguistic variation, the LADO procedure and the problems involved in it. In the next stage of the course, the students attend three lectures by guest speakers.

The first guest speaker is a linguist who is responsible for LADO-procedures at the Dutch immigration and naturalization service (www.ind.nl). She discusses the various aspects and steps of the procedure. Central in this lecture is the question of whether it is possible to locate a speaker on the basis of his or her *second* language. More specifically, is it possible to differentiate the L2 English of speakers from various parts of Africa, to make the differences between these L2 varieties of English explicit and relate them to the L1's of the speakers? With Simo Bobda et al. (1999) as background reading the students have to analyze a speech fragment from an African English speaking informant and to derive the geographical origin of this speaker, using online databases of English accents and other resources.

The second guest speaker is a linguist who works for the *Taalstudio*, a Dutch company that makes (among others) counter expert reports for asylum seekers in LADO procedures (www.taalstudio.nl). She discusses the debate on whether native speakers can be reliable language analysts in LADO (cf. Cambier 2010), the guidelines for LADO (cf. the annex in Zwaan et al. 2010) and the requirement for a scientifically responsible approach to LADO and the research needed for that (cf. McNamara et al. 2010).

LADO can be considered as a forensic application of linguistics. To make the students familiar with other forensic applications of linguistics there is a third speaker, a phonetician who works for the Dutch forensic institute (http://www.forensischinstituut.nl/). He discusses the role of language analysis in criminal investigations, not only to locate people but also to derive a speaker profile and to establish whether two speech fragments recorded on two distinct occasions can be traced to the same speaker. Here the data in MIMORE become relevant again to make judgements on speech fragment origin, identity and speaker's profile explicit.

For the final session of the course, the students have to analyze a public debate on the racist anti-islam pamphlet *De ondergang van Nederland, land der naieve dwazen* (The decline of the Netherlands, country of naive fools). This pamphlet was published in 1990 under the name of Mohamed Rasoel, and the public debate is about whether the Dutch author Gerrit Komrij could be the real author. The students have to evaluate the argumentation in this debate, which is mainly about style. They get a brief introduction in measuring style and the use of stylistic analysis tools, such as the demonstrator Stylene of the University of Antwerp (http://www.clips.ua.ac.be/cgi-bin/stylenedemo.html).

3. Conclusion

From the course evaluations it is clear that this course opens up a whole new world for the students, not only the world of linguistics but also the relevance of linguistics for problems in the real world. They learn to look at language variation in a new way, ask fundamental scientific questions about it, get familiar with resources that show the actual variation, and learn to question and evaluate the validity of such resources. They then learn to analyze real language fragments, using the online tools, databases and maps available in MIMORE. It becomes clear that such tools and resources are indispensible for a serious attempt to locate a speech fragment and its speaker. Unfortunately, the Dutch language area is one of the few language areas in the world for which such extensive and detailed resources are available.

References

Barbiers, S., H.J. Bennis, G. de Vogelaer, M. Devos and M.H. van der Ham, 2005. *Syntactische Atlas van de Nederlandse Dialecten/Syntactic Atlas of the Dutch Dialects Volume I.* Amsterdam : Amsterdam University Press.

Barbiers, S., 2006. De Syntactische Atlas van de Nederlandse Dialecten. In: *Taal & Tongval*: 18, 7-40.

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

Barbiers, S, L. Cornips and 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: Beal, J.C., K.P. Corrigan, H.L. Moisl (ed.). *Creating and digitizing language corpora. volume 1: Synchronic databases.* Hampshire: Palgrave Macmillan. 54-90.

Barbiers, S., J. van der Auwera, H.J. Bennis, E. Boef, G. De Vogelaer and M.H. van der Ham, 2008. Syntactische Atlas van de Nederlandse Dialecten Deel II / Syntactic Atlas of the Dutch Dialects Volume II. Amsterdam : Amsterdam University Press, 2008.

Barbiers, S. 2015. European Dialect Syntax. Towards an infrastructure for documentation and research of endangered dialects. In M. Jones (ed.) *Endangered Languages and New Technologies*. Cambridge: Cambridge University Press. 35-48.

Cambier-Langeveld T. 2010. The validity of language analysis in the Netherlands. In K. Zwaan et al (red.), 21-34.

Dunbar, R. 1998. The social brain hypothesis. *Evolutionary Anthropology* 6 (5), 178-190.

Corver, N., M. van Koppen, H. Kranendonk and M. Rigterink, 2007. The noun phrase: Diversity in Dutch DP design (DiDDD). *Scandinavian Dialect Syntax 2005, Nordlyd, 34,* 73-85. Kristine Benzen & Østein Alexander Vangsnes (Eds.)

Corver, N., M. van Koppen and H. Kranendonk, 2013. De nominale woordgroep vanuit dialect-vergelijkend perspectief: Variaties en generalisaties. *Nederlandse taalkunde*, 18(2), 107-138.

De Schutter, G., T. Goeman, B.L. van den Berg, T. de Jong 2005. *MAND Morfologische Atlas van de Nederlandse Dialecten Deel I / MAND Morphological Atlas of the Dutch Dialects Volume I.* Amsterdam: Amsterdam University Press.

De Wulf, C., J. Goossens, & J.Taeldeman, 2005. Fonologische Atlas van de Nederlandse Dialecten. Deel IV. De consonanten. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Foulkes, P & Wilson, K. 2011. Language analysis for the determination of origin: An empirical study. *Proceedings of the 17th International Congress of Phonetic Sciences*, 17, 691-694.

Goeman, T., 2006. De Morfologische Atlas van de Nederlandse Dialecten (MAND); zero en bewaard gebleven morfologische informatie. Taal & Tongval Themanummer 18, 66-92.

Goeman, A., 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, J., J. Taeldeman and G. Verleijen, 1998. Fonologische Atlas van de Nederlandse Dialecten. Deel I. Gent: Koninklijke Academie voor Nederlandse Taal en letterkunde.

Goossens, J., J. Taeldeman and G. Verleijen, 2000. Fonologische Atlas van de Nederlandse Dialecten. Deel II. De Westgermaanse korte vocalen in open lettergreep. Deel III. De Westgermaanse lange vocalen en diftongen. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

McNamara, T., M. Verrips, C. van den Hazelkamp 2010. LADO, validity and language testing. In K. Zwaan et al. (eds.), 61-72.

Muysken, P. 2010. Multilingualism and LADO. K. Zwaan et al. (eds.), 89-98.

Patrick, P.L. 2010. Language variation and LADO. K. Zwaan et al. (eds.), 73-88.

Pauwels, J. 1958. *Het dialect van Aarschot en omstreken*. Brussel: Belgisch Interuniversitair Centrum voor Neerlandistiek

Simo Bobda, A, H.G. Wolf, L. Peter, 1999. *Identifying regional and national origin of English-speaking Africans seeking asylum in Germany.* In: *Forensic Linguistics* 6, 300–19.

Smidt van Gelder, N. 2012. *Wao kûmst dich vanaâf? Taalanalyse in asielprocedures: het vermogen van native speakers om sprekers van een ander dialect te herkennen*. Unpublished MA thesis Utrecht University.

Taeldeman, J. 2006. De Fonologische Atlas van de Nederlandse dialecten (FAND). Opzet, uitwerking en operationaliteit. *Taal & Tongval* Themanummer 18, 116-147.

Zwaan, K., M. Verrips and P. Muysken (eds.), 2010. *Language and Origin: The Role of Language in European Asylum Procedures: Linguistic and Legal Perpectives.* Nijmegen: Wolf Legal Publishers.