European Dialect Syntax. Towards an infrastructure for documentation and research of endangered dialects.

Sjef Barbiers (Meertens Instituut and Utrecht)

1. Introduction

Dialects have not hitherto played a very prominent role in the field of endangered language documentation and research - and for understandable reasons. Given the large number of endangered languages, the work that needs to be done in this field is already overwhelming. Moreover, as even a so-called ‘small language’ area such as Dutch is fragmented into over 200 dialects, including dialects in the endangered language enterprise would lead to an explosion of varieties to be documented.[1] However, this chapter argues that there are compelling reasons to include them. To make the documentation and analysis of dialects feasible, an online research infrastructure is needed, where linguists can store and access the relevant data and where they can co-operate in the description and analysis of these data. This chapter describes the European Dialect Syntax project (2005-2112) (hereafter, Edisyn), which attempts to establish a documentation and research infrastructure for the (endangered) dialects of Europe (and beyond). Its main focus is on syntactic variation. The new technology used by Edisyn provides access to large amounts of dialect data that were previously not available to the linguist.

2. Dialects as endangered languages

There are two main reasons to consider, and to treat, dialects as endangered languages. First, it is commonly argued in Linguistics that no principled difference exists between dialects and languages in terms of their linguistic properties or complexity. The distinction between dialects and languages is usually based on political, social, cultural and economic criteria. The only real linguistic difference is that official languages have usually been subject to processes of standardization and its concomitant superficial and often invented grammar rules. Viewed from this perspective, it could even be argued that the documentation of dialects should be given priority over that of languages, as dialects arguably present more ‘natural’ systems and are therefore more interesting scientifically.

Second, there is a general and global consensus about the urgent need to collect, digitize and document dialect data. All over the world, local dialects are rapidly changing and/or disappearing in the wake of urbanization, increasing mobility, the use of (social) media, the influence of supralocal varieties (Wolfram and Schilling-Estes 1995) and language contact more generally. It may very well be that today’s oldest generation will be the last to retain local dialects (cf. Trudgill 2011).

Recording dialects, transcribing and documenting them and making them digitally available will probably not prevent them from changing and disappearing, but it will contribute to the preservation of our cultural heritage, the possibility of linguistic research and the increasing awareness of the inherent variability of language. Recordings of dialects exist in many language areas. However, these are usually only known about and available locally.

Reliable information on the number of dialects in the world, the number of speakers per dialect and their degree of endangerment is not available. Some (online) resources include dialects, for example, the *UNESCO Atlas of the World’s Languages in Danger*, *Ethnologue* and the *Wikipedia* list of endangered languages, however the information provided by these resources is often incomplete and somewhat dated.[2] This is because of the nature, and sheer size and complexity of the empirical domain. The world does not consist of a collection of discrete and stable dialects, but rather of a large number of dialect continua with fluid borders (cf. Bloomfield 1935: 51). The speakers of these dialects are often not able to identify which dialect they speak. This, in turn, makes it difficult to establish which dialects are endangered. In short, it is probably impossible to provide a complete, correct and up-to-date overview of the world's dialect situation.

Given the size and complexity of the issue, it is clear that a 'top-down' approach to documenting these endangered dialects is unlikely to succeed. Thanks to the technological developments of the past decade it is now possible to combine a 'bottom-up' approach with a more general availability of data. Every research and documentation group should select a moderately-sized dialect area and should collect data and information on the dialects in this area in a systematic way. The resulting data should be made available in an online research infrastructure that is accessible to all interested parties and that makes it possible to search, organize, visualize and analyze the data. The goal of the Edisyn project has been to set up such an infrastructure for the dialects of Europe (and beyond), with a mainly syntactic focus.[3]

3. The role of syntax in dialectology

Dialectology is traditionally concerned with phonetic and lexical variation. The goal of dialectology is to describe the geographical distribution of this variation and to establish the historical relations between dialects. Data resulting from this type of research are often elicited by asking informants to translate lists of words into a given dialect. For several reasons, morphosyntactic variation is more of a *terra incognita* (cf. Barbiers & Goeman 2013). First, dialect speakers themselves are usually much less aware of morphosyntactic variation than they are of phonetic and lexical variation. A famous example of this is given in Pauwels (1958). Pauwels presented the speakers of the Brabantish dialect of Aarschot with a negative concord construction involving two negative words to express a single negation. Speakers would usually deny that this construction was possible in their dialect by using the very construction that was being studied (1):[4]

(1) Interviewer:
Could you say in your dialect:
Hij wil **nie** eten **nie**.
he wants not eat not
'He does not want to eat.'
Dialect speaker:
Nee, dat kunnen we **nie** zeggen **nie**.
no, that can we not say not
'No, we can not say that.'

Such low awareness of features has sometimes even led to the conclusion that syntactic variation is virtually non-existent across dialects (cf. Kloeke 1927).

Second, it is much more difficult to investigate morphosyntactic variation than lexical and phonetic/phonological variation. While word lists and oral histories resulting from free or (semi-)guided conversations are usually sufficient to obtain a detailed picture of the lexical and phonetic/phonological properties of dialects, this is much less the case for morphosyntactic variation, which is more difficult to detect.

For example, many Dutch dialects display the comparatively rare linguistic feature of complementizer agreement (where complementizers such as *dat* 'that' and *als* 'if' take a suffix if expressing plurality). It took quite some time before this phenomenon was discovered because, in many dialects, the plural suffix is a schwa and phonetically-driven schwa insertion between two consonants is a quite frequent phenomenon in Dutch (cf. van Haeringen 1939). Only careful comparison of minimal pairs (2) can demonstrate that the schwa in (2b) is present to express plurality.

(2) a. Ik denk dat ze loop-t.
I think that she walks
b. Ik denk dat-e ze loop-e
I think that.pl they walk.pl

Moreover, morphosyntactic variation may only reveal itself in low-frequency, complex sentences. For example, dialectal and colloquial Dutch show a lot of morphosyntactic variation in long relative and WH-clauses (cf. SAND Volume 1; Barbiers, Koeneman & Lekakou 2009; Boef 2013) and in verb clusters (cf. SAND Volume 2; Barbiers 2005). Long relative- and WH-clauses vary with respect to the form and presence of the relative and WH-pronouns involved and the presence of one or more complementizers (3,4). Verb clusters show word order variation, such that we find four different orders in the Dutch dialects (5).

(3) Different ways to express the meaning 'Who do you think I have seen?' in varieties of Dutch.
a. **Wie** denk je **dat** ik gezien heb?
who think you that I seen have
b. **Wie** denk je **wie** ik gezien heb?
who think you who I seen have
c. **Wat** denk je **wie** ik gezien heb?
what think you who I seen have
d. **Wie** denk je **die** ik gezien heb?
who think you REL I seen have[5]
e. **Wie** denk je **wie** **(of)** **(dat)** ik gezien heb?
who think you who (if) (that) I seen have

(4) 'This is the man that I think I have seen.'
a. Dit is de man **die** ik denk **dat** ik gezien heb.
this is the man REL I think that I seen have
b. Dit is de man **die** ik denk **die** ik gezien heb.
this is the man REL I think REL I seen have
c. Dit is de man **die** ik denk **wie** ik gezien heb.
this is the man REL I think who I seen have
d. Dit is de man **die** **da** ik denk **die** **da** ik gezien heb.
this is the man REL that I think REL that I seen have

(5) 'I think that everyone should well be able to swim'.
a. Ik vind dat iedereen goed **moet kunnen zwemmen**.
I think that everyone well must can.inf swim.inf
b. Ik vind dat iedereen goed **moet zwemmen kunnen**.
I think that everyone well must swim.inf can.inf

c. Ik vind dat iedereen goed **zwemmen moet kunnen**.
I think that everyone well swim.inf must can.inf
d. Ik vind dat iedereen goed **zwemmen kunnen moet.**
I think that everyone well swim.inf can.inf must

If such complex constructions are found in corpora at all, their low number makes it impossible to establish the range and limits of any variation. Consequently, syntactic constructions need to be examined extensively and in detail using a sophisticated methodology (cf. section 3).

A third cause for the neglect of syntax in dialectology relates to the sociology of the field of linguistics. From the 1950s onwards, syntax was considered the preserve of generative linguistics, the primary goal of which was not to describe language variation but rather to discover the universal principles of natural language (cf. Barbiers 2013). However, during the 1970s and 1980s, developing a general theory of syntactic variation became a more and more prominent goal (cf. Chomsky 1981). From the 1980s onwards, the syntactic variation of dialects has entered the generative scene. Indeed, many projects established to collect data on dialect syntax have a generative origin.

4. Methodology for collecting dialect syntax data

A sophisticated methodology for the collection of data on dialect syntax was developed in the first three large-scale projects on dialect syntax: the ASIS project on Northern-Italian dialects which began in the early 1990s (cf. Benincà & Poletto 2007); the SAND project on Dutch dialects which began in 2000 (cf. Barbiers & Bennis 2007; Barbiers, Cornips & Kunst 2007) and the SADS project on Swiss German dialects which also began in 2000 (cf. Bucheli & Glaser 2001). These all form part of the Edisyn project.[6] This section gives a brief outline of some of the methodological requirements for large-scale research on dialect syntax. For a more detailed account of the main methodological considerations in Edisyn, see Cornips & Poletto (2005).

4.1. Selection of interview locations

Since it is difficult to know in advance which locations within a given language area should be included in a large-scale survey of dialect syntax, when the interview locations are being selected, existing knowledge about dialectal variation in that area should be combined with some general principles. The first principle is to overlay a grid on the map and select one or more locations from each cell in the grid. The number of locations per grid cell and the size of the grid cells of course depend on practical constraints such as the resources, manpower and time available for the task. The grid ensures that the distribution of locations over the language area is even, which is crucial for the investigation of the relation between linguistic features and geographical distribution and the way this distribution came about (think of, e.g. settlement history, language contact, political developments, geographical boundaries). It is also crucial for the implementation of visualization techniques that extrapolate the areal distribution of linguistic features from the individual locations (cf. Wattel & van Reenen 1994). A second principle would be that the number of locations should be higher in transitional areas, since these usually reveal more, different and less stable variation. Third, dialect areas known for their great diversity should have more interview locations. Fourth, isolated locations should be included in the sample as they may have developed quite independently from the rest of that dialect area. Examples of such locations could be islands or, alternatively, locations that are socially, culturally, economically or religiously isolated from their environment.

4.2 Informants and the interview setting

One of the complications in research on dialect syntax is that dialect speakers often speak a regional and/or a standard variety in addition to their dialect. The selection of speakers and the methodology of data collection should be such that the influence of such varieites is minimized as much as possible.

In the SAND-project, which had as its major goal to map the geographic distribution of (morpho-)syntactic properties, informants had to meet the following requirements: (i) aged between fifty-five and seventy; (ii) born and raised in the interview location and living there without any interruption longer than seven years; (iii) same requirement for the parents of the informants; (iv) no higher education and/or a normative attitude towards the dialect; (v) lower middle class;[7] (vi) active user of the dialect in at least one domain outside of the family. As a result of applying these criteria, the informant group will be relatively homogeneous, and thus linguistic variation due to social factors other than geographical proximity is factored out as much as possible.

There were two main reasons to work with informants aged between 55 and 70. First, these older speakers grew up in a time when the position of dialects was much stronger than today and most local dialects had not yet merged into regiolects. Secondly, it has been shown that there is a typical pattern for dialect use in different age groups. It tends to peak during adolescence, then goes down reaching its lowest level around the age of 45 and then goes up again, reaching a new peak around the age of 70 (cf. Downes 1984:191).[8] As a consequence of the choice of this age group, one could say that the maps in SAND volumes I and II, although they are based on data collected in the early 2000s, depict the geographic distribution of syntactic variables in the Dutch dialects between 1930 and 1950. A different choice of age group would almost certainly yield different geographic patterns.

Clearly, if the goal is to document and analyse language variation and change as complete as possible then other age groups should be included. More generally, such investigation should include all variation arising from other factors than geography, as is done in sociolinguistics, and this makes the task even bigger.[9]

Where possible, the interview should be carried out using the local dialect. In order to obtain spontaneous speech, two or more dialect speakers, both of whom meet the criteria described above, are encouraged to start a conversation in the local dialect on topics that they would normally discuss in that dialect, such as family life, local celebrations, customs and other aspects of everyday life. When syntactic constructions need to be tested systematically, one of the dialect speakers can be trained to conduct the interview by presenting the test sentences in the local dialect and asking the other dialect speaker if this sentence can be said in this way in his dialect. Limiting the role of the researcher in the interview setting is desirable as this minimises the likelihood of accommodation to the standard language.

The different Edisyn projects described in this chapter have shown that documentation and research of dialects requires both systematic data testing and spontaneous conversation. Spontaneous conversation cannot tell us if a construction is systematically possible or not in a given dialect. On the other hand, however, systematic testing tells us little about optionality and the relative frequency of two or more variants (cf. Fernández-Ordóñez 2010).

There are various tasks that can be used in the systematic testing of morphosyntactic constructions. One is an indirect judgement task in which informants are asked if certain sentences proposed in the local dialect are common in that dialect, with commonality usually expressed on a point scale. Using such a scale is more useful than simply asking if a sentence is 'good' or 'bad', as this may trigger normative behavior. A second task involves sentence repetition, which is particularly useful if the informant's response involves unconsciously changing the syntax of the original test sentence. Further tasks include translation, picture tasks and cloze tests.

It should be noted that, even with these sophisticated methodologies, the resulting data only scratch the surface of the morphosyntactic variation present in the dialects. For example, for each of the seven doctoral dissertations that were written on the basis of the SAND-project (Zeijlstra 2004; van Craenenbroeck 2004; van Koppen 2005; de Vogelaer 2005; Haslinger 2007; Neuckermans 2008; Boef 2013) it was necessary to go back to the informants and test many more sentences. The advantage of course was that there was already an extensive network of informants and a database of syntactic variables, so the students knew where to go and what to look for.

5. Online infrastructure for dialect syntax research and documentation

5.1 Large dialect syntax projects

To date, eleven large-scale dialect syntax projects have been completed and thirteen are still running. For a complete overview and descriptions of the individual projects, see dialectsyntax.org. This website also provides a manual for dialect syntax projects, advising on organizational, methodological and technological aspects of such projects. Seven out of the eleven projects completed have resulted in the creation of a database, which can be searched using the Edisyn search engine, a tool developed at the Meertens Institute (meertens.knaw.nl/edisyn/searchengine). The search engine offers the possibility of searching with strings, Parts-of-Speech-Tags and English glosses and mapping data sets on Google maps. The searchable databases include: (a) SAND (Dutch dialects) (meertens.knaw.nl/sand); (b) ASIT (Italian dialects) (asit.maldura.unipd.it); (c) Cordial-Sin (Portuguese dialects) (clul.ul.pt/en/resources/212-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects); (d) The Nordic corpus of Scandinavian dialects (tekstlab.uio.no/nota/scandiasyn/index.html); (e) FRED (English dialects) (www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/); (f) EMK (Estonian dialects) (murre.ut.ee/home); (g) The Slovenian dialect syntax database (meertens.knaw.nl/edisyn/searchengine).

These databases are of mixed types. While the SAND, ASIT and Slovenian databases involve the elicitation and translation of test sentences, the Portuguese, Scandinavian, English and Estonian collections consist of corpora of conversations and stories. The geographic distribution of syntactic phenomena in these dialect databases can be mapped by selecting and analyzing the relevant data using the Edisyn research tool.

The plan is that in the future, databases will be added from projects that are currently running. These include the Scandinavian judgement database (tekstlab.uio.no/nota/scandiasyn/) and databases on dialects of Basque (Basdisyn; basdisyn.net), Spanish (COSER; lllf.uam.es:8888/coser), Occitan (DADDIPRO; dialectsyntax.org/wiki/Projects_on_dialect_syntax#DADDIPRO_on_Occitan_Dialects), Breton (Arbres; arbres.iker.univ-pau.fr), Alemannic (SynAlm; ling.uni-konstanz.de/pages/home/synalm/), Welsh (SAWD; lion.ling.cam.ac.uk/david/sawd), American English (YGDP; microsyntax.sites.yale.edu), Hessian (SyHD;

www.syhd.info), and Malagasy.

5.2 Infrastructure

An online research and documentation infrastructure for dialects should meet a number of requirements. First, the databases and tools included in such an infrastructure should not be stored on one central server. Rather, they should constitute a distributed network of databases, searchable using a common search engine (preferably via the Internet) and analysable with using a cartographic tool in order to visualize the geographical distribution of one or more syntactic property. The advantage of such a decentralized infrastructure is that every research group involved is able to maintain and update their own database independently.

Second, the infrastructure should be open access. This ensures that language researchers, educators, policy makers and the communities thata provided the data all will have access to these resources. Before publishing the data in open access, researchers should ask their informants for written permission. In the ideal case, this is done before the collection of the data, as getting permission afterwards is more difficult and more work. For those cases in which the informant can not be traced anymore a disclaimer can be added to the website stating that if any rights of the informant or his family are violated by the publication of the data, the data will be removed or made inaccessible upon request. Third, it is very important that every database be enriched with standardized metadata, so that the database can be selected on the basis of its properties (cf. CLARIN: Common Language Resources and Technology Infrastructure; clarin.eu). These metadata can include, for example, information on the language area and the dialects, dates of the recordings and profiles of informants.

Fourth, the sound recordings of each interview, if they exist, should be made available in the database. This is important because these are the raw data that every researchers should be able to verify. Publications on each dialect should be directly linked to these data whenever relevant, giving rise to so called enhanced publications that make the research results verifiable.

A fifth requirement is that the sound recordings are aligned with both phonetic (IPA) and orthographic transcriptions. Since this is a huge task, normalized orthographic transcriptions that that retain the sounds that may be relevant for morphosyntactic variation can be used instead of phonetic transcriptions, but preferably only temporarily. Transcriptions are necessary in order to search the sound files by phonetic or orthographic strings.

As a sixth requirement, English glosses should be added and aligned with the transcriptions in order to make it possible to search all the databases using (strings of) English words. A complete translation of the sentence in English should be included so that it is clear what the sentence means. English glosses are needed to make data from different dialect families accessible to the international research community. It is useful to also add glosses in the standard language associated with a particular dialect family, as this makes the comparison of these dialects much easier.

The seventh requirement is that enrichment of the data should include Parts-of-Speech-tagging (POS) in order to facilitate searching the databases using (strings of) tags. For example, if we want to know which dialects in Europe have complementizer agreement (cf. section 3) then it is necessary to interrogate the database for the sequence *C Infl Pron*. The problem here is that the amount of material is often too large to do this tagging manually, while the amount of material per dialect is too small to train an automatic tagger. Moreover, different research groups/language areas tend to use distinct sets of POS-tags, which makes it impossible to search multiple databases using a single set of tags. A common, standardized and well-defined tag set is therefore essential. Syntactic annotation is also advantageous in order to make it possible to search the database for syntactic constituents (such as prepositional phrases) and to investigate how these vary cross-dialectally. Syntactic annotation presupposes POS-tagging. Once this is available, the computer can carry out a considerable part of the syntactic annotation task.

Finally, each sentence in a dialect database should be geo-referenced so that the data can be used automatically as input for a cartographic tool. Search results should be easy to save and to export to statistical tools.

5.3 The Edisyn research infrastructure

The Edisyn infrastructure was set up to meet the above requirements. However, due to practical constraints, most of the databases included only meet part of them. The goal of setting up a distributed network of databases accessible via the Internet proved to be too ambitious. Most of the research groups involved did not have the necessary technical and financial resources or the expertise to make their databases available in the required way. As a temporary solution, it was decided to store a version of each database on the Meertens server, with the exception of the Scandinavian database which is searchable through a web service.

Currently the Edisyn search engine meertens.knaw.nl/edisyn/searchengine/) allows researchers to search one or more of the seven databases using strings, English glosses and POS-tags. However, not all of the databases include English glosses (or translations). All the databases include POS-tags, but these differ from database to database. An ISOCAT-certified POS-tag set has been created so that the databases may be searched using one common tag set (cf. Kunst & Wesseling 2011). Searching using syntactic annotation is not yet available. For those databases that include them, sound recordings are not yet accessible via the Edisyn search engine. However, these can be found by searching the individual databases.

Searches using the Edisyn Search engine yield lists of sentences that contain the properties that were searched, geographic coordinates, names of the locations and POS-tagging. These results can be plotted on a Google Map which may then be turned into a static map.

5.4 The usability of dialect syntax databases

The availability of large-scale dialect syntax databases is crucial in order to document the linguistic features of dialects that are changing rapidly and that may well disappear sooner or later. The tools they provide make the tasks of dialectologists and dialectometrists easier and more interesting and, moreover, they greatly enhance the empirical basis of syntactic research.

Theoretical syntactic frameworks such as generative grammar are currently shifting away from the methodology of idealization of data in the search for the the universal syntactic properties of natural language, and now take into account the full range of syntactic variation that can be found in colloquial language. In other words, it is seeking to understand syntactic variation in its full complexity, that is as a result of the interaction that occurs between fixed syntactic principles and factors at other linguistic levels and at cognitive and social levels.

These databases allow for statistical testing of potential correlations between syntactic properties. Examples of such potential correlations are features such as rich agreement and pro drop, agreement as a precondition for displacement, auxiliary doubling and the loss of the simple past. The data available in many of the databases are fine-grained, systematic and extensive enough to investigate such correlations at the level of individual members of a paradigm rather than at the level of an entire language, an approach necessitated by the current Minimalist hypothesis that there are no parameters defined over entire languages/dialects and that there is only parametrization at the level of individual lexical items and phonological spell-out.

Since the growing network of dialect syntax databases will extend beyond Indo-European (e.g., Basque, Malagasy), it is also possible to distinguish between correlations that hold within language families and those that hold across them. This is important because certain correlations may be due to common historical origins and developments, while others may be due to intrinsic and perhaps universal properties of natural language.

The network also allows for comparison of dialect families. For example, while many dialects of Dutch have so called long WH-doubling (6a), none of them has short WH-doubling (6b) (cf. Barbiers, Koeneman and Lekakou 2009). In certain Italian dialects, exactly the opposite holds (6c,d) (cf. Poletto and Pollock 2004).[10] The question is then whether this systematic difference between two dialect families can be derived from some other difference.

(6)a. **Wat** denk je **wie** ik gezien heb? Dutch
what think you who I seen have
'Who do you think I have seen?'
b. ***Wat** zag je **wie**? Dutch
what saw you who
'Who did you see?'
c.***Cossa** galo dito **chi** che el ga invidà? Paduan
What has-he said who that he has invited
'Who did he say that he invited?'
d. **Cossa** invitito **chi**?! Paduan
What invite-you who?
'Who did you invite?!'

Since the syntactic data in the dialect syntax databases are geo-referenced, the relationship between geographical patterns and grammatical systems can be investigated. For example, verb cluster interruption (the occurrence of a non-verbal constituent between the verbs in a clause-final verb cluster in Dutch) is possible with six different types of syntactic constituents in a central area located in the south west of the Dutch language area (cf. Barbiers et al. 2008, map 30b).[11] (7a, b) illustrate this feature for definite and bare plural objects.[12]

(7) a. Ik zei dat Willy moest **de auto** verkopen.
I said that Willy must.PAST the car sell
'I said that Willy should sell the car.'
b. Ik weet dat Jan wil **varkens** kopen.
I know that John wants pigs buy
'I know that John wants to buy pigs.'

The types of syntactic constituents with which verbal cluster interruption is possible decreases when one moves to the east and to the north. Indeed, in many non-central locations, verb cluster interruption is impossible (the verbs in a verb cluster are always adjacent). A pilot study shows that when asked about this, speakers in such non-central areas judge the rarest type of interruption (that which occurs in the smallest area), to be 'worse' than a type that is less rare etc. The most common type of verb interruption is deemed to be relatively (though not completely) acceptable. This is intriguing and gives rise to questions such as whether this correlation is due to a speaker's familiarity with particular dialect features, the frequency with which they encounter this feature, or whether indeed speakers have inherent intuitions about markedness.[13] More generally, the databases make it possible to investigate whether certain geographic patterns and clusterings are the result of factors external to language or due to inherent properties of linguistic systems.

6. Conclusion and future prospects

Recent technological developments have made it possible for large amounts of dialect data to be made available, searchable and analysable online. This will be of considerable benefit for the documentation of the world's dialects, most of which are endangered. The Edisyn project has shown that setting up a sophisticated infrastructure of databases for research on dialect syntax is technologically feasible, albeit with a few limitations, which can often be overcome by intensive co-operation between research groups. Data form the core of these research infrastructures and it is vital that research groups throughout the world make available their dialect data by digitizing, transcribing and enriching them. This will also make clear which data are missing from which dialects, thereby giving rise to new data collection initiatives.

The Internet offers new possibilities for collecting large amounts of data from large amounts of speakers in a large number of locations via online written or spoken questionnaires that systematically test properties of the dialects (cf. Boef 2013). As an increasing number of households in Europe now have access to a computer, this method of data collection will become more and more important

It is also vital that existing research and documentation infrastructures should be integrated with one another. An example of one such integration is the MIMORE tool (meertens.knaw.nl/mimore) that makes it possible to search three Dutch dialect databases at the same time, (SAND on morphosyntactic variation, GTR on phonological, phonetic and morphophonological variation and DIDDD on variation in nominal groups). On a larger scale, the integration of infrastructures such as Edisyn with those of LLMap (linguistlist.org), SSWL (Syntactic Structures of the World's languages; sswl.railsplayground.net) and WALS (World Atlas of Linguistic Structures; wals.info)) is also highly desirable.[14]

Notes

[1] Clearly, 'counting' the exact number of dialects in any given language area is an impossible task. However, this approximate figure is derived from the results of two large-scale dialect projects in the Dutch language area, MAND and SAND, both of which are discussed below.

[2] http://www.unesco.org/culture/languages-atlas/;http://www.ethnologue.com/web.asp. http://en.wikipedia.org/wiki/Lists_of_endangered_languages

[3] The Edisyn project was funded by the European Science Foundation. It is based at the Meertens Instituut (Royal Netherlands Academy of Arts and Sciences).

[4] The (un-)grammaticality of this negative concord construction is independent of the choice of main verb. Thus, in the relevant dialect of Aarschot all main verbs can occur with doubled *nie*, while in the majority of the Dutch varieties doubled *nie* (or *niet*) is categorically excluded.

[5] The Dutch relative pronoun *die*, homophonous with the distal demonstrative, is only compatible with plural and with singular common gender nominals.

[6] Cf. http://asis-cnr.unipd.it/ for ASIS, http://www.meertens.knaw.nl/sand/zoeken/index.php for SAND and http://www.ds.uzh.ch/dialektsyntax/eckdaten.html for SADS.

[7] In most parts of the Dutch language area, the higher social classes rarely use dialect. Clearly, this may be different in other language areas.

[8] According to Holmes (1992), this pattern is due to changing social pressure across the life span.

[9] Gender may also cause linguistic differences. However, given the difficulties involved in finding consultants who meet the requirements mentioned above, the SAND project includes both male and female informants, and assumes that gender has little bearing on morphosyntactic variation.

[10] The phenomenon is termed 'doubling' because that which is expressed with one Wh-element in Standard Dutch and Italian is expressed by two Wh-elements in the dialects.

[11] For clarity, verb particles that occur inside verb clusters have been excluded here.

[12] Other types of constituents that can interrupt the verb cluster include mass noun objects, indefinite objects, manner adverbs and PP complements.

[13] Questions of this type will be investigated in the Maps and Grammar project (2013-2018) funded by the Dutch research organization (NWO).

[14] The data from SAND are already available in SSWL.

References

Barbiers, S., Word order variation in three-verb clusters and the division of labour between generative linguistics and sociolinguistics. In Cornips, L. & K. Corrigan (eds.). *Syntax and Variation. Reconciling the Biological and the Social*, Amsterdam/Philadelphia: John Benjamins, 2005, 233-264.

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

Barbiers, S. and H.J. Bennis, The Syntactic Atlas of the Dutch Dialects. A discussion of choices in the SAND-project, *Nordlyd* 34 (2007), 53-72. http://septentrio.uit.no/index.php/nordlyd.

Barbiers, S., L. Cornips, J.P. Kunst, 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.P. Corrigan and H.L. Moisl (eds.), *Creating and digitizing language corpora. Volume 1: Synchronic databases*, Hampshire: Palgrave Macmillan, 2007, 54-90.

Barbiers, S., O.N.C.J. Koeneman and M. Lekakou, Syntactic doubling and the structure of wh-chains, *Journal of Linguistics* 45 (2009), 1-46.

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

Barbiers, S and T. Goeman, Research results from on-line dialect databases and dynamic dialect maps, to appear in F. Hinskens and J. Taeldeman (eds.) *Handbook of Language and Space* (Chapter 34). Berlin: Mouton de Gruyter, 2013.

Benincà, P. and C. Poletto, The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects, *Nordlyd* 34 (2007) (1) http://septentrio.uit.no/index.php/nordlyd

Boef, E., *Doubling in relative clauses: Aspects of morphosyntactic microvariation in Dutch*. Dissertation Utrecht University. LOT Dissertation Series 317, 2013.

Bloomfield, L., *Language*, George Allen & Unwin: London, 1935.

Bucheli, C. and E. Glaser, The Syntactic Atlas of Swiss German Dialects: empirical and methodological problems, in S. Barbiers, L. Cornips and S. van der Kleij, *Syntactic Microvariation*. Meertens Institute Electronic Publications II, 2001. http://www.meertens.knaw.nl/books/synmic/index.html

Chomsky, N., *Lectures on Government and Binding*, Dordrecht: Foris Publications, 1981.

Cornips, L. and C. Poletto, On standardising syntactic elicitation techniques, PART I. *Lingua* 115 (7) (2005), 939-957.

Craenenbroeck, J. van, *Ellipsis in Dutch dialects*, Diss. Universiteit Leiden, Utrecht: LOT-dissertations 96, 2004.

Downes, W., *Language and Society*, London: Fontana, 1984.

Fernández-Ordóñez, I., La Grammaire dialectale de l'espagnol à travers le Corpus oral et sonore de l'espagnol rural (COSER), in *Corpus* 9, special issue *La Syntaxe du Corpus/Corpus Syntax* edited by M. Olivieri, 2010, 81-114.

Haeringen, C.B. van, Congruerende voegwoorden, *Tijdschrift voor Nederlandse Taal- en Letterkunde* 58 (1939), 161-176.

Haslinger, I., *The Syntactic Location of Events: Aspects of Verbal Complementation in Dutch*, dissertation Tilburg University, LOT *Dissertation* Series 169, 2007.

Kloeke, G.G., *De Hollandse expansie in de zestiende en de zeventiende eeuw en haar weerspiegeling in de hedendaagsche Nederlandsche dialecten, proeve eener historisch-dialect-geographische synthese*, Noord- en Zuid-Nederlandse dialectbibliotheek 2, 's-Gravenhage: Nijhoff, 1927.

Koppen, M. van, *One Probe-Two Goals: Aspects of agreement in Dutch dialects*, Diss. Universiteit Leiden, Utrecht: LOT dissertations 105, 2005.

Kunst, J.P. and F. Wesseling, The Edisyn search engine, *Oslo Studies in Language*, 3 (2: J. B. Johannessen (ed.), Language Variation Infrastructure), 63-74, 2011.

Neuckermans, A., Negatie in de Vlaamse dialecten volgens de gegevens van de Syntactische Atlas van de Nederlandse dialecten (SAND), Doctoraat Universiteit Gent, 2008

Pauwels, J.L., *Dialect* van *Aarschot en omstreken*, Tongeren: Belgisch Interuniversitair Centrum voor Neerlandistiek, 1958.

Poletto, C. and Pollock. J.-Y., On Wh-clitics, Wh-doubling in French and some North Eastern Italian dialects., *Probus* 16 (2004), 241–72.

Trudgill, P., *Sociolinguistic typology: Social determinants of linguistic complexity*, Oxford: Oxford University Press, 2011.

Vogelaer, G. De, *Subjectsmarkering in de Nederlandse en Friese Dialecten*. Dissertation Universiteit Gent, 2005.

Wattel, E. and P. Th. van Reenen, Visualisation of extrapolated socio-geographical data, Rapport WS-429, Dept. of Mathematics Vrije Universiteit Amsterdam., 1994

Wolfram, W. and N. Schilling-Estes, Moribund dialects and the endangerment canon: The case of the Ocracoke Brogue, *Language* 71(4) (1995), 696-721.

Zeijlstra, H., *Sentential Negation and Negative Concord*, Diss. Universiteit van Amsterdam, Utrecht: LOT-dissertations 101, 2004.

[1] Clearly, 'counting' the exact number of dialects in any given language area is an

[2] http://www.unesco.org/culture/languages-atlas/; http://www.ethnologue.com/web.asp. http://en.wikipedia.org/wiki/Lists_of_endangered_languages

[3] The Edisyn project was funded by the European Science Foundation. It is based at the Meertens Instituut (Royal Netherlands Academy of Arts and Sciences).

[4] The (un-)grammaticality of this negative concord construction is independent of the choice of main verb. Thus, in the relevant dialect of Aarschot all main verbs can occur with doubled *nie*, while in the majority of the Dutch varieties doubled *nie* (or *niet*) is categorically excluded.

[5] The Dutch relative pronoun *die*, homophonous with the distal demonstrative, is only compatible with plural and with singular common gender nominals.

[6] Cf. http://asis-cnr.unipd.it/ for ASIS, http://www.meertens.knaw.nl/sand/zoeken/index.php for SAND and http://www.ds.uzh.ch/dialektsyntax/eckdaten.html for SADS.

[7] In most parts of the Dutch language area, dialect use is largely absent in the higher social classes. Obviously, this may be different in other language areas.

[8] According to Holmes (1992), this pattern is due to changing social pressure across the life span.

[9] Gender may also cause linguistic differences. However, given the difficulties involved in finding informants who meet the requirements mentioned above, the SAND project included both male and female informants, and assumed that gender had little bearing on morphosyntactic variation.

[10] This phenomenon is termed 'doubling' because that which is expressed with one WH-element in standard Dutch and Italian is expressed by two WH-elements in the dialects.

[11] For clarity, verb particles occurring inside verb clusters have been excluded here.

[12] Other types of constituents that can interrupt the verb cluster include mass noun objects, indefinite objects, manner adverbs and PP complements.

[13] Questions of this type will be investigated in the 'Maps and Grammar' project (2013-2018) funded by the Netherlands Organization for Scientific Research (NWO).

[14] The SAND data are already available within SSWL.