Bridging the Gap between First Language Acquisition and Historical Dialectology with the Help of Digital Humanities

Leonie Cornips[1,2], Mark Kemps Snijders[1], Martin Snijders[1], Jos Swanenberg[3], Folkert de Vriend[1]

1: Meertens Institute
Joan Muyskenweg 25, Amsterdam, The Netherlands
2: Maastricht University
Grote Gracht 90-92, Maastricht, The Netherlands
3: Tilburg University
Warandelaan 2, Tilburg, The Netherlands
{Leonie.Cornips, Mark.Kemps-Snijders, Martin.Snijders, Folkert.de.Vriend}@meertens.knaw.nl,
A.P.C.Swanenberg@uvt.nl

Abstract

It is remarkable that first language acquisition and historical dialectology remain strange bedfellows although in historical linguistics there is the common assumption that language change in the past is due to the process of non target like transmission of linguistic features between generations i.e. between parents and children. Both disciplines remain isolated from each other due to, among others, different methods of data-collection and different types of resources with empirical data. The aim of this paper is to demonstrate that the common assumption in historical linguistics, mentioned above, can be examined in detail with the help of Digital Humanities projects like the CLARIN-NL project COAVA (Cognition, Acquisition and Variation tool).

1. Aim of the paper

It is remarkable that first language acquisition and historical dialectology remain strange bedfellows although in historical linguistics there is the common assumption that language change in the past is due to the process of non target like transmission of linguistic features between generations i.e. between parents and children. Both disciplines remain isolated from each other due to, among others, different methods of data-collection and different types of resources with empirical data. The aim of this paper is to demonstrate that the common assumption in historical linguistics, mentioned above, can be examined in detail with the help of Digital Humanities projects like the CLARIN-NL project COAVA (Cognition, Acquisition and Variation tool).

2. Language acquisition and transmission

In general, in acquisition studies it is not controversial to assume that the child is cognitively equipped such that she explores the linguistic possibilities within a specific language and stabilizes on a language that is target-like, i.e. equivalent to that of the adults in her linguistic community. However, when this exploration takes too long and language external and/or language internal factors start to interfere with this process one may expect to find 'deviations' from adult language to arise in the language acquired by children (Cornips & Hulk 2008). In that case, the transmission process between generations of speakers i.e. between parent(s) and their children, is not target-like anymore which will result in emerging language variation and change. According to Kroch (2001): "Language change is by definition a failure in the transmission across time. Failures of transmission seem to occur in the course of language acquisition. Given a set of assumptions of UG, successful acquisition of a language's syntax clearly depends on the interaction of its structural properties with the character of the learner, so that as we learn more about the latter, we have a hope of better understanding diachrony." (see also Labov 2007).

With respect to language acquisition studies, people have long wondered how much of human knowledge is learned and how much is present at birth. We know from the experimental studies by Spelke that infants as young as three and four months, much like adults, understand that the world is composed of physical objects that are solid, substantial, and continuous in time and space, and interact with one another by contact and force transmission. The 'Spelke object' is what the infant has innately. The words for these objects constitute a relatively large proportion of children's early vocabularies. Thus words can be subdivided into basic level vocabulary such as 'dog' or 'tree' and superordinate or subordinate vocabulary like 'animal' or 'mammal' and 'terrier' or 'retriever', respectively (Bloom 2001). As lexical semantics distinguishes word form and word meaning, we will regard the word or lexical item as the form, and the concept as the meaning.

In cognitive linguistics, lexical concepts are categories that give structure to our knowledge of the world, linguistic features included (Geeraerts 1986: 187). These categories are hierarchically structured, with a central role for the most salient objects. An explanation of the lexical stability of basic level vocabulary might be that concepts at the basic level (basic level objects) are concepts that are deeper entrenched than others (Geeraerts e.a. 1994: 138142). For instance the subcategories of basic level objects are supposed to be less salient, less entrenched, and therefore their vocabulary (hyponyms) show a high degree of lexical variation. Another instance is the use of metaphors in lexicalization procedures, a creative process that results in, again, a high degree of lexical variation. Because of conceptual saliency basic level objects, however, get simplexes that are geographically widely spread (Rosch 1978). If lexical variation could be translated to the type-token-ratio (the relative degree of lexical variation) and the geographical distribution of lexical items, one could measure lexical variation accurately and make more specific hypotheses on the entrenchment and saliency of concepts.

Since children connect lexical items to Spelke objects so young already, basic level vocabulary constitutes an excellent starting point for bridging the gap between language acquisition and historical linguistics. We will focus on historical dialectology in this respect.

3. Historical dialectology: dictionaries of Brabantic and Limburgian dialects

In historical dialectology, much attention is paid to detect the largest differentiations between dialects through space and time. The variability of lexical variation in a relatively small language area can be very different. For specific concepts such as 'blue titmouse', 'thunder-shower' or 'pointy chin-beard' lexicographers find a sometimes overwhelming number of different words and wordings (Swanenberg 2004, 2010). Some of those lexical items are geographically restricted to a small dialect area or even one location only. For other, more generic concepts like 'bird', 'sun' or 'nose' there's hardly any or no lexical variation at all. Such concepts often are regarded as basic level objects and their lexicalization constitutes basic level vocabulary.

So research on the basis of the dialect dictionaries of the Brabantic and Limburgian dialect areas (in Belgium and the south of the Netherlands) shows that these language varieties exhibit an overwhelming amount of variation at the lexical level for most parts of the vocabulary. Many concepts have tens or even hundreds of different lexical items in relatively small geographical spaces, such as the Brabantic or Limburgian dialect area. These words are often complex: compounds and collocations that sometimes are periphrastical or even metaphorical. The arachnid daddy-long-legs has for instance 68 different lexical items in the Brabantic dialect area (Swanenberg 2010). Among those are:

hooispin 'hay spider'
hooiwagen 'hay wagon'
hooipaard 'hay horse'
wegwijzer 'guide'
horlogewerker 'watchmaker'
mieke langbeen 'Mary longleg'
scheper langpoot 'shepherd longpaw' etc.

However, the dialect dictionaries show remarkably little lexical variation regarding words for basic level objects. In fact the words for basic level objects usually are cognates of the words in other Germanic languages and dialects or even other Indo-European languages (Swadesh 1971: 283). Basic level vocabulary, as presented in Table 1 (see next column), in other words mainly consists of simplexes (Berlin 1992: 26-31), free words that are etymologically opaque and have a long history.

So, the lexical variation in a dialect area can be quite variable, depending on the conceptual level of an object, amongst other factors. This makes it worthwhile to examine for lexical items if there is a relation between (i) relative moment of acquisition of a lexical item and (ii) lexical variation throughout geographical space.

Eng.	**Dutch**	**Ger.**	**Fris.**	**Swe.**	**Latin**
sun	zon	Sonne	sinne	sol	sol
nose	neus	Nase	noas	näsa	nasus
fish	vis	Fisch	fisk	fisk	piscis
wine	wijn	Wein	wyn	vin	vinum
father	vader	Vater	heit	fader	pater
rose	roos	Rose	roas	ros	rosa
red	rood	rot	read	röd	ruber

Table 1: Examples of basic level vocabulary in various languages

4. COAVA

In the CLARIN-NL demonstration and curation project COAVA (Cognition, Acquisition and Variation Tool) tools are being developed that will enable exploring the lexical characteristics of language acquisition data and historical dialect data in two distinct resources. The tools and data in COAVA will offer support for the type of research sketched in the previous sections.

4.1 The Resources in COAVA

The two resources with empirical data that are being used in COAVA are child data in CHILDES and the dialect data from the Dictionaries of the Brabantic and Limburgian Dialects. Both resources until now have only been examined in isolation from each other.

For CHILDES only the data from the monolingual files of Dutch child language (Dutch and Flemish) are used. The Dutch CHILDES datasets are available in the CHAT format and in XML format together with a user interface for browsing, searching and other tools at the CHILDES website (http://childes.psy.cmu.edu/). CHILDES is the child language component of the TalkBank system. TalkBank is a system for sharing and studying conversational interactions (http://childes.psy.cmu.edu/). These files contain longitudinal data of children.

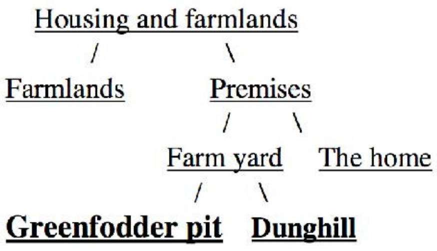
The second resource contains the data used for creating the Dictionary of the Brabantic dialects (WBD; completed in 2005) and the Dictionary of the Limburgian Dialects (WLD; completed in 2008). These data were digitized in the project Digital databases and digital tools for WBD and WLD (D-Square) (De Vriend, F & J. Swanenberg 2006). This resource contains lexical variation for a large part of the southern Dutch dialects from a period in time in which the vocabulary of the traditional dialects in the Dutch language area was disappearing at a rapid pace. The data for the Brabantic and Limburgian dialects were collected between 1880-1980. The vocabulary in these resources has two characteristics that make it stand apart from standard vocabulary: they are oral vocabularies and they are geographically differentiated. The data are onomasiologically arranged: every database table in the resource deals with a certain conceptual field (e.g. 'birds', 'the miller', etc.). The resource is in MySQL format and is available at the D-Square website together with a taxonomy in XML format (http://dialect.ruhosting.nl/d2/).

4.2 The COAVA tools

In the COAVA project explorative research into both resources will be enabled by offering the researcher specialized search interfaces. For implementing such search interfaces SOLR is used. SOLR is technology that came available in the open source community in recent years. With the SOLR technology we are able to build extensive faceted search interfaces on top of each of the two resources. Faceted search interfaces provide users fine-grained utilities that give them extended control, adaptability and flexibility (with regard to their constructed queries and retrieved result sets). Therefore, multiple web based facetted search interfaces are being developed. On the server side Apache SOLR (http://lucene.apache.org/solr/), an open source search engine development tool, is used to create Restful search services for both resources.

For the dialect resource access to the data will also be provided through a taxonomy of senses. Figure 1 shows part of this taxonomy.

Figure 1: Partial taxonomy for the dialect resource



Clicking on an end leaf of the taxonomy, like the sense *groenvoerkuil* ("greenfodder pit"), takes the user to all dialect data for that sense.

Search results for each resource will be further supplemented with information that is relevant in the context of the research sketched in the previous sections.

For the CHILDES data information about the age of the child who utters a specific lexical item i.e. noun is important. Therefore for each search result in CHILDES the age of the child will be provided.

For the dialect resource it is the information about the amount of variation in the geographical space that is of special importance. This information will be visualized with the use of automatically generated dialect maps using cartographic software developed at the Meertens Institute. Also a measure for the amount of geographical variation will be developed in the project.

4.3 Bridging the gap

In an attempt to bridge the gap between first language acquisition and historical dialectology the COAVA project will also focus on the lexical characteristics and variation of nouns that appear in both resources.

The nouns first need to be located in both resources and tagged for their part of speech.

For the dialect resource this is fairly easy since we can focus on the senses used for the titles of the lemmas. These are in standard Dutch and they contain nouns, verbs or adjectives. No other parts of speech exist in the dialect resource (cf. Figure 1).

For the CHILDES data locating the nouns in the target child utterances is more challenging. Not all CHILDES data are tagged for their part of speech. Therefore we are currently investigating to what extent we can make use of automatic tagging procedures like those offered by the mor tools in CLAN.

Next, the nouns found in CHILDES need to be mapped onto the nouns in the dialect resource. To enable this mapping the child acquisition forms first need to be lemmatized. After lemmatization of the CHILDES nouns these can be mapped onto the nouns in the dialect resource.

Finally, the mapped nouns will enable linking from the nouns in CHILDES to the nouns in the dialect data and vice versa. This will enable exploring if for these nouns there is a relation between their relative moment of acquisition (early or late) and their variation in geographical space.

4.4 Compliance with CLARIN

In addition to developing tools for supporting the type of interdisciplinary research stated above, Digital Humanities also plays an important role in making resources and tools widely accessible to all researchers in the Humanities and Social sciences. To this aim the dialect resource will be curated by converting it to XML. Also CLARIN specific guidelines and practices with respect to persistent identification, (CMDI) metadata, metadata harvesting and access policies will be applied to all relevant material in the project. The COAVA project thus also serves as an example of how different layers in an eScience infrastructure interact for a specific user community or research question.

5. Acknowledgements

The project reported on in this paper was funded by CLARIN-NL (www.clarin.nl).

6. References

Berlin, B. (1992) Ethnobiological classification. Principles of categorization of plants and animals in traditional societies. Princeton.

Bloom, P. (2001) How children learn the meanings of words. Massachusetts: MIT.

Bol, G. and F. Kuiken (1988) Grammaticale analyse van taalontwikkelingsstoornissen. Doctoral thesis, University of Amsterdam.

Cornips, L. and A. Hulk (2008) Factors of success and failure in the acquisition of grammatical gender in Dutch. In: Second Language Research 24 (3), 267-296.

De Vriend, F & J. Swanenberg (2006). D-kwadraat: digitale databanken en digitaal gereedschap voor WBD en WLD. In: Nederlandse Taalkunde, 11(4), 2006: 366372.

De Vriend, F., L. Boves, H. van den Heuvel, R. van Hout, J. Kruijsen & J. Swanenberg (2006). A unified structure for

Dutch Dialect Dictonary Data. Proceedings of The Fifth international conference on Language Resources and Evaluation (LREC 2006) ISBN 2-9517408-4-0.

De Vriend, F. de (2011). COAVA. CLARIN-NL Kick off meeting Call 2, Utrecht, 09-02-2011.

Francopoulo, G. George, M. Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). Lexical Markup Framework (LMF). Proceedings of the Fifth Language Resources and Evaluation Conference (LREC) ISBN 2-9517408-4-0.

Geeraerts, D. (1986) Woordbetekenis. Een overzicht van de lexicale semantiek. Leuven.

Geeraerts, D., S. Grondelaers & P. Bakema (1994) The Structure of Lexical Variation. Meaning, Naming, and Context. Berlijn/New York.

Gillis, S. & A. Schaerlakens (red.) (2000). Kindertaalverwerving. Een handboek voor het Nederlands. Groningen: Martinus Nijhoff.

Kampen, J. van (1997) First Steps in Wh-movement. Delft: Eburon.

Kroch, A. (2001). Syntactic Change. In The Handbook of Contemporary Syntactic Theory, M. Baltin and C. Collins (eds), 699-730. Malden, Mass.: Basil Blackwell.

Labov, William. 2007. Transmission and diffusion. Language 83.344-387.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2007). The TalkBank Project. In: J.C. Beal & K.C. Corrigan & H. Moisl (eds.) Creating and digitizing language corpora. Volume 1: Synchronic databases. Palgrave Macmillan, Hampshire, pp. 163
180.

Nerbonne, J. (2007). Crosstalk in Humanities Computing, International Journal for Humanities and Arts 594 Computing, Vol. 1, Page 85-96.

Nerbonne, J. and W. Heeringa (2010). Measuring dialect differences. In: P. Auer and J. E. Schmidt (eds.) Language and Space. An international Handbook of Linguistic Variation. Volume 1: Theories and Methods, De Gruyter Mouton, Berlin and New York, 550-567.

Rosch, E. (1978) Principles of categorization. In: E. Rosch en B.B. Lloyd (Eds.), Cognition and categorization. Hillsdale.

Swadesh, M. (1971) The origin and diversification of language. Edited post mortem by J. Sherzer. Chicago.

Swanenberg, J. (2010). Als het beestje maar een naam heeft: De verscheidenheid van lexicale variatie. In: J. De Caluwe & J. Van Keymeulen (Eds.), Voor Magda. Artikelen voor Magda Devos bij haar afscheid van de Universiteit Gent. Gent, 561-568.

Swanenberg, J. (2004) Origins of lexical variation. In: B. L. Gunnarsson e.a. (Eds.), Language variation in Europe. Papers from ICLaVE 2. Uppsala, 378-390.

Wijnen, Frank & Edith Kaan (2006). Dynamics of semantic processing: The interpretation of bare quantifiers. Language and Cognitive Processes, 21 (6), 684-720