Available online at www.sciencedirect.com

ScienceDirect

Lingua 178 (2016) 32–45

Lingua

www.elsevier.com/locate/lingua

# The relationship between first language acquisition and dialect variation: Linking resources from distinct disciplines in a CLARIN-NL project[☆]

Leonie Cornips [a,*], Jos Swanenberg [b], Wilbert Heeringa [c], Folkert de Vriend [d]

[a] Meertens Institute (KNAW) & Maastricht University, The Netherlands
[b] Tilburg University, The Netherlands
[c] Groningen University, The Netherlands
[d] Meertens Institute, The Netherlands

Received 8 October 2014; received in revised form 19 November 2015; accepted 19 November 2015
Available online 8 January 2016

**Abstract**

It is remarkable that first language acquisition and historical dialectology should have remained strange bedfellows for so long considering the common assumption in historical linguistics that language change is due to the process of non-target transmission of linguistic features, forms and structures between generations, and thus between parents or adults and children. Both disciplines have remained isolated from each other due to, among other things, different research questions, methods of data-collection and types of empirical resources. The aim of this paper is to demonstrate that the common assumption in historical linguistics mentioned above can be examined with the help of Digital Humanities projects like CLARIN. CLARIN infrastructure makes it possible to carry out e-Humanities type research by combining datasets from distinct disciplines through tools for data processing. The outcome of the CLARIN-NL COAVA-project (acronym of: Cognition, Acquisition and Variation tool) allows researchers to access two datasets from two different sub disciplines simultaneously, namely Dutch first child language acquisition files located in Childes (MacWhinney, 2000) and historical Dutch Dialect Dictionaries through the development of a tool for easy exploration of nouns.

*Keywords:* CLARIN infrastructure; Digital tools; Lexical variation; Dialectal lexicography; Child language acquisition; Corpus linguistics

## 1. Cognition, acquisition and variation tool: a CLARIN project

Digital Humanities programmes like CLARIN make it possible to carry out research by combining datasets from distinct disciplines through the use of tools for data processing. This paper will present the results of a CLARIN-NL project, i.e., the cognition, acquisition and variation project (COAVA).[1] In COAVA, a tool has been developed for easily searching nouns in two large datasets coming from two distinct disciplines: a dataset of child language utterances (the discipline obviously

[☆] We like to thank Hans Verhulst (Tilburg University) for correcting our non-native English and the three anonymous reviewers for providing valuable comments on an earlier version of this paper.

[*] Corresponding author at: Joan Muyskenweg 25, PO Box 94264, 1090 GG Amsterdam, The Netherlands. Tel.: +31 20 4628529.
*E-mail address:* leonie.cornips@meertens.knaw.nl (L. Cornips).

[1] See http://www.meertens.knaw.nl/coavasite/.

http://dx.doi.org/10.1016/j.lingua.2015.11.007

being language acquisition research) and a dataset of lexical variation in dialects (historical dialectology). These two resources containing empirical data have been linked together. The datasets in question are the Dutch monolingual child data (data on child language production) in CHILDES (MacWhinney, 2000) and the digital databases of the Dictionaries of the Dutch dialects of Brabant and Limburg (southern Dutch language area). Before COAVA was available, these resources could only be examined in isolation. The linkage of the two datasets has made it possible to examine the general research question whether there is a correlation between nouns produced by children at an early age (looked at from the perspective of acquisition, dealing with such notions as entrenchment and frequency) and the size of variation in these nouns over a large geographical area (seen from the perspective of lexical dialectology, dealing with such notions as lexical variation, salience and lexical complexity). In order to address this overarching research question, we will in our (modest) case study establish: (i) the age at which young children first produced various nouns, as indicated in the CHILDES datasets, (ii) the size of the (the amount of) geographical variation of these nouns in the datasets of the Dictionaries of the Dutch dialects of Brabant and Limburg. The point of departure in this investigation is that nouns produced at an early age will hardly show any geographical variation. We have developed a tool to examine a number of different nouns that are listed both in the CHILDES datasets and in the dialect database. For these nouns, measures of lexical dialect variation have been developed in COAVA, which through the use of the same tool can be correlated with the age of first production of the noun. Thus, the tool developed in COAVA has enabled us to connect different datasets and make new comparisons possible.

In addition, we will apply a second measurement, involving the lexical complexity factor (operationalized as the number of syllables), as nouns that are acquired early (generally referring to basic level objects) are likely to consist of one syllable while nouns acquired later (mostly expressing subordinate concepts) are more likely to consist of multiple syllables. Together, these measures enable us to examine whether there is a significant correlation between a child's first day of production (in the Childes dataset) and the measure (size) of lexical complexity. The assumption about a late age of acquisition of a linguistic phenomenon making it vulnerable to variation and change is taken up seriously in the COAVA project. The focus in the project, as mentioned above, is on nouns.

This paper is organized as follows. First, the two distinct datasets will be described. Second, the more technical details of the tools that help process the data will be introduced and we will explain how we go about measuring lexical variation. Third, the theoretical backgrounds of the COAVA project will be presented, more specifically how language change may be due to non-target transmission of linguistic features between adults and children. The focus is on the loss of the neuter gender in the definite determiner in Dutch. Finally, we will discuss a preliminary case study in order to show how the developed tool to examine different nouns listed both in the CHILDES datasets and in the dialect database works, and to find out whether the assumption (hypothesis) that nouns that are acquired early show hardly any geographical variation is confirmed or refuted. The outcome of our investigation is that the hypothesis is confirmed for one special subcase but falsified for the majority of cases. In this way, it is demonstrated how the digital data and digital tools that resulted from COAVA can be fruitfully used in linguistic research.

### *1.1. CHILDES*

The language acquisition data are taken from the CHILDES project. CHILDES is the child language component of the TalkBank system. TalkBank is a system for sharing and studying conversational interactions of very young children and one or more adults (MacWhinney, 2000). In the COAVA project the Dutch monolingual first language acquisition transcriptions of conversations from this database were used, namely the files of Antwerp, Bol, Gillis, Groningen, Schaerlaekens, Van Kampen, and Wijnen. This subset consists of 193,380 child utterances. The files contain longitudinal production data of children, which makes it possible to examine the children's developmental path towards the target adult language. The Dutch CHILDES datasets are available in the CHAT standard, both in the format suited for the CLAN tool, as well as in an XML format, complete with a user interface for browsing, searching and available tools at the CHILDES website (http://childes.psy.cmu.edu/). The database and tools developed in COAVA have enabled us to find out at what age children produce certain nouns. Moreover, COAVA reveals the frequency of these nouns as well.

Table 1 presents an example of the first day of production of a particular noun and its frequency, being the number of occurrences in the databases. In this case it concerns the noun *bird* and the subordinate terms *owl*, *blackbird*, *sparrow* and *titmouse*:

The first day of production of a noun and its frequency in the CHILDES corpora can automatically be charted in COAVA, as is illustrated in Chart 1 below.

### *1.2. Dialect dictionaries*

The second resource used in our investigation contains databases consisting of the collections of raw lexical data as included in two large regional dialect dictionaries *Woordenboek van de Brabantse Dialecten* 'Dictionary of the Brabant