The relationship between first language acquisition and dialect variation

Cornips, Leonie; Swanenberg, Jos; Heeringa, Wilbert; de Vriend, Folkert

Published in:
Lingua

DOI:
10.1016/j.lingua.2015.11.007

Publication date:
2016

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

Table 1
Example of first day of production and frequency of nouns in child language.

noun	First day	frequency
'bird' *vogel*	641	81
'owl' *uil*	680	11
'blackbird' *merel*	909	5
'sparrow' *mus*	821	4
'tit(mouse)' *mees*	869	1

dialects' (WBD, 1967–2005) and *Woordenboek van de Limburgse Dialecten* 'Dictionary of the Limburg Dialects' (WLD, 1983–2008). The data for the Brabant and Limburg dialects were collected largely between 1880 and 1980. Hence, these dictionaries contain lexical variation in a large part of the southern Dutch dialects extending over a period in which the vocabulary of the traditional dialects was disappearing at a rapid pace. They describe thematically (and also alphabetically via indexes) the dialects of Limburg and Brabant, both located in the south of the Dutch language area i.e., comprising the provinces of Antwerp, Flemish-Brabant, Brussels and Limburg in Belgium, and the provinces of Noord-Brabant and Limburg in the Netherlands. The vocabulary in the two dictionaries has two characteristics that make it stand out in comparison to the standard Dutch vocabulary. First of all, they are oral vocabularies and second, they are

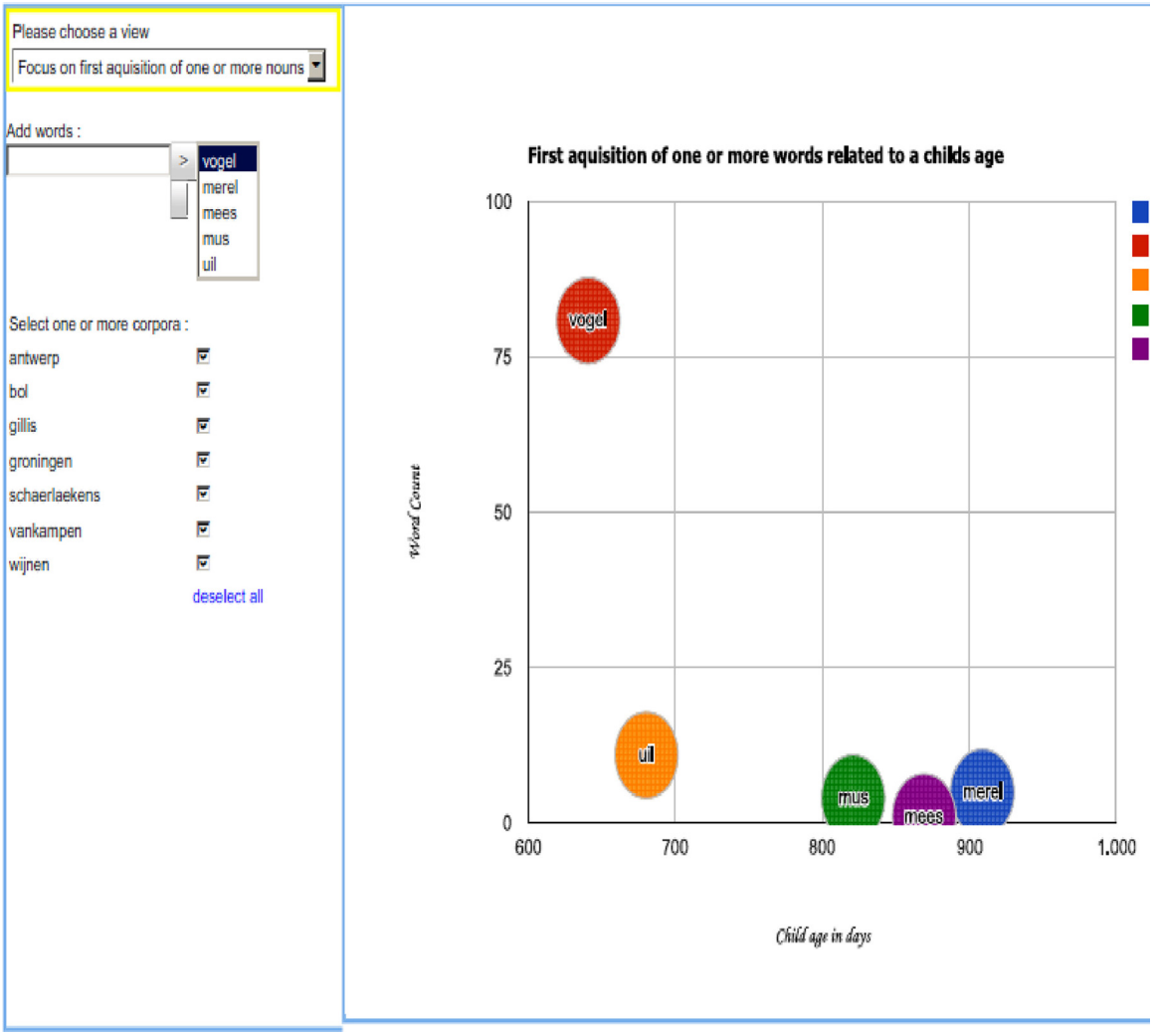


Chart 1. Example of first day of production of a noun and its frequency in the CHILDES corpora (automatically generated by the COAVA-tools).

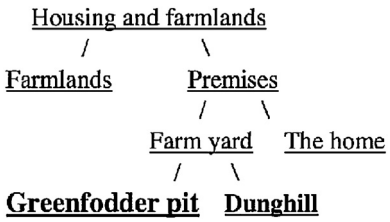


Fig. 1. Example of a partial taxonomy as employed in the dialect dictionaries (De Vriend et al., 2006).

geographically differentiated. The data are onomasiologically arranged, which means that every fascicle in the dictionaries deals with a certain conceptual field (examples being 'birds', or 'the miller').

The Dictionary of the Brabant dialects (henceforth: WBD) contains 1,365,593 reports of nouns (i.e. replies in questionnaires, entries in publications, etc.) for 5544 concepts; the Dictionary of the Limburg Dialects (henceforth: WLD) contains 1,277,247 reports of nouns for 7016 concepts; together this makes 2,642,840 reports of nouns for 9137 concepts.

The total number of concepts is de-duplicated' since the two dictionaries often contain the same concepts e.g., concepts occurring in both dictionaries are only counted once.

Access to the data is acquired via semantic taxonomies in which concepts are arranged hierarchically, as in Fig. 1.

The databases contain linguistic information (dialect form, 'dutchified' headword i.e., dialectal headword written as if the word were a standard Dutch word, lexical meaning), geographical information (locality, dialect area, province) and information on the source (inquiry forms or monotopic dictionaries and the date of documentation) (De Vriend et al., 2006). The most typical way for the user to access the data is through the use of the browsable taxonomy of concepts. The databases are thus approachable via search tools but also via a thematic taxonomy.

The dictionary data and the hierarchical structures of chapters, paragraphs and lemma's in the onomasiological dictionaries were digitized in the project *Digital databases and digital tools for WBD and WLD* (D-Square) (De Vriend and Swanenberg, 2006). 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/).

Clicking on an end leaf of a taxonomy, such as *groenvoerkuil* 'greenfodder pit' in the example given in Fig. 1, takes the user to all dialect data available for that concept. Search results for each resource are further supplemented with information about the variation in the geographical space. This information can be visualized with the use of automatically generated dialect maps using cartographic software developed at the Meertens Institute and applied to these data in COAVA. Map 1 is an example of such a dialect map. A measure for the proportions of geographical variation has also been developed in COAVA, as will be discussed in Section 4.

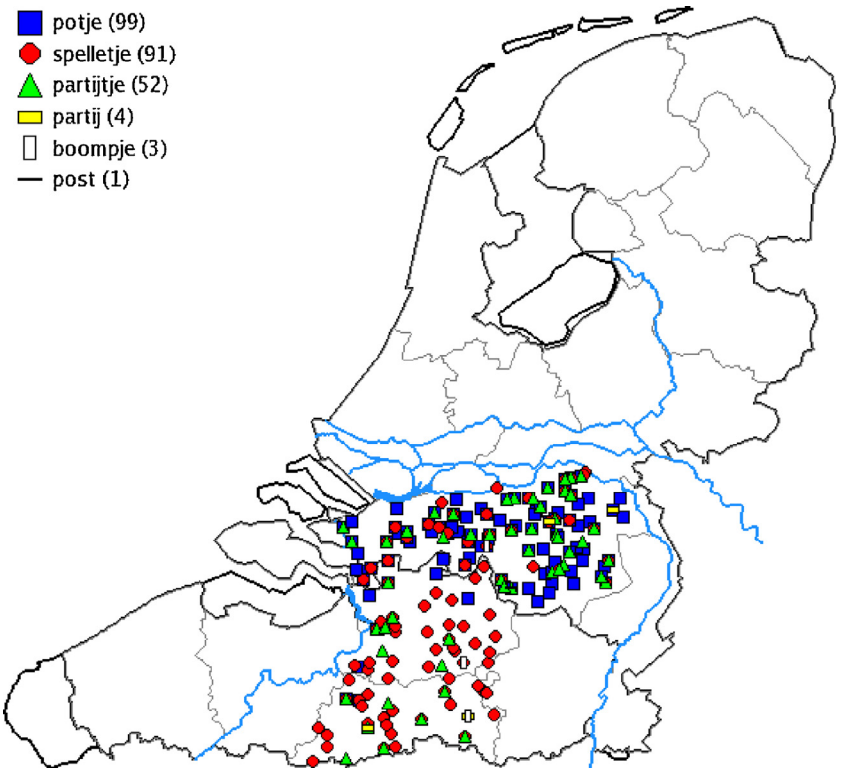
1.3. Linking both resources

Within the COAVA project a tool was developed enabling search queries for specific nouns in both resources simultaneously (the Childes child acquisition database and the Dialect Dictionaries). Searching nouns in the Dialect Dictionaries is fairly easy since they are organized by concept. These concepts are in standard Dutch and consist of nouns, verbs or adjectives. Other parts of speech are practically absent in this dialect resource (cf. Fig. 1).

Locating the nouns in CHILDES, and therefore locating the nouns in the target child utterances, however, presented more of a challenge. Not all CHILDES data were tagged for their parts of speech. To resolve this we made use of the automatic lemmatizing and tagging procedures offered by the *mor* tools in CLAN.[2] The nouns found in CHILDES were mapped onto the nouns in the dialect resource. To enable this mapping the nouns in the dialect resource were manually checked for their parts of speech and only the nouns were tagged. Subsequently, mapping of the nouns made it possible to link the nouns in CHILDES to the nouns in the dialect data, and vice versa. It also enabled us to explore whether there is a relation between the relative moment of production (early or late) of these nouns and their variation in geographical space.

To implement the COAVA search interfaces, we used a technology known as SOLR (http://lucene.apache.org/solr/), which has recently become available in the open source community. The SOLR technology allowed us to build extensive faceted search interfaces by applying multiple filters on top of each of the two resources, which make it possible for users to explore a collection of information. Faceted search interfaces provide users with fine-grained utilities affording them

[2] We like to thank Steven Gillis (University of Antwerp) for letting us use his lexicon for these procedures.



Map 1. The geographical distribution of the noun *spelletje* 'playgame'.

extended control, adaptability and flexibility (with regard to their constructed queries and retrieved result sets). Multiple web-based faceted search interfaces were thus developed.

In the next sections we will discuss some of the theoretical backgrounds of the COAVA project in order to introduce the research questions, measures of lexical variation, and the case study we will present in the final part of this paper.

2. Children as agents of diachronic change

In diachronic studies on the effects of language contact situations the general hypothesis is that there is a strong relation between language acquisition and diachronic language change. The child is cognitively equipped to explore the linguistic possibilities within a specific language and stabilizes on a language that is target-like, i.e., closely equivalent to that of the adults in his or her linguistic community. However, if the children's onset of acquisition is delayed, this can lead to incomplete acquisition resulting in language variation and ultimately in language change (Meisel, 2011:121). Thus, children acquiring their first language, and certainly children acquiring a second language, are frequently regarded as the principal agents of diachronic change although "the causes and the precise nature of the processes of change are (. . .) far from clear" (Meisel, 2011:121–123). Also Labov (2007:346) argues that: "The continuity of dialects and languages across time is the result of the ability of children to replicate faithfully the form of the older generation's language, in all of its structural detail [. . .]". Thus, variation and change occurs where children do not 'copy' their parents'/caretakers' language but instead orient to their peers, for example.

The research on both monolingual and bilingual child language acquisition shows that certain linguistic phenomena are acquired early while others are acquired relatively late, for instance in early school years (cf. Unsworth et al., 2011). Different internal and external factors have been considered responsible for structures being acquired 'late'. According to Unsworth et al. (2011, and references cited), internal factors may involve linguistic properties or cognitive and processing prerequisites. Linguistic properties may range from elements or structures being linguistically complex to their being underspecified or underdetermined by the grammar and therefore requiring more input and/or higher levels of processing abilities in the analysis and integration of their linguistic properties and lexical or pragmatic conditions. Prototypical external factors influencing acquisition are socio-economic factors, parental education and schooling.

A good example of 'late' acquisition of a linguistic phenomenon in Dutch is the neuter gender of the definite determiner. Standard Dutch makes a binary distinction between common and neuter nouns. This gender distinction is morphologically

Table 2
Errors in the use of common definite determiner *de* instead of *het* with neuter nouns by monolingual children between 3;2 and 7;10 years (taken from Blom et al., 2008:314).

Age range child L1	N of children	Neuter nouns: *de* instead of *het*	
3;2–3;10	7	88%	37/42
4;0–4;11	17	56%	54/93
5;1–5;11	15	31%	27/87
6;2–6;11	11	29%	31/108
7;1–7;10	14	24%	29/122

visible in the determiner if it has the features singular and definite: neuter nouns take the article *het*: common nouns take the article *de* (both of them corresponding to 'the' in English). In an experimental setting, Blom et al. (2008:314) show that even seven-year-old monolingual children still use the common definite determiner *de* with neuter nouns instead of the required determiner *het* in 24% of the cases (bilingual children show a higher percentage of the overuse of the common definite determiner with a neuter noun, see Cornips and Hulk, 2008) (Table 2):

What is important for this paper is that this late acquisition of the neuter determiner *het* in Dutch results in long-lasting variation between *het* and *de* with neuter nouns among even older children between 10;5 and 12;11 years (cf. Cornips et al., 2006). In theory, this type of variation may lead to the loss of neuter gender in Dutch, and hence to the loss of grammatical gender altogether and thus to language change. The fact that grammatical gender in Dutch, being a late acquisition in monolingual Dutch children, is a vulnerable distinction is evidenced by the fact that it has disappeared completely in Dutch lexifier contact languages such as Negerhollands (Muysken, 2001:165), Berbice Dutch, Afrikaans (Donaldson, 1993; Ponelis, 2005), Curaçaos-Dutch (Joubert, 2005), Surinamese-Dutch (Cornips, 2005), and Indisch-Dutch (De Vries, 2005).

The neuter gender in Dutch is clearly a linguistic phenomenon that is interesting from the point of view of language acquisition or transmission influencing language variation, confirming the words of Kroch: ''Language change is by definition a failure in the transmission across time. Failures of transmission seem to occur in the course of language acquisition'' (2001:700).

3. Transmission, lexical variation and geographical distribution

In the COAVA project, we take the age of acquisition of a noun as coinciding with the child's first production of it, as indicated in the Childes databases.

In research on language acquisition, it is assumed that the first few years of life are the crucial time for children to acquire a language. It is also claimed that there is a relation between the age of acquisition of a noun and its level of entrenchment (Deane, 1992:194–195). In particular, the earlier the age at which an item (for example a noun) or structure has been acquired by the child, the deeper this item/structure will be entrenched. Entrenchment pertains to how frequently an item has been invoked and thus to the thoroughness of its mastery and the (relative) ease with which it is activated or retrieved (Langacker, 1991:45). Van Berkum (1996) examined the relative distribution of *de*- and *het*-words in computerized databases and found that in running texts, the estimate is roughly 2:1, respectively. The fact that the determiner *de* is much more frequent in the input than the determiner *het* may explain why the determiner *de* is overused by the children until a relative late age.

In the acquisition process, nouns are not learned indiscriminately. Some nouns are learned before others. Object names constitute a relatively large proportion of children's early vocabularies. They can be subdivided into basic level objects such as *dog* or *tree* and superordinates or subordinates like *animal* or *mammal* and *terrier* or *retriever*, respectively (Bloom, 2001). According to Waxman and Leddon: ''Within the first year of life, infants will begin to establish systematic links between words and the concepts to which they refer. On the conceptual side, they will begin to form categories of objects that capture both the similarities and differences among the objects they encounter. Most of these early object categories will be at the basic level (i.e., dog) and the more inclusive global level (i.e., animal). Infants will begin to use these early object categories as an inductive base to support inferences about new objects that they encounter'' (2011:181). But clearly, across languages, infants' earliest lexicons tend to show a ''noun advantage, with nouns referring to basic level object categories (e.g., cup, dog) being the predominant form'' (Waxman and Leddon, 2011:181).

In cognitive linguistics, concepts can be defined as 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. As lexical semantics distinguishes word form and word meaning, we will regard the noun as the form, and the concept as the meaning. Since children connect names to objects, basic level object vocabulary constitutes

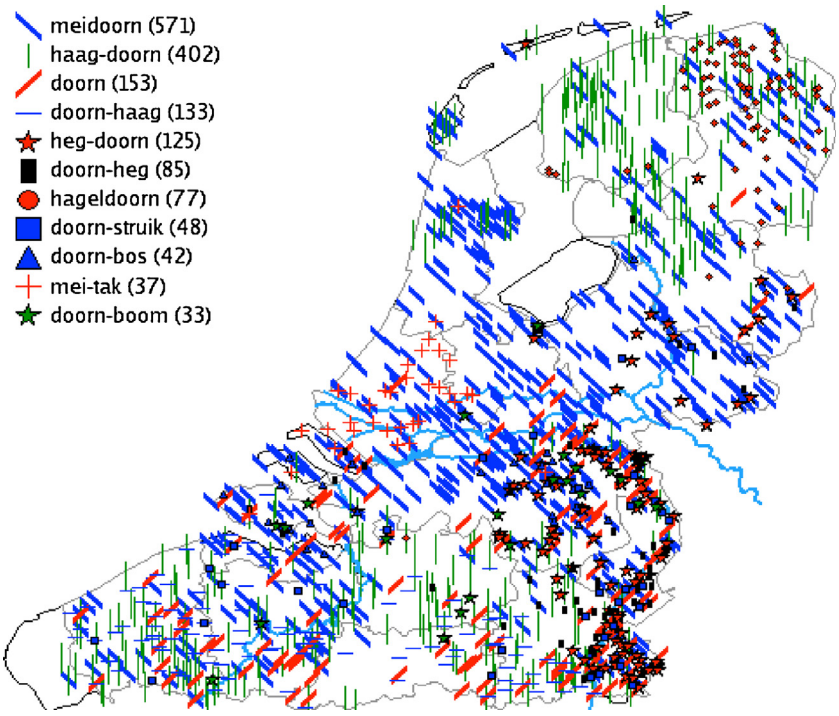
Table 3
Variation of basic concepts in Swadesh list (1971).

German	Dutch	English	Icelandic	Swedish	Latin
Sonne	Zon	Sun	Sól	Sol	Sol
Wein	Wijn	Wine	Vín	Vin	Vinum
Nase	Neus	Nose	Nef	Näsa	Nasus
Fisch	Vis	Fish	Fiskur	Fisk	Piscus
Vater	Vader	Father	Faðir	Fader	Pater
Rose	Roos	Rose	Rós	Ros	Rosa
Rot	Rood	Red	Rauður	Röd	Ruber

an excellent starting point for investigating the relation between language acquisition and language variation and change. The names for these objects, i.e. nouns, constitute a relatively large proportion of children's early vocabularies. An explanation of the lexical stability of basic level vocabulary (see below and Table 3) might be that the concepts this vocabulary denominates, concern basic objects that are deeper entrenched in human cognition than other objects (Geeraerts et al., 1994:138–142). The subcategories of basic level objects are supposed to be less salient, less entrenched, less frequent and as a result children's vocabularies show a high degree of lexical variation where hyponyms of basic level objects are concerned. What is important for our case study later (see Section 4) is that because of their conceptual salience, basic level objects are referred to by simplexes like nouns as 'fish', 'sun', e.g., simple words without affixes that are geographically widely spread (Rosch, 1978; see Table 3 below).

One of the aims of the COAVA project was to develop tools enabling us to search nouns in the Dutch monolingual acquisition child data files in CHILDES (cf. MacWhinney, 2000) on age of first production and frequency of use. This would make it possible to compare children's vocabularies – with nouns specified for age of production and frequency – with the vocabularies in the dataset of Dialect Dictionaries.

Let us now turn from child acquisition of nouns to the lexical variation and change component of the COAVA-project, and more particularly, to the geographical spread of dialect nouns. Before we do so, however, it is important to note that the Dutch language area shows a bewildering amount of geographical variation at the level of lexicon, phonology and morphosyntax. Map 2 illustrates lexical variation throughout the Dutch area for the lemma *meidoorn* 'hawthorn', which



Map 2. The geographical distribution of the noun *meidoorn* 'hawthorn'.

shows as many as eleven different realizations in the Dutch speaking parts of the Netherlands and Belgium. And these eleven realizations are merely the most frequent ones[3]:

In historical dialectology, much attention is paid to detecting the largest differentiations between dialects in space and over time. The extent of the variability of nouns in a relatively small geographical area can range considerably. Previous research on the basis of the Dialect Dictionaries of the Brabant and Limburg dialect areas (in Belgium and the south of the Netherlands) has revealed that language varieties spoken in these areas exhibit an overwhelming amount of variation for most parts of the vocabulary. Many concepts are expressed in dozens or even hundreds of different nouns in relatively small geographical spaces. These nouns are often complex, e.g. periphrastic or metaphoric compounds and collocations. Thus there are as many as 68 different words for the daddy-long-legs spider in the Brabant dialect area alone (Swanenberg, 2010), including *hooispin* 'hay spider', *hooiwagen* 'hay wagon', *hooipaard* 'hay horse', *wegwijzer* 'way-pointer', *horlogewerker* 'watchmaker', *mieke langbeen* 'Mary longlegs', *scheper langpoot* 'shepherd longlegs', etc. Lexicographers have found other examples of an overwhelming number of different nouns and wordings for specific concepts such as 'blue titmouse', 'thunder-shower' or 'pointy chin-beard'. The occurrence of some of these nouns may even be restricted to one village or city only (Swanenberg, 2004, 2010).

While the variation in words for subordinates can be staggering, the Dialect Dictionaries show remarkably little lexical variation in the vocabulary for basic level objects. In fact, the nouns for basic level objects tend to be cognates of the corresponding nouns in six Germanic languages and dialects or even other Indo-European languages (Swadesh, 1971:283). Basic level vocabulary consists mainly of simplexes (Berlin, 1992:26–31), free nouns that are etymologically opaque and have a long history. Thus, there is hardly any or no lexical variation at all with respect to more generic concepts like 'fish', 'sun' or 'nose' (onomasiological homogeneity).

The Swadesh list is a classic compilation of basic concepts developed for the benefit of historical comparative linguistic research. This type of vocabulary consists of old nouns with obscure etymologies, with related cognates spread over relatively large areas. Usually basic vocabulary is not complex; most of the words in Swadesh' list are nouns and simplexes.

In Table 3 we have included a set of seven words in five Germanic languages and Latin. All of these words are basic concepts likely to be acquired by children at an early age: the sun, body parts, food, colours, etc. Very little variation is found in denoting these concepts. Most of the words are cognates (i.e., the nouns are etymologically related).

In the next section, we will present measures to express the lexical variation of a concept (as in Map 2) in the dialect landscape.

4. Measures of variation

In order to test our hypothesis that 'early-acquired/produced' nouns will show less variation than nouns acquired/produced at a later age, we need to be able to compare lexical variation, in other words we need a measure to express the lexical variation of a concept in the dialect landscape. In the COAVA project we defined such a measure. The measure is used to add structure and comparability to the lexical variation in the datasets. If lexical variation is translated, for instance, to the type-token-ratio (the relative degree of lexical variation) and the geographical distribution of nouns, it is possible to measure lexical variation accurately. This will be demonstrated in the case study (see Section 5 below). Three measures are considered here: diversity, heterogeneity and entropy.

4.1. Diversity

Diversity is measured as the number of reports of nouns (dialect varieties) per concept, in other words, the cue validity (Geeraerts et al., 1994:156). Cue validity is the ratio between the frequency with which a cue is associated with a category, and the total frequency of the cue in the material. The cue validity of a concept (the referent) with regard to a noun (the name) is the ratio between the frequency of the noun and the frequency of the concept. Thus, cue validity may in this lexicological application be regarded as a specific instance of type-token-ratio, TTR (Watkins and Kelly, 1995). The type here is the concept, and the tokens are the nouns used in certain dialect varieties. TTR has been described as a measure of vocabulary flexibility (Johnson, 1944, as cited in Hess et al., 1986), lexical diversity (Miller, 1981, 1991, as cited in Hess et al., 1986) and vocabulary diversity (Retherford, 2000).

We illustrate this by an example. In Fig. 2a (see below) two matrices are shown, each representing a dialect landscape, where the cells of the matrices represent 16 dialect varieties. In the left matrix each cell has a different letter, meaning that for a particular concept each dialect has a different noun. In the right matrix the same noun was reported for all dialects.

[3] Map taken from: http://www.meertens.knaw.nl/pland/woordenboekartikel.php?term=Meidoorn.

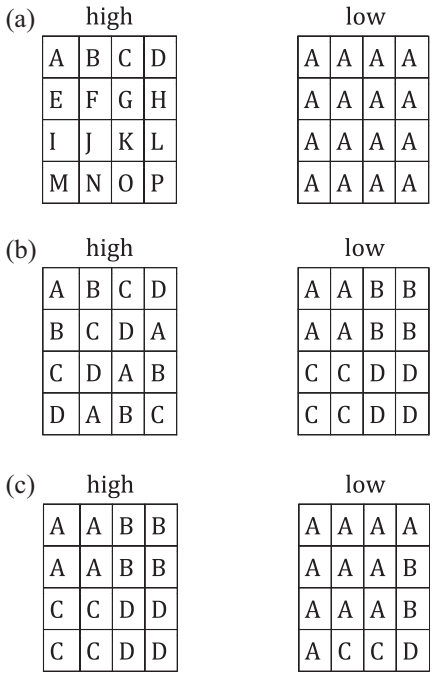


Fig. 2. (a) Measures of lexical variation: ***diversity*** (TTR). (b) Measures of lexical variation: ***heterogeneity*** (SID). (c) Measures of lexical variation: ***entropy*** (ETP).

The TTR for the left matrix is 16 types divided by 16 tokens is 1. The TTR for the right matrix is 1 type divided by 16 tokens is 0.0625. High TTR values indicate strong diversity which is maximal when the number of types and tokens is the same.

In order to subdue the differences in numbers of informants in the lexicographical dialect-research (the dictionaries are largely based on inquiry forms) we use Guiraud scores instead of simple TTR (Guiraud, 1960). The Guiraud score *G* is the number of types divided by the square root of the number of tokens. In the example given above, this would yield a score of 4 (high diversity) and 0.25 (low diversity), respectively.

4.2. Heterogeneity

When looking at the dialect landscape, we can count the number of different lexemes for a particular concept, which we measured using the Guiraud score *G*. However, if we take a closer look we may find that the different lexemes are geographically mixed, or conversely, that the different lexemes define nicely coherent areas. In order to capture this kind of variation, we need a second measure indicating the level of geographical heterogeneity (Geeraerts and Speelman, 2007). Heterogeneity can be accurately measured using a Silhouette index (Rousseeuw, 1987).

Fig. 2b on the left shows four lexemes completely mixed in the dialect landscape. In the dialect landscape on the right, lexemes A, B, C, and D define four coherent dialect groups. The geographic distances between dialect locations with the same lexeme will be small (small intra-cluster distance), and the average distance between dialect location pairs with different lexemes will be large (large inter-cluster distance). The silhouette index is a combined measure of the intra-cluster distance and the inter-cluster distance. For each dialect location *i* where lexeme *j* is used, the silhouette distance s_i^j is equal to:

$$s_i^j = \frac{b_i^j - a_i^j}{\max\left\{a_i^j, b_i^j\right\}}$$

a_i^j is the average geographic distance to the other dialect locations with lexeme *j*. b_i^j is calculated as follows. For each group – except for the group which contains group *i* – the average geographical distance between the dialect locations in the group and dialect location *i* is calculated. Now b_i^j is the group with the smallest group mean compared to dialect location *i*. Therefore, b_i^j is the average geographical distance between dialect location *i* and the geographically closest group where lexeme *j* is not used.

s_i^j ranges from −1 to +1. Values close to 1 indicate that dialect location *i* is surrounded mainly by locations with the same lexeme. If s_i^j equals 0, dialect locations neighbouring dialect location *i* partly use the same lexeme, and partly use a different lexeme. Values close to −1 indicate that dialect location *i* is mainly surrounded by dialect locations with different lexemes.

The silhouette distance for a group *j* is the average silhouette distance of the dialect locations *i* using lexeme *j*:

$$S_j = \frac{1}{m_j}\sum_{i=1}^{m_j} s_i^j$$

Finally, given *K* groups, the global Silhouette index is equal to the average of the silhouette distances of the *K* groups:

$$S = \frac{1}{K}\sum_{j=1}^{K} S_j$$

In our example $K = 4$. Since s_i^j varies between -1 and 1, S_j and S also vary between -1 and 1. We normalize S between 0 and 1 in three steps:

We calculate:	for $S = -1$ this gives:	for $S = 1$ this gives:
$S - 1$	−2	0
$(S - 1)/2$	−1	0
$((S - 1)/2)^* - 1 = SID$	1	0

If *SID* equals 0, the lexical variation around dialect locations on average is minimal. If *SID* equals 1, the lexical variation around dialect locations is maximal on average, i.e. all dialect locations have neighbouring dialects with different lexemes.

4.3. Entropy

The idea behind entropy as a measure of lexical variation is shown in Fig. 2c.

Imagine a dialect landscape where lexemes A, B, C, and D are used for a certain concept. In the landscape on the left, the four lexemes define four equal proportions. Each lexeme has the same relative frequency. Here, entropy is high. It is hard to predict for an arbitrary dialect location which lexeme will be used in the dialect of that location. However, in the landscape on the right, lexeme A is dominant: it occurs in 11 dialect locations. It is very likely that dialect speakers in a location arbitrarily selected from the landscape will use lexeme A.

Entropy is a measure of unpredictability of information content. Entropy *H* is measured as:

$$H = -\sum_{i=1}^{n} p_i \cdot {}^2\log(p_i)$$

In the formula p_i stands for the probability of lexeme *i* for a given concept in a dialect land scape, approximated by taking the relative frequency of lexeme *i* for the given concept in the dialect landscape. The entropy value of the landscape on the left equals 2000: the entropy value of the landscape on the right equals 1372. We normalize entropy values between 0 and 1. Given *n* different lexemes, the maximum entropy value is ${}^2\log(n)$. Therefore, the normalized Shannon entropy (1948) is:

$$\mathrm{ETP} = H/{}^2\log(N)$$

ETP varies between 0 and 1. For the landscapes in Fig. 2c we obtain normalized entropy values of 1 (high entropy) and 0.686 (low entropy) respectively.

Lexical variation is defined in terms of these three measures. As we have seen, the size of variability of nouns in a relatively small geographical area can be very different, depending, among other things, on the conceptual level of an object i.e., on whether we are dealing with a basic, superordinate or subordinate concept. This makes it worthwhile to examine if the relative moment of a child's first production of a noun signals its embedding or entrenchment and, hence, its frequency and whether there is a relation between this moment of first production and the lexical variation and spread throughout geographical space of the noun in question. This relation will be addressed in the modest case study below, which aims to illustrate how the tools developed in the COAVA project can be employed. The hypothesis tested in this case study is the following: nouns produced at a young age are more frequently used than later produced nouns, and, hence, are more deeply entrenched and this will be reflected in lexical stability throughout a geographical area. Further, lexically more complex nouns, as determined by the number of syllables, are acquired later than lexical less complex nouns.

Let us now test our assumption by the tools developed within the COAVA project by our case study.

Table 4
Pearson's correlation coefficients between the day that children produce a noun the first time on average, four measures of lexical variation, the number of syllables and CELEX word frequencies in Brabant dialects. Correlation coefficients given in bold are significant at the 0.05 level.

Child's first day	TTR	Giraud score *G*	SID	ETP	Numb. syll.	CELEX freq.	Log. CELEX freq.
Child's first day	0.143	0.101	0.112	−0.265	0.241	−0.144	−0.067
TTR		**0.971**	**0.496**	**0.376**	0.167	−0.023	−0.261
Giraud score *G*			**0.554**	**0.487**	0.141	0.026	−0.178
SID				0.107	0.107	0.107	0.224
ETP					0.172	0.126	−0.075
Numb. syll.						0.005	**−0.282**

5. Age of acquisition versus lexical dialect variation: a modest case study

In this section we test our hypothesis that nouns which are produced early i.e. produced before other ones by children show hardly any dialect variation, whereas nouns which are produced later show more variation across dialect localities in the dialect landscape. We will illustrate how the COAVA tool works and may help researchers in data processing and analysis. We will first consider whether the measures of lexical dialect variation i.e. TTR, Giraud score *G*, SID and ETP correlate significantly with the age of the child when it first produces the noun. Second, we will examine whether lexical complexity that in general distinguishes basic level objects from subordinates correlates significantly with the child's first day of production of the noun (early or late). Third, we consider whether noun-usage frequencies determine both the age at which a noun is acquired by children and lexical variation. Finally we perform a multiple linear regression analysis in order to find significant predictors, when we consider the geographic measures, the number of syllables and word-usage.

In Section 1 we introduced the data sources used, namely the CHILDES data set and raw lexical data of the dialect dictionaries. In this small case study we use 51 concepts which are found both in the CHILDES data set and in the dialect dictionary. We took into account that the 51 concepts need to be 'culturally neutral', because when one compares acquisition data for the late 20th century children with dialect data that largely reflect an older, predominantly rural life style, the concepts ideally should not belong to cultural or economic domains that severely changed through time (the salience of certain agricultural concepts in the original dialect environment would be different than it is for present-day children). As to the dialect data, we restrict the analysis to 179 localities in the dialect data from one resource, namely WBD.[4] For each of the locations lexemes are found for each concept; note that the measures experience difficulties if there are many empty cells in the matrices (De Vriend et al., 2007: 'raw data reduction').

Using the set of 51 nouns we calculated the Pearson's correlation coefficients between the day that children produce a noun the first time in the Childes file on average and four measures of lexical variation in Brabant dialects. The results are shown in Table 4. None of the measures of lexical dialect variation correlates significantly with the 'Child's first day', that is age of first production (see Chart 1 in Section 1.1). The lexical dialect variation measures significantly correlate to each other, except for SID and ETP.

Let us turn now to the issue of lexical complexity, which has been operationalized by the number of syllables, i.e. one-syllable, two-syllable and three-syllable nouns. The motivation to examine lexical complexity is that it distinguishes basic level objects from subordinates. Whereas basic level vocabulary (Swadesh, 1971, see above) consists of simplexes i.e. most often one-syllable words with a wide geographic range (lexical stability), secondary vocabulary consists of more complex (multiple syllables) and geographically confined terminology. For each of the three groups a fairly large number of nouns is available in the data set, 22 one-syllable words, 19 two-syllable words and eight three-syllable words. The results are shown in the column headed by 'Numb. Syll' of Table 4. The number of syllables does not significantly correlate with 'Child's first day' nor with any of the lexical measures.

One may consider whether frequently used words are acquired earlier by children than less frequently used words, and whether word frequency determines lexical variation. We used word frequencies as found in the CELEX corpus.[5] The results are shown in the eighth column of Table 4. No significant correlation is found with 'Child's first day', the lexical measures and the number of syllables. Frequencies range from 5 to 4475, except for *gezicht* 'face' which has a frequency of 18,973, and has a proportionally strong influence in the analysis when using raw frequencies. There we experimented also with logarithmic CELEX frequencies. The results are shown in the ninth column. They significantly correlate with the number of syllables, which means that words with low complexity are more frequent than nouns which are more complex.

[4] We choose only one of the two dictionaries, building upon earlier studies (De Vriend et al., 2007).

[5] Lexical database, which comprises general lexicons for British English, German and Dutch (http://wwwlands2.let.kun.nl/members/software/celex.html).

Table 5
Results of a multiple linear regression analysis in order to find the predictors of time of the child's first production.

	t Value	Sig.
Giraud score *G*	1.752	
SID	−0.329	
ETP	−2.917	$p < 0.01$
Numb. syllables	2.072	$p < 0.05$
Log. CELEX frequency	0.139	

Finally we performed a multiple linear regression analysis with Giraud score *G*, SID, ETP and logarithmic CELEX frequency as predictors of Child's first day. By using raw Child's first day values, the data would not meet the assumptions made by linear regression models. This was solved by a logarithmic transformation of this variable. Given the high correlation between TTR and Giraud score *G* (see Table 4) and the superiority of Giraud score *G* (see Section 4) we do not include TTR in the model.

The results are shown in Table 5. Both entropy and the number of syllables are found to be significant predictors of (logarithmic) time of the child's first production. Although we did not find either entropy or the number of syllables significantly correlate with Child's first day, they become significant when combined in a multiple linear regression analysis.

We conclude that the COAVA tool is a valuable tool to test hypotheses related to datasets from distinct disciplines which have been difficult or even impossible to address using traditional linguistic research methods. Our point of departure is the assumption that nouns produced at an early age will hardly show any geographical variation. Using this tool we found that entropy is a predictor of age of acquisition: the lower the entropy, the higher the age of acquisition (or the other way around). This means: the later a noun is learned by a child, the better it can be predicted which noun is used in a particular local dialect in the dialect landscape, i.e. the more likely it is that there is one dominant form in the dialect landscape (see Fig. 2c). We also considered lexical complexity, operationalized by counting the number of syllables of words and found this to be a predictor of age of word acquisition as well.

Our findings do not tell us anything about causal relationships, or whether word frequency is found to be a hidden variable that determines age of acquisition, lexical variation and complexity. We do indeed find a correlation between complexity and logarithmic word frequency, but word frequency does not correlate with age of acquisition and any of the lexical measures.

6. Conclusion

In this paper, we demonstrated how hypotheses can be tested by use of the COAVA tool which was developed within a CLARIN-NL project. The CLARIN infrastructure enables research combining datasets from distinct disciplines. In particular, the CLARIN-NL COAVA-project allows researchers to access two datasets simultaneously, namely the Dutch child language acquisition files located in CHILDES, and databases of major dialect dictionaries. Furthermore, tools for visualization of search results and a measure for lexical variation were developed. The latter takes a variety of different factors into account: Diversity, Lexical complexity, Heterogeneity and Entropy.

Our point of departure was that the earlier in life nouns are (first) produced, the less geographical variation they will show. The developed a tool enabled researchers to define (first) time and frequency of production of nouns in child language. We reported on a small case study, testing our assumption, in which 51 different nouns in the CHILDES datasets were selected along with their lexical variants in the dialect database. For these 51 nouns, entropy and the complexity of words (i.e. the number of syllables) turned out to be significant predictors of the age of first production of the noun. One might wonder whether word frequency determines Child's first day, lexical variation and complexity. We did indeed find a significant and inverse correlation between logarithmic word frequency and lexical complexity, but (logarithmic) noun frequency does not correlate with either Child's first day and any of the lexical variation measures. Our data does not suggest that noun frequency is a hidden variable determining both Child's first day and lexical variation. Therefore, more research is required to find out whether our findings establish causality between Child's first day and lexical variation, and if so, to find the direction in which this causality goes.

This paper has demonstrated how the availability of big data sets from different sub-disciplines of linguistics makes it possible to investigate research questions that have been difficult or even impossible to address using traditional linguistic research methods.

Acknowledgments

The project reported on in this paper was funded by CLARIN-NL (www.clarin.nl). CLARIN is committed to establish an integrated research infrastructure of language resources and its technology. The project is carried out by a research team of scholars from the Meertens Institute (Royal Netherlands Academy of Sciences), Maastricht University, Tilburg University, and Groningen University. The time invested to write this paper by co-author Cornips was supported by a Fellowship Grant from The Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS).

References

Berlin, B., 1992. Ethnobiological classification. In: Principles of Categorization of Plants and Animals in Traditional Societies. Princeton UP, Princeton.

Blom, E., Polisenská, D., Weerman, F., 2008. Articles, adjectives and age of onset: the acquisition of Dutch grammatical gender. Second Lang. Res. 24 (3), 297–332.

Bloom, P., 2001. How Children Learn the Meanings of Words. MIT, Massachusetts.

Cornips, L., 2005. Het Surinaams-Nederlands in Nederland. In: Van der Sijs, N. (Ed.), Wereldnederlands: Oude en jonge variëteiten van het Nederlands. SDU, The Hague, pp. 131–147.

Cornips, L., Hulk, A., 2008. Factors of success and failure in the acquisition of grammatical gender in Dutch. Second Lang. Res. 24 (3), 267–296.

Cornips, L., Van der Hoek, M., Verwer, R., 2006. The acquisition of grammatical gender in bilingual child acquisition of Dutch (by older Moroccan and Turkish children). The definite determiner, attributive adjective and relative pronoun. In: Van de Weijer, J., Los, B. (Eds.), Linguistics in The Netherlands 2006. John Benjamins, Amsterdam/Philadelphia, pp. 40–51.

De Vriend, F., Boves, L., Van den Heuvel, H., Van Hout, R., Kruijsen, J., Swanenberg, J., 2006. A unified structure for Dutch Dialect Dictionary Data. In: Proceedings of the Fifth international conference on Language Resources and Evaluation (LREC), .

De Vriend, F., Swanenberg, J., 2006. D-kwadraat: digitale databanken en digitaal gereedschap voor WBD en WLD. Nederlandse Taalkunde 11 (4), 366–372.

De Vriend, F., Swanenberg, J., Van Hout, R., 2007. Dialectgebieden in Brabant. Geografische clustering op basis van de ruwe lexicale gegevens van het Woordenboek van de Brabantse Dialecten. Taal en Tongval. Themanummer 20, 83–110.

De Vries, J.W., 2005. Indisch-Nederlands. In: Van der Sijs, N. (Ed.), Wereldnederlands: Oude en jonge variëteiten van het Nederlands. SDU, The Hague, pp. 59–77.

Deane, P.D., 1992. Grammar in Mind and Brain: Explorations in Cognitive Syntax. Mouton de Gruyter, Berlin/New York.

Donaldson, B., 1993. A Grammar of Afrikaans. Mouton de Gruyter, Berlin.

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

Geeraerts, D., Grondelaers, S., Bakema, P., 1994. The Structure of Lexical Variation. Meaning, Naming, and Context. Mouton de Gruyter, Berlin/New York.

Geeraerts, D., Speelman, D., 2007. De structuur van lexicale onzekerheid. Taal en Tongval. Themanummer 20, 47–61.

Guiraud, P., 1960. Problèmes et méthodes de la statistique linguistique. Presses Universitaires de France, Paris.

Hess, C.W., Landry, R.G., Sefton, K.M., 1986. Sample size and type-token ratios for oral language of preschool children. J. Speech Hear. Res. 29, 129–134.

Joubert, S.M., 2005. Curaçaos-Nederlands. In: Van der Sijs, N. (Ed.), Wereldnederlands: Oude en jonge variëteiten van het Nederlands. SDU, The Hague, pp. 31–58.

Kroch, A., 2001. Syntactic change. In: Baltin, M., Collins, C. (Eds.), The Handbook of Contemporary Syntactic Theory. Basil Blackwell, Malden, MA, pp. 699–730.

Labov, W., 2007. Transmission and diffusion. Language 83, 344–387.

Langacker, R.W., 1991. Concept, Image, and Symbol. The Cognitive Basis of Grammar. Mouton de Gruyter, Berlin/New York.

MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk, 3rd ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Meisel, J., 2011. Bilingual language acquisition and theories of diachronic change: Bilingualism as cause and effect of grammatical change. Biling.: Lang. Cognit. 14 (2), 121–145.

Muysken, P., 2001. The origin of creole languages. The perspective of second language learning. In: Smith, N., Veenstra, T. (Eds.), Creolization and Contact. John Benjamins, Amsterdam, pp. 157–173.

Ponelis, F., 2005. Nederlands in Afrika: Het Afrikaans. In: Van der Sijs, N. (Ed.), Wereldnederlands: Oude en jonge variëteiten van het Nederlands. SDU, The Hague, pp. 15–30.

Retherford, K., 2000. Guide to Analysis of Language Transcripts, 3rd ed. Thinking Publications, Eau Claire, WI.

Rosch, E., 1978. Principles of categorization. In: Rosch, E., Lloyd, B.B. (Eds.), Cognition and Categorization. Lawrence Erlbaum Associates, Hillsdale, pp. 27–48.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. 20 (1), 53–65.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Swadesh, M., 1971. The Origin and Diversification of Language. Edited post mortem by J. Sherzer. Aldine Atherton, Chicago.

Swanenberg, J., 2004. Origins of lexical variation. In: Gunnarsson, B.L., et, al. (Eds.), Language Variation in Europe. Papers from ICLaVE 2. UU, Uppsala, pp. 378–390.

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

Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A., Tsimpli, I., 2011. Bilingual acquisition of Greek voice morphology and Dutch gender: what do they have in common? In: Danis, N., Mesh, K., Sung, H. (Eds.), Proceedings of the 35th annual Boston University Conference on Language Development (BUCLD 35). Cascadilla Press, Somerville, MA, pp. 590–602.

Van Berkum, J.J.A., 1996. The Psycholinguistics of Grammatical Gender: Studies in Language Comprehension and Production Ph. D. dissertation. Max Planck Institute for Psycholinguistics.

Watkins, R.V., Kelly, D.J., 1995. Measuring children's lexical diversity: differentiating typical and impaired language learners. J. Speech Hear. Res. 38, 1349–1355.

Waxman, S.R., Leddon, E.M., 2011. Early word learning and conceptual development: everything had a name, and each name gave birth to new thought. In: Goswami, U. (Ed.), The Wiley-Blackwell Handbook of Childhood Cognitive Development. Wiley-Blackwell, Malden, MA, pp. 180–208.

WBD, 1967–2005 Woordenboek van de Brabantse Dialecten. Assen/Groningen/Utrecht, Van Gorcum/Gopher.

WLD, 1983–2008 Woordenboek van de Limburgse Dialecten. Assen/Groningen/Utrecht, Van Gorcum/Gopher.