Part 2: **Language variation from a nonlinguistic perspective**

A Variation in speakers' cognitive capability

Karlien Franco and Dirk Geeraerts

1 Botany meets lexicology: The relationship between experiential salience and lexical diversity

Abstract: In this paper, we explore the relationship between the experiential salience of natural concepts (i.e. the degree to which the concept are well-known to language users, because they occur frequently in the everyday environment of the speakers) and the structure of the lexicon. More specifically, we focus on whether the amount of lexical dialect variation found in names for naturally occurring plants in ecologically consistent geographical regions in the northern part of Belgium, is influenced by the frequency of these plants in these regions. In contrast with previous research in linguistics, which has mostly focused on the relationship between the (textual) frequency of constructions in language use and language variation and change, we confront non-linguistic referential data with linguistic dialect data.

In practice, we use the distribution of naturally occurring plants in the language area, as described in the standard reference work on plant distribution in the northern part of Belgium (Van Landuyt et al. 2006), to determine whether more frequent plants show less variation. The linguistic data come from the digitized databases of the Dictionaries of the Brabantic, Limburgish and Flemish dialects (WBD; WLD; WVD). We consider several measures that quantify the amount of lexical variation per plant. First, we take into account the number of unique lexemes per plant per (ecologically consistent) region. Second, we use the type-token ratio per plant per region, with the number of types equal to the number of unique lexemes per plant and the number of tokens calculated as the total number of records per plant. Third, we use the measure of internal uniformity per plant per region, which quantifies the degree of lexical standardization in the names for each plant.

The results for the three measures per plant per region diverge. Overall, they show that, although plant frequency alone does not cause complete lexical standardization in a particular region, more frequent plants do show a smaller amount of lexical variation.

Karlien Franco, KU Leuven, Belgium, karlien.franco@kuleuven.be
Dirk Geeraerts, KU Leuven, Belgium, dirk.geeraerts@kuleuven.be

https://doi.org/10.1515/9783110610895-005

AU: Address has been removed in Affiliation detail "Blijde-Inkomststraat 21 PO Box 3308, B-3000 Leuven", please confirm.

1 Background

This paper focuses on variation in the names given to plants that occur naturally in the northern part of Belgium where Brabantic, Limburgish and Flemish dialects of Dutch are spoken (i.e. Flanders and Brussels). While a variety of work on plant name variation exists about these dialects (see Brok 2003), most of this research focuses on a small set of plants, the names that occur for these plants in particular locations and an etymological interpretation of the name (see for instance Pauwels 1933, Brok 1991, 2006). In this paper, we take a different approach. More specifically, we use the semantic field of plant names as a case study to investigate whether lexical diversity, i.e. variation in the number of names that exist for a particular plant, correlates with the degree of salience of the plant for language users.

Two contrasting hypotheses can be envisaged concerning the influence of concept salience on lexical diversity. On the one hand, previous research concerning the influence of semantic features on lexical geographical variation across the Limburgish dialects indicates that more salient concepts show a smaller amount of lexical diversity (Geeraerts & Speelman 2010, Speelman & Geeraerts 2008). For example, the concept KNOKKELKUILTJES 'the little dents between the knuckles of the hand' shows more lexical heterogeneity than the more salient concept KEEL 'throat'. While these studies focused on concepts from the semantic field of the human body, Franco, Geeraerts & Speelman (2015) showed that the influence of salience on lexical geographical variation also holds in other semantic fields.

Further evidence for this hypothesis comes from Swanenberg (2000), who relies on notions identified in the Cognitive Linguistics research paradigm to analyze variation in the naming and classification of birds. He shows, for instance, that the degree of preponderance of a particular type of bird in comparison to related bird types has an influence on the names that are used for the bird. Types of birds that are more familiar for the language users (like VELDLEEUWERIK 'skylark', Galerida cristata) are, for instance, named more frequently with hyperonymous names that actually refer to the category as a whole (like *leeuwerik* 'lark', Alaudidae), than less salient ones.

On the other hand, more salient concepts can sometimes also show more lexical variability. The degree of familiarity of a concept has been shown to influence differences in categorization between languages or dialects (for a recent study, see Bromhead 2011). As a result, naming differences occur as well. According to Goossens (1964), for instance, the global applicability of a lexeme[1] correlates with

1 The global applicability of a lexeme ('globaliteitstoepasselijkheid', Goossens 1964:8) refers to the usage of one lexeme for a concept in a particular region, while the concept is conceptualized in a more detailed way in other areas. This contrasts with the local applicability of lexemes

the lack of familiarity of the concept. He argues, for instance, that one reason for the survival of two different names for the two handles of a scythe in the dialects spoken in the central part of Limburg in Belgium, is the high frequency of usage of the instrument in this region. As a result, language users categorize the different parts of the instrument in a more detailed way, by discerning the upper from the lower handle. In the rest of the south-eastern part of the Dutch language area, his dialect map only shows one name to refer to both the upper and lower handle. In this region, the less familiar concept SCYTHE shows less lexical diversity.

Overall, even though these two diverging hypotheses concerning the relationship between salience and lexical diversity can be distinguished, a positive correlation between salience and the amount of lexical diversity (i.e. more salient concepts show more lexical variation) seems to apply primarily to cases like the example of the scythe mentioned above, in which differences of categorization are involved. As the linguistic data that we use were collected at the level of the concept (viz. by the use of questionnaires in which the dialect name for a particular concept is elicited), categorization differences are diminished. For this reason, we expect to find a negative correlation between salience and lexical diversity (i.e. more salient concepts show less lexical variation).

To test this hypothesis, we operationalize local plant salience as the frequency of the plant in the geographical area of the language user, under the assumption that plants that naturally occur more frequently in a specific region are more familiar for the people living in that region. Additionally, the fact that some plants that are infrequent in a particular region but relatively frequent across the entire language area (i.e. locally infrequent, but globally frequent) are probably better known than plants that are infrequent everywhere, may result in a higher degree of salience for the first group of plants. For this reason, we also take into account the global frequency of the plant in the northern part of Belgium.

Interestingly, research in linguistics on the relationship between language variation and frequency has mostly focused on the (textual) frequency of constructions in language use (Divjak & Caldwell-Harris 2015: 54). Schmid (2007:119), for instance, argues that "the frequency of occurrence of concepts or constructions in a speech community has an effect on the frequency with which its members are exposed to them." Rather than relying on corpus data to determine the frequency or entrenchment of plant names or plant concepts in the speech community, in this paper, we aim to determine whether the referential frequency of the plants in everyday life affects language variation as well.

('fragmatoepasselijkheid'), which entails that in some regions conceptual and, thus, lexical differentiation occurs for a concept that is conceptualized as a whole in other regions.

Consequently, we rely on what we will refer to as the degree of *experiential* salience of a plant: the likelihood of language users encountering a particular plant in their everyday environment. To gauge experiential salience, we rely on measures of the referential frequency of the plants. However, other factors that affect experiential salience, like whether or not a plant has medicinal applications, or whether or not a plant is poisonous and to be avoided, can be envisaged as well (see Discussion).

We assume that experiential salience is related to the degree of onomasiological salience of a plant, in the sense that language users probably refer more frequently to concepts they often come into contact with. As a result, experiential salience may affect variation in the names that are given to plants: experientially more salient plants are expected to show less lexical diversity. For example, salient plants, like the common aspen (Populus tremula), which grows frequently throughout the language area under scrutiny, has fewer dialectal variants in the dictionaries that we use (viz. 40) than less frequent plants like the common cowslip, which occurs with 217 different names. The geographical distribution of these plants is shown in Figure 1 and 2. The magnitude of the dots is proportionate with the frequency of the plant in that location (i.e. in that so called hour square, see below) in the period 1972–2004. The squares reflect the distribution of the plant for the period 1939–1971 (also in hour squares).

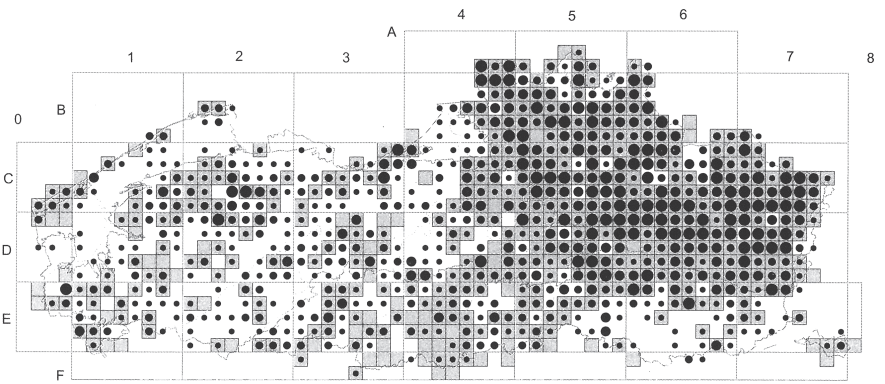


Figure 1: Geographical distribution of the common aspen (Populus tremula), a very frequent plant (Sevenant et al. 2006: 688).

This paper is structured as follows. Section 2 elaborates on the referential plant frequency data and on the linguistic data that are used in this study. In section 3, the results concerning the correlation between local and global plant

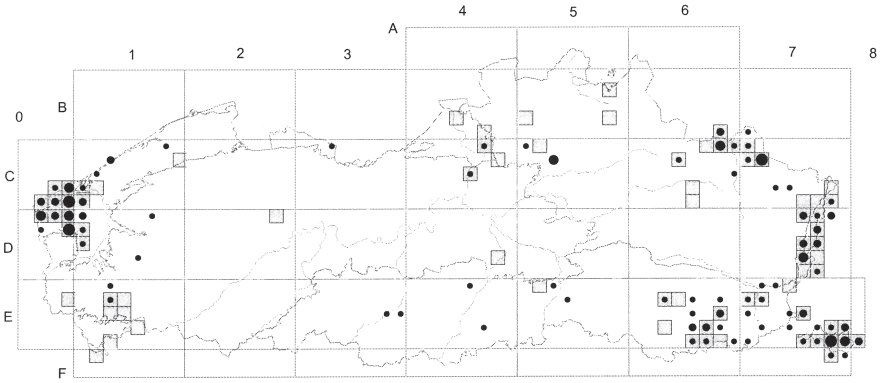


Figure 2: Geographical distribution of the common cowslip (Primula veris), a very infrequent plant (Sevenant et al. 2006: 712).

frequency and lexical diversity are provided. Section 4 provides a discussion of these results, followed by an overview of the restrictions on the present study and some suggestions for future research. Section 5 ties it all together in a conclusion.

2 Data

2.1 Referential and linguistic data

2.1.1 Referential data

As explained in section 1, we use frequency data of naturally occurring plants to gauge the familiarity of the plant in the language area under investigation. These referential data come from the *Atlas van de flora van Vlaanderen en het Brussels Gewest* (Van Landuyt et al. 2006), the standard reference work concerning the distribution of plants in the northern part of Belgium. The data are also available online (http://flora.inbo.be/).

The frequency of plants in the atlas is calculated as follows. The focus area of the atlas (i.e. the northern part of Belgium) is divided into kilometer squares of 1x1 kilometer. These kilometer squares are grouped into hour squares of 4x4 kilometers (see Figure 3). For each hour square, trained field workers investigated at least one quarter of the kilometer squares. The field workers were

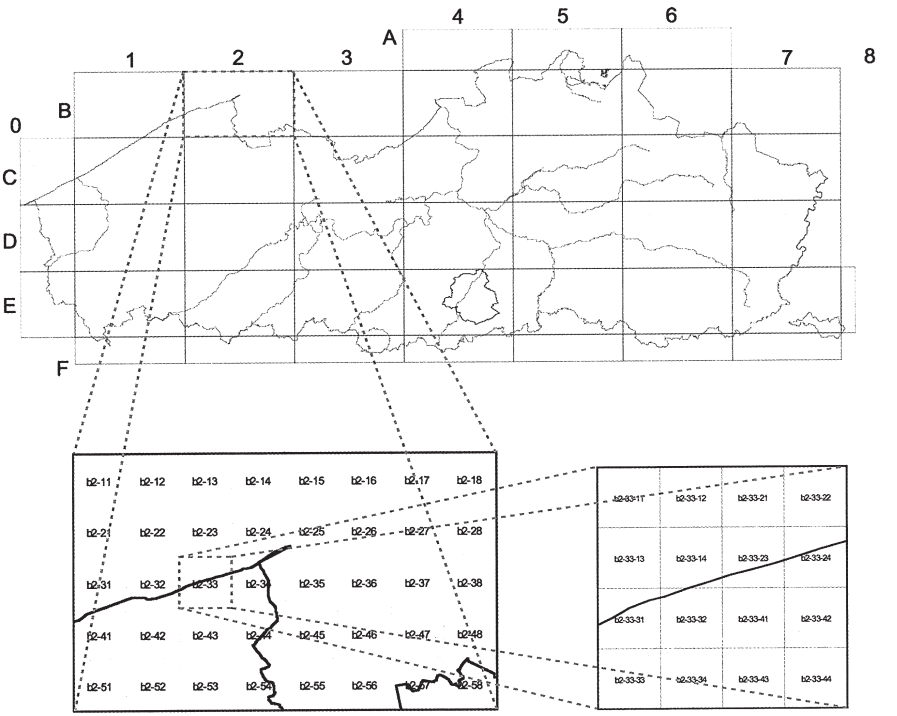


Figure 3: Hour and kilometer squares in the northern part of Belgium (Van Landuyt et al. 2006: 34).

asked to record which plants they encountered while walking through the kilometer square.[2]

We adopt two types of measures of plant frequency that are available in the atlas. On the one hand, we take into account the global frequency of a plant in the northern part of Belgium, expressed as the absolute number of hour and kilometer squares where the plant was encountered. On the other hand, we also use the relative number of investigated kilometer squares in which the plant was found per ecological region to gauge the local salience of a plant. The division of the northern part of Belgium into ecological regions is based on a simplified version of the ecologically coherent districts described in Sevenant et al. (2002). In the atlas, six ecological regions are distinguished: the

2 Some of the data in the atlas also come from secondary sources. However, for the most part, the frequency data relies on the information provided by the field workers (Van Landuyt et al. 2006:34–37).

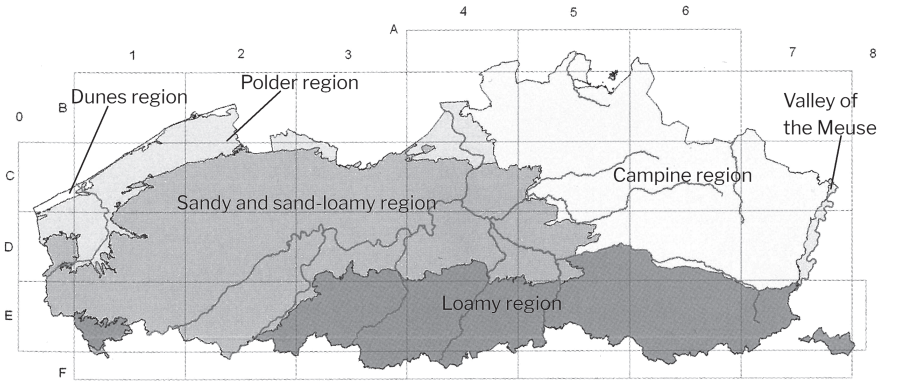


Figure 4: Ecological regions in the northern part of Belgium (Van Landuyt et al. 2006: 87).

Dunes region, the Campine region, the Loamy region, the region of the valley of the river Meuse, the Polder region and the Sandy and sand-loamy region (see Figure 4).

Because the atlas not only contains different measures of plant frequency (viz. local and global plant frequency), but also data from different periods,[3] we use four measures of plant frequency in total: one measure of local plant frequency and three measures of global plant frequency. The measure of local plant frequency is provided as a proportion in the atlas, i.e. the number of kilometer squares in which a plant was encountered divided by the total number of reliably investigated kilometer squares in a particular ecological region (Van Landuyt et al. 2006: 99). The measures of global frequency, however, are supplied as absolute values, i.e. the total number of kilometer or hour squares in which a plant was found in the northern part of Belgium.

1. local plant frequency:
   the relative number of investigated kilometer squares in which a plant was encountered per ecological region between 1972 and 2004 (*local relative frequency km squares 1972–2004*)

3 Due to historical developments, the atlas contains data from two different periods (1939–1971 and 1972–2004; see Van Landuyt et al. 2006: 9–31, 35). As the data collection process has remained the same since 1939 and as we have no obvious theoretical reasons to only rely on data from one period, we include data from both periods in the analysis.

2. global plant frequency:
   a. the absolute number of kilometer squares in which the plant was encountered throughout the northern part of Belgium between 1972 and 2004 (*global absolute frequency km squares 1972–2004*)
   b. the absolute number of hour squares in which the plant was encountered throughout the northern part of Belgium between 1939 and 1971 (*global absolute frequency hour squares 1939–1971*)
   c. the absolute number of hour squares in which the plant was encountered throughout the northern part of Belgium between 1972 and 2004 (*global absolute frequency hour squares 1972–2004*)

As the amount of kilometer squares in the northern part of Belgium is very large and as not all kilometer squares were investigated by the fieldworkers, most plants seem to be relatively infrequent when kilometer square calculations are used (although some plants are locally very frequent, see Van Landuyt et al.: 69–80). As a result, global frequency per hour square is probably a better measure of plant frequency. However, all four plant frequency measures are highly correlated in the data set ($.85 \leq r \leq .98$; $p < 0.001$).

2.1.2 Linguistic data

The linguistic data used in this study come from three related sources. We use the digitized databases of the *Flora* chapter of the Dictionaries of the Brabantic, Limburgish and Flemish dialects (WBD, WLD, WVD). These onomasiological dictionaries contain the lexemes that are used in a large number of locations throughout the three dialect areas (N = 1033 locations). We focus on the data from locations in the Belgian part of the Dutch language area and exclude data from The Netherlands, because the referential plant frequency data only contain information about plants in the northern part of Belgium.

Furthermore, we only include data form these databases that were elicited through large-scale questionnaires that were distributed systematically in every location of each dialect area (i.e. in the Brabantic, Limburgish and Flemish dialect areas), and exclude data from small-scale local dictionaries or other sources with a limited scope.[4] In practice, for the Brabantic and Limburgish data,

4 The databases, which serve as the source material for the dictionaries, also contain data from other sources, such as local dictionaries or questionnaires with a smaller geographical range. The Brabantic and Limburgish questionnaires were distributed between 1960 and 1982. The Flemish questionnaire data were elicited later: between 1998 and 2000.

we only use data that were collected on the basis of the questionnaires of the *Nijmeegse Centrale voor Dialect- en Naamkunde*. For the Flemish data, we also limit the dataset to only include data collected through the questionnaires that were sent out systematically throughout the dialect area by the lexicographers. Even though these Flemish questionnaires are not identical to the Brabantic and Limburgish ones, they are equivalent. The Flemish questionnaires include, for instance, questionnaires on plants in general (number 104, distributed in 1998), on grass (number 112, distributed in 1999) and on trees and shrubbery (number 115, distributed in 1999). The Limburgish and Brabantic data mostly come from questionnaire N 82 (1981; plants in general and trees and shrubbery), and from questionnaire N 92 (1982; names for plants and herbs). We restrict our attention to plants that occur in all three databases.

As the three dictionaries have been collaborating since 1990 (Kruijsen 1996) to achieve consistency and alignment of the databases, we believe that restricting our attention to the data that were collected through the large-scale questionnaires and that occur in all three dictionaries, ensures maximal comparability between the sources. Moreover, the analysis requires that the data were collected in a systematic way, because counting the number of different lexemes per plant concept is only feasible if the data were collected in the same locations for each concept. By only relying on the questionnaire data, we ensure that the geographical scope of the data that we use is as systematic as possible.[5]

Table 1: Number of concepts and number of records per ecological region.

ecological region	number of concepts	number of records
Dunes region	84	1887
Polder region	101	9636
Sandy and sand-loamy region	114	22755
Loamy region	132	5738
Campine region	118	692
Valley of the river Meuse	65	99

5 However, as the analysis will show, the amount of data is still relatively small for a number of plants and can differ between plants. Two explanations can be envisaged. First, it is possible that the questionnaires were not distributed as systematically as expected throughout the language area. However, another explanation may be that the small amount of data also reflects a lack of familiarity of the plant concepts (see Geeraerts & Speelman 2010 and Speelman & Geeraerts 2008): perhaps the respondents did not reply to questions of the questionnaire about plants that they were unfamiliar with, because they did not know the name for the plant in their local dialect.

Overall, the data set contains 137 different concepts. The number of concepts and the total number of records per ecological region is shown in Table 1. This table reveals large differences between ecological regions. On the one hand, this can be explained by the fact that the surface area of the ecological regions differs. The Dunes region, for example, is a rather narrow strip of land in the west of the northern part of Belgium. As a result, the number of locations in this region is relatively small. On the other hand, differences in the number of concepts and records per ecological region can also be explained by the fact that, overall, a large proportion of the data come from the WVD. This dictionary contains 30 666 records for the plant concepts under scrutiny, while the WLD and WBD combined only contain 10 203 records. As the data from the WVD mostly span the Dunes region, the Polder region and parts of the Loamy region and of the Sandy and sand-loamy region (see Figure 5), it is not surprising that the number of records is the largest in these regions.

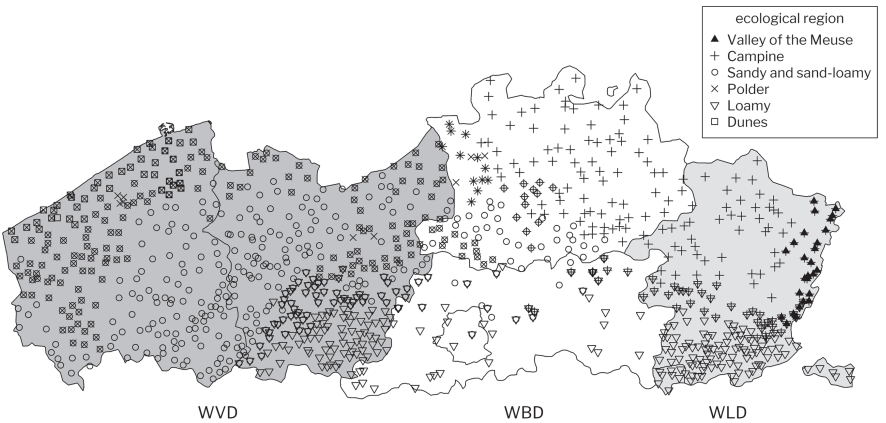


Figure 5: Dialect boundaries as represented by the WBD (white), WLD (light grey) and WVD (dark grey) and ecological regions in the northern part of Belgium.

Figure 5 further shows that the data is relatively sparse in the south of the center of the northern part of Belgium, which is covered by the WBD. It also indicates that some locations belong to more than one ecological region. This has to do with the fact that the ecological regions are defined at the level of the municipality in Sevenant et al. (2002), even though the borders of ecological regions sometimes run through a municipality. For example, the municipality of Bruges belongs to three different ecological regions: the western part of Bruges belongs to the Dunes region; the central, largest part of this municipality is part of the

Polder region; the eastern part of Bruges is included in the Sandy and sand-loamy region.

2.2 Calculating lexical diversity per concept

To operationalize the amount of lexical diversity that is found for the plant concepts in the dataset, we compare the influence of plant frequency on three measures of lexical richness. Each measure is calculated per plant per ecological region.

The first measure, *number of different lexemes*, is computed by counting the number of lexemes that occur per plant per ecological region. The number of different lexemes ranges from 1 to 92, but most concepts have a relatively low value for this variable (mean = 8.14, sd = 11.49). We include this simple calculation of lexical diversity because it is also used in other studies that were mentioned in section 1 on the relationship between concept salience and lexical heterogeneity in the WLD (Franco, Geeraerts & Speelman 2015, Geeraerts & Speelman 2010, Speelman & Geeraerts 2008).

However, a strong positive correlation between the number of different lexemes and the number of records that occur in the data set per concept exists (see Figure 6; $r = 0.91$, $p < 0.001$). As noted before, we only use data from the

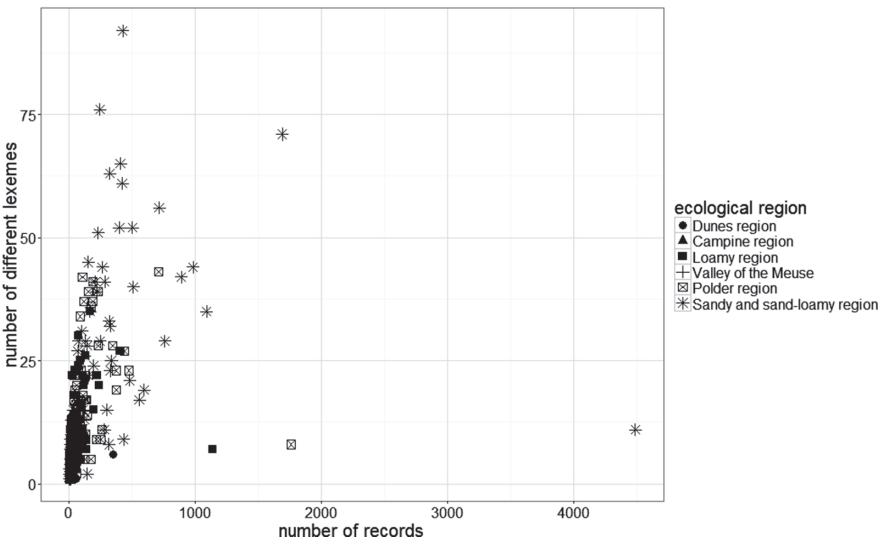


Figure 6: Correlation between number of records and number of different lexemes.

questionnaires in the dictionary to ensure that the data was collected as systematically as possible. As Figure 6 indicates, however, the number of records per concept differs strongly: the number of records ranges from 1 to 4487, with mean 66.46 and standard deviation 240. Most of the concepts with a large number of records come from the Sandy and sand-loamy region (indicated with *). The concepts on the bottom right side of the plot represent the concept OAK in three different regions. From left to right, they are based on data from the Loamy region, from the Polder region and from the Sandy and sand-loamy region.

A second measure of lexical diversity that is included in the analysis, is the type-token ratio (TTR) per plant per ecological region (see for example Tweedie & Baayen 1998). We use it to account for differences in the number of records (i.e. the number of tokens) that are available per concept, which can affect the number of different lexemes (i.e. the number of types) that are found for each concept per region. The type-token ratio approaches 0 when a small number of types is available, given the number of tokens. It is equal to 1 when the number of types is equal to the number of tokens.

TTR decreases when more tokens for the same number of types occur per concept, with values close to 1 expressing a large amount of lexical variation and figures close to 0 indicating that the concept shows a small amount of lexical diversity (left panel of Figure 7). For example, the ratio is close to 0 when for a total of 1000 tokens, only 90 different lexical items are found (.09), while it is close to 1 when the same number of unique lexical items occurs for 100 tokens

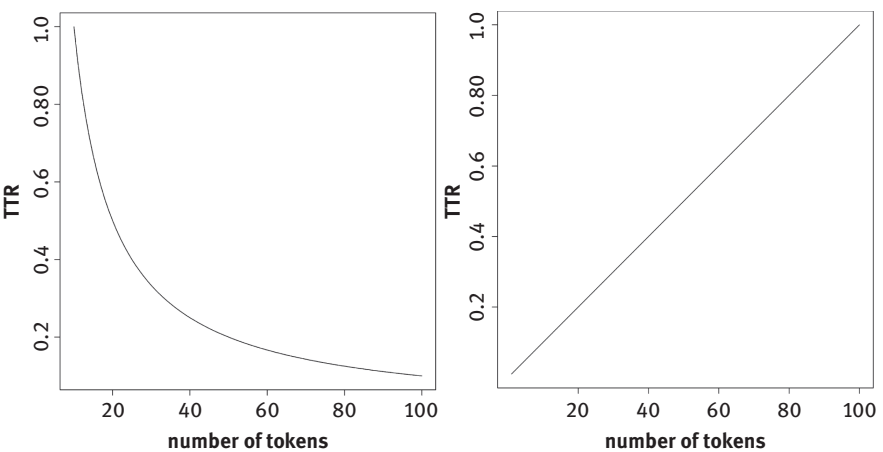


Figure 7: Type-token ratio for increasing numbers of tokens (left panel) and increasing numbers of types (right panel).

(.9). TTR is also smaller when fewer types for the same number of tokens occur per concept (right panel of Figure 7), again with low values for a small amount of lexical diversity and with values close to 1 demonstrating a large amount of lexical variation for the concept. For instance, TTR is high (.9) when 90 unique lexical items occur for a total of 100 observations (i.e. a lot of lexical diversity: almost one new lexeme for every additional observation), while it is low (.1) when 10 unique lexical items occur for the same amount of tokens (i.e. little lexical variation).

However, TTR is also sensitive to the amount of observations per concept ($r = -0.87$, $p < 0.001$), probably because the dataset contains a relatively large proportion of concepts that have the same number of types and tokens (viz. 28.2%). For all these concepts, a limited number of records is available in the data. For example, the aspen (Populus tremula) occurs only once in the data from the Campine region; the forget-me-not (Myosotis arvensis) occurs once in the data from the Meuse valley. The plant with the largest number of types and tokens and TTR = 1 is the common corn-cockle (Agrostemma githago) in the Loamy region (11 types, 11 tokens).

A third measure we use is the measure of internal uniformity, which was first used in Geeraerts, Grondelaers & Speelman (1999; also see Speelman et al. 2003) to determine the degree of uniformity in the usage of lexical variants in a speech community. Maximal uniformity (or standardization) occurs when everyone uses a single variant to describe a particular concept in the same situation. In our dataset, a complete lack of uniformity in an ecological region would occur when a different lexical item is used for every observation for the plant in that ecological region. In this paper, we calculate this measure to determine whether plants that are more frequent in a particular region also show a higher degree of lexical standardization, in the sense that one lexical variant takes precedence over its competing heteronyms. However, since the ecological regions often span more than one dialect area, other factors, like dialect boundaries, probably influence the degree of standardization as well. (Note that, in this paper, we are interested in a descriptive form of lexical standardization, whereby one lexical item becomes the preferred variant to refer to a particular concept, resulting in highly homogeneous (and, thus, standardized and uniform) language use. We are not concerned with a normative reading of the term 'standardization', which assumes that the preferred variant is prescriptively imposed upon a community of language users.)

The quantification of uniformity takes into account both the number of alternatives that occur in the data to express a certain concept (uniformity is smaller if more alternatives for the same concept exist) and the relative frequency of each of these variants (uniformity is higher if there is a clearly dominant term for a

concept). For concept Z in data set Y, internal uniformity $I_Z(y)$ is calculated as follows:

$$I_Z(Y) = \sum_{i=1}^{n} F_{Z,Y}(x_i)^2$$

In this formula, x_1 to x_i are the lexemes that are used to express concept Z in data set Y and $F_{z,y}$ is the relative frequency of lexeme x in data set Y for concept Z. The measure of internal uniformity ranges from 0 to 1, with 0 indicating a complete lack of uniformity (i.e. a lot of lexical diversity) and 1 indicating complete uniformity (i.e. a lack of lexical diversity). The correlation between this operationalization and the number of records per concept is lower, but still significant ($r = -0.665$, $p < 0.001$): concepts with more records in the dataset show a smaller amount of uniformity.

To match the linguistic and the referential data, we restrict our attention to plant concepts that are available in the questionnaires from all three dictionaries (see above). We exclude plant concepts that do not refer to actual plants, like BLOEMKNOP 'bud', or that are too general, in the sense that they do not refer to a particular type of plant, like MOS 'moss'. Then, we assign each location in the dictionary data to the ecological regions that were distinguished in the atlas. For this procedure, we rely on Sevenant et al. (2002), which contains an overview of the municipalities in Belgium per ecological region. However, we make some adaptations to the description of Sevenant et al. (2002) to obtain the simplified version of the ecological regions that is used in the atlas. In a next step, we add both the global plant frequency information and local plant frequency per plant per ecological region to the dataset, on the basis of the scientific names of the plants that are provided in the Dictionary of the Limburgish Dialects (WLD: 25–30). Finally, we calculate the number of different lexemes, the type-token ratio and internal uniformity per plant per ecological region.

For example, the dataset contains the three measures of lexical diversity (columns 5–7 in Table 2) for the wood anemone (Anemone nemorosa) in five ecological regions (viz. the Campine, Dunes, Loamy, Polder and Sandy and sandloamy region; column 3).[6] It also includes the local frequency of this plant in these five regions, expressed in percentages (column 8), and the global frequency of the plant (measured in three ways in columns 9–11, see above) in the northern part of Belgium. In contrast with the measure of local frequency (i.e. per ecological region), global plant frequency is the same in each ecological region, as it is a measure of the frequency of the plant in the northern part of Belgium. In the

6 Note that we have no information about the wood anemone in the ecological region of the valley of the Meuse, because no linguistic data is available for this plant from locations belonging to this ecological region.

Table 2: Wood anemone (Anemone nemorosa) in the final dataset.

1	2	3	4	5	6	7	8	9	10	11
plant	**scientific name**	**ecological region**	**number of records**	**number of diff. lexemes**	**TTR**	**internal uniformity**	**local rel. freq. kmsq. '72–04**	**global abs. freq. kmsq. '72–'04**	**global abs. freq. hoursq. '39–71**	**global abs. freq. hoursq. '72–04**
wood anemone	Anemone nemorosa	Campine	1	1	1	1	9.80	2031	409	507
wood anemone	Anemone nemorosa	Dunes	12	5	0.417	0.222	0.00	2031	409	507
wood anemone	Anemone nemorosa	Loamy	90	25	0.278	0.117	48.60	2031	409	507
wood anemone	Anemone nemorosa	Polder	72	23	0.319	0.118	0.40	2031	409	507
wood anemone	Anemone nemorosa	Sandy- & sandloamy	206	41	0.199	0.199	23.90	2031	409	507

analysis, we aggregate over all the regions and over all the plants (N = 614). We test whether the measures of lexical diversity (columns 5–7) correlate with the measures of plant frequency (columns 8–11).

2.3 Correlating plant frequency and lexical diversity

To test whether plant frequency has a significant influence on the diversity in the names for plants in the data set, we use Spearman's rank correlation tests. More specifically, we test whether the plant frequency measures (l*ocal relative frequency km squares 1972–2004, global absolute frequency km squares 1972–2004, global absolute frequency hour squares 1939–1971* and *global absolute frequency hour squares 1972–2004*) correlate significantly with each of the three operationalizations of lexical diversity (*number of different lexemes*, *type-token ratio* and *internal uniformity*). We also calculate the correlation coefficient. This coefficient ranges from -1 to 1, with negative values representing a negative correlation between the variables and positive values indicating a positive correlation. When the coefficient is 0, no correlation between the variables is found.

Note that the interpretation of the correlation coefficient differs for the number of different lexemes per concept and TTR on the one hand, and for internal uniformity on the other hand. A positive correlation coefficient for plant frequency and the former measures indicates that more frequent plants show *more* lexical diversity. However, a positive correlation coefficient for plant frequency and the latter measure shows that internal uniformity correlates positively with plant frequency and, thus, that more frequent plants show *less* lexical diversity.

3 Results

This section presents the results of the analysis. All the analyses were carried out with R 3.2.4 (R Core Team 2016). In 3.1, we correlate the four measures of plant frequency from the atlas with the three measures of lexical diversity, calculated per plant per ecological region. In 3.2 the relationship between global and local plant frequency is scrutinized. Although we expect to find negative correlations between number of different lexemes and TTR and the plant frequency measures, and a positive correlation between these measures and internal uniformity, we consistently find the opposite effect for number of different lexemes and for the measure of internal uniformity. An explanation for these findings is provided in the discussion (Section 4).

3.1 The relationship between plant frequency and lexical diversity

Table 3 provides an overview of the p-value and Spearman's rank correlation coefficient for each combination of the measures of plant frequency and of lexical diversity per plant. The table indicates that a significant correlation (alpha = 0.05) exists between plant frequency and lexical diversity in all the cells. However, as the absolute values of the coefficients are never larger than 0.261, the correlation between plant frequency and lexical diversity is not very strong. Furthermore, plant frequency does not always correlate with lexical diversity in the way that was expected. More specifically, for number of different lexemes, which is shown in the second column of the table, positive correlations are found, while internal uniformity, in the fourth column, shows significant negative correlations. This means that more variation is found for plants that are more frequent, both locally and globally.

Table 3: Correlation between measures of plant frequency and measures of lexical diversity per plant.

	number of different lexemes	type-token ratio (TTR)	internal uniformity
local relative frequency km squares 1972–2004	0.261 $p < 0.001$	–0.256 $p < 0.001$	–0.191 $p < 0.001$
global absolute frequency km squares 1972–2004	0.241 $p < 0.001$	–0.261 $p < 0.001$	–0.156 $p < 0.001$
global absolute frequency hour squares 1939–1971	0.233 $p < 0.001$	–0.223 $p < 0.001$	–0.155 $p < 0.001$
global absolute frequency hour squares 1972–2004	0.240 $p < 0.001$	–0.256 $p < 0.001$	–0.158 $p < 0.001$

For the third column in the table, which provides the results for TTR, all the measures of plant frequency show a negative correlation with lexical diversity. This is in accordance with what was expected: more frequent plants show a smaller amount of lexical diversity. However, as suggested in section 2.2, TTR is sensitive to the number of tokens per concept, in the sense that TTR is high for concepts with the same number of types and tokens, even when only a small number of records is available for these concepts. As about one third of the concepts in the data set have a TTR value of 1, inspecting whether the negative correlation persists when only plants with a TTR value lower than 1 are included in the analysis may offer some more insight into the relation between plant frequency and

TTR. Table 4 shows Spearman's rank correlation coefficients and the p-values for this subset of the data. Even though the correlation coefficients are slightly lower than in Table 3, the significant negative correlations persist: more frequent plants show a smaller amount of lexical diversity.[7]

Table 4: Correlation between four measures of plant frequency and TTR per plant for concepts with TTR smaller than 1.

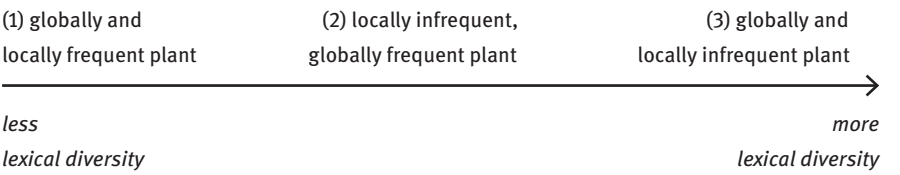
	correlation coefficient and p-value for Spearman's rank correlation
local relative frequency km squares 1972–2004	–0.220 $p < 0.001$
global absolute frequency km squares 1972–2004	–0.261 $p < 0.001$
global absolute frequency hour squares 1939–1971	–0.206 $p < 0.001$
global absolute frequency hour squares 1972–2004	–0.255 $p < 0.001$

3.2 The relationship between the global and local frequency of a plant

Concerning the relationship between the four measures of plant frequency that were used, Tables 3 and 4 show that both the local and global frequency of a plant correlate with lexical diversity. By solely relying on these measures, we cannot determine whether local and global frequency have the same effect on lexical diversity. As explained in section 1, we assume that plants that are only infrequent in a particular region are still more salient overall than globally infrequent

7 Additionally, we checked whether significant correlations are also found for plant concepts with at least 50 records. This data set is smaller (N = 137) and, probably as a result, some of the plant frequency measures lose their significance. A significant correlation is still found between TTR and global absolute frequency hour squares 1972–2004 ($p < 0.05$, $r = -0.18$). Near-significant negative correlations, which would probably reach significance in a larger data set, still occur between TTR and global absolute frequency km squares 1972–2004 ($p < 0.1$, $r = -0.16$), and between internal uniformity and local relative frequency km squares 1972–2004 ($p < 0.1$, $r = -0.15$). All the correlation coefficients have the same sign as in the larger data set, but the absolute values are lower. Overall, these results suggest that the relationship between plant frequency and lexical diversity is not solely dependent on the amount of data per concept.

plants and, thus, show less lexical diversity.[8] In sum, we expect to find that lexical diversity follows the following pattern:



To determine whether this relationship holds, we build three mixed-effects linear regression models with as a response variable the number of different lexemes (model 1), TTR (model 2) and internal uniformity (model 3) per plant per ecological region. Since the dataset contains multiple observations for each ecological region and for most of the plants, we use these factors as random effects in the models. We include frequency category per plant as a fixed-effects predictor in each of these models (N = 336).[9] This variable has three possible levels, depending on the global and local frequency of the plant:

1. very frequent plants, i.e. plants that occur in at least 2/3 of the hour squares that were investigated between 1939 and 1971 and that are available in at least 70% of the kilometer squares of the region under scrutiny (N = 106), e.g. the common nettle (Urtica dioica) in all ecological regions;
2. plants that are globally frequent, but infrequent in a particular region, i.e. plants that occur in at least 2/3 of the hour squares that were investigated between 1939 and 1971, but that are only available in less than half of the km squares in a particular region (N = 51), e.g. the common bent (Agrostis capillaris) in the Polder region;
3. plants that are globally and locally infrequent, i.e. plants that occur in less than 1/3 of the hour squares that were investigated between 1939 and 1971 and that are only available in less than half of the km squares in a particular region (N = 179), e.g. the sweetscented bedstraw (Galium odoratum) in all ecological regions.

8 As an anonymous reviewer points out, differences in experiential salience may also be found when a plant is locally frequent but globally infrequent. However, as only two plants in our dataset would belong to this category (viz. the wild privet (Ligustrum vulgare) and the goldmoss stonecrop (Sedum acre), two plants that are typically found near the sea and, thus, grow frequently in the Dunes area, but only rarely occur naturally in the rest of the northern part of Belgium), we did not take this category into account.

9 Because we are mostly interested in the extreme cases in this part of the analysis, we do not include all the plants in the models. More specifically, plants that are relatively 'neutral' regarding global or local frequency, i.e. plants that are neither locally, nor globally very frequent or infrequent, are not assigned to any of the frequency categories.

Table 5 shows the output of the three regression models. At the top of the Table, the random effects (all adjustments to the intercept) are shown with their corresponding standard deviation, and the residual error. Each model has the same random effects structure, with a random intercept for plant and a random intercept for ecological region. In each of the models, this random structure was statistically validated before including the fixed-effects predictor.[10] The bottom of the page shows the model diagnostics. Marginal and conditional R^2 show the proportion of variance explained by the fixed effects alone, and the proportion explained by the combination of the fixed and random factors, respectively.[11]

The middle part of Table 5 shows the estimate and p value for the fixed-effects predictor *frequency category*. In models 1 and 2, a higher value for the response variable indicates a larger amount of lexical variation per plant (operationalized as number of different lexemes and TTR, respectively). In these models, we would therefore expect positive estimates for the locally and globally infrequent plants, in comparison to the reference level (globally and locally frequent plants), which is captured in the intercept. However, the results are not completely in line with this expectation. For number of different lexemes, the amount of variation decreases for less frequent plants. However, this unexpected negative trend is probably connected to the fact that for less frequent plants, a smaller amount of records is available per plant. In fact, there is a significant positive correlation between the number of responses per plant and the three plant frequency categories ($H = 31.645$, $p < 0.001$). The globally frequent plants have 160 records on average (sd = 497); for locally infrequent plants, the mean number of records per plant is 93 (sd = 259); for globally infrequent plants, the average number of records is only 26 (sd = 60). As the number of lexemes and the number of records per plant per region are highly correlated (see 2.2), it is not surprising that for the less frequent plants, a smaller number of different lexemes is found.

In model 3, higher values for the response variable signify a smaller amount of variability. We therefore expect negative estimates for the locally and globally infrequent plants. However, in this model, we find the opposite effect as well: less frequent plants show a significantly higher amount of internal uniformity. In sum, only the results for TTR are as expected: both locally and globally infrequent

10 Ideally, we would have liked to use a random intercept for each plant per ecological region. However, the data do not support models with a random structure this complex. Instead, we use a separate random intercept for plant and ecological region and verify that intercept-only models with this random structure perform better than models without one or both of these random intercepts.

11 Marginal and conditional R^2 were calculated using sem.model.fits() from the piecewiseSEM-package (see https://jonlefcheck.net/2013/03/13/r2-for-linear-mixed-effects-models/, accessed 05.05.2017).

Table 5: Output for the random and fixed effects for mixed-effects linear regression models with as response variables the number of different lexemes per plant (model 1), TTR per plant (model 2) and internal uniformity per plant (model 3) per plant frequency category (reference level: globally frequent plants). Marginal R2 shows the proportion of variance explained by the fixed effects alone. Conditional R^2 depicts the proportion of variance explained by the fixed and random factors.

	model 1 nr. of different lexemes			model 2 TTR			model 3 internal uniformity		
random effects									
			std. dev			**std. dev**			**std. dev**
plant	intercept		6.240	intercept		0.210	intercept		0.133
ecological region	intercept		6.131	intercept		0.204	intercept		0.188
Residual			8.593			0.201			0.250
fixed effects									
	estimate	**std. error**	**p value**	**estimate**	**std. error**	**p value**	**estimate**	**std. error**	**p value**
intercept (glob. freq.)	11.004	2.875	< 0.01	0.434	0.093	< 0.01	0.4701	0.0842	< 0.01
locally infrequent	−1.549	2.158	NS	0.111	0.058	<0.1	0.1183	0.0564	< 0.05
globally infrequent	−7.035	1.808	< 0.001	0.243	0.054	< 0.001	0.1617	0.0443	< 0.001
model diagnostics									
marginal R^2		0.068			0.087			0.043	
conditional R^2		0.542			0.707			0.481	

plants have a significantly higher estimate than the frequent plants. This means that the less frequent a plant is, the higher its TTR value and, thus, higher the amount of variation in the names for the plant.

4 Discussion

Overall, the results of our analyses show that a correlation exists between plant frequency and lexical diversity. Although we aimed to show that experientially more familiar plants show less lexical diversity, the results are not completely in line with this expectation. One explanation for this finding is that the correlation between the measures of lexical diversity and the number of records that are available per plant influences the results to a certain extent (see 2.2). Although for TTR and internal uniformity, some (near-)significant correlations are still found when only plants with a relatively high number of records are included in the analysis, this is not the case for different number of lexemes (see section 3.1 and footnote 7). Consequently, the correlation between lexical diversity and number of records especially affects the results for the measure of number of different lexemes per concept: obtaining a higher number of different lexemes when more data is collected, is expected (although this number is likely to stabilize when enough tokens are available).

Interestingly, the results for TTR and internal uniformity differ, even though both of these measures take the number of tokens per concept into account. Before identifying some suggestions for future research in 4.2, section 4.1 will outline two explanations for these diverging results. On the one hand, TTR and internal uniformity can be different because they measure conceptually different phenomena. On the other hand, the measures were calculated per ecological region, but an ecological region may include different dialect regions.

4.1 TTR versus internal uniformity

The results for the TTR measure are as expected (less lexical diversity is found for more frequent plants and locally infrequent plants show less lexical variation than globally infrequent plants). Furthermore, the correlation persists even when only concepts are included in the analysis for which TTR is smaller than 1 (see Table 4). The results for internal uniformity show the opposite trend. Because the measures of lexical diversity are calculated at the level of the ecological region, the relationship between internal uniformity and TTR can probably be explained in terms of the degree of standardization per ecological region.

Table 6 shows the difference between the two measures. The number of tokens is comparable for the four plants, great mullein (Verbascum Thapsus) in the Loamy region, bitter dock (Rumex obtusifolius) in the Polder region, black locust (Robinia pseudoacacia) in the Sandy and sand-loamy region and forget-me-not (Myosotis arvensis) in the Dunes region. The number of different lexemes decreases from top to bottom (see Appendix 1 for an overview of the lexical items used per plant). Table 6 confirms that while the TTR measure cannot distinguish row 2 from the third one, the measure of internal uniformity can. The latter is sensitive to the number of lexemes that occur per concept *and* to the number of tokens per lexeme (i.e. type). It is low for concepts which show a smaller amount of standardization (i.e. one lexical item takes precedence over its competing dialectal heteronyms), like the bitter dock in the Polder region, and higher for plants with a larger degree of standardization, like the black locust in the Sandy and sand-loamy region.

As a consequence, even though plant frequency has an influence on the number of lexemes per concept, as indicated by the results for TTR, it does not necessarily ensure that one lexeme becomes the preferred lexeme over its competing synonyms throughout the ecological region. While for more frequent plants, the number of different variants decreases for the same amount of tokens,

Table 6: A comparison of number of different lexemes, TTR and internal uniformity.

	plant name, ecological region	number of records	distribution of types	number of different lexemes	TTR	internal uniformity
1	great mullein (Verbascum Thapsus), Loamy region	26	$lexeme_{1...18}$ occur once $lexeme_{19...22}$ occur once	22	0.846	0.050
2	bitter dock (Rumex obtusifolius), Polder region	38	$lexeme_{1,2}$ occur once $lexeme_3$ occurs 3 times $lexeme_4$ occurs 4 times $lexeme_5$ occurs 10 times $lexeme_6$ occurs 19 times	6	0.158	0.338
3	black locust (Robinia pseudoacacia), Sandy and sand-loamy region	26	$lexeme_{1,2,3}$ occur once $lexeme_4$ occurs 23 times	4	0.154	0.787
4	forget-me-not (Myosotis arvensis), Dunes region	52	$lexeme_1$ occurs 52 times	1	0.019	1

this does not mean that every language user chooses the same name in the same situation (i.e. ecological region). Geographical variation within an ecological region, for example, is not neutralized by the high natural frequency of a plant. In fact, if a plant has both a low value for TTR and for internal uniformity, this means that, while the plant does not have a large number of different lexemes given the number of available tokens, the number of records per lexeme per plant per region does not differ a lot and the tokens are distributed over the different lexemes in a relatively homogeneous way.

By inspecting the frequency of the lexemes for globally frequent plants with both a low value for TTR and for internal uniformity, we can confirm whether this explanation holds. Table 7 shows the five plants with the lowest value for internal uniformity and TTR < 0.2.[12] Tables 8 and 9 show the frequency of the lexical items that are used for the lesser burdock and the broadleaf plantain in the Sandy and sand-loamy region, (row 1–2 in Table 7) which will be discussed in more detail below. The distribution of the lexemes for the other plants in Table 6 is comparable to these plants (see Appendix 2): all five plants have about 3–5 lexemes that are very frequent in comparison to the other words for the concept.

Table 7: Overview of the five plants with the lowest value for internal uniformity and TTR < .2.

plant	ecological region	number of records	number of different lexemes	TTR	internal uniformity
broadleaf plantain (Plantago major)	Sandy and sand-loamy	218	39	0.179	0.079
lesser burdock (Arctium minus)	Sandy and sand-loamy	420	61	0.145	0.100
blackberry bush (Rubus fruticosus)	Sandy and sand-loamy	500	52	0.104	0.106
English plantain (Plantago lanceolate)	Sandy and sand-loamy	141	28	0.199	0.111
lesser burdock (Arctium minus)	Polder	226	39	0.173	0.112

12 The five plants in Table 6 come from the Sandy and sand-loamy or Polder region. Most of the data for the plants come from the same dictionary (WVD). They were all counted in at least 76% of the hour squares in the entire region of the atlas between 1972 and 2004, which confirms that they are globally frequent.

Table 8: Frequency of lexical items for the lesser burdock in the Sandy- and sand-loamy region (N = 420).

lexical item	N	lexical item	N	lexical item	N
kleef	2	plakkerbollen	2	plakbollen	4
klitkruid	2	plakkersbezetjes	2	plakdistel	4
wier	2	plakkerstruik	2	plakkers-, plakkertjeskruid	4
bommetjes	2	plakmadammetje	2	plakmadammetjes	4
bot	2	plakt-de-baard	2	distel	6
distelknoop	2	reit	2	klit	6
distelstekker	2	smijtdodde	2	distels	6
distelvinken	2	smijters	2	plakker	6
doppers	2	speenkruid	2	klis(se)bol	8
dotsjes	2	stekelharen	2	soldate-, soldatenknop(je)	8
everzwijnkruid	2	stekeltjes	2	klis(se)kruid	10
haakbloemen	2	stekers, stekertjes	2	stekkers, stekkertjes	12
klauwkruid	2	stekker	2	plakkruid	14
kleeftebollen	2	stekkertjeskruid	2	plakkers, plakkertjes	14
klissenstok	2	sterkerbol	2	kleefte	20
klister	2	toorvel	2	klissen	26
knopkruid	2	weerhaakjes	2	soldate(n)knoppen	28
mottebollen	2	zoete distel	2	kleef-, klevekruid	34
mouwenkruipers	2	grote klis	4	klis	116
pieker	2	kleefbollen	4		
piekertjes	2	klissebollen	4		

For the lesser burdock in the Sandy and sand-loamy region (N = 420), for example, *klis* occurs 116 times (see Table 8). Four other lexemes occur more than 15 times (*kleefte*; *klissen*; *soldate(n)knoppen* and *kleef-, klevekruid*). The other lexemes are less frequent. Overall, the tokens of these plants are distributed in a relatively homogeneous way over the different lexemes. Plotting the geographical distribution of the lexemes on a map indicates that more than one lexeme occurs in some locations: the language users know more than one local dialect word to refer to the concept (Figure 8). *Klis* is used throughout the ecological region. Other variants sometimes occur in locations where *klis* was found as well, or in locations close to towns with *klis*. Interestingly, these other variants also have a more limited geographical distribution than *klis*.

Furthermore, other factors can be envisaged that determine which lexeme is used in which location. For example, it may be the case that the geographical distribution of the variants within the ecological regions reflects dialect boundaries and, thus, does show some degree of standardization, albeit on a different level than per ecological region. In this case, one would be able to find a number of relatively small geographical areas where a particular variant is used. An example of this can be found if the variants for the broadleaf plantain that occur more than

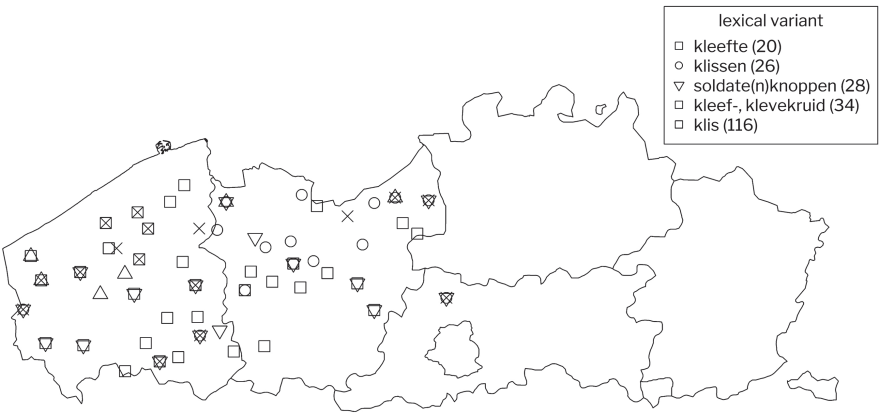


Figure 8: Geographical distribution of lexemes with N ≥ 15 for lesser burdock in the Sandy and sand-loamy region.

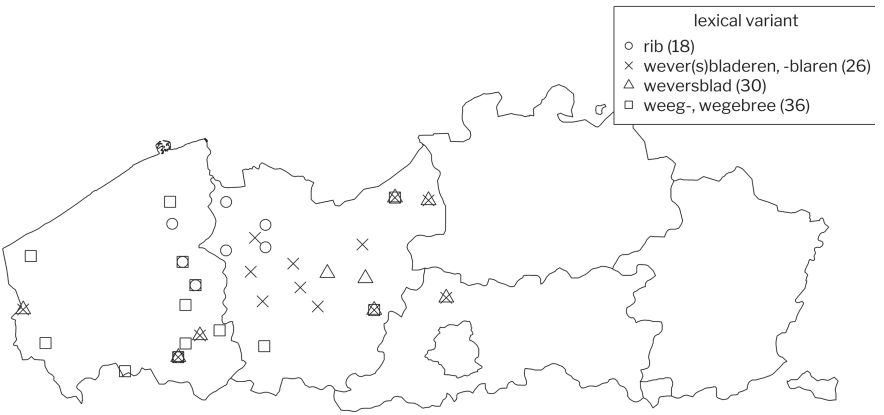


Figure 9: Distribution of lexemes with N ≥ 15 for broadleaf plantain in the Sandy and sand-loamy region.

15 times in the data are plotted on a map (Figure 9, also see Table 9). Even though these variants are relatively frequent in comparison to the other lexemes for this concept, they all seem to only be used in a particular geographical area of the Sandy and sand-loamy region.[13]

13 The distribution of the lexemes also seems to reflect the traditional dialect borders that have been distinguished in the Flemish dialects, as for instance, presented on the traditional map of Daan (1969).

Table 9: Frequency of lexical items for the broadleaf plantain in the Sandy- and sand-loamy region (N = 218).

lexical item	N	lexical item	N	lexical item	N
bree	2	varkensblad	2	zwijnegras	2
zwijnsoren	2	varkensblaren	2	grote weegbree	4
boterblad	2	varkensgras	2	kattestaart	4
breedblad	2	weegiebladen	2	weeg-, wege(s)bladen, -blaren	4
breedbladige weegbree	2	weegweeblad	2	wegaard(s)blad	4
breedbladweegbree	2	weewaarsblad	2	wegbree	6
dokke	2	weeweeblad	2	weewaarsbladen	8
dokkeblaren	2	weeweegbree	2	honde-, hondsrib	10
grote smart	2	wegaardsblaren	2	brede weegbree	14
honderibben, hondsribberen	2	wemel	2	rib	18
keunoren	2	weversbloemen	2	wever(s)bladeren, -blaren	26
papbladen	2	wilgebladen	2	weversblad	30
platen	2	zevenblaren	2	weeg-, wegebree	36

The diverging results for the models for TTR and internal uniformity per plant frequency group (Section 3.2) can be interpreted in a similar way. The analysis showed that the predicted value for TTR and for internal uniformity is smaller for the very frequent plants than for the locally and globally infrequent plants. The smaller values for TTR are in line with what was expected: a high value for global frequency can reduce the amount of diversity in the names for locally infrequent plants. The results for internal uniformity seem to contradict this finding. However, it is possible that the unexpected higher degree of uniformity of frequent plants is again related to the fact that there is no uniformity within the ecological region: the tokens for these plants may be distributed among the different lexemes that occur for the plants in a relatively homogeneous way. Additionally, since the number of records per plant also correlates with the frequency of the plant, a smaller number of tokens (and, thus, types) is available for the infrequent plants. This results in a seemingly more homogeneous distribution of the variants in the ecological regions (high degree of internal uniformity) and in a higher value for TTR.

4.2 Suggestions for future research

The analysis also showed that the absolute value of the correlation coefficients is relatively low (it is never higher than 0.261). This indicates that other factors than referential plant frequency probably influence the amount of lexical diversity

found in names for plants. For example, the poisonousness, usefulness or folkloric salience of a plant also influence how familiar the plant is. Furthermore, the number of tokens per plant may serve as an operationalization of familiarity of a plant as well (see Geeraerts & Speelman 2010, Speelman & Geeraerts 2008).

However, an additional explanation for the low correlation coefficients in the analysis is that the plants that are included in the dictionary data are overall relatively frequent. For example, the mean value for relative frequency per ecological region per plant for all the plants in the online database of the atlas is 12.46%. The mean value for this measure in the data set that was used for this paper is 37.78%. Of course, it is not surprising that only dialect data for relatively frequent plants is at our disposal. On the one hand, some of the plants in the atlas are probably so infrequent that they are not known to laymen. As a result, it may be the case that the lexicographers are not aware that these plants exist. On the other hand, if they are aware of the plants, it is possible that they are not interested in the names for these plants in local dialects, because they expect that asking for the names for these plants will not provide them with enough data. As was shown above, even for the relatively frequent plants that are available in our dataset, some plants are not represented by a large number of records in the linguistic data, which may have to do with the fact that these plants are unfamiliar for language users.

Aside from the fact that collecting dialect data for less frequent plants could corroborate the findings of this paper further, there are some restrictions on the present study that should be addressed in follow-up research. First, for the analysis, we lumped together all the data from the three dictionaries that were used. Although these data were not collected in exactly the same period, we did not control for diachronic differences between the sources: because most of the data come from the dictionary of Flemish dialects and because we aggregate over all the plants and ecological regions, we expect that this diachronic noise does not bias the analysis to a large degree.[14] Further, since the editors of the three dictionaries probably did not always make the same decisions about how to group different phonological variants into one lexeme, the data set may contain false heteronyms, lexemes that are treated as separate headwords in one dictionary, while they are treated as the same word in another one. For example, in the WLD, the phonological variant *bosbessen* 'bilberry' is grouped under the lexeme *bosbes*, while in the WBD, related phonological variants like *bosbeize*, *bosbeze* and *bosbieseme* are grouped under *bosbezen*, *bosbezen* and *bosbezem*, respectively. To cope with this difficulty, it would be necessary to compare the group-

14 We also executed the analysis on the Flemish data alone and obtained very similar results.

ing of the phonological variants in all the dictionaries. However, as this paper aimed to take a more aggregated approach towards variation in plant naming, we assumed that the dictionaries are similar enough to be compared and that this kind of noise would be filtered out due to the aggregative approach that we employed. Therefore, an interesting addition to this study would be to extend the scope to other dialect or language areas to investigate whether the findings are stable in other datasets and outside the region of the northern part of Belgium.

Third, other lines of investigation can be envisaged as well. The response variable, lexical diversity, can be operationalized in other ways than was done in this paper. For example, we could consider Guiraud's score, a transformation of the type-token ratio that is less dependent on the number of tokens per observation, as an alternative operationalization of lexical diversity. Additionally, although we only briefly mentioned how the geographical spread of the variants can differ, including this as a measure of diversity may offer further insight into the structure of the variation.

Extensions of the predictor variable, experiential frequency, are possible as well. For instance, a valuable addition to this study would be to further investigate the relationship between the experiential salience of naturally occurring plants and the number of records that are available in the data. Furthermore, additional explanatory variables, like geographic features or dialect boundaries within the Flemish, Brabantic and Limburgish dialects, or operationalizations of plant frequency based on folkloric information (e.g. usefulness or poisonousness of plants) could be included in the analysis. Moreover, comparing lexical data across different time periods can reveal whether the degree of lexical diversity decreases for plant names over time, and whether this is influenced by plant frequency.

Finally, in this paper we aimed to investigate whether experiential salience, in the form of referential frequency, influences lexical diversity. Other semantic fields can be envisaged in which this correlation can be tested. For example, rather than focusing on flora (or fauna), it would be interesting to expand the scope to a semantic field that is more prone to cultural differences, like the field of artifacts. Using other semantic fields will also allow for a comparison between concepts that occur naturally or that are conceived in a social environment.

5 Conclusion

In this paper we linked referential data to linguistic data to test whether the referential frequency of a plant, which was used to gauge experiential salience,

correlates with the amount of lexical variation that is found in the names for the plant. The analysis showed that some significant correlations exist: overall, plants that occur more frequently in a particular area seem to show a smaller degree of lexical diversity. However, the correlation is not strong enough for plant frequency to cause complete lexical uniformity within an ecological region and other factors play a role as well. Furthermore, a small-scale investigation of locally infrequent, but globally frequent plants revealed that the global frequency of a plant can cause a decrease in naming variation. However, more data is necessary to corroborate this finding. Overall, we were able to show that the everyday environment of a language user can influence the amount of lexical variation for a concept and that using referential data to study lexical variation can provide further insight into factors that influence language variation in a speech community.

Appendix 1: Lexical items for plants in Table 6, and forget-me-not (Myosotis arvensis) in the Dunes region

Appendix 1.1: Distribution of lexemes for the great mullein (Verbascum Thapsus) in the Loamy region

lexical item	N	lexical item	N	lexical item	N
gele kaars	1	toorts	1	zoklappen	1
gele thee	1	toppen	1	kalverwortel	1
kattenkop	1	wilde zokken	1	kaars	2
koningskaars	1	wolplant	1	paaskaars	2
lammetjesblaren	1	wolvenstaart	1	wilde tabak	2
lammetjesoren	1	zokjes	1	wolharen	2
maagdenkaars	1	zokken	1		
stalkaars	1	zokkenblaren	1		

Appendix 1.2: Distribution of lexemes for the bitter dock (Rumex obtusifolius) in the Polder region

lexical item	N	lexical item	N
wilde zuring	1	schape-, schaap(s)zurkel	4
Dokke	1	wilde zurkel	10
Paardezurkel	3	zurkel	19

Appendix 1.3: Distribution of lexemes for the black locust (Robinia pseudoacacia) in the Sandy and sand-loamy region

lexical item	N
acajou	1
robinia	1
valse acacia	1
acacia	23

Appendix 1.4: Distribution of lexemes for the forget-me-not (Myosotis arvensis) in the Dunes region

lexical item	N
vergeet-mij-niet(je)	52

Appendix 2: Frequency of lexemes for five plants with lowest value for internal uniformity and TTR < .2

Appendix 2.1: Distribution of lexemes for the blackberry bush (Rubus fruticosus) in the Sandy and sand-loamy region

lexical item	N	lexical item	N	lexical item	N
braambeien	1	hut bramen	1	braambeierstruik	3
braamberen	1	karrebezen	1	braambes(se)struik	4
braambezie	1	karrelbezie‘nstruik	1	braambezi‘n, -bezies	4
bramel	1	kattebeierboom	1	braambeier	5
kruip	1	moerbezen	1	braambeiers-, braambeier(en)hut	6
barstebeier	1	mondebeiers	1	bramers	6
bezenstruik	1	paters	1	braambees	7
braambeeshut	1	stekelbraam	1	braambeziestruik	12
braambeinen	1	struik braambezen	1	braambezelaar	16
braambessentronk	1	wilde frambozen	1	braamhut	20
braambezenbos	1	braam-, bramenhul	2	braambeiers	33
braambezenhul	1	braambees-, braambezetronk	2	braambees-, braambeze(n)struik	49
braambezietronk	1	braambessen	2	braam	58
braambreien	1	braambezebeier	2	braambezen	58
braamgewas	1	braambrei(en)struik	2	braam-, brame(n)struik	80
bramels	1	bramelhut	2	bramen	94
doorntakken	1	bramerstruik	2		
hul bramen	1	braambes	3		

Appendix 2.2: Distribution of lexemes for the English plantain (Plantago lanceolate) in the Sandy and sand-loamy region

lexical item	N	lexical item	N	lexical item	N
bagweeblad	1	wever(s)kruid	1	honde-, hondstong	4
dokken	1	hondsribberen	2	wegaard(s)bladen, -blaren	5
kattestaart	1	konijneneten	2	wegbree	5
keuneblad	1	weeg-, wegeblad	2	wever(s)blaren	7
kleine wegbree	1	weewaarsbladen, -blaren	2	smalle weegbree	14
papbladen	1	wegaard(s)blad	2	weeg-, wegebree	18
ribbeplaten	1	keunoren	3	honde-, hondsrib	27
stokjes	1	smalle rib	3	rib	28
vettekerte?	1	weegbladen, -blaren, wegebladen, -blaren	3		
weeweeblad	1	weversblad	3		

Appendix 2.3: Distribution of lexemes for the lesser burdock (Arctium minus) in the Polder region

lexical item	N	lexical item	N	lexical item	N
distel	2	plakdistel	2	klevers	4
kleef	2	plakker	2	klis(se)bol	4
klitkruid	2	plakpotten	2	klis(se)kruid	4
wier	2	reit	2	klissebollen	4
bommetjes	2	smijtbollen	2	plakbollen	4
distelvinken	2	smijtdodde	2	stekkers, stekkertjes	4
doppers	2	soldate-, soldatenknop(je)	2	distels	6
dotsjes	2	stekers, stekertjes	2	kleef-, klevekruid	10
kleeftebollen	2	stekmadammetjes	2	plakkers, plakkertjes	10
klissebloem	2	sterkerbol	2	soldate(n)knoppen	10
klister	2	zoete distel	2	klissen	22
pieker	2	grote klis	4	kleefte	24
piekertjes	2	kleefbollen	4	klis	64

References

Brok, Har. 1991. *Enkele bloemnamen in de Nederlandse dialekten: Etnobotanische nomenclatuur in het Nederlandse taalgebied (Publikaties van het P. J. Meertens-instituut voor dialectologie, volkskunde en naamkunde van de Koninklijke Nederlandse akademie van wetenschappen 18)*. Amsterdam: Meertens Institute.

Brok, Har. 2003. *Publicaties over plantennamen in Nederland*, Nederlandstalig *België en Frans-Vlaanderen* (Werken van de Koninklijke commissie voor toponymie en dialectologie. Vlaamse afdeling 24). Tongeren: Michiels.

Brok, Har. 2006. *Stinkend-juffertje en duivelskruid: Volksnamen van planten*. Amsterdam: Salome.

Bromhead, Helen. 2011. Ethnogeographical categories in English and Pitjantjatjara/Yankunytjatjara. *Language Sciences* 33(1). 58–75.

Daan, Jo. 1969. Dialecten. In Jo Daan & Dirk P. Blok, *Van randstad tot landrand. Toelichting bij de kaart: Dialecten en naamkunde (Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse akademie van wetenschappen te Amsterdam 37*), 7–43. Amsterdam: Noord-Hollandsche uitgeversmaatschappij.

Divjak, Dagmar & Catherine L. Caldwell-Harris. 2015. Frequency and entrenchment. In Ewa Dąbrowksa & Dagmar Divjak (eds.), *Handbook of Cognitive Linguistics (Handbooks of Linguistics and Communication Science = Handbücher zur Sprach- und Kommunikationswissenschaft 39)*, 53–74. Berlin: De Gruyter Mouton.

Franco, Karlien, Dirk Geeraerts & Dirk Speelman. 2015. The influence of semantic features on lexical geographical variation. In Johannes Wahle, Marisa Köllner, Harald R. Baayen, Gerhard Jäger & Tineke Baayen-Oudshoorn (eds.), *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics. Quantitative Investigations in Theoretical Linguistics (QITL). Tübingen, Germany, 4–6 November 2015*. art.nr. 11.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Institute.

Geeraerts, Dirk & Dirk Speelman. 2010. Heterodox concept features and onomasiological heterogeneity in dialects. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 23–40. Berlin/New York: De Gruyter Mouton.

Goossens, Jan. 1964. Enkel- en veeltoepasselijkheid van betekenaars op de taalkaart. In *Taalgeografie en semantiek. Lezingen gehouden voor de dialectencommissie der koninklijke Nederlandse academie van wetenschappen op 27 december 1962 door dr. J. Goossens en dr. Jan van Bakel. (Bijdragen en mededelingen der dialectencommissie van de koninklijke Nederlandse akademie van wetenschappen ter Amsterdam XXVIII)*, 3–27. Amsterdam: Noord-Hollandse uitgevers maatschappij.

Kruijsen, Joep. 1996. De Nijmeegse dialectlexicografische projecten. *Trefwoord* 11. 93–107.

Pauwels, Jan. 1933. *Enkele bloemnamen in de Zuidnederlandsche dialecten* (Noord- en Zuid-Nederlandsche dialectbibliotheek 5). The Hague: Nijhoff.

Pickl, Simon. 2013. Lexical Meaning and Spatial Distribution: Evidence from Geostatistical Dialectometry. Literary and Linguistic Computing: Journal of the Association for *Literary and Linguistic Computing and The Association for Computers and the Humanities* 28(1),63–81.

R Core Team. 2016. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Schmid, Hans Jörg. 2007. Entrenchment, salience and basic levels. In Dirk Geeraerts & Hubert Cuyckens. *The Oxford Handbook of Cognitive Linguistics*, 117–138. Oxford: Oxford university press.

Sevenant, Marjanne, Jan Menschaert, Martine Couvreur, Anne Ronse, Moira Heyn, Joks Janssen, Marc Antrop, Maarten Geypens, Martin Hermy & Geert De Blust. 2002. *Ecodistricten: ruimtelijke eenheden voor gebiedsgericht milieubeleid in Vlaanderen. Studieopdracht in het kader van actie 134 van het Vlaams Milieubeleidsplan 1997–2001*. Commissioned by the het Ministry of the Flemish Community. Administration of the Environment, Nature and Land Use.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-Based Linguistic Uniformity as a Generic Method for Comparing Language Varieties. *Computers and the Humanities* 37(3). 317–37.

Speelman, Dirk & Dirk Geeraerts. 2008. The role of concept characteristics in lexical dialectometry. *International Journal of Humanities and Arts Computing* 2. 221–242.

Swanenberg, Jos. 2000. *Lexicale variatie Cognitief-semantisch benaderd: over het benoemen van vogels in Zuid-Nederlandse dialecten*. Doctoral dissertation. Nijmegen: Radboud University.

Tweedie, Fiona J. & Harald R. Baayen. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32. 323–352.

Van Landuyt, Wouter, Ivan Hoste, Leo Vanhecke, Paul van den Bremt, Ward Vercruysse. & Dirk de Beer. 2006. *Atlas van de Flora van Vlaanderen en het Brussels Gewest*. Brussels: Research Institute for Nature and Forest, National Botanic Garden of Belgium & Flo.Wer.

WBD = Swanenberg, Jos & Har Brok. 2002. *Woordenboek van de Brabantse Dialecten. Deel III, 4.3, Flora*. Assen: Van Gorcum.

WLD = Kruijsen, Joep & Har Brok. 2002. *Woordenboek van de Limburgse dialecten. Deel III, 4.3, Flora*. Assen: Van Gorcum.WVD = De Pauw, Tineke, Jacques Van Keymeulen & Har Brok. 2002. *Woordenboek van de Vlaamse dialecten. Deel III, 3: Flora*. Tongeren: Michiels.