Degrees of semantic control in measuring aggregated lexical distances

Kris Heylen, Tom Ruette

University of Leuven

1 Introduction

The goal of the current study is to show how aggregated lexical variation can be studied by means of corpus-based techniques, which differ in their amount of semantic control. In the current variationist field, one finds many studies of phonological or morphological variation on the basis of corpora[1]. Remarkably at first sight, though, studies of lexical variation in corpora are rare, especially in comparison with dialectology, where the study of lexical variation is part of the main research goal. In contrast to the other corpus-based variationist studies, however, the dialectological account of lexical variation is very much restricted to elicited data, as stored in well-known dialect atlases. Therefore, the current study sets out to show how this void of corpus-based studies of lexical variation can be filled, while taking into account possible issues with lexical semantic complexity.

In the introduction to the paper, we would like to point out two things. First, we will explain why there is a plethora of studies on phonological and morphological variation and a scarcity of studies on lexical variation. Second, we will shed a different light on what can be understood under *lexical variation* from a corpus-linguistic point of view.

1.1 The stigma of lexical variation

The stigmatization of studies on lexical variation can partially be attributed to the well-debated article of Lavandera (1977). Before we give a brief account of its criticism, the context of the first generation variationist studies must be sketched. After William Labov performed his famous New York City experiment (Labov, 1966), a methodological consensus in the form of the *Principle of Accountability* (Labov, 1972) was established that the only valid sociolinguistic variable consisted of a set of variants that do the same "thing" (Chambers & Trudgill, 1980, p. 91). This is obviously a functional-semantic restriction on what may

be considered to be a sociolinguistic variable. With this in mind, the Lavandera critique is sparked off by an article of Sankoff (1972), in which she extends the Labovian phonological variable to a syntactic variable. The simple conceptual jump of Sankoff (1972) that if allophones may consitute a variable, two seemingly identical syntactic alternants may do so too, is problematic: where the Labovian work correlates the choice between "meaningless" variants and a socio-stylistic distribution of the options, Sankoff tries to link a socio-stylistic distribution of options to the choice between variants that are constrained in usage: e.g. she finds that the choice for "que"-deletion in Montreal French is not socio-stylistically motivated, but rather syntactic. Lavandera's criticism zooms in on this point, and proposes that only semantically equivalent options may constitute a variable, thus effectively excluding every linguistic phenomenon, except morpho-phonological variables.

It is now clear why the study of lexical variation is so problematic in the sociolinguistic field — perhaps even more problematic than syntactic variation. The main issue is that a lexical sociolinguistic variable is allowed to consist only of words that are equivalent in all perspectives, i.e. meaning, except for their socio-stylistic distribution. For words, this is a (quasi) impossible task. The meaning of words is highly contextual, even highly individual, and — according to the latest Cognitive Linguistic insights in lexical semantics (Geeraerts, 2010) — completely encyclopaedic (Taylor, 1989, Chapter 5). From the point-of-view of Lavandera, therefore, lexical variation can simply not be studied in a sociolinguistic way. Interestingly enough, this is also one of the reasons why dialectology refrains from corpus-based studies. Instead, dialectologists have elicited lexical preferences from their subjects by using questionnaires or naming tasks, to keep strict control on the meaning component. In order to gain even more control, dialectologists primarily elicited concrete notions, which can be shown to the subject, or can be described in sufficient detail.

In the current study, we argue in favor of a non-elicited approach to lexical variation: a large-scale corpus study of aggregated lexical variation. The idea of aggregation is central here. Whereas typical variationist studies zoom in on the socio-stylistic distribution of a single variable, we will aggregate the variational patterns of many lexical variables, as is common in dialectometry(Goebl, 1982; Grieve *et al.* , 2011). Although we loose detail in the behavior of individual variables, we do gain an insight in the overall variational patterns that play in the lexicon. Moreover, because we combine the patterns of many variables, subtle meaning differences between the variants of a single variable — which were the reason for Lavandera (1977) to abolish non-phonological variation — are averaged out, and become less important. Now, optimistically speaking, only the problem of finding a large set of lexical alternation variables remains. For this we will employ a (semi-) automatic way of modelling semantics in corpora: on the one hand, we want to identify concepts and lexical variants that can refer to them. On the other hand, we want to check whether lexemes refer to the same concept across varieties and not to another one.

1.2 What is lexical variation?

Although the previous paragraph dealt exclusively with the sociolinguistic notion of lexical variation, we would like to show two more approaches to lexical variation, which can be found in the literature. These methods differ from the variationist approach because they assign a different value to what we will call "semantic control" in the variable. In total, we distinguish three types of lexical variation: the use of different lexemes, the use of different lexemes to express the same thing (cf. alternation variables), and the different uses of a single lexeme. In what follows, we assume that we are looking at the lexical variation between two subcorpora.

The use of different lexemes A very straightforward way of looking at lexical variation is merely the observation that two subcorpora use different words. Although the two subcorpora might have considerable overlap in the lexical types, the frequency distribution of these types might point towards a difference between the subcorpora. Most famously, this approach has been put to work by the work of Douglas Biber (Biber, 1988, 1995; Biber & Barbieri, 2007). In his seminal 1988 book, Biber compared the frequency distribution of certain words, e.g. first person pronouns, across texts from different registers, and showed how frequency distributions turned out to be correlated to the register. From a semantic point-of-view, this is the most uncontrolled approach, and only applicable — for these purposes — to function words. Nevertheless, it is a very popular and widespread method, at the basis of many applications such as authorship attribution and document classification.

The use of different lexemes to express the same thing The second approach is derived from the variationist view on linguistic variation, ignoring the warning words of Lavandera (1977). It is claimed that lexical variation exists in the options that language users have to name a certain concept. This type of lexical variation is historically — in a European pre-structuralist tradition — known as *onomasiological* variation. Just like Labov (1978) suggests in his reply to Lavandera, we adopt a slightly relaxed attitude: perhaps it is true that the options are not exactly identical in their meaning at all levels, but it is not unthinkable that these lexical options are substitutable in many cases. This appeal to common sense is also present in a more recent paper by Edmonds & Hirst (2002).

The different uses of a lexeme A third and last possible way of defining lexical variation is different from the previous two approaches, because it does not compare (orthographically) different lexemes, but looks at the different uses of a single lexeme. Words may have different meanings[2] depending on the context. The most obvious and extreme example is of course polysemy. On a more subtle level, and with relation to the variationist perspective of this paper, we might find that a certain word is used to express A in one situation, but B in another

situation, with *A* and *B* only slightly different. As an example, the word *ketel* is used to refer to all sorts of pots in Limburgian (Dutch) dialects, whereas it's use is far more restricted in the other Dutch dialects. This type of lexical variation has been extensively studied by Justyna Robinson (e.g. Robinson, 2010), and can be grasped under the label *semasiological* variation.

On the basis of these three notions of lexical variation, we will calculate the lexical distance between subcorpora. The underlying idea is that a low amount of variation implies a small lexical distance. How exactly we will quantify this distance is explained below. To figure out the influence of semantic control on lexical distance measurements, we perform a corpus-based study of three registers in two national varieties of Dutch. The results show how different the outcomes of the semantically controlled measurements are: the approaches do not really agree with each other. The conclusion of the paper will be that a combination of two specific approaches might be the most trustworthy solution.

The remainder of this paper consists of Section 2 in which the compilation and structure of the corpus is described. Section 3 introduces the Semantic Vector Space model, which we will take as a starting point for studying lexical variation in the three ways that were described above. The actual lexical variables that we will be using to measure the lexical distance between the subcorpora of our corpus are introduced in Section 4. Section 6 then overviews the results of these three different ways of controlling the semantics of lexical distance measurements, and results are discussed in the final Section.

2 Corpus

Our corpus consists of texts that were gathered from Usenet posts, popular newspapers, quality newspapers and official government announcements (legalese). For each of these text types, we have texts written by people from the Netherlands and from Belgium. Moreover, we only gathered texts that were published between 1999 and 2004. As we will measure the distances between the parts of this corpus by counting lexical items, every subcorpus needs to big enough to supply reliable frequencies. Table 1 gives an overview of the sizes of the subcorpora. With almost 2 billion words, we can be quite certain that the frequencies for the lexical items in the corpus are representative of their actual usage.

	Usenet	Popular news	Quality news	Legalese	Total
BE	22 million	905 million	373 million	70 million	1.4 billion
NL	26 million	126 million	161 million	115 million	428 million
Total	48 million	1 billion	499 million	185 million	1.8 billion

Table 1: Overview of subcorpora and their sizes in words

However, word derivations or inflections could also introduce an error in the frequency counts. This can be solved by not just counting the occurrences, but

to count the root form, possibly controlled by the part-of-speech. For that reason, all texts in the corpus were automatically lemmatized and annoted for part-of-speech by the current state-of-the-art dependency parser for Dutch, which is Alpino (Bouma *et al.* , 2001). A further cause of mistakes in the frequency counts may be due to polysemy of the lexical items. How we have dealt with that problem in the current study — and how we will deal with it in further research — is explained in Section 4.

The actual texts of the corpus were either downloaded from the internet — Usenet and legalese — or obtained from the publishers — the popular and quality newspapers. The newspapers were requested and processed by the University of Twente and Groningen for the Netherlandic material, and by the University of Leuven for the Belgian material. The Usenet articles were downloaded from the Usenet archive online at Google Groups[3] by means of a series of Python scripts which removed meta-information (e.g. headers) and duplicated content (e.g. quotes). The legalese consists of the downloaded texts from the "Staatsblad" in Belgium[4] and The Netherlands[5]. All the Usenet and legalese texts were downloaded during 2010 and 2011. The corpus is not freely available due to copyright restrictions, but anybody is free to request and download the same materials.

3 Semantic Vector Space Models

The three types of lexical variation from the introduction 1 can now be formulated corpus linguistic terms: (1) the subcorpora use different lexemes. (2) the subcorpora use different lexemes for the same concept. (3) the subcorpora use the same lexemes differently, i.e. with a different meaning. For the large-scale corpus-based operationalization of these three approaches and the level of semantic control they require, we turn to a statistical approach developed in Computational Linguistics. There, so-called Semantic Vector Spaces (SVS's) have become the mainstay of processing semantics in large corpora. These models capture semantics in terms of frequency distributions of words over documents and of words co-occurring with other words. They have been applied to a wide variety of computational linguistic tasks – from Information Retrieval (Baeza-Yates & Ribeiro-Neto, 1999) and Question answering (van der Plas *et al.* , 2010) to automated essay scoring (Landauer & Dumais, 1997) or the modeling of human behavior in psycholinguistic experiments (Lowe & McDonald, 2000). In recent years, Semantic Vector Spaces have also seen applications in more traditional domains of linguistics like diachronic lexical studies (Sagi *et al.* , 2009), or, as in our case, the study of lexical variation (Peirsman *et al.* , 2010).

Broadly speaking, Semantic Vector Spaces can be used to model two types of semantics: text semantics and word semantics . Each type of semantics comes with its own specific SVS implementation (see Turney & Pantel (2010) for a general overview). In this paper we will use both a text-oriented SVS, for our analysis

of the first type of lexical variation, and a word-oriented SVS, for the two other types.

SVS models for text semantics try to capture the semantic content of documents by recording which words occur in each document and how often. Documents that contain the same words and with similar frequencies are then said to have the same semantic content. In practice, these models construct a so-called term-by-document matrix, in which each document is assigned a vector that captures the frequency distribution over all words in a given vocabulary (i.e. the *terms* in SVS parlance). Usually, the vocabulary is restricted to words that are of interest to a certain domain. In our case, we will restrict the vocabulary to the set of lexical items that forms the basis of comparison for our 3 types of measuring lexical variation (see section 4) For the vector comparison, SVS's use a geometrical metaphor (hence *Spaces*): the term-by-document frequencies can be seen as co-ordinates defining a point in a high-dimensional term space. Points (documents) closer together in the space contain the same terms and are said to be semantically more related. The resulting document distances can then be used to classify or cluster the documents. We will therefore call this type of SVS the *document classification* approach and we will use it to operationalize our first type of lexical variation. By constructing a word-by-subcorpus matrix, we will measure to what extent our regionally and stylistically stratified subcorpora use the same words (see section 5.1).

Let us now turn to the SVS models for word semantics. They are also based on a frequency distribution matrix but instead of the semantics of documents, the focus lies on the semantics of the words. To model word semantics, these SVS's record the co-occurrence frequencies of a set of target words with a large set of context words. The hypothesis is that words occurring in similar contexts, i.e. that are surrounded by the same context words, will have a similar meaning. For example, the semantic similarity of *clinic* and *hospital* can be induced from the fact they both co-occur with words like *doctor, nurse, operation, treatment*, etc. In practice, most models define context as the words occurring in a given window around the target words. In this study we set the window to 5 words to the left and right. As most models, we work with a restricted vocabulary of possible context words: we used the 4000 most frequent words, excluding the top 30, which were all function words. The raw co-occurrence frequencies were weighted with Point Wise Mutual information to increase the weight of more informative words, i.e. those that co-occur only with a limited set of (semantically related) target words. Using the same geometrical metaphor as before, the target words then become the points in a high dimensional space of context words. Target words are close together in the space if they share relatively high co-occurrence frequencies with the same context words and therefore they are likely to be semantically similar. Following SVS standards, the cosine was used as a proximity measure. Computing the cosine similarity between all pairs of target word vectors results in a target-word-by-target-word similarity matrix.

A word-based Semantic Vector Space will be the input for the operationaliza-

tion of the two other types of lexical variation in our study. Firstly, the word-by-word similarity matrix can be subjected to a cluster analysis that groups the target words into sets of near-synonyms, i.e. lexemes referring to the same ''thing". In our onomasiological measurment of lexical variation, we can then analyze whether the subcorpora differ with respect to their lexical choices, given the concepts. Secondly, we can construct target word vectors for each subcorpus separately and then calculate the similarity between the vectors. This will tell us for each target word whether it is used in the same contexts and with the same meaning in the different subcorpora. In our semasiological measurment of lexical varition, we can then assess to what extent the subcorpora tend to use lexemes differently, i.e. with a different meaning. In the next section, we will specify which target words we will use as the lexical variables in our variation study.

4 Variable set

In most studies that aggregate a number of linguistic variables to analyze underlying variational dimensions, the variable set is usually limited. In dialectometry, most studies aggregate the variables that are available in dialect atlases. In sociolectometric research (e.g. Geeraerts *et al.* , 1999; Soares da Silva, 2010), lexical variable from random lexical fields were chosen. And in stylometric studies, a collection of so-called functional variables is gathered from grammars and stylebooks.

Although quite similar in method and approach, there is a difference in research question between dialectometric or stylometric studies and sociolectometric studies. Whereas dialect- and stylistic research sets out to point out a specific regional or functional difference between certain language varieties, our study does not presuppose a variational dimension. In other words, we do not have an a priori dimension of variation that we want to point out, but rather, we want to discover these dimensions bottom-up. Therefore, in our study, it would be wrong to analyze a dataset that is biased towards a certain pattern.

This is exactly why previous (lexical) sociolectometric studies have limited themselves to the analysis of two (or a small number of) lexical fields. It is manually feasible to get a representative, or even an exhaustive list of concepts that belong to the same lexical field, and for this limited amount of concepts it is possible to find most or all the words that can refer to each one of these concepts. As such, there is no variational pattern pre-programmed in this set of variables under investigation. The only way to discover a certain pattern is by investigating how the items in the variables are distributed over subcorpora that differ along the dimension under investigation.

However, the manual variable collection of this approach make it unscalable to a study that has the ambition to investigate the variational patterns of the lexicon in general. Given that the vocabulary is vast, only a tiny portion and a very

specific part of the lexicon is analyzed when the variable set is collected manually. Ideally, the variable set should consist of a set of words that is representative for a sizeable part of the vocabulary, and such a quantity of variables should be gathered in an automatic and bottom-up way.

So, what exactly is the task that we give to this automatic approach? We want (a) to find words that refer to the same concept, and (b) to find a large number of concepts that come from different parts of the vocabulary. The first part ensures that our variable set contains (onomasiological) lexical variation, the second part takes care of the representativity of the feature set. Here, the word-based Semantic Vector Space model outlined in the previous section comes into play.

The word-by-word similarity matrix is submitted to a clustering algorithm known as *Clustering by Committee* (CBC) (Pantel & Lin, 2002). CBC was designed to describe each sense of a target word by means of a cluster of words, that are semantically very related to the sense of the target word under consideration. One phase of the algorithm consists of finding clusters of semantically related words, so-called *committees*, and the outcome of this phase seems to comply roughly with our desired feature set. Because the underlying method is (still) somewhat imprecise, we will manually filter out those committees that are representative of a single concept — cf. step (a) of the task. With relation to step (b) of the task, we can point out that CBC is completely frequency related, and thus there is a bias in the retrieved clusters towards frequent concepts. Nonetheless, these concepts come from diverse lexical fields.

There are 476 variants (lexemes), contained in 218 variables (concepts), that were manually retained from the automatically generated alternation variables. We restricted the variants to nouns only, because Vector Space Models appear to be most successful for referential items. In Table 2, a selection of variables is presented. The two other types of lexical variation require a less strict variable selection: in principle, the document classification approach would have allowed to analyze frequency differences between subcorpora for all lexemes, and in the semasiologal approach, we could have compared target word vectors between subcorpora for a much larger selection of lexemes. However, our explicit aim is to compare the three approaches to lexical variation, so that we restricted ourselves to the same selection of 476 lexemes in all approaches.

5 Degrees of semantic control

In this section, we arrive at the heart of our research goal: what is the influence of different degrees of semantic control on an aggregated study of lexical variation. More specifically, we compare an approach with no semantic control at all (Section 5.1) to two approaches that apply a different type of semantic control: an approach that accounts for the difference in usage of a single word (semasiological approach, Section 5.2), and an approach that accounts for the differ-

CONCEPT	Items
MANNER	wijze, manier
GENOCIDE	volk_moord, genocide
POLL	peiling, opiniepeiling
MARIHUANA	cannabis, marihuana
PUTSCH	staatsgreep, coup
MENINGITIS	hersenvliesontsteking, meningitis
DEMONSTRATOR	demonstrant, betoger
AIRPORT	vliegveld, luchthaven
COLDNESS	koude, kou
TORTURE	marteling, foltering
VICTORY	zege, overwinning
HOMOSEXUAL	homo, homoseksueel
SAXOPHONE	sax, saxofoon
INTERNETPROVIDER	provider, internetprovider, internetaanbieder
AIRCONDITIONING	airconditioning, airco
RELIGION	religie, godsdienst
THE OTHER SIDE	overkant, overzijde
EXPLOSION	explosie, ontploffing
RESTROOM	toilet, wc
INJURY	kwetsuur, letsel
BLAST	windstoot, ruk_wind
LAST MINUTE	nippertje, valreep
XENOPHOBIA	vreemdeling_haat, xenofobie
PASSER-BY	voorbijganger, passant
AIR STRIKE	luchtaanval, bombardement
FIGHTING SPIRIT	vechtlust, strijdlust
GOVERNMENT FORCES	regeringsleger, regeringstroepen
CAR	auto, wagen
PROFIT FORECAST	winst_verwachting, winst_prognose

Table 2: Example variables

ences in naming a concept with a different word (onomasiological approach, Section 5.3).

5.1 Document classification

The first approach is based on a straightforward document classification algorithm and refrains from any semantic control, cf. the first type of SVS introduced above. Technically speaking, the frequencies of the features in a subcorpus constitute an identifying vector for the subcorpus, and the similarity between two subcorpora is measured by means of the cosine similarity metric. This metric measures the cosine of the (hyperdimensional) angle between two vectors: if the cosine is close to 1, the angle between the two vectors is small, and the two subcorpora are considered to be very similar. The cosine metric applied to two subcorpora V_1 and V_2, on the basis of their identifying vectors $\vec{x}$ and $\vec{y}$ is formally described in Equation 1.

$$cos(V_1, V_2) = cos(\vec{x}, \vec{y}) = 1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \quad (1)$$

Because the cosine metric merely works with the (raw) frequencies of the input features and does not take into account that there are groups of features that are semantically related, we call this a non-semantically controlled approach. This is the most rudimentary approach, and from its typical application in document classification, we know that it should primarily pick up on referential or content-related differences between the subcorpora.

Calculating the similarity between all pairs of subcorpora that we have avaible yields a similarity matrix. This similarity matrix can easily be converted to a dissimilarity matrix, by subtracting it from 1. The resulting distance matrix can be visualized by means of Multidimensional Scaling (MDS). In Figure 1 and Figure 2, one can find the two- and three-dimensional solution of the non-metric MDS implementation `isoMDS` in the statistical program `R`, with the `MASS` package loaded.

The two-dimensional solution shows a strong group of Belgian and Netherlandic newspapers, except for the Belgian popular newspapers. The two national varieties of Usenet also practically overlap. Dominating the first (horizontal) dimension, we find the Belgian and Netherlandic legalese subcorpora, on the right side of the plot. The second (vertical) dimension is not very outspoken: it seems to largely set apart the Usenet subcorpora. Because of this unclear dimension, and despite the already very low stress value, we calculate a three-dimensional solution in Figure 2, to see if this clears up dimension 2.

The three-dimensional solution preserves the distinction between legalese and the other subcorpora on dimension 1, and it confirms the idea that dimension two sets apart Usenet. The third dimension, which is admittedly also not very outspoken, now seems to tear apart the Netherlandic and Belgian subcorpora. The Netherlandic subcorpora are consistently “lower” than the Belgian

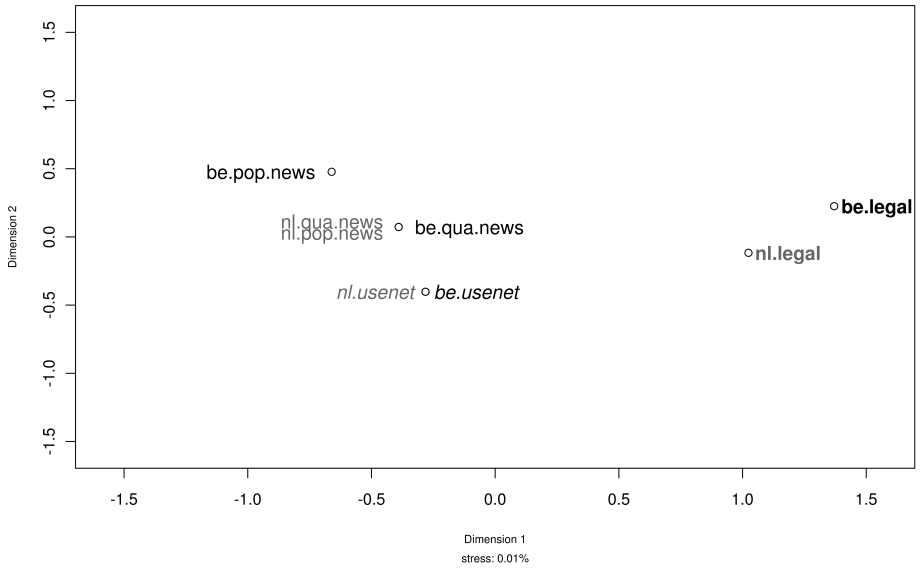

Figure 1: 2D Multidimensional Scaling visualization of lexical distances, without semantic control

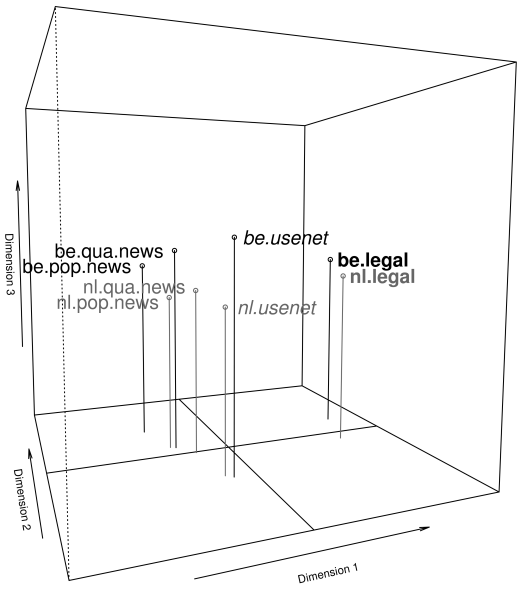

Figure 2: 3D Multidimensional Scaling visualization of lexical distances, without semantic control

subcorpora.

5.2 Semasiological approach

In the semasiological approach, we take each of the 476 lexemes in our variable set and construct a word-based SVS with a separate target word vector for each subcorpus. Calculating the cosine similarity between the vectorso results in a distance matrix between subcorpora. Subcorpora are similar if they tend use a particular lexeme in the same way, i.e. in the same contexts and with the same meaning. They are different if they use the lexeme with a different meaning. Aggregating over all lexemes, we can assess to what extent the subcorpora show variation in word usage and word meaning in general. Given the lexemes L_1 to L_m, then the global dissimilarity D between two subcorpora V_1 and V_2 on the basis of L_1 up to L_m can be calculated as:

$$D_{cos}(V_1, V_2) = \sum_{i=1}^{m} (Dcos_{L_i}(V_1, V_2)) \tag{2}$$

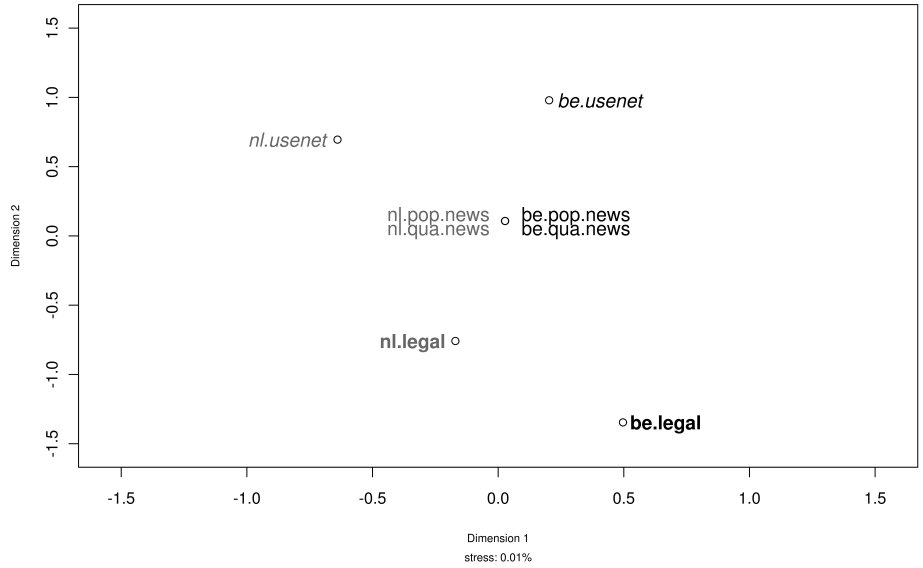

Figure 3: 2D Multidimensional Scaling visualization of lexical distances, with semasiological control

Looking at the MDS solution for subcorpus distance matrix, we see that the semasiological approach immediately promotes the national distinction in the variable set. The two-dimensional visualization should be interpreted as follows: especially in Usenet and legalese, a word is not used the same in Belgium and The Netherlands. Remarkably, all the newspapers do agree on how to use

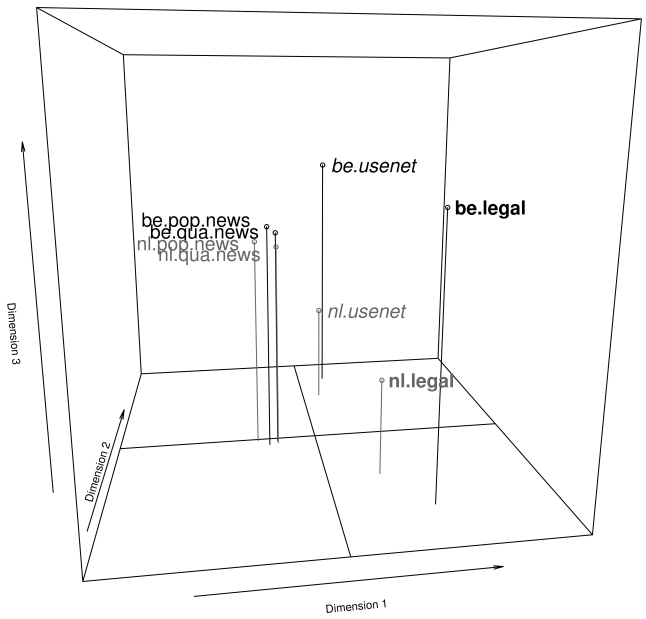

Figure 4: 3D Multidimensional Scaling visualization of lexical distances, with semasiological control

a certain word. This might point out that the deviating behavior of Usenet and legalese has to do with their specificity and technicality.

The three dimensional solution surprisingly changes the order of the dimensions: suddenly, dimension 1 distinguishes the newspapers from Usenet and legalese, and dimension 2 distinguishes legalese from Usenet. It is only at dimension 3 that the country distinction — so obvious in the two-dimensional solution — reveals itself.

5.3 Onomasiological approach

Finally, we arrive at the onomasiological approach. Here, the distance metric is informed about the groups of semantic related input features by means of recalculating the frequency of a single feature relative to the sum of the frequencies of the semantic group to which it belongs. As such, the distance metric is sensitive to lexical variation as could be measured as the well-known sociolinguistic alternation variable.

Given two subcorpora V_1 and V_2, a group of semantically related words L and x_1 to x_n the exhaustive list of words in L, then we refer to the absolute frequency F of the usage of x_i for L in V_j with[6]:

$$F_{V_j,L}(x_i) \tag{3}$$

Subsequently, we introduce the relative frequency R:

$$R_{V_j,L}(x_i) = \frac{F_{V_j,L}(x_i)}{\sum_{k=1}^{n}(F_{V_j,L}(x_k))} \tag{4}$$

Now we can define the (City-Block) distance D_{CB} between V_1 and V_2 on the basis of L as follows (the division by two is for normalization, mapping the results to the interval [0,1]):

$$D_{CB,L}(V_1, V_2) = \frac{1}{2}\sum_{i=1}^{n} |R_{V_1,L}(x_i) - R_{V_2,L}(x_i)| \tag{5}$$

To calculate the dissimilarity between subcorpora on the basis of many groups of semantically related words, we just sum the dissimilarities for the individual groups. In other words, given a set of groups L_1 to L_m, then the global dissimilarity D between two subcorpora V_1 and V_2 on the basis of L_1 up to L_m can be calculated as:

$$D_{CB}(V_1, V_2) = \sum_{i=1}^{m}(D_{L_i}(V_1, V_2)W(L_i)) \tag{6}$$

The W in the formula is a weighting factor. We use weights to ensure that groups of words which have a relatively higher frequency (summed over the size of the two subcorpora that are being compared[7]) also have a greater impact on the distance measurement. In other words, in the case of a weighted calculation, semantic groups that are more common in everyday life and language are treated as more important.

In Figure 5, we see again the distinction between the legalese and the rest of the subcorpora, but this time, the grouping of the subcorpora is extremely tight. Therefore, a three-dimensional solution in Figure 6 might reveal some more variation in the patterning of the subcorpora.

Indeed, the first dimension still singles out the legalese subcorpora, but now, the second dimension puts the Belgian subcorpora at the front, and the Netherlandic subcorpora at the back. The third dimension seems to pull down the Usenet subcorpora, distinguishing them from the newspapers and legalese.

6 General Discussion

In the above experiment, we have compared three ways of measuring the lexical distance between subcorpora. The first approached neglected any kind of semantic control of the lexical input features. Figure 2 shows how this *document classification* approach separates – as expected – first the registers: dimension 1 singles out the legalese, and dimension 2 splits Usenet from newspaper articles. It is only at the third dimension that a very weak country distinction, most obvious in the Usenet material, props up.

The second and the third approach presented two different perspectives on semantic control of lexical variation. On the one hand, the difference in use of

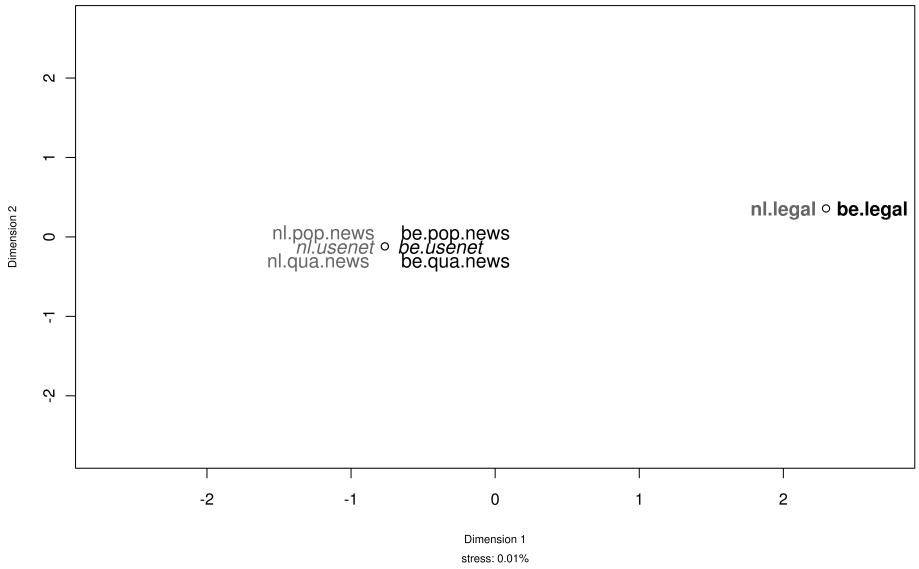

Figure 5: 2D Multidimensional Scaling visualization of lexical distances, with onomasiological control

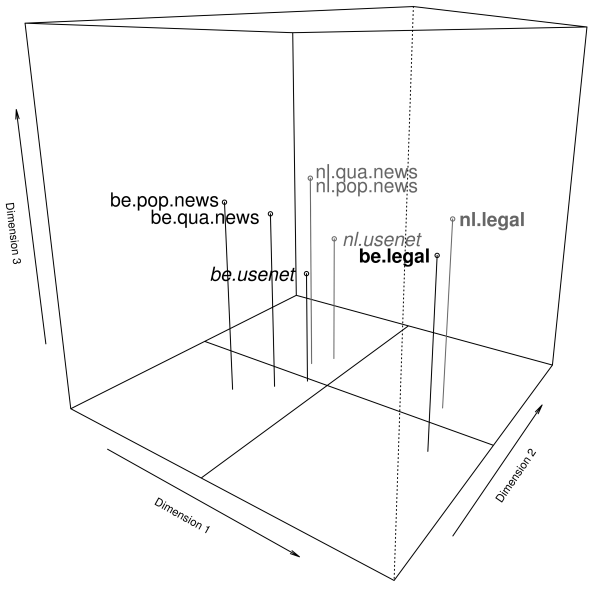

Figure 6: 3D Multidimensional Scaling visualization of lexical distances, with onomasiological control

an individual word across the different subcorpora was used as the basis for a *semasiological* distance metric. This showed (Figure 3) that on average there is a national pattern in words that are used differently. However, this national pattern is not homogeneous across the three registers that are present in our corpus. The newspaper articles show surprisingly similar use of words, whereas the more specific subcorpora of legalese and Usenet drift apart.

On the other hand, the choice of a specific word for expressing a certain concept was the basis for the *onomasiological* distance metric. Figure 6 shows how neatly legalese is split of first, how then second the national distinction becomes clear, and how finally the Usenet subcorpora are separated. And all these distinctions are due to the fact that another word is used to refer to a concept.

There are two main conclusions to be drawn here. First, we have pointed out that semantic control (or the lack thereof) has a profound influence on the outcome of an aggregated study of lexical variation. Although the three approaches agree on the special position of legalese, it is unclear how the variational dimensions relate to each other after that. At least, the three approaches do not agree with each other, although every single approach is valid in its own right.

Second, assuming that we discard the approach without any semantic control whatsoever — because one can hardly call its input lexical variation —, an approach that combines both the semasiological and onomasiological control is probably needed. The semasiological approach comes first to identify which words are actually comparable across the subcorpora: if a certain word is used completely different in Belgium than in The Netherlands, it makes no sense to use it in the onomasiological approach. Indeed, the onomasiological approach assumes that the words in the variable are interchangeable, but a word with a large semasiological range does not comply with that expectation.

Of course, in an ideal situation, this boils down to a scrutiny of every single observation as to verify whether the occurrence is relevant. This is obviously not feasible for a large-scale corpus-based study with thousands of observations. Luckily, further developments in the SVS domain will soon be able to model the usage-based meaning of a single token, allowing us to evaluate and correct the accurateness of the frequency counts. Furthermore, the current study did not consider an account of the individual variables in order to complement the interpretation of the analyses. We considered this to be outside the scope of the current paper, which wanted merely to compare the outcomes of three aggregating techniques. However, the introduction of quantitative methodologies that allow an insight in the behavior of individual variables is part of further research.

Notes

1. if a corpus is taken to be a naturalistic sample of language use
2. The link between meaning and use will be discussed below.
3. `http://groups.google.com`
4. `http://www.ejustice.just.fgov.be/cgi/welcome.pl`

5. https://www.officielebekendmakingen.nl/staatsblad
6. The following introduction to the City-Block distance method is taken from Speelman *et al.* (2003, Section 2.2).
7. The size of the two subcorpora is not the actual amount of words in the two subcorpora, but the sum of all profiles in these two subcorpora with a frequency higher than 30.

References

Baeza-Yates, Ricardo, & Ribeiro-Neto, Berthier. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber, Douglas, & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes*, **26**(3), 263–286.

Bouma, Gerlof, van Noord, Gertjan, & Malouf, Rob. 2001. Alpino: wide-coverage computational analysis of Dutch. *Pages 45–59 of:* Daelemans, Walter, Sima'an, K., Veenstra, J., & Zavrel, J. (eds), *Computational Linguistics in the Netherlands 2000. Rodolpi, Amsterdam*.

Chambers, Jack K., & Trudgill, Peter. 1980. *Dialectology*. 2., 15. print edn. Cambridge textbooks in linguistics. Cambridge University Press.

Edmonds, Philip, & Hirst, Graeme. 2002. Near-synonymy and Lexical choice. *Computational Linguistics*, **28**(2), 105–144.

Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Berlin: Mouton De Gruyter.

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

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Oesterreichische Akademie der Wissenschaften.

Grieve, Jack, Speelman, Dirk, & 2011, Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, **23**, 193–221.

Labov, William. 1966. *The social stratification of English in New York City*. Center for Applied Linguistics.

Labov, William. 1972. Some principles of linguistic methodology. *Language in Society*, **1**(1), 97–120.

Labov, William. 1978. Where does the linguistic variable stop? A response to Beatriz Lavandera. *Working papers in sociolinguistics*, **44**, 5–22.

Landauer, T., & Dumais, S. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, **104**, 411–240.

Lavandera, Beatriz. 1977. Where does the sociolinguistic variable stop? *Working papers in sociolinguistics*, **40**, 6–24.

Lowe, Will, & McDonald, Scott. 2000. The direct route: Mediated priming in semantic space. *Pages 675–680 of: Proceedings of the 22nd Annual Conference of the Cognitive Science Society.*

Pantel, Patrick, & Lin, Dekang. 2002. Discovering word senses from text. *Pages 613–619 of: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002).*

Peirsman, Yves, Geeraerts, Dirk, & Speelman, Dirk. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, **16**(4), 469–491.

Robinson, Justyna. 2010. Awesome insights into semantic variation. *In:* Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds), *Advances in Cognitive Linguistics.* Berlin/New York, Mouton de Gruyter.

Sagi, Eyal, Kaufmann, Stefan, & Clark, Brady. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. *Pages 104–111 of: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics.* Athens, Greece: Association for Computational Linguistics.

Sankoff, Gillian. 1972. Above and beyond phonology in variable rules. *In:* Shuy, R. W., & Bailey, C. (eds), *New ways of analyzing variation in English.* Washington, D.C.: Georgetown University Press.

Soares da Silva, Augusto. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. *In:* Geeraerts, Dirk, Kristiansen, Gitte, & Peirsman, Yves (eds), *Advances in cognitive sociolinguistics.* Berlin/New York, Mouton de Gruyter.

Speelman, Dirk, Grondelaers, Stefan, & Geeraerts, Dirk. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, **37**, 317–337.

Taylor, John R. 1989. *Linguistic categorization: Prototypes in Linguistic Theory.* Oxford: Clarendon Press.

Turney, Peter, & Pantel, Patrick. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.

van der Plas, Lonneke, Bouma, Gosse, & Mur, Jori. 2010. Automatic Acquisition of Lexico-semantic Knowledge for QA. *Pages 271–287 of:* Huang, Chu-Ren, Calzolari, Nicoletta, Gangemi, Aldo, Lenci, Alessandro, Oltramari, Alessandro, & Prevot, Laurent (eds), *Ontologies and Lexical Resources for Natural Language Processing*. Cambridge University Press.