Stable Lexical Marker Analysis: a corpus-based identification of lexical variation

Dirk De Hertog, Kris Heylen, Dirk Speelman

Abstract

Research questions that deal with mutual intelligibility and that investigate language attitudes in pluricentric languages rely on a correct assessment of the loci of divergence, differences in word choice being one of the most salient. Quantitative corpus-based methods can aid researchers to identify this lexical variation. This paper will focus on the language-independent method of Stable Lexical Marker Analysis (SLMA, Speelman et al. 2008) to find variety-specific words in representative corpora. The method is based on the keyword-analysis approach (Scott, 1997) but allows a graded rather than a categorical assessment of markedness and includes a mechanism to circumvent topical bias in corpora. The paper discusses further improvements to SLMA in order to deal with gradedness and offers a quantitative and qualitative analysis of results from a case study on the identification of lexical markers for Netherlandic and Belgian Dutch.

1. Introduction

The lexical variation as observed between the regional varieties of a pluricentric languague, can be seen as a sociolinguistic variable in the Labovian sense (Geeraerts 2009). The profile based approach to lexical variation (Geeraerts, Grondelaers, Speelman 1999; Silva 2006) focuses on the onomasiological choices for a given concept across two language varieties. All possible synonyms to refer to a concept together with their relative frequencies constitute the profile of that concept. After quantifying the distance between the observed frequencies for each concept, the average distance over a set of concepts is taken to represent the onomasiological difference between the language varieties under investigation.

In the study of the lexical variation in pluricentric languages, considerable time and effort has to be put in the first step of identifying relevant lexical variables, i.e. the sets of words that are the loci of potential divergence between varieties in the lexicon. Instead of manually selecting variables that are known to display variation, quantitative corpus-based methods allow extracting interesting lexical differences automatically, in a truly usage-based way and on a large scale. This offers the possibility to include variation that is too subtle to be easily discovered by manual inspection. This variation can then undergo linguistic analysis, an example of which is the research of the *into*-construction for British and American English by Wulff, Stefanowitsch and Gries (2007).

The method discussed here is a further refinement of the keyword-analysis based approach (Scott 1997), called Stable Lexical Marker Analysis (SLMA), that was first proposed in Speelman, Grondelaers and Speelman (2006). Keyword analysis is the identification of words that are representative for a chosen corpus. The corpus is effectively an aggregate of texts representive for two language varieties. Pluricentric languages, languages that have two official varieties, such as British English and American English, Portuguese and Brazilian, Netherlandic Dutch and Belgian Dutch, can be compared to each other to identify relevant lexical variation. The frequency information for words is analysed in parallel, using a statistical hypothesis test based on frequencies that captures information regarding a word's affiliation to a specific language variety. The word *diaper* for instance is typical for American English and will exclusively occur in American English texts, the word *nappy* is used in British English to refer to the same concept. The keyword analysis will identify the variants to be keywords for their respective variety of English.

A keyword analysis makes the binary decision whether a word is a keyword or not. There are two problems with this kind of approach. The first problem is that a binary categorisation is strongly simplified take on linguistic reality; it does not show the more graded scale which is appropriate when representing the markedness of a word vis à vis a language variety. More concretely in the context of determining whether a word belongs to a language variety, several possibilities of the occurrence of the word in the variety exist, that can be formulated in terms of this graded scale of markedness. The degree of markedness is for instance due to the spread of a word in a certain region; some words start off as generally accepted in one variety of a language, after which it is gradually accepted by users of the other language variety as well. Each of these statuses can be connected to frequency information obtained from the corpora. A word can be exclusive to one of both varieties and it is expected then to occur exclusively in that context. These words are unknown to the language users of the other variety. A second group of words are those words that are highly marked for a given variety, but that are nonetheless known to the users of the other variety. These words are marked by a highly significant difference in occurrence throughout the corpora. Finally the third group of words are those that are more variety-neutral.

A second problem with the keyword analysis has to do with the use of aggregated frequency over the entire corpus. Speelman (2006, 2008) and Gries (2009) point out that on top of frequency information, also information regarding distribution has to be included in the analysis when dealing with compiled corpora. Topical bias sometimes causes inflated frequency counts of certain words in a part of the corpus and as such the count does not reflect the actual status the word has in the language; its widespread use is a corpus-induced artefact. An example of topical bias it the temporary popularity of a word due to for instance a special event happening in the (local) news. Extensive coverage about an electoral period in America, might not coincide with an electoral period in England, which would cause 'election' to be identified as a typical American word. This is evidently not the case.

The problems with the keyword-analysis have been addressed by Speelman (2006) with the introduction of the Stable Lexical Marker Analysis (SLMA). To control for topical bias introduced in a part of the corpus, the method checks a word's consistency of use throughout the corpus. This is operationalized by subdividing the corpus and performing repeated hypothesis tests. A word can turn out to be a signficant *keyword* for a variety in all tests, in no test, or any number in between. This automatically introduces a graded scale of markedness. However, the procedure, though more fine-grained than one single all-or-nothing keyword-analysis, still uses repeated binary categorisation tests. The consequence of this operationalisation is a measure that results in extreme values for words and a limited graded scale. Although there is a continuous scale in principle, only a handful of the possible values actually occur. The operationalisation proposed in this paper abandons this paradigm further by including a more direct means of comparing frequencies, while it retains the benefits from integrating the anti-topical bias mechanism.

The remainder of the paper is structured as follows. In the first part the SLMA-method is explained technically, along with the improvements that have been introduced in the years after its inception. Then the material that is used to identify variety-specific words is described in more detail. In a third section the results are quantitatively scrutinized, by means of a compiled reference list both for Netherlandic and Belgian Dutch. The quantitative analysis will first show how the new implementation fares with regard to the earlier implementation, and then the results will be discussed in their own right. It is followed by a qualitative analysis, in the form of an error analysis of a sample of the obtained results. The final section sums up the findings of the third section.

2. The Stable Lexical Marker Analysis method

SLMA was developed in the cross section between corpus linguistics (Kilgariff 2001) and variational linguistics in the tradition of Labov (1972), and is used to identify so called *lexical markers* of different language varieties. It is conceptually based on the keyword-analysis introduced by Scott (1997). A keyword analysis uses frequency information of a word from two different corpora to assess whether a word is associated to one of them. The analysis uses a statistical test, the chi-square test, to verify the hypothesis that the distribution of the word is different for both corpora. If the p-value associated with the chi-square test is lower than a certain threshold (mostly .05 is chosen), it is unlikely that the difference in distribution can be attributed to chance and the word is identified as a keyword of the corpus. The stable lexical marker analysis method builds further on this idea. It also relies on statistical hypothesis -testing by comparing a word's frequency distribution in two corpora representative of two language varieties. There are two main differences. The first difference is the choice of hypothesis-test. The log likelihood ratio was chosen because it has been shown to provide a better p-estimate for somewhat lower values (Dunning, 1993). The second difference is based on the insight that a straightforward comparison between two corpora, based on traditional keyword analysis (Scott 1997) suffers from topical bias. The marker analysis score is calculated specifically to reflect the dispersion of a word, and hence the consistency and stability of its difference in usage between language varieties. To make it more concrete: two corpora (A and B), each of which is representative for a language variety might be divided into 8 parts: {A_1, A_2, ... A_8} and {B_1,B_2,... B_8}. The next step is a pairwise comparison between all of the A-members and all of the B-members: {A_1, B_1} , {A_1, B_2} , ... {A_8, B_8}. In each pairwise comparison, statistical hypothesis testing determines which words are lexical markers that occur significantly more frequently in the A-corpus as compared to the B-corpus. A scoring scheme is applied so that a word gets credit for each pairwise comparison in which it is a lexical marker. If a word obtains a maximum score over all pairwise comparisons, it is called a stable lexical marker. For the example above, there are 64 possible combinations between group A and group B so the maximum score is 64 and the minimum score is 0. This way, the analysis provides a ranking that assigns the highest scores to the words that most consistently occur with a significantly higher frequency in corpus A as compared to corpus-B. The formula to obtain the score is given below.

$$w = \sum_{a=1}^{n} \mathrm{S}_{\mathrm{AB}}$$

Where S_{AB} is a significant comparison between corpus A and B, n is the number of comparisons.

The original implementation of SLMA suffers from a sensitivity to extreme frequency counts, and even log-likelihood cannot deal well with low frequency words. Words with a relatively high frequency count are often falsely categorised (for our purposes) as markedly different, and for words with a relatively low frequency count, the method lacks the power to make a well-founded decision. An example of a high frequency lexical variable that would be attributed a high SLMA-score are the alternations of *toward* for British English and *towards* for American English. The word is used in both varieties and is not as extremely marked for either variety of English as the method would suggest. This contrasts with the choice for *nappy* or *diaper* as these words are variety-exclusive and would correctly score highly on the markedness-scale. Relatively low frequencies could nonetheless result in a score lower than expected.

In a first step to overcome these problems, a more fine-grained measure of markedness has been incorporated that on top of repeated significance testing, also takes into account the actual size of the difference in occurrence. This is called the effect size in statistical terms and takes the form of odds ratios that are averaged over each pairwise comparison of the subcorpora-frequencies that reaches significance. The averaged odds ratios capture the odds to which a word is associated to a corpus, as opposed to a statistical hypothesis test that simply states that a difference of occurrence exists throughout the corpora. A hypothesis test does not further distinguish between a difference of for instance the frequency pairs (1,100) and (40, 70). An effect size on the other hand, in the form of odds ratios would show the difference to be a hundred to one and seven to four. A further log transformation of the odds ratios improves the ease of interpretion of the results with regard to markedness. The scale after the log transformation ranges from high negative values to high positive values. Higher values mean stronger association. Evidently, negative association to one corpus implies positive association to the other one.

In a second step, the p-values used for the hypothesis test underlying the method have been calculated using Fisher's exact test (Pedersen 1996) for low-frequency words, for which the approximation of p-values of log-likelihood is not trustworthy. The formula to obtain the score becomes:

$$w = \log\left(\frac{\mathrm{FWa}\ /\ \mathrm{FWm}}{\mathrm{FWb}\ /\ \mathrm{FWn}}\right)^{*} \textstyle\sum_{a=1}^{n} \mathrm{S}_{\mathrm{AB}}$$

Where FW_{a} ,FW_{b} ,FW_{m} ,FW_{n} are respectively the frequencies of the word in corpus A and B, the frequencies of all words in corpus A and B, S_{AB} is a significant comparison between corpus A and B, n is the number of comparisons.

The method is applicable to various corpora for which associated lexical variables need extraction. The case study in this paper concentrates on extracting variables for Netherlandic and Belgian Dutch corpora.

3. Corpora, statistical analysis and reference material

Dutch is a pluricentric language with two standard varieties: Netherlandic Dutch spoken in the Netherlands and Belgian Dutch in Flanders.

A collection of comparable Netherlandic (400 million words) and

Belgian (1.3 billion words) national newspaper material from the period 1997-2005 has been used to test the method's performance in identifying the lexical peculiarities of each variety . Although we are aware that each language variety is represented by only one genre in this way, the newspaper material at least ensures us that standard varieties of the respective regions are scrutinized.

The corpora have been divided in 13 (the Netherlandic Dutch material) and 16 parts (the Belgian Dutch material), resulting in partitions of about 50 million words for the Netherlandic material and 100 million for the Belgian Dutch material.

Frequency information of words from the corpora has been used as input both for the original and adapted SLMA-method, which have been explained in section 2. The words obtain an SLMA-score and are ranked by the calculated continuous value that shows typicality for one variety at its head and for the other at its tail. In our implementation, positive scores show positive association to Belgian Dutch corpus material, while negative scores show positive association to Netherlandic Dutch corpus material.

The results are compared to reference lists of known variety-specific words: for Belgian Dutch the *Referentiebestand Belgisch Nederlands* was used. For Netherlandic Dutch we used the regional labeling in the *Prisma Handwoordenboek Nederlands*. Both lists have been manually compiled under supervision of prof dr. Willy Martin (Vrije Universiteit Amsterdam) and prof. dr. Willy Smedts (KULeuven). The Belgian Dutch material has been gathered by consulting lexicographical sources, corpora and informants and contains 1389 words. The labeling of the Netherlandic Dutch material has been carried out by language specialists. The Netherlandic list contains 2293 words. For Netherlandic Dutch only those words that according to the dictionary are labeled as Netherlandic Dutch on a lemma level have been included in the analysis.

For the qualitative analysis of words identified as marked by our statistical analysis but not included in the reference lists, we consulted the Prisma dictionary alongside online resources, such as google.

4. Belgian Dutch and Netherlandic Dutch lexical markers

The results of the analysis will be discussed in this section in various ways. First, a quantitative analysis will compare the performance of the old and new implementation of SLMA vis-à-vis the refefence materials. Both the the ability to identify lexical markedness and the coverage of words in the reference lists will be analysed. In a second section a qualitative analysis of typical examples will show the benefits and caveats of the new method's assessment of markedness.

4.1. Quantitative analysis

In a first quantitative assessment, we inspect the scores that the old and new SLMA implementations attribute to the words in our two reference lists. We expect the known Netherlandic and Belgian words to be separated by their SLMA-scores. High positive scores should be Belgian Dutch words, high negative scores Netherlandic Dutch. Neutral scores show neutrality of a word with regard to language variety.
The score distributions of both implementations are visually presented with boxplots in figure 1 and 2 respectively.

The boxplots contain a lot of quantitative information. The bold horizontal line in the middle of the box represents the median value. The pluses in the boxes are the average scores. The box itself contains half of the total amount of words in the list. The bottom line is the first quartile of data, the top line the third quartile. The dashed vertical lines flowing from the centre box, also called the whiskers, signify the largest and smallest SLMA-scores that lie within a 1.5 interquartile range from the box. Outliers are depicted by means of small circles and lie outside the range of the whiskers by definition.

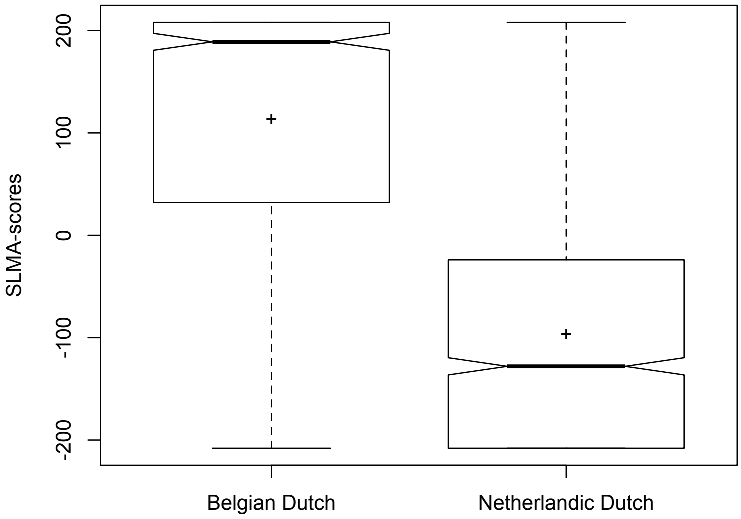

Figure 1. Boxplots of traditional SLMA-score of marked Belgian and marked Netherlandic words

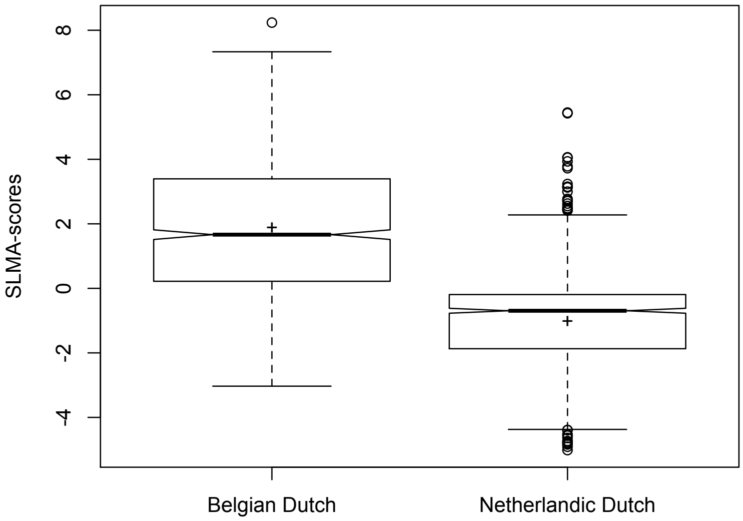

Figure 2. Boxplots of new SLMA-score of marked Belgian and marked Netherlandic words

The boxplots immediately show that the second method offers a more nuanced picture with regard to markedness. The range of possible values for the second method shows that it is less prone to attribute extreme values; whereas the range of possible values for the first figure ranges from -208 to 208 for words belonging to either variety, the second figure shows that the scores given to marked words for one variety do not take the extreme values attributed to words marked for the other variety. For example, Netherlandic Dutch words are not attributed very high positive scores in the second figure. This is the case however when we look at the first figure.

There is also a clear tendency for the earlier implementation to categorise the words as highly marked or not. This can be seen in the high median values and in the absence of a second whisker: three quarters of the Belgian words score higher than 180, three quarters of the Netherlandic material less than -110. The results in other words show how the all-or nothing fashion of a keyword-analysis is still present, albeit less pronounced. The second boxplot distributes the scores more evenly and with it makes a more fine-grained decision on markedness. The more even distribution also shows in the average values lying closer to the median value in the second boxplot. The fact that the first method shows a tendency to attribute very high scores to most words in the list, a desired attribute when dealing with middle range frequency words, is due to the nature of the underlying hypothesis test. For words with a high frequency profile the image of markedness is distorted, as the significance tests prove positive quite easily when confronted with a lot of *evidence*, in other words, with higher frequencies. The second method balances this by integrating relative frequency counts together with the significant dispersion the word shows in the subcorpora. The better results prove the benefits of implementing the effect-size by means of averaged odds ratios and show that relative frequency contains useful information when dealing with the markedness of a word with regard to a language a variety. It has to be said as well, that both methods not only contain information on markedness but also on prevalence in the language itself. Words with a lower frequency automatically have a lower score.

A logistic regression shows that both methods model the data better than the null-model and both models have a Wald p-value lower than .001. However, Nagelkerke's R^2 for the early SLMA-method is 0.53, whereas the newer method has an R^2 of 0.62 , proving a higher adequacy to model the data. An analysis of variance of the residual deviance of both regressions, with a p-value below .001, also shows that the difference between the models cannot be attributed to chance.

The second boxplot shows that the Belgian Dutch words exhibit a higher SLMA-score than the Netherlandic Dutch words. On average the Belgian Dutch words have a score of 1.89, while the Netherlandic Dutch words have a mean value of -1.01. Half of the words are clustered together around the mean values. The smaller upper half of the Netherlandic Dutch center box, compared to the equal size of the two parts of the Belgian box, show that more Belgian Words are assessed as neutral by the method than the other way around. About 25 percent of the words seem to be wrongly assessed by the method for each language variety. Finally it can be seen that the number of outliers are much higher for the Netherlandic material than they are for the Belgian material.

We can speculate that the higher SLMA-scores for the Belgian words and the many outliers for the Netherlandic words are caused by the greater size of the Belgian corpus.

4.2. Qualitative analysis

The coverage of the corpus material with regard to the reference lists is discussed before zooming in on the words from the reference lists that obtained an SLMA-score. First an explanation is sought why some words are not present in the corpora.

The reference list of Belgian Dutch words counts 1389 words. 260 of those are not covered by the material. Examples of Belgian Dutch words not found in the corpus such as *tempeest, turfkantoor, bedpan*, *paardenoog* and *vaderkensdag* will show there are several reasons for this.

Tempeest is the old Belgian Dutch word for *thunder storm*. However, as such it is hardly used anymore. Belgian Dutch wordlists are known to include classic examples of Belgicisms. They are known to the literate reader, but hardly present in actual language use. *Turfkantoor* is an office where horse betting is done. The highly specific context in which it is used, makes it known only to a select audience. In the newspaper material we have at our disposal the term does not occur however, showing that the popularity of concept, probably due to the low amount of news articles covering the act of betting on horses, is rather low. *Bedpan*, with the same meaning as *bedpan* in English, is also rather inpopular, due to the fact that the object itself is hardly used anymore. *Paardenoog*, a rather informal variant of *spiegelei*, a *fried egg*, will not easily be found in a newspaper context because it is known by Belgian language users to have a more appropriate Standard Dutch equivalent. *Vaderkensdag*, *Father's day* in its turn is probably the word in the list which strikes the reader of Dutch to be most dialectic in nature. Again the status of the word makes it unlikely to be used in a formal context, and even in colloquial Dutch, it seems unlikely to us to be frequently used.

In short, the reasons for these words not to occur in the newspaper material is one of low usage in formal written language. The cause of this could be on the one hand that the actual word is not frequently used in the language, because of its archaic status, or because of a highly specific context often not known to the general public. On the other hand it could be that the word is not used much in a written and formal context. Often the more Standard Dutch equivalent has taken its place and is anchored more deeply than the marked equivalent.

For the Netherlandic Dutch words, of the 2293 words covered by the reference list, 1096 are not found in the corpora. The coverage for Netherlandic Dutch words, taking this specific reference list, is therefore much lower than the coverage for the Belgian material. Similar reasons brought up for the Belgian material can explain some words not being covered. *Buuf* an informal designation for female neighbour is unlikely to occur in a written context. Words such as *elfstedenkoorts*, the hype that surrounds a rare yet very popular ice skating event in the Netherlands, are not found due to the event not having taken place during the period for which we have Netherlandic Dutch newspaper material. Moreover, the reference list of Netherlandic Dutch is more up to date than the Belgian Dutch reference list. A word such as *polderblindheid*, blindness induced by a monotone landscape, is a fairly new word (every occurrence found in google is accompanied by a definition, showing that the meaning of the word is not naturalised yet) and is probably therefore not covered.

Finally some seperable verbs do not seem to be presented for both language varieties. The lemma found in the reference list, is not found in the corpus as a whole, but only in its seperated form. Parsers are meant to overcome this problem, but are not succesful in identifying all relevant verbs. Examples (English translation in square brackets and other language equivalent in round brackets) are *aanwippen* [bringing a short visit to someone] (*binnenwippen*) and *aanplempen* [to fill up].

In the discussion about those words that are covered by the corpora a distinction will be made between lexical items that are marked and identified by the method as such, that seem to be unmarked but are identified by their SLMA-score as belonging to a variety, that are marked but are wrongly classified by the method as marked for the other variety, words that are neutral and that show up with a neutral SLMA-score, items that according to the attributed SLMA-score are marked but are neutral according to the reference lists and finally words that have neutral scores but appear to be marked.

The quantitative analysis already showed a large number of words to be correctly classified by the method's scoring mechanism and hence the method's ability to capture information concerning markedness. Words in the reference list that are effectively identified by our method as marked for Belgian Dutch (English translation between square brackets, Netherlandic alternative for the same concept between round brackets) are: *werkonbekwaam* [unable to work] (*arbeidsongeschikt*) , *werkkledij* [working clothes] (*werkkleding*, *beroepskleding*) and *plaasteren* [to plaster] (*pleisteren*, *stukadoren*). For Netherlandic Dutch obeservations include (Belgian alternative between round brackets): *vluchtstrook* [emergency lane] (*pechstrook*), *korenwolf* [common hamster] (*gewone hamster*) and *sappelen* [to tire out] (*afbeulen*). These words will not immediately be understood by language users of the other variety.

Other words are not included by the reference material but show as marked by our method. The following enumeration makes clear that the reference lists are incomplete and that a further analysis of the words would make them eligible candidates for list-inclusion in order to improve its coverage. A large number of words in this category are proper names of locations, streets and local celebrities. A second group of words show a slightly different spelling in either variety, e.g. *tornooi* and *toernooi* [tournament], but also more systematic differences such as an apparent different use of hyphens in collocations, or a different use of suffixes as in *schuimig* and *schuimachtig* [foamy]. A third group of words are exclusive for the language variety for which the method shows marked scores. They could also be marked due to a different meaning for the same surface form, or again because the concept is more popular in one of both regions.

Some examples of exclusive words for Belgian Dutch are *wachtbekken* and *carpoolparking*. *Wachtbekken*, an area used as a natural water buffer to prevent flooding of other areas, has the Netherlandic

Dutch equivalent *bufferbekken*. *Carpoolparking* has as a Netherlandic Dutch counterpart *park-and-ride*.

An example of a more popular concept is that of *trekpaard*. *Brabants trekpaard* is a collocation marked for Belgian Dutch simply because more of these animals roam Flanders.

Finally some of the words are striking at first glance because no dictionary records any difference in meaning. When looking at the meaning of these words on the internet it becomes apparent that they are different nonetheless. This group of words shows most clearly how an automatic identification method can help in retrieving interesting differences between the two varieties that are yet either unknown, or not recorded yet.

The word *uitdrijving* in Belgian Dutch is used not only to describe the act of exorcism, but also the forced eviction of tenants. In the Netherlands the first meaning is shared, the second one is not strictly the same, it is only used when people are evicted from a place on a grand scale, and a third one is giving birth to someone. A more subtle difference in meaning can be found with *verdringen*, literally *to set aside*. The Dutch seem to set aside objects in a very literal way; they actively make room and conquer the occupied space. The Belgians first of all set aside mental activities, emotions and memories. Another word that to the unknowing reader seems unmarked is *tijdelijkheid*. A Flemish person would say it is the equivalent of *temporarilarity* in English, for the Dutch it means *fitting for the time setting in which it originated* and is mostly said about architecture.

A closer look at words that are included in the reference lists, but whose SLMA-score suggests that they are marked for the other language variety unanymously point to the inability of the method to deal with homonymy and polysemy.

Belgian Dutch words exhibiting negative SLMA-scores are for instance *schoon, doctorandus* and *noemen*. *Schoon* is often used in the meaning of *beautiful* in Belgian Dutch and *clean* in Netherlandic Dutch. *Doctorandus* in the Netherlands is a person that has obtained his master's degree, while in Flanders it is a person who is pursuing his doctoral degree. *Noemen* then in Netherlandic Dutch is used when you give a name to something or someone, while in Belgian Dutch it has both this meaning and the meaning of being called a certain name, *heten* is the Standard Dutch equivalent. The popularity of concept can then further explain the inclination towards negative scores. Words marked as

Netherlandic Dutch by the reference list are *boom*, *syndicus* en *lijstduwer*. *Boom* is homonymous in the meanings of *tree* and the English *boom*. Our reference list designates the latter meaning as more Netherlandic Dutch. A *Syndicus* in Flanders can be any of the following persons: the janitor of a building, an official representative of the judicial executor, or a provisional trustee. In the Netherlands it is a civil servant charged with giving advice to local authorities. Finally *lijstduwer* is a politician that is supposed to attract votes for his party. In Belgium it is an election candidate that is always mentioned at the very bottom of the voting list, and as such one could say he supports the list. In the Netherlands the candidate could be positioned anywhere in the list, but he is mostly ranked in a way that it is unlikely that he were ever to be elected.

These examples actually show that these words are indeed also marked for the variety the method suggests. The designated concept is more widespreadly used in one of both varieties however, which results in a a marked score for that variety.

Most words are found to be neutral: horloge [watch], eindredactie [final editing], verbrokkelen [to crumble]. It is theoretically possible that some of these words turn out to be marked as well, though it is difficult to verify. A manual inspection of a limited sample did not turn up any of these cases.

Then finally there are those words that are neutral according to the method but are marked according to the reference lists. Examples for Belgian Dutch are: *schachtendoop* [initiation] (*ontgroening*), *aanklagen* [to indict] (*ten laste leggen*) and *verwikkeling*, which has a rather specific meaning in Belgian Dutch: a complication that occurs on top of an initial disease. The meaning in Netherlandic Dutch is synonymous to *verwarring, moeilijkheid* [complication, infiltrated by]. Two of these words have a rather low frequency in the Belgian Dutch corpus; *schachtendoop* occurs 24 times and *aanklagen* has a frequency of 5. This results inevitably in a very low SLMA-score as there are only few comparisons that will reach significance when this total frequency is further divided among the subcorpora. The reason that *verwikkeling* has a low SLMA-score cannot be attributed to its low frequency (90 in fact). It is however again a polysemous word used differently in the two language varieties. The frequency of the two uses is equally high which results in a neutral score. Marked Netherlandic Dutch words have a neutral SLMA-score for the same reasons. *Dieplader* has a very low frequency and *gecoiffeerd* is a polysemic word with the Belgian Dutch meaning of *cut hair* and the Netherlandic Dutch meaning of *to flatter*, *to praise*.

Table 1. Overview of categorisation possibilities for SLMA-scores with regard to markedness according to the reference lists

SLMA-score	Reference Lists	Belgian Dutch	Netherlandic Dutch	Reasons for divergence
Marked	Marked for the same variety	werkonbekwaam	vluchtstrook	-
	Marked for the other variety	doctorandus	syndicus	Polysemy, homonymy: one of both concepts is more popular in a variety
	Neutral	uitdrijving	tijdelijkheid	Not included in reference lists
Neutral	Neutral	horloge, verbrokkelen		-
	Marked for a variety	verwikkeling, aanklagen	gecoiffeerd, dieplader	Polysemy, homonymy: both concepts are equally popular in the varieties
				low frequency words

5. Conclusions

This paper explained how SLMA can be used as an automatic way to extract marked items for a language variety. The method has been technically discussed and the results have then been analysed. The quantitative analysis used known reference lists to assess whether the method correctly classified known marked words. It has been shown that it is mostly successful, and that the newer implementation of SLMA provides a more nuanced scoring mechanism than the older one. The qualitative analysis showed that the corpora do not cover all the words found in the reference lists due to several reasons; the concept can be unpopular, hardly known, archaic, or simply not used in a written context. The words that have an unexpected SLMA-score are shown to be polysemous or homonymous. Finally the analysis also showed that the method identifies words that are not yet covered by dictionaries (or at least not by our reference lists), and have a markedness that may yet be unknown to language users from the different regions.

Future work would benefit from an extension across the lines of implementing word meaning in order to disambiguate polysemous and homonymous words. A second step in the form of the identification of synonyms would further contribute to the implementation of word meaning. Peirsman (2010) shows promising results in the identification of synonyms across language varieties by using vector space models. Vector space models calculate semantic relatedness between two words on the basis of the contexts in which those words occur. The profile based approach to lexical variation can then use the results of an automatic identification of lexical items in combination with the automatic detection of synonyms to constitute a word's profile.

References

Dunning, Ted
1993 Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1): 61-74.

Geeraerts, Dirk, Stefan Grondelaers, and Dirk Speelman
1999 *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen [Convergence and divergence in Dutch vocabulary. An enquiry into clothing and football terms].* Amsterdam: Meertens Instituut.

Geeraerts , Dirk
2009 Lexical variation in space. In *An International Handbook of Linguistic Variation*, Peter Auer and Jürgen Erich Schmidt (eds.), 821-837 Berlin/New York: Walter de Gruyter.

Gries, Stefan
2009 Dispersions and adjusted frequencies in corpora: further explorations. In *Corpus-linguistic applications Current studies, new directions*, Stefan Th. Gries, Stefanie Wulff, and Mark Davies (eds.), 197-212. Amsterdam/New York: Rodopi.

Kilgarriff, Adam
2001 Comparing corpora. *International Journal of Corpus Linguistics* 6 (1): 97-133.

Labov, William
1972 *Socioloinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Pedersen, Ted
1996 Fishing for exactness. *Proceedings of the South-Central SAS Users Group Conference, Texas:* 188-200.

Scott, Mike
1997 PC Analysis of Key Words -- and Key Key Words. *System 25*(1): 1-13.

Silva, Augusto Soares da
2006. Convergência e divergência no léxico do Português Europeu e do Português Brasileiro: resultados do estudo sobre termos de futebol e de moda [Convergence and divergence in the lexicon of European and Brazilian Portuguese: results of an enquiry into football and fashion terms]. In *Textos Seleccionados do XXI Encontro Nacional da Associação Portuguesa de Linguística,* 633-646. Lisboa: Associação Portuguesa de Linguística.

Speelman, Dirk, Stefan Gondelaers, and Dirk Geeraerts
2006 A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch. In *Corpus Linguistics around the World,* Andrew Wilson, Dawn Archer and Paul Rayson (eds.), 195-202. Amsterdam: Rodopi.

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts
2008 Variation in the choice of adjectives in the two main national varieties of Dutch. In *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems,* G. Kristiansen and R. Dirven (eds.), 205-233. Berlin/New York: Mouton de Gruyter.

Peirsman, Yves, Dirk Geeraerts and Dirk Speelman
2010 The Automatic Identification of Lexical Variation between Language Varieties*. Journal of Natural Language Engineering* 16 (4): 469-491.

Wulff, Stefanie, Anatol Stefanowitsch, & Stefan Th. Gries.
2007 Brutal Brits and persuasive Americans: variety-specific meaning construction in the into-causative. In *Aspects of meaning construction,* Günter Radden, Klaus-Michael Köpcke, Thomas Berg, & Peter Siemund (eds.), 265-281. Amsterdam/Philadelphia: John Benjamins.