Putting Meaning on the Map: Integration of geographic and semantic variation in multivariate models of language use

Natalia Levshina
University of Marburg
Research Centre 'Deutscher Sprachatlas'
natalevs@gmail.com

Abstract

The paper provides an overview of the recent developments in (dia)lectally enriched lexical and constructional semantics. These developments reflect the quantitative turn that has been taken by dialectology and linguistics in general, as well as the active integration and cross-fertilization of different linguistic disciplines. A vivid manifestation of the latter development is multifactorial models, which reveal an interplay of various semantic, pragmatic, geographical, social and other factors of language use. The paper focuses on the onomasiological (from meaning to form) and semasiological (from form to meaning) studies that take into account geographic variation. It also discusses the achievements of the aggregate lexical (dia)lectometry based on the onomasiological profile approach.

1. Recent trends in the studies of dialectal variation

Like most linguistic disciplines nowadays, contemporary dialectology has recently been undergoing two major changes. On the one hand, it has taken a rigorously quantitative turn. On the other hand, we are witnessing a rapid integration of dialectology with the neighbouring disciplines.

The quantitative turn has manifested itself in the move from 'traditional' dialectology to dialectometry (see Szmrecsanyi, this vol.), which measures linguistic distances between geographic locations. These distances are computed by aggregating the overlap across many individual variables (e.g. regional phonetic or lexical variants) in the speech of informants in different locations, as it was done in a pioneering study by Séguy (1971). However, the differences between geographic variants are often a matter of degree. For instance, [zɛn], a variant of Dutch *zijn* “to be”, is more similar to [zɛ'n] than to [zɪnt], and all three variants are more similar to one another than to [bɪn] (see the data in Heeringa & Nerbonne 2001). To take such continua into account, dialectometrists compute acoustic/articulatory distances between phonemes or Levenshtein distances between the geographic variants of lexemes (e.g. Nerbonne et al. 1996). The distances between the locations are next added up across the words or phonemes and can then be represented visually and analyzed. One of popular methods is a cluster analysis of the locations, which allows to identify dialect regions (Figure 1).

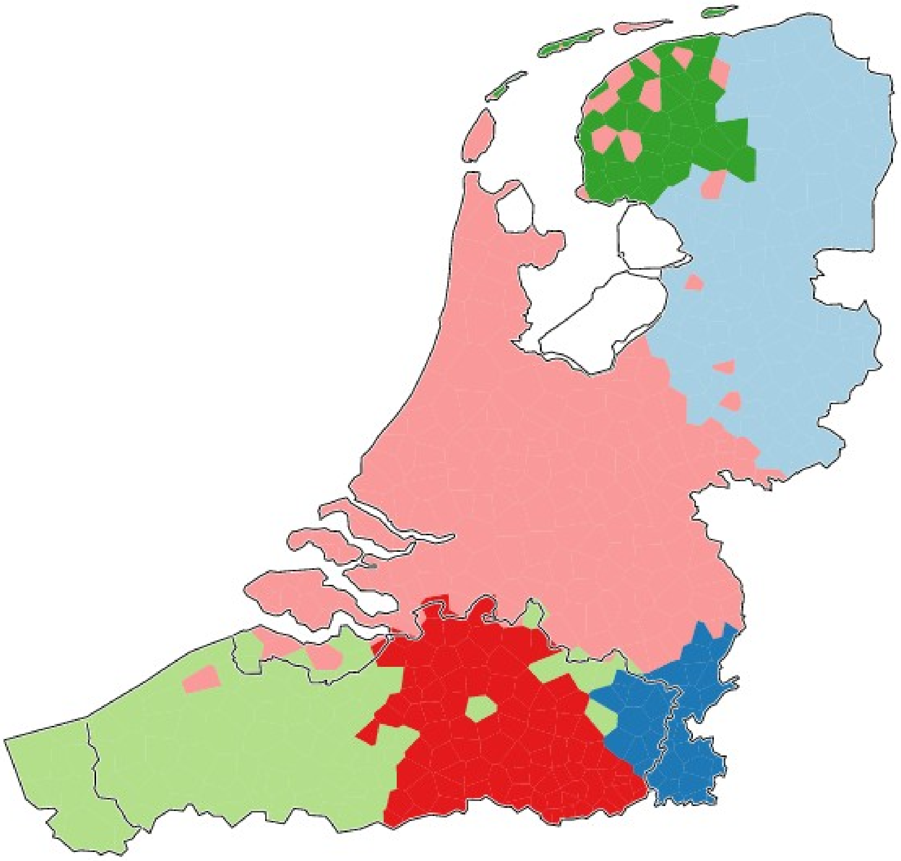


Figure 1. Six dialect groups in th Dutch-speaking area (Prokić et al. 2012). The clustering is based on the diferences and similarities in the pronunciations of 562 words in 613 locations.

This method is currently being developed in several directions. First, in most dialectometric studies, which are based on data aggregation, one actually loses the information about the individual variables that enter the analysis. Several methods to overcome this drawback have been developed, e.g.

Prokić 2007; Grieve et al. 2011; Prokić et al. 2012. For instance, Grieve's approach, which is based on spatial correlation measures and factor analysis, allows one to represent different dimensions of dialectal variation, retaining information about the initial linguistic variables, whereas Prokić et al. (2012) identify the regional shibboleths (distinctive variants) of Dutch and German dialect areas. Another pertinent question is how to incorporate the social characteristics of individual speakers in dialectometric models, as it was done in traditional dialectology. A possible solution has been offered by Wieling (2012), who uses mixed-effect regression modelling with a few demographic variables as fixed effects.

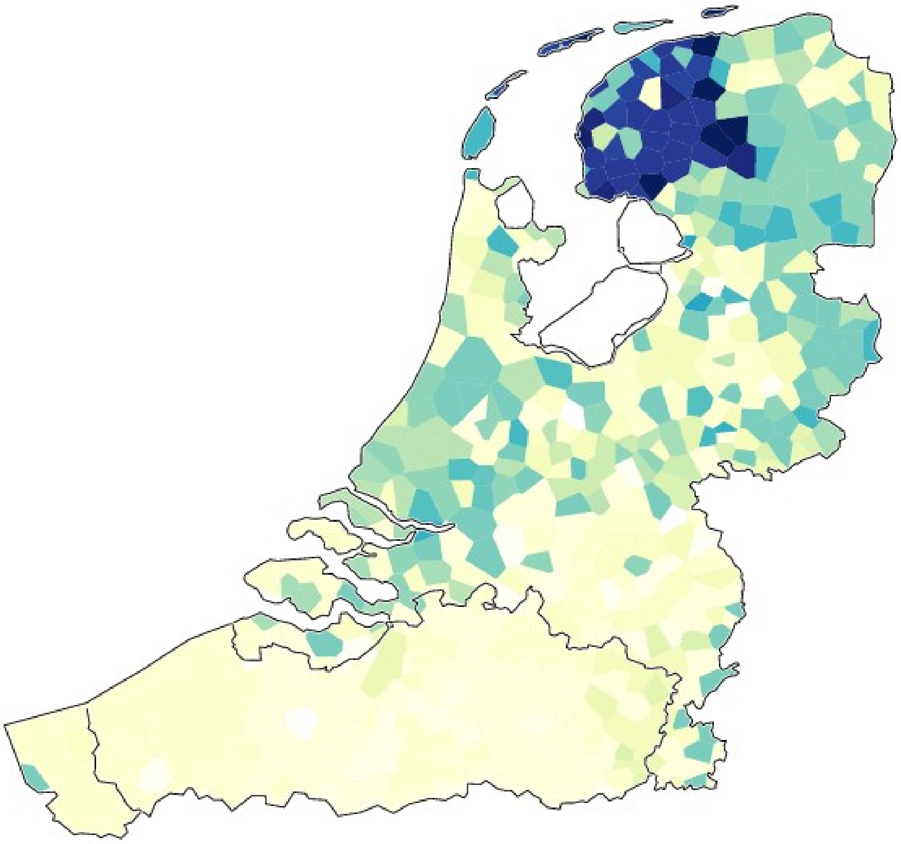
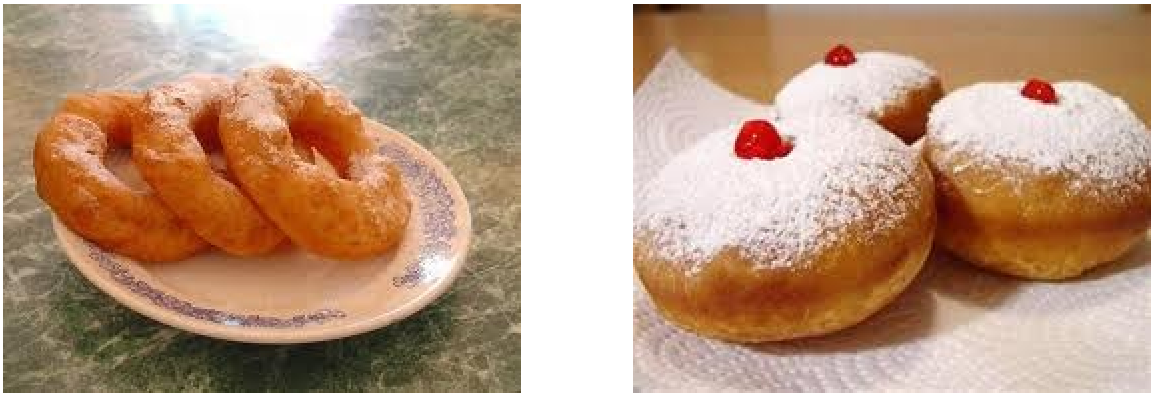


Figure 2. Dutch dialect areas based on the pronunciation of the word *vrijdag* "friday" (Prokić et al. 2012). The word *vrijdag* is the most distinctive of the Frisian area (the dark region on the map).

On the other hand, the boundaries between dialectology and neighbouring disciplines are also becoming more flexible. Probably, the most natural and fruitful connection is between dialectology, typology and areal linguistics (cf. this volume). Provided sufficient amounts of data are available, quantitative dialectometric models can be easily extended beyond one language. For instance, Gooskens and Heeringa (2004) consider the relationships between several geographic varieties of Frisian, on one hand, and other Germanic languages, on the other hand. This enables them to model the dialect continuum in Frisian and also to interpret the distances between the Frisian varieties and other Germanic languages in terms of genetic relationships and language contact. Another interdisciplinary example is perceptual dialectology, which integrates dialectology, social psychology, psycholinguistics and sociolinguistics. Basically, it is 'folk dialectology', which deals with beliefs, attitudes, stereotypes that ordinary speakers have about dialects (e.g. Preston 1999; Berthele 2010).

This article will focus on yet another recent development. Although there has been substantial progress in the dialectal studies of phonetic, and some morphological and syntactic alternations, which did not involve (obvious) semantic differences between the geographic variants, it is clear that there exists substantial cross-dialectal variation in the meaning of words and constructions. For instance, in Pskov dialects the noun *pivo* denotes any kind of beverage, whereas in contemporary literary Russian it means “beer” (Lukjanova 2010). The dialectal variant reflects a more ancient meaning, which has become specialized. On the other hand, the same concept can be denoted in different dialects by various words and constructions. An example is the famous difference in the way of naming a doughnut with a hole in Moscow (*pončik*) and Saint Petersburg (*pyška*). Note that one has chances to hear the word *pončik* in Saint Petersburg, as well, because the word also refers to round fried pies, sometimes with filling (see Figure 3).



pončik (Moscow) / *pyška* (Saint Petersburg)

pončik (both cities)

Figure 3. Semantic variation of *pončik* and *pyška* in Moscow and Saint Petersburg.

Such semantic differences are often subtle and probabilistic. They involve varying degrees of entrenchment and salience of semantic features and lexical units, rather than crisp and clear distinctions (cf. Geeraerts et al. 1994). The question is then, how to disentangle the semantic and geographic factors in a robust empirical model of language use? There exist several solutions, depending on the specific task and perspective of the study. They can be classified into three main groups, which are outlined below (examples are provided in the following sections):

1. Onomasiological studies of near-synonymous words and constructions. That is, they model the speaker's choice between different naming opportunities, establishing the role of the above-mentioned factors in predicting the use of an expression in a given context. The weights of these factors are compared across the lects with the help of logistic regression or other statistical techniques. Such analyses normally represent integrative quantitative models of language variation, which include geographic, social, semantic and pragmatic factors. This is why the approach is often referred to as multifactorial grammar. These studies are based on comparable corpora, which contain balanced samples from different language varieties, commonly referred to as “lects” – an umbrella term for any geographic and social language varieties, including idiolects.
2. Semasiological studies, which focus on the variation in the use of a word or construction in different lects. These analyses differ in the number and nature of the variables which serve as a *tertium comparationis*. Some of them are based on collocational analysis, as the so called distinctive collexeme analysis developed by Gries and Stefanowitsch (2004), which compares the slot fillers of a construction in different lects. Some other methods are multifactorial and involve large sets of heterogeneous variables.
3. Aggregate lectometric studies, which measure cross-lectal lexical differences across the concepts that belong to a specific lexical field. One of the best-known approaches is based on so called onomasiological profiles, which consist of the lexical terms that designate a specific concept, and their relative frequencies in two or more language varieties.

The above-mentioned methods are based on several fundamental assumptions. First, corpus data are seen as a reliable source of evidence about our knowledge of a language because they represent non-elicited usage in natural settings. Second, they are based on the distributional hypothesis, which says that natural language semantics can be modelled with the help of contextual features found in large amounts of text data. This idea is expressed in John Firth's famous maxim “You shall know a word by the company that it keeps" (1957: 11). One of the early implementations of the distributional approach in lexical semantics is the distributional analysis of verbal semantics by Apresjan (1966).

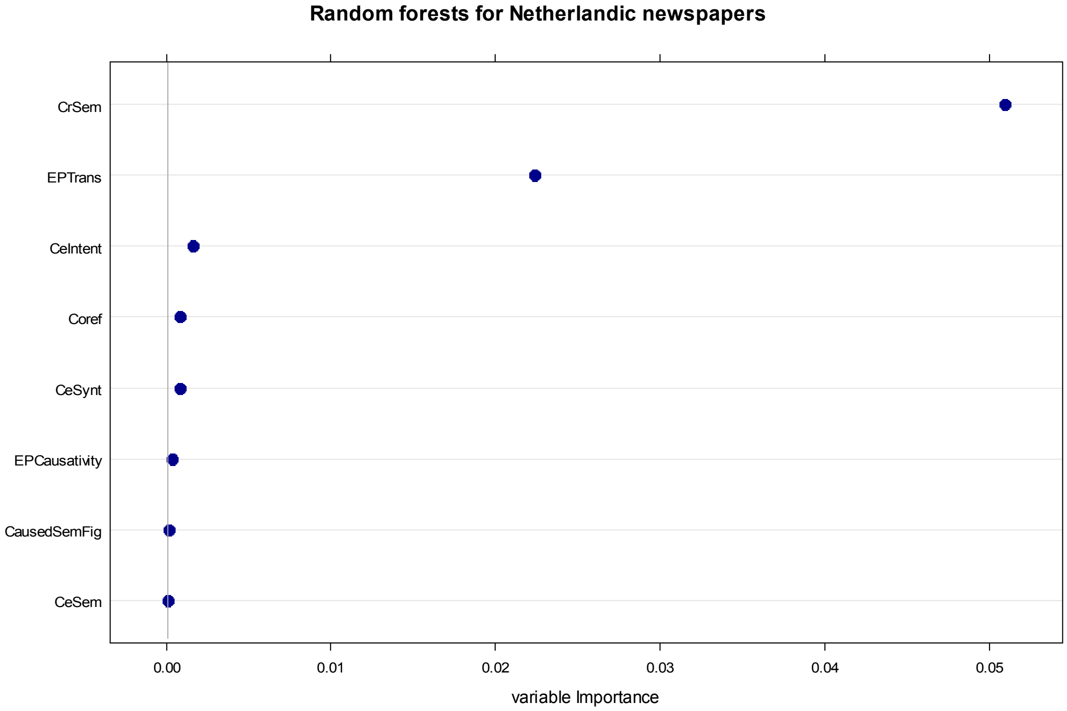
The structure of the remaining part of the paper is as follows. The next section provides examples of lectally enriched onomasiological studies of specific words and constructions. Section 3 describes some quantitative models of semasiological variation of words and constructions in different lects. Section 4 offers an overview of studies based on aggregation of cross-lectal differences in the lexical choices. Finally, Section 5 provides a summary and suggests some paths for further investigation.

2. Modelling the speaker's choice: onomasiological multifactorial grammar

The multifactorial onomasiological studies that model the speaker's choice between two or more constructions deal mainly with the national varieties of English and Dutch. They usually focus on abstract syntactic constructions. A well-known example is a series of papers on the English dative alternation by Joan Bresnan and her colleagues (Bresnan et al. 2007; Bresnan & Hay 2008; Grimm & Bresnan 2009; Bresnan & Ford 2010). The alternation involves two variants: the double object construction, as in *Mary gives John the book*, and the prepositional *to*-dative, as in *Mary gives the book to John*. In Bresnan et al. 2007, the authors show that the use of one or the other construction can be predicted with high precision by contextual factors. The multivariate analyses (logistic regression) provide clear evidence of a 'harmonic alignment' of distinct but related features. Namely, discourse givenness, animacy, definiteness, pronominality and relative length of the postverbal participants (the recipient and the theme) are aligned with the position that follows the verb first (the recipient in the double object construction, and the theme in the prepositional dative construction), whereas non-givenness, inanimacy, etc. are observed in the final position in both constructions.

The follow-up studies of Bresnan et al. 2007 show that although the main division of pragmatic labour between the constructions holds across many national varieties of English, there are subtle yet significant differences in the effect size of different variables. This suggests that the speakers of different varieties may be more or less sensitive to particular contextual cues. For instance, a comparison of size effects for the dative alternation with the verb *give* in American and New Zealand English (Bresnan & Hey 2008) demonstrated that New Zealand speakers are more sensitive to animacy than Americans, and that this difference is a result of a gradual language change.

Other examples of slightly different constraints between lects have been observed in the use of the *'s*- vs. *of*-genitive in several varieties and registers of English (Szmrecsanyi 2010), in the presence or absence of presentative *er* “there” in Belgian and Netherlandic Dutch (Grondelaers et al. 2002), and in the semantics of the Dutch analytic causatives with auxiliaries *doen* “do” and *laten* “let” in Netherlandic and Belgian (Flemish) newspapers (Levshina 2011). Figure 4 shows the effects of semantic and syntactic variables in predicting the choice between the auxiliaries for the Netherlandic and Belgian newspaper data. The effect sizes were measured with the help of random forests (cf. Tagliamonte & Baayen 2012). The further the point from the origin, the more important the variable.



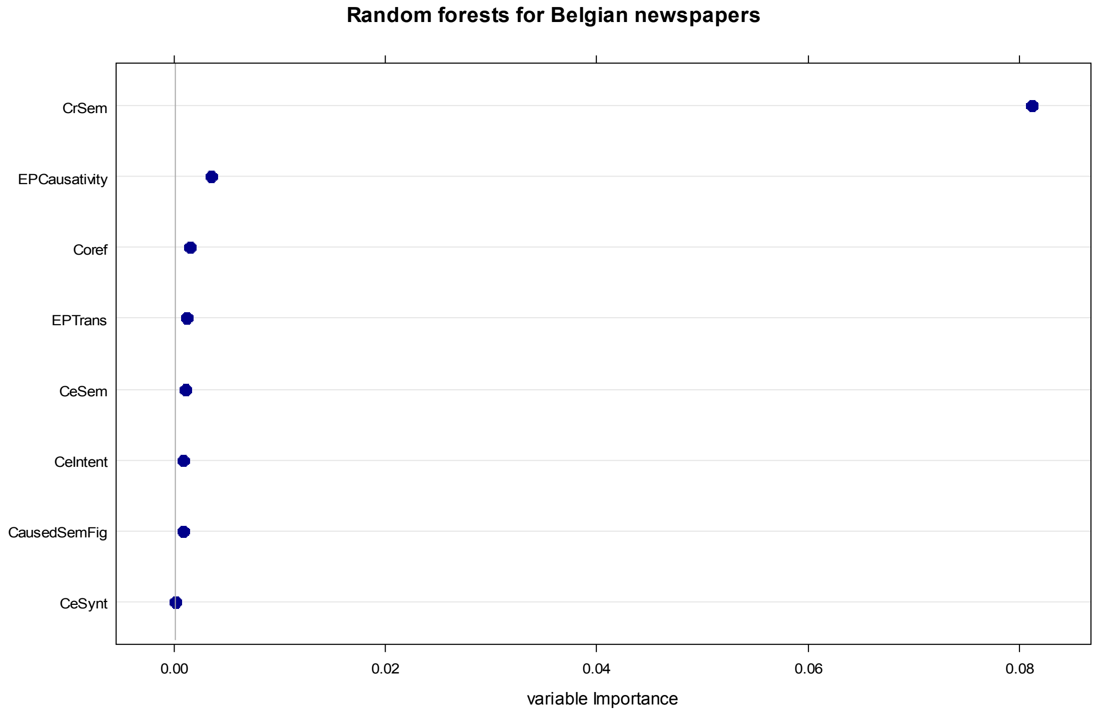


Figure 4. Different effect sizes of contextual variables in the choice between two Dutch causative verbs in the newspaper register. Above: Netherlandic data; below: Belgian (Flemish) data.

The plots show that the effect of transitivity of the effected predicate *EPTrans* (e.g. *ontwerpen* "design" in *Hij liet zijn huis ontwerpen* "He had his house designed") plays a much stronger role in the choice between the two constructions in the Netherlandic newspapers than it does in the Flemish ones. A closer inspection, however, shows that this is a result of the high frequency of the collocations of *laten* with the transitive verbs *weten* "know", *zien* "see" and *horen* "hear" in the Netherlandic newspapers. This demonstrates an important role of lexically specific exemplars in constructional variation. Such lexical differences may arise due to many possible factors, for instance, different degrees of importance of particular referents and their characteristics in language communities, or different degrees of routinization of particular linguistic expressions, which makes them more or less readily available to the speakers in the process of communication. The inventory and relative weight of such factors in language use are still to be established.

Quantitative lectally enriched models of lexical variation are, unfortunately, quite rare in contemporary usage-based linguistics. One of the few examples is Glynn's (2009) study of near-synonyms *annoy*, *bother* and *hassle* in British and American English. However, he does not find substantial geographic differences in the division of semantic labour between the verbs.

3. Modelling lectal differences in the semantic structure of words and constructions

3.1. Lectally oriented distinctive collexeme analysis

Distinctive collexeme analysis developed by Gries & Stefanowitsch (2004) is based on the analysis of corresponding constructional slot fillers in near-synonymous constructions (e.g. the double object dative vs. the prepositional dative) in different lectal variants of the same constructions. It has been applied by Wulff et al. (2007) to model the semantic differences between the British and American variants of the V-*into*-V causative, as in *He tricked me into employing him*. The authors compare thousands of verb pairs, which fill in the first and the second verb slots and employ a statistical test (Fisher's exact test) to identify the verbs that are the most typical of one or the other lect. For instance, many verbs distinctive of the British variant of the construction (more precisely, those that occupy the first verbal slot, as *trick* in the example above) designate negative emotions (e.g. *She terrified me into doing it*) and threatening (*He blackmailed me into doing it*). These semantic classes are not typical of the American collexemes. On the other hand, the distinctive American verbs often refer to communication (e.g. *She talked me into doing it*), with a total lack of such distinctive collexemes in the corresponding slot of the British construction. Wulff et al. hypothesize that these and other differences may reflect the varying degrees of entrenchment of specific semantic frames in the two cultures, hence the title of their paper, “Brutal Brits and Persuasive Americans”.

A similar analysis has been carried out in Levshina et al. 2009, where the Netherlandic and Belgian variants of the above-mentioned causative construction with *doen* were compared. The analyses of the constructional slots revealed that the Netherlandic variant seems to be restricted to the so called affective causation, i.e. situations when a stimulus produces a mental reaction, e.g. *Zijn kapsel doet me denken aan een vogelnest* “His hairstyle reminds me of (lit. makes me think) of a bird's nest”. Since Netherlandic Dutch is traditionally considered to be the leader in various processes of language change, and Belgian Dutch is believed to be more archaic, the more limited semantic repertoire of Netherlandic *doen* ties in well with the previous observations of the ongoing qualitative and quantitative shrinking of the auxiliary *doen* in Dutch (e.g. Duinhoven 1994; Verhagen 2000).

3.2. Multivariate comparisons of the semantics of lectal variants

The multivariate data of the type described in Section 2 can also be used to compare the semantic differences between different lectal variants of the same word or construction. One of the examples is Glynn's (In press) analysis of the British and American differences in the polysemy of *annoy*. Like in the above-mentioned study of verbs *annoy*, *bother* and *hassle*, he uses multiple correspondence analysis to explore the relationships between various semantic features and the national varieties, but this time he zooms in only on one specific verb, *annoy*. The analyses show that although the central sense ("anger") is shared by both varieties, there are some lectal differences on the semantic periphery of the verb. More specifically, the American variant often denotes "hurt", which is associated with serious topics, e.g. anxiety, personal relationship troubles or family concerns. The British variant is more often used in the sense "tire/interrupt", often in humorous way, being also associated with such features as familiarity, interruptions and expenditure of energy.

Multivariate methods, such as cluster analysis and multidimensional scaling, also allow one to compare the general semantic organisation of lectal variants. For instance, in an analysis of corpus data coded for 35 semantic, morphological, syntactic and other variables, Levshina (2012) found that the Netherlandic causative *doen* (see above) is not only less frequent and less semantically diverse than its Belgian Dutch counterpart, but it also has a different structure, with a tight cluster that corresponds to affective causation, and only sporadic uses of the other senses. This is shown in Figure 5, which represents the individual exemplars of the construction in the common semantic space created with the help of a multidimensional scaling analysis. Note that the closer the exemplars on the map, the more semantic features they share. The results suggest that the general schema of the Netherlandic causative has become weaker than that of the Belgian variant. This finding is in line with the process of *doen* gradually becoming obsolete.

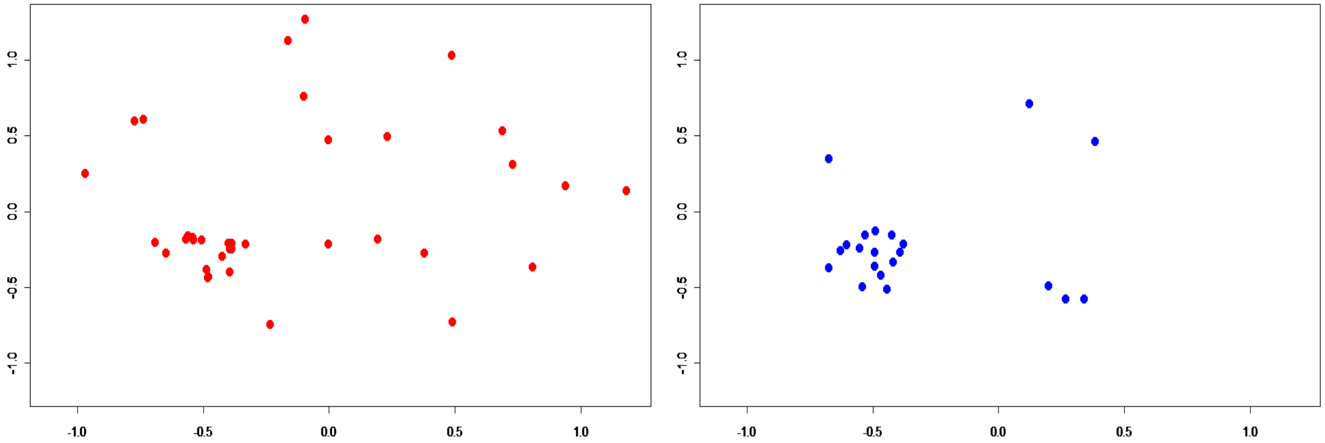


Figure 5. The semantic space of the causative *doen* in Dutch with the Belgian (left) and Netherlandic (right) exemplars (spoken data). The Belgian causative auxiliary demonstrates greater semantic variation and productivity than its Netherlandic counterpart, which is almost exclusively represented by the expression *doen denken aan* "remind, make think of" (the dense cluster on the left).

4. Aggregate lexical lectometry: onomasiological profiles

The onomasiological profiles approach was developed in Geeraerts et al. 1999 and tested on two lexical fields – clothing and soccer terms. The data represented several lectal dimensions: geographic (two Belgian vs. two Netherlandic locations), temporal (1950, 1970 and 1990) and register-related (magazines and newspapers, on the one hand, and shop window labels [for clothes], on the other hand). An onomasiological profile is formed by different naming choices for a concept available to speakers of a language. For instance, the most common names for a dress in Dutch are *jurk*, *kleed* and *japon*. Next, the relative frequencies of use of each lexical variant are calculated in each lect. Such profiles are created for every concept (e.g. DRESS, SHIRT, OFFSIDE) in each lexical field, and the aggregate uniformity indices between the lects are computed. A comparison of the uniformity measures revealed a strong lexical convergence between the Belgian and Netherlandic varieties, as the uniformity measures steadily increased from 1950 to 1990. The method also allowed to measure the distances between different strata within each geographic variety. For instance, it has been shown that the uniformity index between the Belgian magazines (the standard variety) and shop-window data (a substandard stratum) is lower than that between the corresponding Dutch sub-varieties. This suggests a difference in the structure of the dialect repertoires in the Netherlands and the Dutch-speaking part of Belgium (cf. Auer 2005), which relates to the belated standardization of Belgian Dutch. A more advanced quantitative approach has been recently proposed by Ruette and Speelman (In press), who use the data from Geeraerts et al. 1994. Their approach is based on Individual Differences Scaling, which is a form of Multidimensional Scaling. It allows for capturing the aggregate differences, but also keeps the possibility open to investigate the behaviour of the individual variables, similar to the recent dialectometric studies that were discussed in the beginning of the paper.

The study by Geeraerts et al. (1999) has been replicated for European and Brazilian Portuguese by Soares da Silva (2010). The uniformity indices show that the varieties are diverging in the clothing vocabulary. The results also indicate a greater difference between the Brazilian standard and substandard strata than between those in European Portuguese.

5. Summary and perspectives

The above-mentioned studies demonstrate that semantic and geographic variation can be integrated in fine-grained and robust models of language use. Although most cross-lectal differences in the use of constructions and words are quite subtle, they are nonetheless very revealing, often being signals of ongoing processes of language change, or suggesting interesting lectal differences in the entrenchment of specific semantic frames.

So far, the approaches presented here have been used only for a few languages. Most languages (as well as words and constructions) are still awaiting their turn. Of course, the above-mentioned methods are highly labour-intensive because they require large representative data sets and advanced statistical techniques. However, these approaches are an important step towards a more complete and realistic model of language variation.

The future directions of research, in my view, are mostly related to “catching up” with the achievements of quantitative dialectology and dialectometry in phonetics and morphology. Unfortunately, the study of lexical and syntactic variation in non-elicited texts requires very large data samples, which should be collected for many locations and individuals. On the other hand, it is crucial to explore a much larger repertoire of constructions and lexical fields, in order to be able to see the general tendencies. To achieve these goals, we should use computational linguistic tools to automate the semantic annotation process. Finally, the studies should go beyond one language and take into account areal effects and genetic relationships. All this will make lexical and constructional dialectology a more challenging, but also a more fruitful enterprise in the future.

References

Apresjan, Juri.1966. Analyse distributionelle des significations et champs semantiques structurés. *Langages* 1: 44–74.

Auer, Peter, 2005. Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In: Nicole Delbecque; van der Auwera, Johan and Dirk Geeraerts (eds.), *Perspectives on variation: Sociolinguistic, historical, comparative,* 7–42. Berlin: Mouton de Gruyter.

Berthele, Raphael. 2010. Investigation into the folk's mental models of linguistic varieties. In: Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 265–290. Berlin/New York: Mouton de Gruyter.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. In John R. Firth (ed.), *Studies in Linguistic Analysis*, 1–32. Oxford: Blackwell.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Krämer and Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan and Marilyn Ford. 2010. Predicting Syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1): 168–213.

Bresnan, Joan and Jennifer Hay. 2008. Gradient Grammar: An Effect of Animacy on the Syntax of give in New Zealand and American English. *Lingua* 118(2): 245–59.

Duinhoven, Anton M. 1994. Het hulpwerkwoord doen heeft afgedaan [The auxiliary verb doen has had its day]. *Forum der Letteren* 35(2): 110-131.

Geeraerts, Dirk, Gitte Kristiansen and Yves Peirsman (eds.). 2010. *Advances in Cognitive Sociolinguistics.* Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers and Peter Bakema. 1994. *The structure of lexical variation. Meaning, naming, and context*. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat [Convergence and divergence in the Dutch lexicon]*. Amsterdam: Meertens Instituut.

Glynn, Dylan. In press. Sociolinguistic Cognitive Semantics. A quantitative study of dialect effects on polysemy. *Review of Cognitive Linguistics.*

Glynn, Dylan. 2009. Synonymy, lexical fields, and grammatical constructions. A study in usage-based cognitive semantics. In: Hans-Jörg Schmid and Susanne Handl (eds.), *Cognitive foundations of linguistic usage patterns: empirical studies*, 89-117. Berlin/New York: De Gruyter Mouton.

Gooskens, Charlotte and Wilbert Heeringa. 2004. The Position of Frisian in the Germanic Language Area. In: Dicky Gilbers, Maartje Schreuder and Nienke Knevel (eds.), *On the Boundaries of Phonology and Phonetics*, To honour Dr. Tjeerd de Graaf, 61-87. Groningen, University of Groningen.

Gries, Stefan Th. and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics* 9 (1): 97-129.

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

Grimm, Scott and Joan Bresnan. 2009. Spatiotemporal variation in the dative alternation: a study of four corpora of British and American English. Paper presented at conference Grammar & Corpora 2009, September 2009, Mannheim, Germany.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts 2002. Regressing on *er*. Statistical analysis of texts and language variation. In Annie Morin & Pascale Sébillot (eds.), *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data,* 335–346. Rennes: Institut National de Recherche en Informatique et en Automatique.

Heeringa, Wilbert and John Nerbonne. 2001. Dialect Areas and Dialect Continua. In: David Sankoff, William Labov and Anthony Kroch (eds.), *Language Variation and Change*, Vol. 13, 375-400. New York: Cambridge University Press.

Levshina, Natalia. 2012. Comparing constructicons: A usage-based analysis of the causative construction with *doen* in Netherlandic and Belgian Dutch. *Constructions and frames* 4(1): 76-101.

Levshina, Natalia. 2011. *Doe wat je niet laten kan [Do what you cannot let]: A usage-based analysis of Dutch causative constructions*. PhD diss., University of Leuven.

Levshina, Natalia, Dirk Geeraerts and Dirk Speelman. 2011. Changing the world vs. changing the mind: Distinctive collexeme analysis of the causative construction with *doen* in Belgian and Netherlandic Dutch. In F. Gregersen, J. Parrot and P. Quist (eds.), *Language variation - European perspectives III. Selected papers from the 5th International Conference on Language Variation in Europe*, Copenhagen, June 2009, 111-123. Amsterdam: John Benjamins.

Lukjanova, Svenlana V. 2010. Naimenovanija napitkov v narodnoj reči (na materiale pskovskikh govorov) [Names of beverages in folk speech (on the material of Pskov dialects]. *Vestnik Pskovskogo gosudarsvennogo pedagogičeskogo instituta*. Pskov: PGPU, 33-37.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten and Willem van de Vis. 1996. Phonetic Distance between Dutch Dialects. In: G.Durieux, W.Daelemans and S.Gillis (eds.), *CLIN VI: Proceedings of the Sixth CLIN Meeting,* 185-202. Antwerp: Centre for Dutch Language and Speech (UIA).

Preston, Dennis R. (ed.). 1999. *Handbook of perceptual dialectology.* Amsterdam: John Benjamins.

Prokić, Jelena. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In: *Proceedings of the ACL 2007 Student Research Workshop*, 61-66. Prague.

Prokić, Jelena, Çağrı Çöltekin and John Nerbonne. 2012. Detecting Shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Avignon.

Ruette, Tom and Dirk Speelman. In press. Transparent aggregation of variables with Individual Differences Scaling. To appear in *Literary and Linguistic Computing*.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138): 335–357.

Soares da Silva, Augusto. 2010. Measuring and parametrizing lexical convergence and divergence between European and Brazilian Portuguese. In: Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 41–84. Berlin/New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In: Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 141–166. Berlin/New York: Mouton de Gruyter.

Tagliamonte, Sali A. and R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24 (2): 135-178.

Verhagen, Arie. 2000. Interpreting usage: Construing the history of Dutch causal verbs. In: M. Barlow and S. Kemmer (eds.), *Usage based models of language,* 261-286. Stanford: CSLI Publications.

Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD thesis, Rijksuniversiteit Groningen.

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