BENEDIKT SZMRECSANYI: Grammatical Variation in British English Dialects. A Study in Corpus-Based Dialectometry. Cambridge etc.: Cambridge University Press 2013. 211 pages. (Studies in English Language). € 77,99 (€ 58,67 Ebook)

First of all, we like to congratulate the author for this interesting book. Originally, this book was a habilitation thesis, which he submitted to the University of Freiburg in 2011, but in revised form it has now been published by Cambridge University Press. The author compares dialects in England and Scotland syntactically by using corpus-based dialectometry. In chapter 1 the author writes: "In short, this book explores how and to what extent morphosyntactic variability in traditional British English dialects is structured geographically" (p. 1). And: "The study proceeeds from the fact that we know next to nothing about aggregate morphosyntactic variability in British English dialects." (p. 1).

Morpho-syntactic distances between dialects are measured by using a methodology which is dubbed by the author as CORPUS-BASED DIALECTOMETRY. On the basis of naturalistic corpus data aggregated distances are calculated. Each isogloss suggests another division of the dialect landscape, but when combining – i.e.taking the aggregate of – a multitude of linguistic features, we are able to see the global pattern. The method used by Szmrecsanyi may be considered as the morpho-syntactical counterpart of the methodology of HOPPENBROUWERS & HOPPENBROUWERS (1988), who counted phonetic segments (see Example in Table 2) or phonetic features of segments using a Dutch dialect atlas in which each local dialect is represented by 139 phonetically transcribed sentences (see also HOPPENBROUWERS & HOPPENBROUWERS 2001). Instead of counting phonetic segments or features, Szmrecsanyi counts morpho-syntactic features, and the aggregate distance is

based on the feature frequency differences.

According to Szmrecsanyi, corpora yield a more realistic linguistic signal compared to atlas material which are commonly used in dialectometry. First, he motivates this by claiming that the frequency-based analyses on the basis of corpora "better match the perceptual reality of linguistic input than discrete atlas classifications." (p. 4).

In order to understand this, we illustrate both the analysis of atlas data and the corpus frequency analysis with a small example at the level of the sound components. Assume two local dialects *A* and *B*. For each dialect we have a mini corpus containing realizations for *cellar*, *house* and *milk*. In dialect *A* the realizations are pronounced as [kɛlər], [hus] and [mɔlkə], and in dialect *B* as [kɛldər], [ys] and [mɛlək]. In a typical atlas data analysis, corresponding words are compared to each other, and within the realizations of the words, corresponding phonetic segments are compared to each other. This is illustrated in Table 1.

	A	*B*	difference
cellar	kɛlər	kɛldər	1
house	hus	ys	2
milk	mɔlkə	mɛlək	3
			6

Table 1. Comparing dialect *A* and *B* in the sound components using an approach which is common for atlas data.

Using a corpus-frequency approach we start by deciding about the phenomena found in our corpus which we want to count. In our example we

start by making an inventory of phonetic segments found in the two corpora: k, ɛ, l, ə, r, d, h, u, s, y, m, ɔ. For each of the segments we count its frequency in dialect A and in dialect B. The results are shown in Table 2. The last column shows the frequency differences or differentials, and the sum of the differences is equal to 6. Szmrecsanyi uses a slightly different distance measure, namely, the square root of the squared differences, which is 2.45.

	k	ɛ	l	ə	r	d	h	u	s	y	m	ɔ	total
A	2	1	2	2	1	0	1	1	1	0	1	1	13
B	2	2	2	2	1	1	0	0	1	1	1	0	13
Difference	0	1	0	0	0	1	1	1	0	1	0	1	6

Table 2. Comparing dialect A and B by using a corpus-frequency-based approach. Frequency differences are shown in the last column.

Secondly, the author motivates his preference of corpus-based approaches to atlas-based analyses by saying that atlas data is "non-naturalistic, meta-linguistic, and competence-based in nature" since it relies on "elicitation and questionnaires" (p. 4). Besides, field-workers and atlas-compilers stand between the dialect speaker and the analyst. Below we will come back on the use of corpora and the use of frequency-based methods.

The analyses of Szmrecsanyi are based on data from the *Freiburg Corpus of English Dialects* (FRED). The corpus "samples traditional dialect speech all over Great Britain." (p. 15) The version used by the author contains 368 individual texts which contain 2,437,000 words. The 427 informants sampled in this data set were usually NORMs. The author writes that some

dialect locations are represented by one or two interviews, while others are represented by dozens of interviews. We agree with the author that this is not necessarily a problem, but we would have liked to see that the author had proven that each local dialect is represented by a sufficiently large text sample.

For reasons of statistical robustness local dialects samples within a county are taken together. There are 34 counties in total. Standard British English and Standard American English are added, the data is taken from the *International Corpus of English* and the *Santa Barbara Corpus of Spoken American English* respectively.

The first step is to select the features. The goal was to include as many linguistic features as possible. Literature was consulted in order to identify suitable phenomena. This resulted in 57 morpho-syntactic features, concerning pronouns and determiners (six features), the noun phrase (six features), primary verbs (four features), tense and aspect (seven features), modality (three features), verb morphology (four features), negation (eight features), agreement (seven features), relativization (three features), complementation (six features) and word order and dicourse phenomena (three features). We like it that features are chosen regardless of whether they had previously been reported being geographically distributed. Thirty-one features and their frequencies per county were extracted by means of a Perl script. Twenty-six features were extracted semi-automatically, since manual disambiguation prior to the actual extraction procedure was required.

Given 34 counties and 57 features, frequencies are obtained for all features per county resulting in a 34 × 57 FREQUENCY MATRIX. Distances between counties are calculated by calculating the square root of the sum of the 57 squared feature differentials. Before doing so, the frequencies are normalized over the number of words per local dialect. Subsequently, the

frequencies are log-transformed, which is an improvement compared to the methodology of HOPPENBROUWERS & HOPPENBROUWERS (1988) since this suppresses the effect of large feature differentials and the effect of frequency outliers. Szmrecsany relates this to the weighted identity value (WIV) which was introduced by GOEBL (1983), which we doubt. In Goebl's WIV *similarity* on rare features of two local dialects contributes more to the aggregated similarity than similarity on frequent features. However, when using logarithmic frequencies, *differences* between rare features count stronger than differences between frequent features. It seems to us that this is another kind of effect than the effect obtained by using WIV. Note also that the frequency of features does not have any effect on the weighting of differences when using WIV.

The distances between the counties appeared to be (fairly) normally distributed (Figure 4.1, p. 72). In Sections 4.1.1 and 4.1.2 the 'big picture' is shown. In Section 4.1.1. network maps generated with RuG/L04[1] (been developed in the Groningen school of dialectometry under supervision of John Nerbonne) are shown. In Map C.1 (p. 173) each pair of counties is connected by a line. Darker blue lines represent a small distance, and ligher yellow lines represent large distances. Szmrecsanyi also uses a reverse network map. In this map darker blue lines represent large distances and ligher yellow ones small distances. Pairs of counties which are further than 250 km from each other are not connected. We found this kind of map counter-intuitive and hard to interpret.

In Section 4.1.2 maps from the Salzbug school of dialectometry (which is under supervision of Hans Goebl) are presented. First a beam map (map C.3

1 RuG/L04 is developed by Peter Kleiweg at the University of Groningen and freely available at http://www.let.rug.nl/kleiweg/L04/ .

at p. 174) in which neighbours that are close morphosyntactically are connected by red and thick ('warm and heavy') beams, and neighbors that are distant are connected by blue and thin ('cold and thin') beams. This map is our favourite map, since in this map it immediately becomes clear that there are clusters in the southwest and in the north of England. Another map from the Salzburg school is a honeycomb map (map C.4, p. 174). In this map, each county is represented by a polygon. When two neighboring counties strongly differ, they are separated by a blue and thick ('cold and thin') line, and when they are very similar, their boundary is visualized by a red and thin ('warm and thin') line. In this map the north of England is shown as a relatively homogeneous area.

In Section 5.3 Szmrecsanyi correlates the linguistic distances with no fewer than three different geographic measures: AS-THE-CROW-FLIES DISTANCE, LEAST-COST-WALKING DISTANCE (obtained using Google Maps) and LEAST-COST TRAVEL TIME (again Google Maps turned out to be helpful). Additionally, distances were correlated with LINGUISTIC GRAVITY, which was calculated for any pair of measuring points as the product of their populations divided by the square of least-cost travel time. Thus the author uses (and refers to) to the gravity model as proposed by TRUDGILL (1974), Subsequently, the gravity-scores were log-transformed in order to alleviate the effect of outliers. This endeavour resulted in two important findings. First, linguistic gravity appeared to be the best predictor, which explains 24.1% of the distances between all of the British counties, less than 1% of the distances between the English counties, and 46.5% of the distances between the Scottish Lowlands. Among the ' pure geographic measures' least-cost travel time would be the best, the respective percentages are 7.4%, 0.7% and 39.4%. Finding linguistic gravity means rehabilitation for this measure, given the fact that studies of HEERINGA

ET AL. (2007) and NERBONNE / HEERINGA (2007) failed to detect a significant effect of linguistic gravity. Here we would have liked to see a multiple regression analysis with population products and least-cost-walking distance as predictors of morpho-syntactic distance in order to see the contribution of the two predictors individually. Second, for all of the four measures there was a big difference between England (varying between less than 0.1% and 0.7%) and Scotland (varying between 31.4% and 46.5%). The author also concludes that "the Scottish Lowlands seem to be organized along the lines of a dialect continuum to a much larger degree than England is." But we would have like to read a bit more about this: why does the structure of the dialect landscape in Scotland differ so much from the structure of the English landscape?

Since in many dialectometric studies linguistic distances are correlated with as-the-crow-flies distance, Szmrecsanyi lists the results of eight studies, which comprise lexical, pronunciation and syntactical measurements. When especially comparing the results of Szmrecsanyi for all counties (4% of the variance is explained) or English counties only (1% of the variance is explained), most of the studies report a much higher amount of explained variance, range from 21% to 49%. Only a study of GOOSKENS / HEERINGA (2004) about Norwegian dialects has a percentage that comes close to the percentage found for the complete FRED dataset (5%), since Norway is very mountainous. The explained variance for Scotland only (33%), however, is in line with the existing studies.

The author, of course, is wondering why the shared variance between his measurements and as-the-crow-flies distances is much lower than in most other dialectometric studies, all of them being atlas-based approaches. He then states: "Compared to corpus-based and frequency-centered approaches, atlas-based approaches overestimate the importance of geography." (p. 160). This is

because 1) "... dialect cartography depends on interesting geographic distributions" (p. 161) and 2) atlas data are (typically) categorical, which is a kind of data reduction that "may exaggerate the explanatory power of geography because linguistic contrasts and distinctions appear more pronounced than they actually are." (p. 162). Then the author writes that "an "intelligent" combination of outlier removal, data reduction, and feature selection will boost the explanatory power of as-the-crow-flies distance in the FRED dataset to about $R^2 \approx 25$ percent." But this is, of course, not a clean way of analyzing data. The author writes: "... feature selection and data reduction are essentially a form of academic fraud."

We agree that we should use representative data samples, without subjective feature selection and reduction. But we find it striking that the author is not questioning his own methodology. We think that by using the feature frequency method not all information which is contained by the corpus data is used, and therefore only a weak geographic signal is detected. We will illustrate this by an example, using data from BOLOGNESI / HEERINGA (2002).

BOLOGNESI / HEERINGA (2002) selected a random sample of 200 words from a corpus of about 260,000 words in contemporary Sardinian texts. The speakers of different dialects and languages translated and pronounced the sample words in 54 Sardinian dialects. On the basis of the transcriptions, distances in the sound components were measured between the 54 local dialects with Levenshtein distance. In this data set there is no feature selection, no categorization, no reduction. Meanwhile, the data set has been extended to 77 local dialects (BOLOGNESI / HEERINGA, 2005). Using this extended data set we calculated distances in the sound components with the feature frequency method, the phone frequency method and Levenshtein distance.

When using the feature frequency method, we follow HOPPENBROUWERS / HOPPENBROUWERS (2001) and use the same set of binary features. The authors basically used the features from *The Sound Pattern of English* (SPE) (CHOMSKY / HALLE 1968), an articulation-based system, as starting point, and adapted it so that the distinctions in the Dutch atlas material were represented as well as possible. Thus, they got a feature system of 21 features. Since samples differ in size, relative frequencies are calculated. When calculating the distance, we follow Szmrecsanyi by calculating the Euclidean distance, i.e. the square root of the squared feature differentials.

The Levenshtein distance as a tool for measuring linguistic distances between dialects was introduced by KESSLER (1995). The Levenshtein distance is a numerical value defined as the cost of the least expensive set of insertions, deletions and substitutions needed to transform one string into another (KRUSKAL 1999). For example, when *heart* is pronounced as [hɑrt] in dialect *A* and as [ɛrtə] in dialect *B*, then we find three differences: the [h] in *A* is not found in *B*, the [ɑ] in *A* corresponds with the [ɛ] in *B*, the [ə] in *B* is not found in *A*. Given 200 word realizations per dialect, the Levenshtein distance of a dialect pair is equal to the sum of the differences found in 200 pairs of word realizations.

We applied the feature and the phone frequency method and Levenshtein distance to the set of 200 Sardinian dialects. As to the frequency-based methods, we tried to keep as close as possible to Szmrecsanyi's methodology, which differs in some points with the methodology of HOPPENBROUWERS / HOPPENBROUWERS (1988, 2001). We calculated logarithmic frequencies (*ln*(frequency)+1) and normalized them by dividing them by the number of segments in the corpus. Two local dialects *A* and *B* are compared to each other by calculating the Euclidean distance beween the

frequencies of dialect *A* and the corresponding frequencies of dialect *B*. We correlated the distances with the geographic as-the-crow-flies distances between the 77 locations.[2] The shared variances are shown in Table 3.

	Correlation r	Shared variance R^2
Feature frequency method	0.156	2%
Phone frequency method	0.518	27%
Levenshtein distance	0.662	44%

Table 3. Correlations and shared variance (R^2) between as-the-crow-flies distances and three linguistic distance measures.

The percentage of shared variance found for the feature frequency method (2%) comes close to the percentage found by Szmrecsanyi when using the complete FRED data set (4%). The percentage of shared variance strongly and significantly increases when using the phone-frequency method. Why does this happen? We suggest because of the fact the phone frequency method does not just consider the frequencies of individual features, but considers the frequency of features in combination with other features. For example both [i] and [a] are [+FRONT] and [-ROUND], but [i] is [+CLOSE] and [a] is [-CLOSE]. When counting the number of [i]'s in the corpus, combinations of [+FRONT -ROUND +CLOSE] are counted, and when counting the number of [a]'s, combinations of [+FRONT -ROUND -CLOSE] are counted. However, when simply counting the number of front vowels separated from the number of close vowels, we ignore the way in which a feature co-occurs with another feature. We ignore

2 We thank Roberto Bolognesi for his kind permission to use this data set in this review.

information which is provided by the corpus. By using the phone frequency method, the explanatory power of as-the-crow-flies distances is boosted from 2% to 27% without "outlier removal, data reduction, and feature selection" (cf. p. 163).

Is it possible to use a morpho-syntactical counterpart of the phone frequency method? We suggest to consider sentences as units, just as phonetic segments (or phones) are the units used by the phone frequency method. For each sentence the presence or absence of each of the 57 features can be determined. Thus, an inventory of feature combinations is obtained. Subsequently per dialect the frequencies of the feature combinations can be determined.

Table 3 shows that distances measured with Levenshtein distance correlate significantly stronger with as-the-crow-flies distances than distances obtained with any of the two frequency-based methods. Why? Because the Levenshtein distance utilizes the fact that our corpus is a parallel corpus. For the same set of words realizations are found for each local dialect. With Levenshtein distance each word in dialect *A* is compared to a corresponding word in dialect *B*. Or even more precisely: each segment in a word in dialect *A* is compared with the corresponding segment in the corresponding word in dialect *B*. Information about correspondence is completely ignored when using frequency-based methods. Using this information the explanatory power is boosted further to 44%, again without "outlier removal, data reduction, and feature selection" (p. 163).

All of the dialectometric studies mentioned in Table 8.2 in Szmrecsanyi's book are based on parallel corpora, and the authors utilize the fact that they used parallel corpora to a more or lesser degree. However, when using naturalistic corpora like the FRED data, it is hard to determine

(automatically) which words and sentences in corpus *A* correspond with words and sentences in corpus *B*. We agree that spontaneously spoken and recorded speech is the most natural data, but compared to the use of parallel corpora – the data of Bolognesi & Heeringa (2002) is an example – the lack of considering correspondences between words and sentences has a serious drawback as we showed in our example.

When linguistic distances between the counties have been calculated, we can proceed by classifying the dialects. It is nice to see that the author considers both 'dialect continua' and 'dialect areas'. We like the literature overview about the notion of dialect areas in Section 6.1. In Section 6.3 Szmrecsanyi applies cluster analysis, and on p. 118 he mentions that there exist several alternatives. A bit further he mentions the cophenetic correlation coefficient (CPCC) which measures the agreement between the distances as suggested by the cluster result – a tree in which the counties are the leaves – and the original linguistic distances between the counties. The cluster method which has the largest CPCC is the best, in our opinion, but despite the fact that the author explains the CPCC, he does not use this for deciding which cluster method to use, which surprises us.

Despite our critical remarks – they belong to a review – we like the book. As it is written in the summary, the author utilized "state-of-the-art dialectometrical analysis and visualization techniques".[3] The book is "original both in terms of its fundamental research question" and "in terms of its methodology" (p. I). Questions and (supposed) weaknesses are subject to future research.

3 We do not like each example sentence in the summary, given what a word like 'damn' actually means.

LITERATURE

BOLOGNESI, ROBERTO / WILBERT HEERINGA (2002): De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. In: Gramma/TTT: tijdschrift voor taalwetenschap, 9(1), 45-84.

BOLOGNESI, ROBERTO / WILBERT HEERINGA (2005): Sardegna fra tante lingue. Il Contatto linguistico in Sardegna dal Medioevo a oggi. Cagliari: Condaghes.

CHOMSKY, NOAM / MORRIS HALLE (1968): The Sound Pattern of English. New York: Harper & Row.

GOEBL, HANS (1983): Stammbaum und Welle. In: Zeitschrift für Sprachwissenschaft, 2, 3-44.

GOOSKENS, CHARLOTTE / WILBERT HEERINGA (2004): Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. In: Language Variation and Change, 16(3), 189-207.

HEERINGA, WILBERT / JOHN NERBONNE / RENÉE VAN BEZOOIJEN / MARCO RENÉ SPRUIT (2007): Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied. In: Nederlandse Taal- en Letterkunde, 123(1), 70-82.

HOPPENBROUWERS, COR / GEER HOPPENBROUWERS (1988): De feature frequentie methode en de classificatie van Nederlandse dialecten. In: TABU, Bulletin voort taalwetenschap, 18, 51-92.

HOPPENBROUWERS, COR / GEER HOPPENBROUWERS (2001): De indeling van de Nederlands streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM. Assen: Koninklijke Van Gorcum.

KESSLER, BRETT (1995): Computational dialectology in Irish Gaelic. In: Proceedings of the European ACL, 60-67. Dublin: Association for Computational Linguistics.

KRUSKAL, J. B. (1999): An overview of sequence comparison. In: SANKOFF, D. / J. KRUSKAL (eds.): Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison, 1-44. Stanford: Center for the Study of Language and Information. 2nd edition. 1st edition appeared in 1983.

NERBONNE, JOHN / WILBERT HEERINGA (2007): Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation. In: FEATHERSTONE, S. / W. STERNEFELD (eds.): Roots: Linguistics in Search of its Evidential Base. Studies in generative grammer 96. Berlin and New York: Mouton De Gruyter, 267-297.

TRUDGILL, PETER (1974): Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. In: Language in Society, 3(2), 215-246.

Groningen WILBERT HEERINGA