Commentary: Corpus-based methods

Stefan Th. Gries

1. Introduction

Over the last 25 years or so, linguistics has been changing considerably. The change I am referring to is twofold. On the one hand, there is a new emphasis on empirical data: after the empirical dark ages of generative grammar, more and more linguists now (i) study data more useful than isolated acceptability judgments and (ii) bring both experimental and observational data to bear on the description and explanation of linguistic phenomena as well as on the formation of theories and frameworks. On the other hand, there is also a new awareness that the kind of quantitative methods well established in other social sciences can hugely benefit linguistic research, and a large array of statistical methods – ranging from simple monofactorial tests to complex multivariate procedures – are now more widespread in linguistics than ever before. From my own biased perspective, it seems that in particular the analysis of corpus data has made huge leaps forward: from the virtually exclusively descriptive studies that dominated most of 20th century corpus linguistics, over the first studies using some statistics (e.g. Church and Gale 1990 or Leech and Fallon 1992), over the first multifactorial corpus studies (e.g. Gries 1999, 2001), to the currently most advanced work using mixed-effects regression models (e.g. Bresnan et al. 2007 or, in a diachronic-linguistics setting, Gries and Hilpert 2010).

This influx of empirical data and methods has also begun to have an impact on historical semantics, a field that by its very nature relies on observational data, viz. historical corpora. The papers in this section testify to this fact in multiple ways: they all involve sizeable datasets and recent or even completely novel multifactorial or multivariate analytical approaches, and they do not, in fact, require a commentary to highlight their impressive achievements. In this paper, I will offer some thoughts on directions that work of the kind exemplified here can (or should?) consider to offer even more to historical linguists in general and historical semanticists in particular; in the interest of using particular data sets for exemplification, not all examples will involve historical semantics proper. In Section 2, I will make a case for a greater role of bottom-up approaches (as opposed to top-down approaches). In Section 3, I will consider the use of a wider range of multifactorial or multivariate methods (as opposed to monofactorial approaches). In Section 4, I will offer a small number of additional points, provide several pointers for further reading, and conclude.

2. Bottom-up approaches rather than top-down approaches

It is uncontroversial that while the results of a statistical evaluation of corpus data influence one's interpretation of the data, much of the annotation, coding, and understanding of the temporal structure of the data enters into the analysis long before the results, and these components of one's analysis are still typically determined in a top-down fashion. Hilpert's paper exemplifies how, for instance, a particular clustering method, Variability-based Neighbor Clustering (VNC; cf. Gries and Hilpert 2008, 2010; Gries and Stoll 2009; Hilpert and Gries 2009), allows for identifying temporal stages in historical data that are more meaningful for a given phenomenon than the more globally assigned stages of corpus compilers can ever be, and he shows how such a division of the data into stages can then facilitate the subsequent analysis of diachronic trends. Similarly, Sagi et al. show how semantic broadening and narrowing can be objectively identified and reliably tracked using the fully automated approach of Latent Semantic Analysis, i.e. on the basis of collocate frequencies. While their approach is still experimental in nature (by their own admission), it is an intriguing technique with much potential.

In this section, I want to follow up in this spirit and promote an approach to (historical) corpus data that is more bottom-up / data-driven than is usually found. In Section 2.1, I will briefly mention a few useful exploratory tools; in Section 2.2, I will discuss the notions (and threats) of granularity and dispersion.

2.1. Some exploratory tools

The main characteristic of the VNC algorithm discussed by Hilpert is that it is a clustering approach developed to cluster temporal stages in that it respects the temporal ordering of data. However, historical semantics need not be diachronic semantics since it can be historical but nevertheless synchronic in nature. In such cases, many other methods that do not (have to) respect temporal ordering can be just as useful. For example, hierarchi-

cal agglomerative cluster analyses can be used to, say, determine how many groups of Indo-European languages should be distinguished based on their cognate similarities (cf. Johnson 2008); *k*-means cluster analyses can be used to identify a user-defined number of clusters of elements in a data set; multidimensional scaling can be used to represent a multivariable distance matrix in a two-dimensional coordinate system, etc. That is, there are many ways in which researchers can validate their own ideas against exploratory methods. Rather than analyzing some data with the expectation that there are five clusters in there and a corresponding set of five categories, why not validate that assumption by applying a bottom-up method to the data and see whether such a method returns the same five clusters, or five others, or more or less . . .?

But even if the data in question *are* diachronic such that the temporal order of data points or vectors must be respected, then there are still interesting ways in which data can be explored from a bottom-up perspective. In Hilpert and Gries (2009), we present a heuristic algorithm called Iterative Sequential Interval Estimation (ISIE). This is an algorithm that iteratively moves through successively larger parts of temporally-ordered data to produce estimates about how a trend will develop over time: "the method gives us a range of expected values for the [temporally] next step [. . .] If that next step happens to go beyond the expected values, we have detected a change that merits further attention" (p. 393). Figure 1 exemplifies this approach for the words *in* and *whom* in the TIME corpus: the *x*-axis represents the timeline, the *y*-axis and the black solid lines and points the observed frequencies, and the grey triangles with the dashed lines

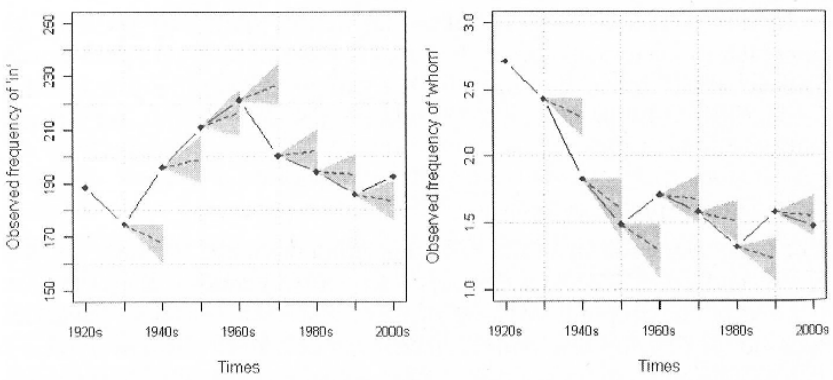

Figure 1. ISIE for *in* and *whom* in the TIME corpus

indicate the algorithm's projection of how the frequency will develop, given the weighted developments in the past. As is clear, the observed frequency of *in* is often in the predicted triangle and the cases where they are not are erratic, once up and once down. The pattern for *whom* is much clearer and many of the (expected) decreases are as predicted.

The final example uses the data provided in Hilpert's Table 1 and Figure 1, the frequencies of *keep* V-*ing* in the TIME corpus. Hilpert argues (correctly) that there is an increase over time (Kendall's $\tau = 0.425$), but it is instructive to characterize the increase in more detail. Hilpert's Figure 1 shows the trend is neither fully monotone or linear: there is an increase until somewhere between 1996 and 2001 (or 2002?) but then the points level off. Here, an approach called regression with breakpoints can be useful. Using the open source programming language and environment R (R Development Core Team 2011), a tool by now widely used in linguistics, I wrote a script to split the temporal data into two parts all possible ways and compute linear regressions for each to determine which split of the data into two stages (i) leaves least deviance in the data unexplained and (ii) corresponds most closely to a non-parametric smoother. The two most useful splits of Hilpert's data are represented in Figure 2, with the best one (according to the two above criteria) in the left panel with *r*- and *p*-values for the two resulting regressions.

In both panels, the *x*-axes represent time, the *y*-axis the normalized token frequencies, the dashed vertical lines the times of the split, and the

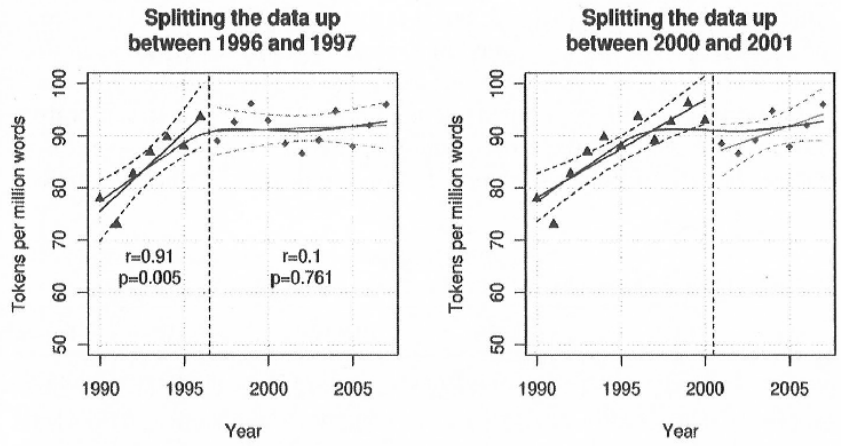

Figure 2. Two regressions with breakpoints on Hilpert's *keep* V-*ing* data

black and grey lines on the left and the right of the split are the regression lines and their confidence intervals for the two splits. The result shows clearly that there was a strong increase between 1990 and 1996 ($r = 0.91$), and a complete leveling-off after that ($r = 0.1$). Again, the point is that the analyst – in this case, me – only chose objective criteria and a well-known algorithm to let the data determine which temporal division is most appropriate, and I submit that such exploratory tools can be extremely useful.

2.2. The many facets of granularity and dispersion

A probably even more important aspect in which bottom-up approaches are very useful is concerned with the notion of granularity, which is related to the above, but also different enough to merit its own little section. By *granularity*, I am here referring to the fact that corpus data can be studied on many different levels of resolution, and in fact even the three papers in this section already use multiple and different levels. For example, Hilpert's data are aggregated, first, at the level of the year and then, after the application of VNC, to the collexeme data at the level of multi-year clusters, and then, after splitting up genres, at the levels of years and genres. By contrast, Geeraerts et al.'s data are aggregated, first, at the level of the text and, then, at the level of three reference points, but from the study *per se* it is not clear whether a different resolution (e.g. one based on a bottom-up method) would have yielded results with a higher degree of discriminatory power. Similarly, Sagi et al's data are, first, aggregated at the level of the document and, then, at the level of four time periods (cf. their Table 1), but it is again not obvious that a bottom-up approach would have resulted in the same four time periods as in their Table 1. In an ideal world, decisions about the level of granularity at which a phenomenon will be studied would be made after some exploratory analysis that suggests the most illuminating level of resolution or at least shows that the distinction chosen is not problematic (which is unlikely

Table 1. Table 2 (Appendix 2) from Hundt and Smith (2009)

	LOB	FLOB	BROWN	FROWN	Totals
Pres. perf.	4196 (10.5%)	4073 (10.4%)	3538 (8.7%)	3499 (8.8%)	15306
Simple past	35821	35276	37223	36250	144570
Totals	40017	39349	40761	39749	159876

to be the case in the three papers in question, but that doesn't invalidate the general point).

Unfortunately, for researchers working with corpora, the above differences are only some of the many choices affecting one's results. The by-subject vs. by-item distinction well-known from experimental studies applies in historical linguists' observational data, too: do we study phenomenon *P* per speaker/writer, per text (which may contain several speakers), per file (which may contain several texts), by genre (which may be represented by several files), ...? Do we study a syntactic phenomenon *P* on the basis of all the, say, verb forms that occupy a particular slot in it, all verb lemmas, only the most frequent *x* (elements/percent)?, ...? And how do we handle cases where we decide on one level of resolution (say, register) but the word *w*, which is fairly frequent in a register we study, is highly underdispersed in that register/corpus, meaning it only occurs, say, in two of the 90 files of that register/corpus, but very frequently? As I have shown elsewhere (Gries 2006, 2008, 2010, to appear), every quantitative corpus result can be strongly affected by these decisions and the dispersion of elements in (parts of) corpora, and when one lets the data decide, the corpus divisions that exploratory methods return often cut across established levels linguists would like to use. (Note that Sagi et al.'s use of *tf.idf* addresses this issue to some extent since this measure is sensitive to dispersion on the chosen level of granularity.)

What all this amounts to is that there are usually no hard and fast answers to the question of which level of granularity/resolution one's data should be studied at, and that decisions in favor of a particular level of granularity can only be made for each phenomenon and for each corpus separately. This means that one should explore at least several levels of granularity above and below the one originally intended to, again, make sure that one does not pick one level for analysis (more or less as a matter of convenience) when the action is in fact somewhere else.

3. Multifactorial/-variate approaches rather than monofactorial approaches

In this section, I want to briefly make a plea for an even wider use of multifactorial methods in (historical) semantics. As before, the three papers in this section are already exemplary in many ways. Hilpert's study explores semantic change on the basis of long vectors of collexeme strength data, which means he takes many lexical dimensions on which *keep* V-*ing* con-

structions can vary into consideration. Sagi et al. do not use the fine-grained resolution of Hilpert's – a syntactically precisely-defined slot in one construction – but use a much larger collection of collocates of their words in question – *dog*, *hound*, *deer*, ... – which makes their data somewhat noisier, but also much more voluminous and, to some extent at least, more comprehensive (given the possibility that more words whose use may reveal something about their target words are included). Last but not least, Geeraerts et al. take as input the manual annotation of concordance lines based on several variables and use a logistic regression to predict uses of *anger*. Thus, the studies in this section are maybe not representative of the typical historical (semantic) study but illustrate beautifully how their methods can and should complement traditional ways of analysis.

To give an additional example, Table 1 represents a case in point from a recent study looking for diachronic change in the present perfect (as compared to the simple past).

Hundt and Smith (2009: 51) summarize these data by saying "[simple pasts] have also decreased over time" and "when it comes to relative frequencies of the PP and the SP in BrE and AmE, we are – again – dealing with stable regional variation rather than ongoing diachronic change." However, as for the former, it is not clear to me which part of this table supports the stated conclusion. As for the latter, one needs to first recognize that Hundt and Smith presented the data as if it was a two-dimensional data set TENSE (present perfect vs. simple past) × CORPUS (LOB vs. Brown vs. FLOB vs. Frown) whereas in fact it is a three-dimensional data set: TENSE (present perfect vs. simple past) vs. VARIETY (BrE vs. AmE) vs. TIME (early (for LOB and Brown) vs. late (for FLOB and Frown)), as represented in Table 2.

With this format, their second conclusion translates into a statement about the absence of significant interactions involving TIME. And in fact they are right: a Poisson regression on these data reveals that TIME has a significant effect (the numbers for the later corpora are lower; $p < 0.001$), but does not participate in a significant interaction with TENSE and/or VARIETY (all p's > 0.38)). Nevertheless, it is not possible to be sure of this without the type of multifactorial analysis the papers in this section – in particular Geeraerts et al. – have conducted. Thus, obviously, multifactorial questions require multifactorial methods. For Hundt and Smith's data, the results of the above Poisson regression are represented in Figure 3: the x-axes represent independent variables and the y-axis and the figures in the bars represent the predicted frequencies from the regression.

Table 2. Redesigned Table 2 (Appendix 2) from Hundt and Smith (2009)

Tense	Variety	Time	Frequency
pres. perf.	BrE	1960s	4196
pres. perf.	BrE	1990s	4073
pres. perf.	AmE	1960s	3538
pres. perf.	AmE	1990s	3499
simple past	BrE	1960s	35821
simple past	BrE	1990s	35276
simple past	AmE	1960s	37223
simple past	AmE	1990s	36250

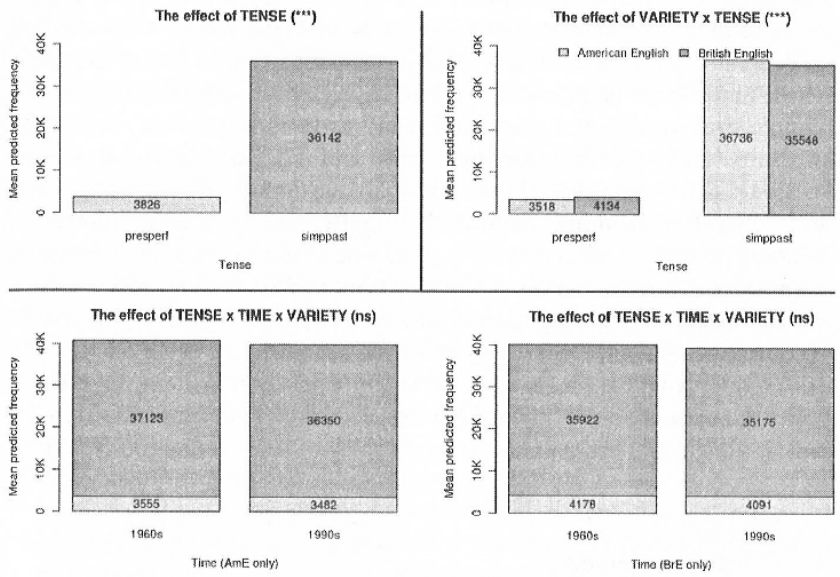

Figure 3. The effects of TENSE (upper left panel), VARIETY x TENSE (upper right panel), and TENSE x TIME x VARIETY (lower panel) of Hundt and Smith's Table 5

Geeraerts et al. laudably use exactly one such multifactorial approach, one of the multifactorial methods most widely used in corpus-linguistic studies: binary logistic regression. However, many other tools are available and should also find their way into the historical semanticist's toolbox. The most straightforward extension is a method that I thought Geeraerts et al. were going to use: a multinomial logistic regression, which conceptually differs from its binary counterpart in that the dependent variable can have more than two levels. (The dependent variable in Geeraerts et al. could have been NOUN (*anger* vs. *ire* vs. *wrath*) ...) Also, Poisson regressions of the type exemplified above are a useful tool for when the dependent variable consists of frequencies. Then, a very important recent development is the use of mixed-effects models (or multi-level models), a class of regression models that allows the user to include both fixed effects (variables whose levels exhaust all the possible levels such as SPEAKERSEX, where levels other than male and female are unlikely to be attested) and random effects (variables whose levels in the analysis are only a sample of those in the population such as SPEAKER or VERB, where we would like to generalize to more than just the few speakers or verbs in our samples). These models can handle samples with dependent data points and uneven sample sizes much better than traditional regressions, provide much more precise results, and are becoming a more and more widespread technique in all areas of linguistics. (It has to be noted, though, that the method is still being developed and fine-tuned.) Finally, there is a large number of alternative approaches out there that may be of use to researchers working with noisy observational data. Classification and regression trees, support vector machines, learning algorithms, and neural networks are a few of the currently hot methods that are worth keeping an eye on (cf. Baayen 2011): once applications are available (or, even better, R packages) that make the applications of these tools easier, historical corpus-based studies will be able to explore even the most complicated data with renewed vigor.

4. Concluding remarks

There are a few final comments I wish to make. Again, these comments must not be understood as a critique of the papers in this section, which already do a lot of the things I would like to see in corpus-based work (in historical semantics/linguistics).

First, there are the notions of interdisciplinarity and methodological pluralism. The papers in this volume and in this section are already interdisciplinary in how they bring together methods and insights from different linguistic disciplines. In this section alone, historical semantics is enriched by methods used in cognitive linguistics (distinctive collexeme analysis), first language acquisition (VNC), sociolinguistics (lectal variables in Geeraerts et al.'s logistic regression), computational linguistics / information retrieval (LSA), so when I advocate even more of this, then my target group is not the present authors. There are many more fields that have exciting methods to offer. I have already mentioned quite a few but as one additional example let me mention work in corpus-based dialectology, where statistical techniques and exciting visualization tools are now used for the bottom-up identification and characterization of dialect continua on the basis of corpus data (cf. Szmrecsanyi and Wolk 2011). Methods like these, which add a geographical perspective to the data, are just waiting to be added on top of the bottom-up methods discussed in this section, and there are many more interesting approaches out there once we look beyond linguistics proper (neighbor-clustering approaches to two- or three-dimensional data are common in the study of ecosystems, for instance).

Given the above, methodological pluralism follows naturally: many studies can benefit from using several of the approaches advocated here together. For example, Gries and Hilpert (2010) first use VNC to arrive at temporal stages of the diachronic development of the third person singular marker in English, and then they use these stages as a predictor in a generalized linear mixed-effects model to explore which linguistic features accompanied and/or drove that change, and similar applications are conceivable even for the papers in this section, as when Geeraerts et al. and Sagi et al. might benefit from the VNC approach to obtain the best temporal divisions in their data, or when Hilpert's data might be explored with the above regression-with-breakpoints approach, etc.

The second recommendation I want to make to researchers can only be made very briefly and programmatically: the more complicated one's data and methods and the more they are borrowed from outside of one's core area, the more one needs to use illuminating visualization tools. For example, few people really know what the coefficients of regressions mean (esp. for logistic and Poisson regressions), and few people understand odds ratios or log odds, etc., which makes it all the more important that graphs are used that provide all and only all the important information in a way that readers who are not (yet) statistically savvy can digest

them. Obviously, this is very subjective, but we should all be aware of this and make the time we spend on developing meaningful and interpretable visualizations a function of the statistical complexity of our data and tools and, wherever necessary, provide not just *p*-values but also effect sizes.

By its very nature, this commentary can only scratch the surface, but I hope to have underscored a point that the three papers in this section already made beautifully. Historical corpus linguists and semanticists have a lot to benefit from being open to what new methodologies have to offer. Quantitative methods are being newly developed and popularized all the time, and staying informed about how these methods can help us along in our research should be a prime objective, especially given that linguistics as a whole is undergoing this move towards empiricism. Standard references such as Baayen (2008), Johnson (2008), or Gries (2009) provide easy entries to a whole new world out there that offers possibilities too exciting to be ignored.

References

Baayen, R. Harald
2008 *Analyzing linguistic data: a practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. Harald
2011 Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11 (2): 295–328.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen
2007 Predicting the Dative Alternation. In: Gerlof Bouma, Ineke Kraemer and Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Church, Kenneth W. and William Gale
1990 Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (2): 22–29.

Gries, Stefan Th.
1999 Particle movement: a cognitive and functional approach. *Cognitive Linguistics* 10 (2): 105–145.

Gries, Stefan Th.
2001 A multifactorial analysis of syntactic variation: particle movement revisited. *Journal of Quantitative Linguistics* 8 (1): 33–50.

Gries, Stefan Th.
2006 Exploring variability within and between corpora: some methodological considerations. *Corpora* 1 (2): 109–151.

Gries, Stefan Th.
2008 Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13 (4): 403–437.

Gries, Stefan Th.
2009 *Statistics for linguistics with R: a practical introduction.* Berlin and New York: Mouton de Gruyter.

Gries, Stefan Th.
2010 Dispersions and adjusted frequencies in corpora: further explorations. In: Stefan Th. Gries, Stefanie Wulff and Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, 197–212. Amsterdam: Rodopi.

Gries, Stefan Th.
To appear Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In: Mario Brdar, Milena Žic Fuchs, and Stefan Th. Gries (eds.), *Convergence and expansion in cognitive linguistics.* Amsterdam and Philadelphia: John Benjamins.

Gries, Stefan Th. and Martin Hilpert
2008 The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3 (1): 59–81.

Gries, Stefan Th. and Martin Hilpert
2010 From interdental to alveolar in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14 (3): 293–320.

Gries, Stefan Th. and Sabine Stoll
2009 Finding developmental groups in acquisition data: variability-based neighbor clustering. *Journal of Quantitative Linguistics* 16 (3): 217–242.

Hilpert, Martin and Stefan Th. Gries
2009 Assessing frequency changes in multistage diachronic corpora: applications for historical corpus linguistics and the study of second language acquisition. *Literary and Linguistic Computing* 34 (4): 385–401.

Hundt, Marianne and Nicholas Smith
2009 The present perfect in British and American English: has there been any change recently? *ICAME Journal* 33: 45–63.

Johnson, Keith
2008 *Quantitative methods in linguistics.* Malden, MA: Blackwell.

Leech, Geoffrey and Roger Fallon
1992 Computer corpora – what do they tell us about culture? *ICAME Journal* 16: 29–50.

R Development Core Team
2011 *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

Szmrecsanyi, Benedikt and Christoph Wolk
2011 Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics* 11 (2): 561–592.