Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Pragmatics 52 (2013) 5–16

journal of PRAGMATICS

www.elsevier.com/locate/pragma

Sources of variability relevant to the cognitive sociolinguist, and corpus- as well as psycholinguistic methods and notions to handle them

Stefan Th. Gries

Department of Linguistics, University of California, Santa Barbara, Santa Barbara, CA 93106-3100, USA

Received 27 October 2011; received in revised form 28 May 2012; accepted 28 December 2012

Abstract

This paper is a plea for sociolinguistics to integrate both theoretical and methodological developments from cognitive linguistics and, even more importantly, psycholinguistics. More specifically, I argue that theoretical advances involving exemplar-based models and new methodological tools from psycholinguistics (regressions, in particular mixed-effects models) and corpus linguistics (in particular, more bottom-up studies) would help further sociolinguistics to a considerable degree. To exemplify at least some ways what such developments would look like, I then discuss three small case studies of instances of constructional variation in usage/corpus data, which showcase how contextual as well as cognitive-/psycholinguistic language-internal and sociolinguistic language-external factors interact and can be explored.

Keywords: Sociolinguistics; Psycholinguistics; Exemplar models; Mixed-effects models; Constructional variation

> The overarching objective of this study is to explore the gradient interaction between language-internal and language external factors as a cognitive and cultural phenomenon that comes within the remit of cognitive sociolinguistics. Szmrecsanyi (2010:143f.)

1. Introduction

The linguist's life is a hard one. Language, especially when studied not on the basis of decontextualized and made-up sentences, is one of the most openly manifest forms of human behavior, but also one of the most complex and multi-faceted domains of scientific inquiry. This is because linguistic behavior is influenced by

- specific aspects of the linguistic system having to do with *linguistic form/structure*, potentially ambiguous and polysemous meanings/functions, and their interrelation;
- *within-individual aspects of cognition* having to do with attention, working memory, perception and learning, general intelligence and linguistic as well as academic attainment, and a variety of 'performance'-related factors; these can be subsumed under the notion of cognitive and/or psychological/psycholinguistic determinants;
- *between-individual aspects of interaction* having to do with social, interactional, and cultural forces; these can be subsumed under the notion of sociocultural/sociolinguistic determinants.

E-mail address: stgries@linguistics.ucsb.edu.

0378-2166/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.pragma.2012.12.011

To make matters worse, all these factors influence linguistic behavior only probabilistically and to different extents at different times, and we have no direct access to the interactions of these probabilistic systems. Much of the data we do have comes from corpora or other data elicited in largely authentic settings, but often these data are

- spotty: we only have samples of the 'population' of a language that are, typically, tiny, unbalanced, and unrepresentative;
- hard to use: corpora only contain distributional data – frequencies of occurrence and of co-occurrence and dispersion – so whatever one wants to study needs to be operationalized in a quantitative/distributional fashion, and not all cognitive/psycholinguistic or sociolinguistic variables of interest are easy to operationalize;
- hard to obtain: linguistic patterns are often fuzzy and, thus, hard to define exhaustively, and corpora and their annotation often contain errors and are often difficult to search.

Against this background, it is easy to understand that the two disciplines of cognitive- and psycholinguistic approaches on the one hand and sociolinguistic approaches on the other hand have so far mainly concerned themselves with exploring 'their' sources of variation – cognitive sociolinguistics as its own dedicated field is a young discipline. The inaugural collections that marked the emergence of the discipline are Kristiansen and Dirven (2008), and Geeraerts et al. (2010). In terms of explicitly marking the birth of a new discipline – or a new merger of existing disciplines – both collections define, motivate, and set the stage for a new fruitful and interdisciplinary endeavor by pointing out correctly, for example, how cognitive linguistics can benefit from a (more) sociolinguistically informed perspective: (i) cognitive linguistics would benefit from recognizing the sociolinguistically time-honored fact that it is useful to study more diverse data than the written production of some standard variety of a language; (ii) contextual features of language use are particularly indispensable for an approach that wants to do justice to calling itself usage-based; (iii) usage in turn takes place in social settings, with social goals, and under social constraints, etc. (cf. the introductory chapter in Geeraerts et al., 2010). At the same time, a convincing case is also made for how sociolinguistics would benefit from a (more) cognitive perspective, and I will in fact be more inclusive and say a more *cognitive/psycholinguistic* perspective, since, in fact, semantic or more general cognitive and psycholinguistic considerations are at the core of sociolinguistics, if only for how they affect choices speakers make in discourse.

I agree with nearly all of the assessments to a degree that I would ask, how could one not agree? I have a few minor qualifications and/or additional thoughts, however. One very general point that bears on the argumentation to follow is concerned with the fact that, using Eckert's (submitted for publication) terminology, it seems to me that the kind of sociolinguistics so far most prominently represented in cognitive sociolinguistics belongs to the first wave of variation study, the one concerned with establishing ''broad correlations between linguistic variables and the macro-sociological categories of socioeconomic class, sex class, ethnicity and age'' (p. 11), which is why my discussion below will focus on this wave, too.

Another point is concerned with the (bi)directionality of the exchange between the fields, (bi)directionality in the sense of who offers whom how much. To be quite honest, and to the limited degree that anyone in general and I myself in particular can evaluate two such huge and diverse fields in their entirety, I think that the main direction of knowledge transfer is, and should be, more – though not exclusively, see below – from cognitive linguistics/psycholinguistics to sociolinguistics than the other way round. Why is that?

One reason for this is concerned with theoretical aspects. While a field called *cognitive sociolinguistics* was not explicitly recognized until very recently, it should be acknowledged that at least some areas of what is called *cognitive linguistics* have for quite some time recognized that factors that can fairly uncontroversially be considered sociolinguistic can play an important role and should be integrated theoretically. For example, as early as 1995, Goldberg discusses the meaning component of constructions in her Construction Grammar as follows:

> ''Meaning'' is to be construed broadly enough so as include contexts of use, as well as traditional notions of semantics. That is, a construction is posited when some aspect of the way in which it is conventionally used is not strictly predictable. It would alternatively be possible to define constructions as ordered triples of form, meaning, and context, as s done by Zadrozny and Manaster-Ramer, 1993. (Goldberg, 1995:229, n. 6)

Similarly, in her 2003 overview article, she states that ''the function pole in the definition of a construction indeed allows for the incorporation of factors pertaining to social situation, such as e.g., register'' (Goldberg, 2003:221). Thus, even though cognitive sociolinguistics was not yet 'officially' recognized as a new discipline, some areas of cognitive linguistics' awareness of, and compatibility with, what are now foundational assumptions of cognitive sociolinguistics predate the discipline by approximately 10 years. Thus, to my mind, the field of cognitive linguistics has recognized that sociolinguistic factors are important for the development of a cognitively and psycholinguistically grounded framework and has, developed its main framework in a way so as to accommodate that recognition, even if there were not yet many concrete studies following this thread. On the other hand, it has always been my impression that sociolinguistics has not yet had much of a similar epiphany and has until very recently been mostly content to focus on language-external between-individual aspects. This comes at the cost of (i) de-emphasizing language-internal, within-individual cognitive processes and (ii) not fully appreciating how cognitive- and psycholinguistic theories have evolved to a point where they can substantially inform and, maybe even more importantly, integrate sociolinguistic work.[1]

A timely recent example is the abstract of William Labov's keynote speech at LAUD 2010. I am quoting two statements of Labov's here, first Labov's statement that that ''[t]he central axiom of sociolinguistics is that the community is prior to the individual''; second his abstract stating that

> Early acquisition of [systematic variation] requires probability matching of the language learner. These probabilities are not assigned to individual words or exemplars, but to abstract categories in combinatorial rules. Change within the community structure involves change in the rules and constraints on the realization of these categories. Although linguistic variation is sensitive to both internal and social constraints, recent research has confirmed the independence of these two groups of factors. There is evidence for separate storage of social information in a sociolinguistic monitor, independent of grammatical information. (Labov, http://www.uni-landau.de/anglistik/LAUD10/abstracts/labov.htm)

As for the first quote, provocatively speaking, nothing could be further from the truth. No doubt, both cognitive/psycholinguistic and sociolinguistic studies have their obvious merits in their pursuit of within-individual and between-individual sources of variation. Nevertheless, I believe it is necessary to recognize that something can by definition only be sociolinguistically relevant if, at some point of time, it has passed through the filter of the human mind, where it was either readied for production or comprehended and interpreted. The reverse may also be true to some extent – what is cognitively/psycholinguistically relevant often arises once it is produced within some socially-defined situation. However, as pointed out above, it seems to me as if theory development in cognitive linguistics and psycholinguistics has proceeded in ways that are much more open to embracing the sociolinguistic side than vice versa, which leads me to the second quote.

This second quote is in fact very interesting because recent work in cognitive- and psycholinguistic approaches has adopted a very different perspective. In particular, more and more cognitive linguists assume a so-called exemplar-based perspective. As the name suggests, this approach assumes, *pace* Labov, that acquisition, representation, and processing of language are in fact very much based on individual exemplars: linguistic knowledge resides in a high-dimensional space such that each exemplar of a linguistic experience is stored in this space at coordinates that characterize its properties on the dimensions of that space. Crucially, memory representation within such an exemplar model is extremely rich and includes phonetic, phonological, prosodic, morphological, syntactic, semantic, discourse-pragmatic, and last but not least, sociolinguistic and contextual characteristics, even specific characteristics of speakers.[2] In other words, co-occurrence information of all aspects of the exemplar, involving both linguistic and extra-linguistic aspects are represented alongside each other. Thus, what is stored and processed are *not* so much ''abstract categories in combinatorial rules'' – categories only arise from the clouds that arise from many points stored in close *n*-dimensional spatial proximity. This also means that, from this perspective, no ''independence of [internal and social] factors'' is assumed: factors of both kinds give rise to dimensions on which exemplars can be located. And, even outside of core sociolinguistics proper, evidence so far suggests that is highly compatible with such an account. For reasons of space, two brief examples from different sub-disciplines must suffice here.

One is Bybee's account of how the WXDY construction (e.g., *What's this fly doing in my soup?*) came to be associated with its incongruity meaning, namely by the association of a particular partially-filled lexico-syntactic pattern with a particular pragmatic utterance context. The other example is concerned with an even more sociolinguistic concept, register/genre. Register as a situationally/communicatively-defined, and therefore sociolinguistic, category has repercussions for very many linguistic phenomena. Depending on whom we talk to, we adjust our articulatory effort (s), lexical choices, syntactic complexity, etc. As Halliday (2005:66) put it, ''[r]egister variation can in fact be defined as systematic variation in probabilities; a register is a tendency to select certain combinations of meanings with certain frequencies.'' Critically, however, Biber's exciting work on multidimensional analysis has shown exactly how something as language-external and sociolinguistic as genre variation can be understood as in part language-internal and cognitively/psycholinguistically motivated. Not only is it extremely tempting to conceptualize the dimensions Biber

[1] One reviewer pointed out correctly that, for instance, the sociolinguistic study of Weiner and Labov (1983) predates some of the first psycholinguistic priming studies by J.K. Bock. That is of course correct, and I can add myself that early sociolinguistic work such as Sankoff and Laberge (1978) was among the first to systematically use switch-rate scatterplots to explore speaker-specific preferences. However, I do think it's fair to say that this early recognition of priming-like effects and speaker-specific effects has not left much of a mark on the next few decades of sociolinguistic work in which such determinants have enjoyed very little recognition and little progress was made regarding statistical sophistication (until recently).

[2] For example, Hay et al. (2006:376) find that stored exemplars of vowels are ''indexed to social information such as gender, age and nationality.''

discovered as dimensions compatible with those in the high-dimensional exemplar space mentioned above, but several of the dimensions that are intended to explain language-external variation are also straightforwardly connected to cognitive, i.e., language-internal, characteristics. For instance, the factor of Biber (1988) that explains most of the variation in his data is characterized as ''marking *high informational density* and exact informational content'', which in turn involves, among other things, ''[production] circumstances dictated by *real-time constraints*, resulting in generalized lexical choice and a generally fragmented presentation of information'' (p. 107, my emphasis). If this is not a straightforwardly psycholinguistic account of an initially sociolinguistic finding, what is? Informational density of discourses is straightforwardly correlated with processing effort, and various operationalizations (e.g., type-token ratios as in Szmrecsanyi, 2005, but also entropy-related measures as in Schnoebelen, 2008 or Frank and Jaeger, 2008) are clear correlates of processing in production and comprehension. While other factors' relations to psycholinguistic notions are less direct, they are still suggestive, and there is a now a growing recognition of the fact that such connections are worth exploring (cf. section 4.4.4.2 in Mendoza-Denton et al., 2003).

By way of an interim summary, I hope I have shown that current cognitive- and psycholinguistic work has a lot to offer to sociolinguistics in the sense of raising awareness that nothing is social without being cognitive, and by providing sociolinguistics with a theoretical framework that has been independently validated in the domains of acquisition, representation, and processing, that results from an interdisciplinary cognitive science endeavor, and that provides interesting explanations, predictions, and possibilities for modeling.

There is a second way in which psycholinguistic work in particular has something to offer to sociolinguistics. In some sense at least, both psycholinguistics and sociolinguistics have been relying on quantitative methods to identify interesting and non-accidental variation, and more so than many other branches of linguistics. However, the set of methods that have been used in both disciplines are not without their individual problems. For example, experimental studies in psycholinguistics have long been using quasi F-ratios and/or F_1/F_2 statistics to ensure that experimental effects are reliable across subjects (F_1) and across items (F_2); cf. Satterthwaite (1946), Cochran (1951), Clark (1973), Forster and Dickinson (1976). However, over the past few years, this approach, which has been the default for decades, has been the subject of some controversy. It is sometimes hard to motivate, it cannot be extended easily to other statistical designs, and can handle neither missing data nor the type of unbalanced designs common in corpus linguistics too well (cf. Baayen, 2008: Ch. 7 for recent discussion).

Sociolinguistic approaches carry similar baggage. For many years now, Varbrul analyses have dominated the field; cf. Cedergren and Sankoff (1974) and Paolillo (2002). However, traditional Varbrul analysis is astonishingly lacking in many respects: the Varbrul treatment of interactions is cumbersome to say the least, Varbrul cannot include continuous covariates without factorization and the accompanying loss of information, it is not able to handle repeated measurements/dependent data points and collinear predictors well, it outputs the results in an idiosyncratic format and use idiosyncratic terminology that makes the result of what is essentially a logistic-regression type of approach more difficult to compare than necessary; cf. Mendoza-Denton et al. (2003) as well as Johnson (2008) for comprehensive discussion in favor of logistic regression as a nowadays more appropriate method.

Although both fields face methodological problems, it is my impression that, on the whole, psycholinguistics is on a better path to remedying these. One way in which this is being done is the adoption of mixed-effects, or multi-level, models, which address many of the problems of both disciplines listed above: subject- and word-specific effects can be handled as can unbalanced designs and missing data, predictors of all kinds can be included, etc. However, while psycholinguists have been testing, using, and refining this approach for quite some time, the valor with which many sociolinguists hold on to Varbrul and do not consider even the more standard logistic regressions, let alone mixed-effects models is perplexing and can only be explained with sociology-of-science arguments. Fortunately, however, some scholars have begun to adopt more diverse and more powerful methods; cf., e.g., Johnson (2009, 2010) or Szmrecsanyi (2006, 2010) for examples.

There is another recent and related methodological development that I personally wish sociolinguists to pick up. In particular, corpus linguists increasingly recognize that sociolinguistic or corpus-linguistic distinctions that are easily and frequently made may in fact not be the most revealing ones in the sense of explaining the variability on the data best. This recognition has led to an increase in bottom-up methods, i.e., methods that do not impose any one researcher's more or less motivated distinctions or variable levels on the data, but let the data decide which distinctions are most meaningful. I will exemplify this in more detail below.

In the next two sections of this paper, I will discuss three different case studies to support the general claim that cognitive linguists and psycholinguists and sociolinguists alike may benefit hugely from taking each other's perspectives into consideration, essentially trying to add to a stance taken most prominently by members of the QLVL group in Leuven (in the inaugural cognitive sociolinguistics volumes mentioned above and, say, in Grondelaers et al., 2002 or Speelman and Geeraerts, 2009). The case studies discussed in this chapter are all concerned with synchronic and diachronic constructional variation, i.e., area 2 of Kristiansen and Driven (2008:4), but with a focus on language-internal variation and methodological advances (cf. Geeraerts et al., 2010:1). In section 2, I will briefly discuss two corpus-based studies of syntactic priming in two alternating construction pairs. After a short general introduction to syntactic priming in section 2.1, section 2.2 exemplifies the point that cognitive linguists, psycholinguists, *and* sociolinguists should keep their eyes more open with regard to how factors from both fields can interact. More specifically, I show how a study of cognitive/psycholinguistic determinants of the dative alternation (exemplified in (1)) reveals that the effects of psycholinguistic determinants interact significantly even with a coarsely-defined language-external sociolinguistic variable such as MODE (speaking vs. writing).

(1) a. He sent her a book.
b. He sent a book to her.

Section 2.3 exemplifies that variation may be due to sociolinguistic variables and how strongly sub-registers can affect psycholinguistic variables. More specifically, I show that the effect of psycholinguistic determinants of particle placement (the alternation exemplified in (2)) can differ depending on the registers that are studied, and that lexically-specific effects should be taken into consideration.

(2) a. He picked up a book.
b. He picked a book up.

Finally, section 3 exemplifies the combined methodological advantages of using cognitive and sociolinguistic determinants, bottom-up approaches, and statistics that are speaker- and lexical item-specific. I show that the suffix choices in the diachronic morphological change of the English third person singular suffix (from, say, *giveth* to *gives*) can be predicted with extremely high degrees of accuracy once all the above advice is heeded.

2. Case studies 1 and 2: priming effects in constructional alternations

2.1. Introduction

Syntactic priming refers to speakers' tendency to reuse syntactic structures they have processed before. For example, speakers who have processed a passive sentence are more likely to describe a transitive scenario with another passive sentence than speakers who have processed an active sentence. It is well-known by now that structural persistence is robust, can be long-lasting and cumulative (exhibiting a logarithmic decay curve), and has been observed from comprehension and production to production, across languages, and in L1 and L2. Furthermore, different alternations are differently responsive to priming (the dative alternation is more susceptible to priming than the voice alternation) and, within an alternation, constructions are differently responsive to priming (ditransitives as in (1)a are more susceptible to priming than prepositional datives as in (1)b). We also know that similarity between primes and targets facilitates priming, where similarity can take on a variety of different forms such as when the same verb is used, when prime and target exhibit other lexical/morphological similarity, or even when similar but not identical patterns are present, such as when non-future uses of the verb *go* facilitate the use of the *going-to* future (cf. Pickering and Branigan, 1998; Gries, 2005; Szmrecsanyi, 2005, 2006; Snider, 2009). Finally, recent studies show that priming of clause-level constructions is lexically-specific, specific not only in the above sense that it is stronger when the verb in prime and target is the same, but also such that verbs are differently strongly associated with particular constructions. To my knowledge, Gries (2005) was the first study to discuss this notion, using distinctive collexeme analysis to operationalize verb-specificity; Jaeger and Snider (2008) approach this topic using a different operationalization, an interesting information-theoretic surprisal measure, apparently unaware of Gries's similar approach.

So far, many different cognitive/psycholinguistic determinants of priming have been discovered, but sociolinguistic factors have received much less attention. This is in part due to the fact that most work on priming has been experimental in nature. However, even though there are some corpus-based studies of priming that are compatible with a discourse- and/or sociolinguistic perspective, there is little in terms of, for example, MODE or more fine-grained register differences. Section 2.2 will explore the dative alternation, in particular the degree to which the mode (speaking vs. writing) plays a role for priming effects; section 2.3 will discuss particle placement with an eye to the role of how register affects the predictability of constructional choices based on priming.

2.2. MODE and the priming of the dative alternation

To explore the role of MODE, the data studied in Gries (2005) were complemented by an additional annotation process. Gries (2005) retrieved altogether 3003 prime-target pairs from the British Component of the International Corpus of

English (ICE-GB), a one-million word corpus of British English from the 1990s, 60% spoken data, 40% written data. Each prime-target pair was then annotated for a variety of variables:

- Mode: speaking vs. writing, which is the admittedly crude proxy of a sociolinguistic variable;
- CPrime and CTarget (the constructions in prime and target): ditransitive vs. prepositional dative;
- VLemmaID and VFormID (whether the verb lemma and the verb from were identical in prime and target): yes vs. no;
- SpkID (whether the speaker/writer was the same in prime and target): yes vs. no;
- Distance: the logged and *z*-standardardized distance between prime and target (measured in the ICE-GB's own parse units);
- Coll: the target verb's preference for a construction based on Gries and Stefanowitsch's (2004) data on the dative alternation: small and large values indicate preferences for ditransitives and prepositional datives respectively;
- Surprise: a value between 0 and 1 with low and high values indicating low and high degrees of surprise (in the sense that the target verb is used in a construction it does not 'like').

I then ran a binary logistic regression with CTarget as the dependent variable and subsequent automatic model selection (using *AIC*) on this data set. The maximal model I began with included all main effects and all two- and three-way interactions involving Mode or Surprise so that both psycho- and sociolinguistic variables were included in interactions. After a model selection process during which insignificant predictors were deleted,[3] the final (minimal adequate) model resulted in a highly significant correlation between the predicted probabilities of the prepositional dative and the observed constructional choices (log-likelihood $\chi^2 = 1308$, $df = 17$, $p < 0.001$, $C = 0.875$), with a percentage of correctly classified constructions of 78.2% (random baseline: 50.9%). Crucially for the purposes of the present article, the final model involves psycholinguistic determinants which interact with sociolinguistic variables. The strongest effects of the model are the following:

- a strong effect of Coll: verbs occur in target constructions that they typically prefer ($p < 0.001$);
- one manifestation of a priming effect: prepositional datives prime prepositional datives and ditransitives prime ditransitives, especially when the verb lemmas in prime and target are identical (CPrime × VLemmaID, $p < 0.001$) and the verb forms in prime and target are identical (CPrime × VFormID, $p < 0.001$);
- another manifestation of a priming effect: prepositional prime prepositional datives especially strongly when the distance between prime and target is small (CPrime × Distance, $p = 0.043$);
- finally and most interestingly, the significant interaction CPrime × Distance × Mode ($p = 0.025$): the closer the distance between prime and target, the stronger the priming effect, but less so for written ditransitives.

In sum, most results are compatible with previous findings, especially with psycholinguistic findings, but the final interaction is somewhat unexpected: The interaction shows that the construction of the prime interacts significantly with both a cognitive/psycholinguistic determinant – Distance – and a sociolinguistic determinant – Mode.

This has two kinds of implications. One is strictly methodological and I mention it only to further bring home the point that statistical methods must be able to properly include continuous variables. Even in a tiny case study like the present one, two continuous variables had strong and significant effects on the alternation studied.

The other type of implication is more general. First, I hope this small case study illustrates the large amount of influence that psycholinguistic variables *can* have and how psycho- or sociolinguistic variables *may* only exert an effect in particular combinations of linguistic and psycho- or sociolinguistic variables. I think it would be beneficial for both psycho- and sociolinguists to recognize this at least as a possibility and let this guide them in their research designs. Second, the sociolinguistic variable included here, Mode: spoken vs. written, is particularly interesting because it may actually be just an umbrella term that unites several different, and ultimately cognitive, aspects of communication (such as those identified in Biber's factor-analytic work, e.g., Biber, 1988; cf. above section 1). This raises a third interesting implication: if Mode indeed were just a variable emerging from the combinations of different communicative dimensions, this would mean that it could be interesting to conduct a more fine-grained exploration of this variable. In fact, especially in corpus-linguistic studies, it may *generally* be useful to explore different levels of granularity, or resolution, at the same time, and the second case study will exemplify this logic in the following section.

[3] *Pace* some authors, model selection based on comparing the predictive power of two models, one with and one without an individual variable to assess whether that variable contributes significantly to the model seems to be the current standard approach to identifying significant; cf. Crawley (2007) for much discussion and exemplification.

Table 1
Design of the ICE-GB.

Mode	Register	Sub-register
Spoken	Dialog	Private, public
	Monolog	Scripted, unscripted
	Mix	Broadcast
Written	Printed	Academic, creative, instructional, nonacademic, persuasive, reportage
	Unprinted	Letters, nonprofessional

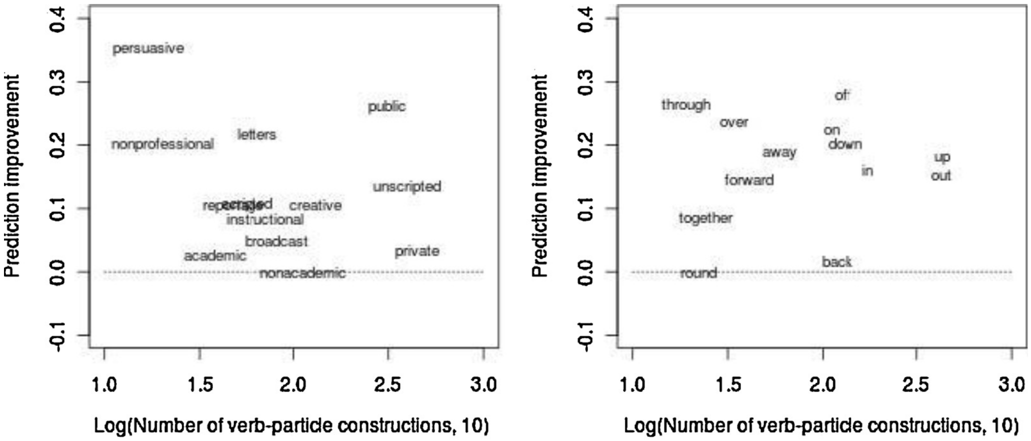

Fig. 1. The percentage of prediction improvement over chance from logistic regressions: left panel: regressions in sub-registers; right panel: regressions for particles.

2.3. Sub-registers and lexical effects on priming of particle placement

The second case study also involves a constructional alternation, this time the one referred to as particle placement and exemplified in (2). 1797 prime-target pairs of verb-particle constructions were retrieved from the ICE-GB and annotated for a set of variables. As before, a variety of variables were annotated:

- CPrime and CTarget (the constructions in prime and target): V-Prt-NP vs. V-NP-Prt;
- VLemmaID and VPartID (whether the verb and the particle lemma were identical in prime and target): yes vs. no (for each);
- Coll: the target verb's preference for a construction based on Gries and Stefanowitsch's (2004) data on the particle placement: small and large values indicate preferences for VPrt-NP and V-NP-Prt respectively;
- Distance: the distance between prime and target (measured in the ICE-GB's own parse units);
- Register: which of 13 sub-registers of the ICE-GB the prime-target pair was found in Table 1.

The data were explored in two different ways. For the first exploration, I computed 13 logistic regressions, one for each sub-register, in which CTarget was predicted on the basis of the other variables and their interactions. In addition, I also computed another set of 13 logistic regressions, one for each of the 13 most frequent particles in all of the data. For this part, the exact results of the regression and the importance of the variables are irrelevant – what is crucial is the stunning amount to which the regressions' predictive accuracies are different for different sub-registers and for different particles. The two panels of Fig. 1 plot how much better than chance the logistic regressions in the sub-registers (left panel) and for the particles (right panel) could predict particle placement (on the *y*-axis, the *x*-axis represents the number of verb-particle constructions that entered into the logistic regressions and it is quite reassuring to see that there is no strong correlation of the prediction improvement with the frequencies of the constructions).

The results are a clear and distinct warning to prematurely generalize without simultaneously taking sociolinguistic and lexically-specific sources of variation into consideration. The left panel shows very clearly that even factors that have been proven to impact particle placement in very many studies are far from helping prediction across the board *even within one and the same register*: in the sub-register of printed non-academic writing, particle placement prediction is at chance level, but in printed persuasive writing, prediction accuracy increases by more than 35%. And within spoken private dialog,

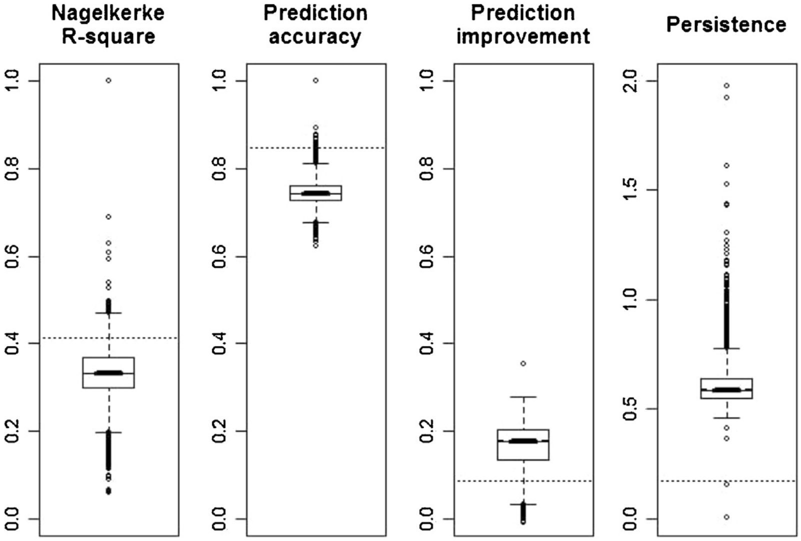

Fig. 2. Results from studying particle placement in all possible corpus parts.

particle placement is very hard to predict, whereas in spoken public dialog, it can be predicted well. The fact that even within something as specific as printed writing or spoken dialog we can observe these huge differences of predictive power shows that sociolinguistic parameters should play an important rule in studies of syntactic alternations in general as well as in studies of psycholinguistic phenomena such as priming.

The results for the particles reveal a similarly high degree of lexically-specific variability. Constructions involving *round* and *back* are hard to predict even with time-proven predictors, constructions involving *off* and *through*, conversely, exhibit a strong prediction improvement.

As a second way of exploring the data, I generated all 8191 corpus samples that can be generated by combining all 13 sub-registers: all 13 individual sub-registers, all 78 pairs that can be formed out of 13 sub-registers, all 286 triples that can be formed out of 13 sub-registers; and so on. For each of these 8191 corpus samples, I ran a logistic regression and stored four parameters:

– the Nagelkerke R^2 of the regression;
– the prediction accuracy of the logistic regression in percent;
– the amount of improvement of the prediction accuracy over chance;
– the effect size of the psycholinguistic priming effect.

That is to say, for each of these four parameters 8191 figures were obtained, which were then summarized in boxplots (and compared to previous results from Szmrecsanyi, 2005, which involved many more predictors and are indicated with horizontal dotted lies) in Fig. 2.

Again, the results indicate just how much variability interactions between socio- and psycholinguistic parameters can involve. The correlation strengths range from below 0.1 (an extremely weak correlation) to 0.7 and higher, with a median significantly lower than Szmrecsanyi's study. Correspondingly, prediction accuracies and prediction improvements are also very variable, indicating, as did the results of Fig. 1, that choosing different sub-register parts of the very same corpus can yield *dramatically* different results. The fourth panel of Fig. 2 makes this point most clearly: the effect sizes (exponentiated regression coefficients) of the psycholinguistically-motivated variable of priming ranges the whole gamut from nearly 0 via 1 (i.e., no effect) up to 2. In other words, depending on which corpus data are inspected, the prime construction can have the expected effect (<1), which it does most of the time, no effect, or even the opposite effect in certain cases. This is a very clear indication of how powerful sociolinguistic determinants can be, not only for predictive accuracies of models in general, but even for very specific psycholinguistic predictors of constructional choices.

3. Case study 3: diachronic morphological change

The final case study in this paper attempts to unite many of the above points: the need to include both cognitive/psycholinguistic and sociolinguistic factors (all of the above sections), to keep an open mind regarding which distinctions

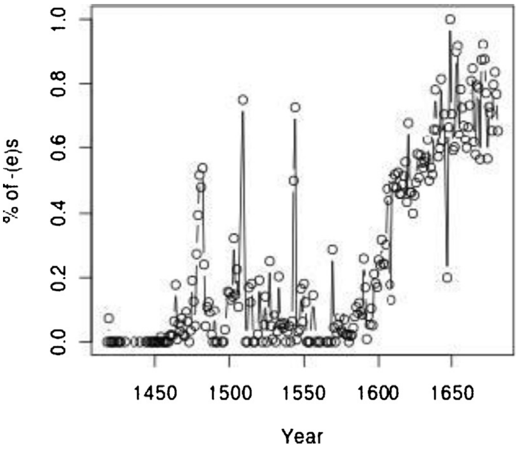

Fig. 3. Results from studying particle placement in all possible corpus parts.

the data might support most (section 2.3), to include lexically-specific effects (section 2.3), and to use statistical tools that can handle diverse data. To reiterate, by diverse data I mean categorical as well as continuous data but also, crucially, interactions. In particular, I would like to draw attention to how Varbrul approaches to interactions can be problematic. First and as mentioned above, interactions cannot be included straightforwardly. Second and more importantly, maybe it is this fact that sometimes leads to researchers' dangerous negligence of interactions.

Consider as one example a study that, like the present case study, involves diachronic change. Jankowski (2004) studies the change of English deontic modality over four time periods and across two varieties of English. In other words, she has two independent variables: Variety (*british* vs. *american*) and TimePeriod (*1902–26*, *1927–51*, *1952–76*, and *1977–2001*). Unfortunately, however, she studies the change over time by conducting *four separate* Varbrul analyses, one for each time period. Thus, since the explanatory variables of Variety and TimePeriod are never exhaustively crossed with themselves *and* all other explanatory variables studied, the analysis can by definition not test all important interactions for significance. For instance, because of this design, Jankowski is not able to test whether Varbrul weights for any explanatory variable (e.g., SubjectReference or VerbType) differ significantly from each other across time periods. Crucially, however, this is exactly what would be necessary to distinguish irrelevant sampling variation from interesting distinctions between the studied time periods: strictly speaking, the way these data were analyzed makes them not speak to the issue of temporal development!

In this case study, I will briefly discuss a study that exemplifies all of the above points on the basis of the change of the English third person singular suffix from 1400 to 1700, based on Gries and Hilpert (2010). We retrieved approximately 21,000 third person verb suffixes from 233 different years from the Corpus of Early English Correspondence (CEEC). The first problem we faced was how to determine which time periods to distinguish because, as is shown in Fig. 3, the ratio of -*(e)s* to -*(e)th* is far from following a simple linear trend.

Given such data, where it is clear that different researchers would recognize very different temporal stages in the data, the first way in which the above advice was heeded was that we used a bottom-up clustering algorithm (VNC; cf. Gries and Hilpert, 2008) to identify outliers as well as the number and duration of temporal stages in the trend from -*(e)th* to -*(e)s*. This data-driven procedure yielded five temporal stages, which were then used as the variable Time in the remainder of the analysis. We then used a generalized mixed-effects model to try to predict the choice of -*(e)s* and -*(e)th* on the basis of a large number of sociolinguistic, psycholinguisic, and phonological independent variables:

- AuthorGender: male vs. female;
- RecipientSameGender: yes vs. no;
- Recipient=CloseFamily: yes vs. no;
- VerbStemWithFinalSibilant: yes (*promise*) or no (*go*);
- FollowingFricative: yes (*walks swiftly*) or no (*walks home*);
- VerbType: lexical vs. grammatical;
- Priming: the last suffix used: -*(e)s* vs. -*(e)th*.[4]

[4] We would have included a variable Dialect if the corpus compilers, who use such a variable in their own study of the same corpus, had not refused to make those data available.

Importantly, we not only included all these variables, but also the interaction of each of these variables with Time so that we, as opposed to Jankowski (2004), would be able to identify where a variable's effect changes over time.

Finally, we also included two so-called random factors in our model: one that results in a little correction of the regression equation for each letter writer, and one that results in a similar correction for each verb lemma. It is these two random intercepts that turn this into a mixed-effects model.

The results were quite astonishing, in particular with regard to the classification accuracy that was obtained. The minimal adequate model resulting from the deletion of all non-significant predictors resulted in a classification accuracy of 94.6%, which not only reflects a highly significant correlation between the predictors involved, but also is significantly better than the classification accuracy obtained from a model without the two random effects (86.4%), which shows that including speaker- and lexically-specific parameters in the analytical process can be extremely useful: not only does it boost the classification accuracy, but it also makes the estimates for all predictors much more precise and, thus, enhances our understanding of the effects.

What about the other predictors and their interactions of predictors with Time? Space precludes an exhaustive discussion of the findings here, but there are several interesting results that could only be obtained by including cognitive/psycholinguistic predictors, sociolinguistic predictors, and their interaction with Time in the model. For example, the psycholinguistic variable of Priming has the expected and strong effect, and it has this effect in all five time periods – i.e., the interaction Priming:Time is not significant. Similarly, AuthorGender had the same effect across the 300 years studied: women were more likely to use the progressive, new form. On the other hand, the interaction RecipientSameGender:Time is significant. Until the end of the 15th century, the old suffix -*(e)th* prevailed regardless of who a writer wrote to, but then during the 16th century, a statistically significant change occurred. when a writer wrote to someone of the opposite sex, the new suffix was much more likely. Similar interactions with Time were observed for the phonological variables, but at later stages: Until the end of the 16th century, the old suffix -*(e)th* prevailed regardless of who a writer wrote to, but then during the first half of the 17th century, verbs whose stem did not end in a sibilant (such as *come*) would already carry the new ending whereas those that did (such as *cause*) would still resist the change.

In sum, this case study brought together different aspects that the previous case studies already alluded to. It revealed the obvious impact of time and some expected effects of cognitive/psycholinguistic as well as sociolinguistic predictors. More importantly, the data also revealed that methodologically more sophisticated approaches – bottom-up approaches to maximize the utility of our temporal/regional etc. stages, including predictors of different types, including interactions of predictors, using regressions that correct for speaker- and/or lexically-specific characteristics, and others – result in analyses that do more justice to the real complexity of the data.

4. Concluding remarks

Methodologically, I hope to have shown that it is essential that methods newer than Varbrul such as logistic regression or mixed-effects models are explored and utilized. They have a lot to offer – in particular the elegant treatment of interactions with and across internal and external factors as well as the possibility to account for idiosyncratic patterns of words, speakers, . . . – and can herald a new era of quantitative analysis in sociolinguistics. That being said, it should also be clear that much work is still required before we have a full-fledged toolkit available.

For example, mixed-effects models are maybe the most promising recent development in statistical modeling, but as far as I can see there are still some open questions. Authors disagree about how to compute *p*-values for random effects and about how model selection should proceed (forward or backwards?), and if backwards, when should the inclusion of random effects be tested (only after all significant fixed effects have been identified or earlier?), and what does a maximal model with random effects even look like (does it include random intercepts *and* slopes for all predictors?). However, it is hopefully only a matter of time until these issues are resolved, and once they are, interesting questions with more theoretical implications arise such as how different factors should be treated in such regression modeling approaches. For instance, it is probably uncontroversial that individual speakers would just about always be entered into such an analysis as a random effect since one would want to generalize to speakers that were not included. On the other hand, it is not immediately obvious how a variable Register would be included: speakers usually command different registers so do we treat it as a random effect, because one wants to generalize to unstudied registers and treat Register as something that is controlled for while one studies, say, language-internal variables? Or does one treat it as a fixed effect, as the studies in the QLVL group have done (with good results)? And how does this problem extend to the variable Variety if we assume that speakers only command one variety? It will be interesting and challenging, to put it mildly, to explore the best ways to study such data . . .

As another example of risks arising from the multifactorial analysis of data, consider the method of Classification and Regression Trees (CART). While I appreciate its power and its non-parametric nature, I think it must not be used in isolation given its ability to overlook patterns in data. Consider the following data (Table 2).

Table 2
A fictitious distribution of a dependent binary variable Var4 as a function of three independent binary variables Var1, Var2, and Var3.

Var1	Var2	Var3	Var4: *x*	Var4: *y*
a	*e*	*m*	6	0
a	*e*	*n*	0	3
a	*f*	*m*	0	0
a	*f*	*n*	1	0
b	*e*	*m*	0	0
b	*e*	*n*	0	1
b	*f*	*m*	0	6
b	*f*	*n*	3	0

Note that when it comes to predicting Var4, Var1 has the best predictive accuracy of the three independent variables: 70%. A CART algorithm such as `rpart` in R (R Development Core Team, 2012) would therefore choose Var1 for its first binary split. However, note as well that the two variables Var2 and Var3, each of which have less predictive power than Var1 (60% and 50% respectively), together have a perfect predictive accuracy! The tree algorithm just 'sees' that Var1 is best for the *first* split, but does not go back to 'reconsider' and, thus, misses that Var2:Var3 would in fact be best.

As a final methodological comment, I also hope to have shown that bottom-up methods should occupy a larger proportion of our evaluative steps. Sociolinguistic (corpus) data can be categorized on many different levels of generalization, and of course even cutting across levels, so that no one division or distinction of the data can be taken for granted in an *a priori* fashion – rather, systematic exploratory data-driven study of the structure inherent in a data set is required (cf. Gries, 2006 for more discussion). In sum, one must exercise a lot of care in the right exploration, and this is of course a learning process for the discipline as well as for each individual researcher (the present author included).

Apart from the above methodological issues, there are also theoretical issues I hope to have touched upon. The main one of these is concerned with the need for more studies that explore the additive and interactive ways in which both language-internal and language-external factors affect variation. Neither discipline can seriously afford anymore to pretend the other perspective does not exist, and the data clearly show that it is only by including them all that we can come to a fuller understanding of the patterns in our data. However, following Mendoza-Denton, Hay, and Jannedy as well as the work they themselves discuss – most notably Pierrehumbert (2001), but also others – I submit that sociolinguistics can benefit more from a psycholinguistic perspective than by just including a few such predictors in their models. Exemplar models are currently the prime candidate for a theory that can *describe* and indeed *explain* how between-individual variation has the structure it has and how the properties of such a system explain language change, by virtue of a probabilistic, high-dimensional representation of individual usage events and their contexts of use (cf. Bybee, 2010:Ch. 1–2). For example, Stefanowitsch and Gries (2008:150) find that ''different constructions differ with respect to the degree to which they exhibit channel-specific collostructional relationships.'' Such a finding falls out naturally from an exemplar-based perspective since speakers' linguistic systems will keep track over time which word-construction pairings they perceive in which contexts, or channels, which in turn will not only lead to speakers developing preferences for word-constructionchannel triples, but also to statistical methods such as collostructional analysis picking up such preferences.

All in all, as cognitive-, corpus-, and psycholinguistic methods and notions help sociolinguistics evolve further, and as psycholinguistics benefits from the recognition of the importance of language-external factors, we all are on our way to a unified treatment of between- and within-individual variation: these are exciting times for cognitive sociolinguistics . . .

Acknowledgements

I thank Stefanie Wulff, members of the audience at LAUD 2010 (in particular Dennis Preston), and two reviewers for comments/feedback; the usual disclaimers apply.

References

Baayen, R. Harald, 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge University Press, Cambridge.
Biber, Douglas, 1988. Variation Across Speech and Writing. Cambridge University Press, Cambridge.
Bybee, Joan L., 2010. Language, Usage, and Cognition. Cambridge University Press, Cambridge.
Cedergren, Henrietta J., Sankoff, David, 1974. Variable rules: performance as a statistical reflection of competence. Language 50 (2), 333–355.
Clark, Herbert H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12 (4), 335–359.
Cochran, William G., 1951. Testing a linear relation among variances. Biometrics 7 (1), 17–32.
Crawley, Michael J., 2007. The R Book. John Wiley and Sons, Chichester.
Eckert, Penelope, submitted for publication. Three waves of variation study: the emergence of meaning in the study of variation.

Forster, Kenneth I., Dickinson, R.G., 1976. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates F_1, F_2, F', min F'. Journal of Verbal Learning and Verbal Behavior 15 (2), 135–142.

Frank, Austin F., Jaeger, Florian T., 2008. Speaking rationally: uniform information density as an optimal strategy for language production. In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the 30th Annual Meeting of the Cognitive Science Society. The Cognitive Science Society, Washington, DC, pp. 939–944.

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

Goldberg, Adele E., 1995. Constructions: A Construction Grammar approach to Argument Structure. The University of Chicago Press, Chicago, IL/London.

Goldberg, Adele E., 2003. Constructions: a new theoretical approach to language. Trends in Cognitive Sciences 7 (5), 219–224.

Gries, Stefan Th., 2005. Syntactic priming: a corpus-based perspective. Journal of Psycholinguistic Research 34 (4), 365–399.

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

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

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

Gries, Stefan Th., Stefanowitsch, Anatol, 2004. Extending collostructional analysis: a corpus-based perspective on ‘alternations’. International Journal of Corpus Linguistics 9 (1), 97–129.

Grondelaers, Stefan, Speelman, Dirk, Geeraerts, Dirk, 2002. Regressing on er. Statistical analysis of texts and language variation. In: Morin, A., Sébillot, P. (Eds.), 6th International Conference on the Statistical Analysis of Textual Data, Cedex, Rennes, pp. 335–346.

Halliday, Michael A.K., 2005. Computational and Quantitative Studies. Continuum, London/New York.

Hay, Jennifer, Nolan, Aaron, Drager, Katie, 2006. From fush to feesh: exemplar priming in speech perception. The Linguistic Review 23 (3), 351–379.

Jaeger, T. Florian, Snider, Neal, 2008. Implicit learning and syntactic persistence: surprisal and cumulativity. In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the Cognitive Science Society Conference, Washington, DC, pp. 1061–1066.

Jankowski, Bridget, 2004. A transatlantic perspective of variation and change in English deontic modality. Toronto Working Papers in Linguistics 23 (2), 85–113.

Johnson, Keith, 2008. Quantitative Methods in Linguistics. Blackwell, Malden, MA.

Johnson, Daniel Ezra, 2009. Getting off the GoldVarb standard: introducing Rbrul for mixed-effects variable rule analysis. Language and Linguistics Compass 3 (1), 359–383.

Johnson, Daniel Ezra, 2010. Rbrul Version 2.0.2. A Function for R. http://www.danielezrajohnson.com/Rbrul.R, (accessed 30.03.12).

Kristiansen, Gitte, Dirven, René (Eds.), 2008. Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems. Mouton de Gruyter, Berlin/New York.

Mendoza-Denton, Norma, Hay, Jennifer, Jannedy, Stefanie, 2003. Probabilistic sociolinguistics. In: Bod, R., Hay, J., Jannedy, S. (Eds.), Probabilistic Linguistics. The MIT Press, Cambridge, MA, pp. 97–138.

Paolillo, John, 2002. Analyzing Linguistic Variation. CSLI Publications, Stanford, CA.

Pickering, Martin J., Branigan, Holly P., 1998. The representation of verbs: evidence from syntactic priming in language production. Journal of Memory and Language 39 (4), 633–651.

Pierrehumbert, Janet, 2001. Exemplar dynamics: word frequency, lenition and contrast. In: Bybee, J.L., Hopper, P. (Eds.), Frequency and the Emergence of Linguistic Structure. John Benjamins, Amsterdam/Philadelphia, pp. 137–157.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. , http://www.R-project.org/.

Sankoff, David, Laberge, Suzanne, 1978. Statistical dependence among successive occurrences of a variable in discourse. In: Sankoff, D. (Ed.), Linguistic Variation: Methods and Models. Academic Press, New York, pp. 119–126.

Satterthwaite, Franklin E., 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin 2 (6), 110–114.

Schnoebelen, Tyler, 2008. Measuring compositionality in phrasal verbs. Unpubl. ms., Stanford University.

Snider, Neal, 2009. Similarity and structural priming. In: Taatgen, N.A., van Rijn, H. (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 815–820.

Speelman, Dirk, Geeraerts, Dirk, 2009. Causes for causatives: the case of Dutch doen and laten. In: Sweetser, E., Sanders, T. (Eds.), Causal Categories in Discourse and Cognition. Mouton de Gruyter, Berlin/New York, pp. 173–204.

Stefanowitsch, Anatol, Gries, Stefan Th., 2008. Channel and constructional meaning: a collostructional case study. In: Kristiansen, G., Dirven, R. (Eds.), Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems. Mouton de Gruyter, Berlin/New York, pp. 129–152.

Szmrecsanyi, Benedikt, 2005. Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. Corpus Linguistics and Linguistic Theory 1 (1), 113–150.

Szmrecsanyi, Benedikt, 2006. Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis. Mouton de Gruyter, Berlin/New York.

Szmrecsanyi, Benedikt, 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In: Geeraerts, D., Kristiansen, G., Peirsman, Y. (Eds.), Advances in Cognitive Sociolinguistics. Mouton de Gruyter, Berlin/New York, pp. 141–165.

Weiner, Judith, Labov, William, 1983. Constraints on the agentless passive. Journal of Linguistics 19 (1), 29–58.

Stefan Th. Gries is professor of linguistics at the University of California, Santa Barbara. Methodologically, he is a quantitative corpus linguist at the intersection of corpus linguistics and computational linguistics, who works on topics in morpho-phonology, syntax, the syntax-lexis interface, semantics, corpus-linguistic methodology, dispersion measures, as well as first and second language acquisition. Theoretically, he is a cognitively-oriented linguist (with an interest in Construction Grammar) in the wider sense of seeking explanations in terms of cognitive processes without being a cognitive linguist in the narrower sense of following any one particular cognitive-linguistic theory.