Journal of Research Design and Statistics in Linguistics and Communication Science

JRDS (PRINT) ISSN 2052-417X
JRDS (ONLINE) ISSN 2052-4188

Article

The Discriminatory Power of Lexical Context for Alternations: An Information-theoretic Exploration

Stefan Th. Gries

Abstract

This paper makes a very exploratory, tentative, and thinking-aloud kind of suggestion for the corpus-based analysis of alternation data. I start from the observation that studies of alternations/choices in particular in corpus linguistics have become increasingly sophisticated in terms of the statistical methods they employ and the number of predictors they involve. While the predictors employed come from many different levels of linguistic analysis – phonology, morphosyntax, semantics, pragmatics/discoursal, textual, psycholinguistic, sociolinguistic, and others – they are usually contextual in nature, meaning they characterize the context of the choice the language user needs to make or has just made. However, one aspect of the context seems to be crucially underutilized when it comes to modeling speakers' choices: the lexical context. In this paper, I build on recent work in computational psycholinguistics to: (a) define a lexical-distribution prototype of each of the (typically, but not necessarily, two) alternants of an alternation; and (b) compute the degree to which each instance of the alternation in question diverges from each of the prototypes. Then, (c) the values that all choices score on the divergences from each of the prototypes are entered as predictors to all others in statistical models to, minimally, serve as a variable that controls for whatever information is contained in the lexical context of an instance of speaker's choice. I exemplify the approach and its sometimes amazing predictive power on the basis of a choice between near synonyms, two morphosyntactic alternations (preposition stranding vs. pied-piping and of- vs. s genitives), and a distinction between the functions of well.

Affiliation

Department of Linguistics, University of California, Santa Barbara, CA, USA
Email: stgries@gmail.com

JRDS VOL 5.1-2 2018 78–106 https://doi.org/10.1558/jrds.38227

equinox
equinoxonline

Keywords: lexical alternations; structural alternations; collocates; KL-divergence

1. Introduction

1.1 General Introduction

A particularly prominent area in corpus-linguistic research is the study of alternations, i.e., the study of what factors are correlated with, or even co-determine, speakers' linguistic choices for one of several ways of saying pretty much the same thing. The probably prototypical examples of this are morphosyntactic alternations that mostly involve different constituent orders and alternations that involve the (lack of) realization of some linguistic expression; well-known examples of the former include the dative alternation (see (1)), the genitive alternation (see (2)), particle placement (see (3)), etc., examples of the latter include *that*-complementation (see (4)) or relativizer omission (see (5)).

(1) a. Sheridan gave Garibaldi the folder.
    b. Sheridan gave the folder to Garibaldi.
(2) a. Delenn was scared by the Vorlon's power.
    b. Delenn was scared by the power of the Vorlon.
(3) a. Londo gave back the jewelry.
    b. Londo gave the jewelry back.
(4) a. G'Kar said the Centauri were a lost race.
    b. G'Kar said that the Centauri were a lost race.
(5) a. Mr. Bester found the telepaths he was looking for.
    b. Mr. Bester found the telepaths that he was looking for.

Typically, the way these kinds of phenomena are studied corpus-linguistically involves retrieving a hopefully reasonable number of matches from some corpus/corpora, annotating them for: (a) the response variable (i.e. which of the, in all above cases, two levels of the response a speaker chose); and (b) predictors that are known or hypothesized to correlate with, and therefore hopefully explain (statistically and linguistically/theoretically), the distribution of the response variable. These days, most studies of this type involve some sort of (mixed-effects) regression modeling, classification/conditional inference trees, random forests, or other kinds of classifiers (see many examples cited below). Trivially, the above kind of classification approach means that, apart from the prototypical cases listed above, many other phenomena can be, and are, studied from a similar methodological perspective such as the choice of one of multiple near synonymous words or the function that a particular expression has.

Given the increasingly statistical nature of linguistics as a discipline, such studies have become more and more sophisticated and the field has uncovered that, for instance, many of the alternations studied so far are correlated with an astonishingly large number of predictors from all sorts of linguistic (language-internal and language-external) levels of analysis:

- *discourse-pragmatic and/or sociolinguistic predictors* involving givenness/newness or inferrability as well as the discourse importance of referents (e.g., Chen, 1986), but also register/genre, mode of production (e.g., speaking vs. writing), speaker sex/class (however operationalized), etc.;
- *semantic predictors* such as overall literalness/idiomaticity of a phrase, animacy and/or concreteness of referents involved in the constructions in question (e.g. Wolk *et al*., 2013), and, for the choice between near synonyms, sometimes very subtle differences in meaning/function that speakers are often not aware of;
- *morphosyntactic predictors* such as weight/complexity (Behaghel, 1909), definiteness (as operationalized in terms of the determiner a noun might take), type of head of an NP (e.g., lexical vs. pronominal) (e.g., Givón, 1983), etc.;
- *phonological predictors* such as (contrastive) stress, rhythmic alternation or segment alternation in general (Gries, 2018b), ease of articulation in particular (e.g., the effect of sibilancy on the genitive alternation, see Rosenbach 2002), etc.;
- *psycholinguistic predictors* involving constructional preferences of words (Gries and Stefanowitsch, 2004) or speakers, priming (Szmrecsanyi, 2006) and *horror aequi* (Rohdenburg, 2003), surprisal (Hale, 2001; Jaeger and Snider, 2008), etc.

In empirical studies, we often find that many of the above kinds of predictors interact with each other, i.e., some predictor (level) may strengthen or weaken or reverse the effect of another predictor (level) on the relevant linguistic choices. For instance, in particle placement there is a strong tendency to prefer the V-Part-DO order when the meaning of the verb phrase is idiomatic (as opposed to when it refers to literal movement of the referent of the direct object to the location or along the path denoted by the particle as in (3)). However, when the DO is pronominal, that overrides that strong preference of idiomatic meanings. Also, with pronominal DOs in general, V-DO-Part is virtually obligatory – unless the pronoun receives contrastive stress, in which even a pronominal DO could follow the particle as in *Mr. Bester took back HER [not HIM]*.

The above classification of predictors into classes is heuristic only and comes with no theoretical commitments: length/weight has been cast as a syntactic or a phonological predictor and is obviously related to givenness/newness and definiteness (given referents are more likely to be expressed with shorter/less heavy and definite NPs) as well as head type (given referents are more likely to be expressed pronominally than new referents), etc. However, what all predictors have in common is that they are contextual in nature: They all have to do with the context in which the linguistic choice was made – either in terms of the situation of production or in terms of characteristics of referents or linguistic expressions in the context of the utterance that have been produced already or that are about to be produced.

In a sense that is trivial: no linguist would deny that context is relevant. This incontestable importance of context notwithstanding it appears to me as if one aspect of the linguistic context of the choice (to be) made is usually not considered, neither as a predictor nor, minimally, as a control variable, and that is the lexical context. Yes, some aspects of the lexical context of a choice are sometimes considered:

- In Szmrecsanyi's (2006) above-mentioned work, he computes the type-token ratio (TTR) of the words in a context window around the linguistic choice under consideration, which can serve as a useful proxy of lexical complexity and, thus, indirectly tell us at least about register/genre;
- in the same monograph, Szmrecsanyi discusses the frequently employed 'usual' notion of syntactic priming (Estival, 1985; Gries, 2005), which he calls α-persistence, but also adds an approach to priming that goes beyond that and which he calls β-persistence, i.e., the fact that a form of the verb *go* is significantly correlated with a preference of the *going-to* future over the *will*-future; even if that use of *go* is as a motion verb (and does, therefore, not involve 'traditional' syntactic priming);
- many studies whose predictors involve semantic characteristics of referents of course require paying intention to the lexical material in the relevant slots, as when, in the case of genitives, possessor and possessum are annotated for animacy and/or concreteness (given that the prototypical *s*-genitive involves an animate/human possessor and an inanimate/concrete possessum, as in *Mr. Garibaldi's peperoni*).

However, these kinds of approaches, while involving some aspects of lexical context, do not use much of the available distributional information: The TTR summarizes the lexical context with only a single number and does not take individual words into consideration once the computation has used

the information of whether a token was a new type in the context window or not. And the other two examples involve a very restricted view of lexical context, namely what may or may not happen in a previous context (has a certain word/construction been observed before or not?) or what is the category of a word in a certain slot. While I do not deny the potential utility of any of these methods – I have used nearly all of them myself many times – the way they incorporate lexical context is simply not particularly 'rich'. From a cognitive-linguistic or usage-/exemplar-based theoretical perspective, one that is informing many alternation studies in the last 15–20 years, one cannot help but wonder whether there's more to it than it seems, and this paper will demonstrate that there is.

In Section 2, I will outline the methodological approach I am proposing; I will first present the overall logic of it before I motivate the main two computational steps. After that, Sections 3 to 6 will present four case studies that follow the general methodology; Section 7 will offer some very tentative thoughts on possible conclusions.

2. Methods

2.1 The Overall Logic and Motivation

The idea to be explored here is to include information about the overall lexical content – not just what happens in a small number of lexical/constructional slots as classified into a usually small number of levels – in, say, a regression model and determine how that lexical information may be correlated with speakers' linguistic choices. However, given the wide range of choice/alternation phenomena and the large number of matches that are often characteristic of alternation studies in contemporary corpora, it is important that such an approach be: (a) generic enough to be applicable to a wide range of phenomena; and (b) scalable up to data sets that involve potentially tens of thousands of concordance lines and even many more collocates. My suggestion boils down to the following steps (here exemplified on the basis of the genitive alternation):

- For each alternant, a prototype is computed based on the distribution of the lexical collocates around the matches with that alternant. More concretely, one computes an *of*-prototype based on the distribution of all collocates around the *of*-genitives and an *s*-prototype based on the distribution of all collocates around the *s*-genitives. (I will explain the computation of the prototypes below.)
- Once a prototype for each alternant has been computed, then each instance of the linguistic choice – i.e. every *of*- and every *s*-genitive

– will be compared to each prototype to determine how much each instance diverges from each prototype. These (two) divergences of each instance to each of the (two) prototypes will be stored in two new variables called, for instance DivFromOf and DivFromS, which therefore express how 'lexically dissimilar' each instance is from either prototype.
- these (here two) new variables encode lexical-context-based information and are available to be used as predictors on top of the usual ones in, say, regression models trying to predict the (lexical, constructional, or other kind of) choice.

The crucial questions are now of course threefold: (a) how to compute the prototype for each alternant; (b) how to compute how much each concordance line/match with its context differs from each prototype; and (c) does this do anything (in terms of classificatory power)?

2.2 Lexically-based Prototypes

As for questions (a) and (b), the approach adopted here is based on recent work in computational psycholinguistics, specifically work by Milin *et al.* (2009), Baayen *et al.* (2011), and Lester (2018) and their information-theoretically-inspired definition of prototypicality in distributional psycholinguistics. Milin *et al.* (2009) explore reaction times to Serbian nouns from a visual lexical decision task and show that the reaction times are significantly correlated with the degree to which a word's morphological frequency profile – how often the noun is attested with each inflectional affix – is different from the overall frequencies of each inflection affix. In other words, in their study, the vector of overall frequencies of each inflectional affix constitutes the prototype, the degree to which an individual noun's frequencies of inflectional affixes differ from the overall frequencies is the divergence from the prototype, and those divergences are significantly correlated with reaction times.

One may wonder what the definition of the overall frequencies as the prototype is based on, which is discussed in detail in both Baayen *et al.* (2011) and Lester (2018). For instance, Baayen *et al.* argue that:

> The probability distributions of the exponents in an inflectional class can be viewed as the prototypical distribution of case endings for that class. The probability distribution of a given word's inflected variants can be viewed as the distribution of a specific exemplar. […] Milin, Filipović Đurdević, and Moscoso del Prado Martín (2009) showed empirically that a greater […] distance from the prototype […] goes hand in hand with longer visual lexical decision latencies. (Baayen *et al.*, 2011: 441)

Lester's view is similar in how he frames this in terms of prototype theory: Discussing Milin *et al.*, he summarizes '[w]ords that matched the average

distribution of nouns from their [inflectional] class, were recognized faster. We refer to this effect as a prototypicality effect. Excusing the homuncular analogy, these lexical prototypes may be thought of as the '"expectations" of the processor' (Lester, 2018: 31).

How is this put into practice? Imagine the miniature corpus result of a genitive concordance shown in Table 1, where the columns L3 to L1 contain schematic left collocates of the genitive choice (in the column MATCH) and the columns R1 to R3 contain schematic right collocates of the genitive choice.

Table 1: Schematic result of a concordance of genitives

L3	L2	L1	MATCH	R1	R2	R3
a	*b*	*c*	***Of***	*d*	*e*	*f*
a	*b*	*g*	***Of***	*b*	*h*	*i*
d	*e*	*f*	***S***	*g*	*h*	*i*
g	*h*	*i*	***S***	*a*	*b*	*c*

Step 1 of computing the prototypes is to convert the above into the frequency table shown in Table 2.

Table 2: Frequency table of the collocates per genitive

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***
of	2	3	1	1	1	1	1	1	1
s	1	1	1	1	1	1	2	2	2

Step 2 is converting these frequencies into per-genitive, i.e. row-wise, percentages, and those two vectors (each of which will sum up to 1) are then, following the above logic, the prototypes; this is shown in Table 3.

Table 3: Percentage table of the collocates: the prototypes of each genitive

	a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***
of	0.167	0.250	0.083	0.083	0.083	0.083	0.083	0.083	0.083
S	0.083	0.083	0.083	0.083	0.083	0.083	0.167	0.167	0.167

That means, the prototypes yielded by this approach are abstract or distributional in nature, they are not concrete examples; they are vectors of collocate percentages. This approach is gratifyingly simple – it basically involves nothing more than frequency lists of concordance contexts and should be usable for anyone who can handle concordancing tools. One important point is that this approach does *not* involve a commitment to a one-dimensional

continuum of genitive use with an *of*-genitive prototype endpoint on one end and an *s*-genitive prototype endpoint on the other and it does so *by design*. While such a commitment might seem obvious or even desirable at first sight, it is in fact not: One needs to bear in mind that there will always be uses that are quite different from *both* the most prototypical *of*-genitive use(s) and the most prototypical *s*-genitive use(s). For instance, these could be cases where the speaker in the corpus used a genitive but most other speakers would have used a N-N compound instead. Thus, one's operationalization must allow for that possibility and therefore not treat the two prototypes as being located on a single continuum.

2.3 Divergences of Cases from Prototypes

As for question (ii), the comparison of each concordance line involving one of the alternants to the prototypes will be made on the basis of the so-called *relative entropy* or *Kullback-Leibler divergence* (KL-) divergence. The KL-divergence is written as D_{KL} (P posterior/data || Q prior/theory) and expresses how much a posterior/data probability distribution of an element diverges from the overall/theoretical overall probability distribution, which also means that (i) D_{KL} is not symmetric (typically, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$). In the present case, P could be the distribution of all the collocates of *one* use of the *of*-genitive while Q could be the distribution of all collocates of *all of*-genitives or the distribution of all collocates of *all s*-genitives. D_{KL} is computed as shown in (6).[1]

$$(6) \quad D_{KL}(P \parallel Q) = \sum_{i=1}^{n} p_i \times log_2 \frac{p_i}{q_i}$$

In the words of Baayen *et al.* (2011: 441), D_{KL} / the relative entropy quantifies how different the exemplar is from the prototype. When the two distributions are identical (i.e., if one were computing $D_{KL}(P||P)$), the log in (6) evaluates to zero, and hence D_{KL} is zero. Another way of looking at the relative entropy measure is that it quantifies how many extra bits are required to code the information carried by one concordance line of an alternant (measured as P) compared to when the overall collocate distribution of the same alternant (measured as Q) is used in its place.

For the above example in Table 1, Table 2, and Table 3, this means that one would compute eight D_{KL}-values, the divergence of each of the four concordance lines in Table 1 (each will be P in two computations) to each of the two prototype rows in Table 3 (each will be Q in four computations). Table 4 breaks down (6) into its different steps and shows the computation of the relative entropy of the first match of the *of*-genitives in Table 1 from the *of*-prototype: The first two rows of Table 4 are the frequency tables for the first *of*-genitive's collocates, the third row is the prototype for *of*-genitives from Table 3, and the

Table 4: Computing the divergence of line 1 of Table 1 to the prototype of the *of*-genitive (from Table 3)

	a	*b*	*c*	*D*	*e*	*f*	*g*	*h*	*i*
line 1 *n*	1	1	1	1	1	1	0	0	0
line 1 %	0.167	0.167	0.167	0.167	0.167	0.167	0	0	0
of protot.	0.167	0.25	0.083	0.083	0.083	0.083	0.083	0.083	0.083
division	1	0.667	2	2	2	2	0	0	0
log2ging	0	-0.585	1	1	1	1	0	0	0
multipl.	0	-0.097	0.167	0.167	0.167	0.167	0	0	0

last three rows show the computation of D_{KL}. If one sums up the last row, one obtains a D_{KL}-value of 0.571 (without rounding, the proper value is 0.5691729).

These computations would be done for the remaining seven combinations of cases and prototypes and added to the data in the form of two variables with four cases each: the variable DivsFromOf (containing the D_{KL}s of each of the four cases to the *of*-prototype) and another variable DivsFromS (containing the D_{KL}s of each of the same four cases to the *s*-prototype). These two variables could then be used as predictors in a regression or classifier to try and see whether the differences from the lexical prototypes discriminate well between the constructional choices.

In the following sections, I will discuss some applications of this method.

3. Case study 1: Six Speed Adjectives

3.1 Introduction

For the first case study, I will use an example that may seem very straightforward, namely the choice of one of six near synonyms: *brisk*, *fast*, *quick*, *rapid*, *speedy*, and *swift*. The reason why this example may seem very straightforward, if not even redundant, is that it has been known for a long time that lexical context seems to strongly co-determine lexical choices. Every corpus linguist knows relevant famous quotes from Firth or Harris on this and every corpus linguist knows the example of *strong tea* and **powerful tea* and there is a lot of work out there on how certain words prefer to collocate with certain other words (e.g., Church and Hanks, 1990; Church *et al.*, 1994). Much of this work has been based on association measures (AMs) and distinctive collocates. Often, this work is slot-based, such as when Gries (2001, 2003b) explores the differences between the two members of *ic-* and *ical*-adjectives with few known semantic differences (such as *electric(al)* or *symmetric(al)*) or when much work in Hunston and Francis's Pattern Grammar establishes correlations between slots in constructions and the meanings of verbs that 'like to go into them'.

Other work, which is ultimately based on much of this, is the work on near synonymy using the method of Behavioral Profiles (Gries, 2010). In this approach, each match of a set of synonyms is manually annotated for a large number of categorical variables from potentially any level of linguistic analysis – phonology, morphology, syntax, and especially semantics – to then determine how similar synonyms are to each other and on what dimensions they differ from each other (most).

The large amount of work on words and their collocations notwithstanding, a lot of times such studies would list the collocates (maybe ranked according to one or more AMs) that go with each synonym of a set and then discuss the semantic differences emerging from that. This kind of work has been very insightful and I am not criticizing it here, but: (a) it is work that often involves a huge amount of very difficult semi-manual annotation of the right slot(s) of the node word that contains the collocate(s); which (b) also means that only one contextual feature may be considered (what happens in slot X?). Also, (c) we usually do not learn much about the actual discriminatory/classificatory power of the collocates themselves. The approach outlined in the previous section will try to address these potential shortcomings.

3.2 Methods

I wrote an R script that retrieved from the British National Corpus World Edition (XML) all instances of the six adjectives and the whole sentences they occurred in (using the lemma/headword annotation and the POS tag 'ADJ'), which resulted in the frequencies of adjectives in Table 5. These frequencies also mean that the two baselines against which to compare any model trying to predict an adjective are 37.6% (the frequency of the most frequent adjective, *quick*) and 27.9% (random proportional guessing).

Table 5: Frequencies of six speed adjectives in the BNC XML World Edition

	brisk	fast	quick	rapid	speedy	swift
Frequency	499 (3%)	4982 (29.8%)	6303 (37.6%)	3520 (21%)	571 (3.4%)	871 (5.2%)

The lexical context was not annotated in any particularly theoretically informed way – all I did was change each word in the context (defined as the same sentence as one of these adjectives) to a combination of the lemma and the POS-tag; in other words, (7)a became (7)b (with the omission of the node adjective *rapid* in question):

(7) a. It is only rapid movements up that become uncontrollable.
b. it~PRON be~VERB only~ADV movement~SUBST up~ADV that~CONJ become~VERB uncontrollable~ADJ .

Then, I performed the above computations on the data:

- I generated a frequency table with the six adjectives in the columns and all collocates ever attested with at least one of the six adjectives in the rows and their co-occurrence frequencies in the cells (i.e., a transposed version of Table 2);
- I converted each column of frequencies into a column of column percentages, which constitute the abstract distributional prototypes of the six adjectives;
- I generated for each of the 16,476 concordance lines a similar percentage table and computed its D_{KL} from each of the six adjective prototypes, i.e. variables that would be called DivFromBrisk, DivFromFast, …, DivFromSwift.

It is vital to realize really how messy and noisy these data are: They contain no syntactic information, no precise morphological information, they are from vastly differently frequent adjectives (recall Table 5), and the sentences in which the adjectives are used are of vastly different lengths. To determine whether these divergences-from-the-prototypes have any discriminatory power (in a statistical, not a psycholinguistic, sense) at all, they were entered into a multinomial regression, specifically a model with the adjective as the six-level response variable and all six divergence vectors as the predictors (without interactions and without allowing for curvature).

Before I discuss the results in the next section, may I invite the reader to pause for a moment and guess what the result might be: how predictive can the combination of divergences from six abstract adjective prototypes really be (maybe expressed in a classification accuracy or an R^2-value of the multinomial model)?

3.3 Results

The six divergence vectors led to a highly significant model (*LR*-statistic= 33,813.92, df=30, $p\approx 0$), to which each divergence vector contributed significantly (all $LR_{deletion}$-statistics>3600, dfs=5, ps≈0). However, these results are not just due to the sample size: Nagelkerke's R^2=0.867, the classification accuracy of the model is 0.853 (significantly higher than either baseline according to exact binomial tests), and the proportional reduction of error (PRE) in 'guessing the right adjective' is extremely high (λ=0.764); see Table 6 for precision and recall scores for each adjective.

A model with pairwise interactions of all divergences did not improve classification accuracy significantly ($p_{binomial}$>0.2). What is the nature of the effects? For considerations of space, I am only showing the effects of two divergences,

Table 6: Precision and recall scores for the six speed adjectives

	brisk	fast	quick	rapid	speedy	swift
precision	0.926	0.852	0.843	0.851	0.898	0.866
recall	0.904	0.834	0.876	0.832	0.865	0.836

namely for *brisk* (because it has the highest precision/recall scores) and for *quick* (because its frequency of occurrence is highest). Figure 1 shows both results: In both plots, the divergences from the corresponding prototypes are on the *x*-axis, the predicted probabilities of adjectives are on the *y*-axis, and the colored curves indicate the predicted probabilities of the six adjectives. Clearly, in the left panel, when the divergences from the *brisk* prototype are smallest (on the left), then *brisk* (red) is predicted overwhelmingly and in the right panel, when the divergences from the *quick* prototype are smallest (on the left), then *quick* (green) is predicted overwhelmingly, and we know from Table 6 that the vast majority of these predictions are correct.

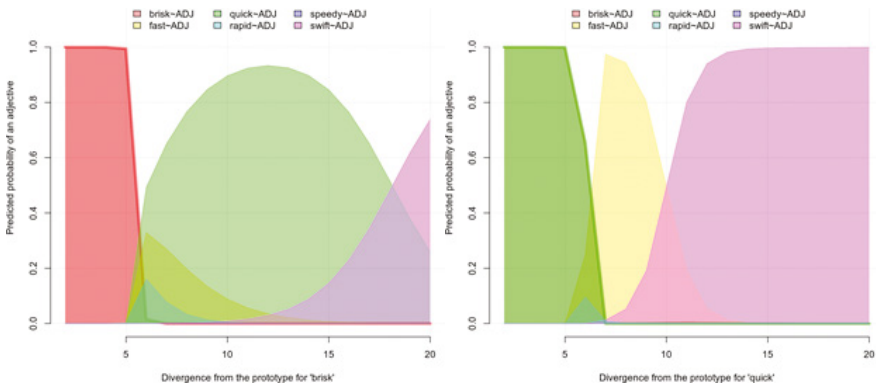

Figure 1: Predicted probability plots for DivFromBrisk (left) and DivFromQuick (right)

3.4 Interim Discussion

The reader should now revisit their estimate of what the results were going to be. On the one hand, one might have expected a relatively good result, given that so many studies have shown that there are reliable associations between node words and collocates that give rise to what Pawley and Syder (1983) called nativelike selection; therefore, a reader may have expected good results. On the other hand, recall again how unbalanced and largely unstructured the data are and how little structured data they actually contain: divergences from six prototypes defined on the basis of no semantics, no syntax, no morphology, no association strength, no dispersion, no keyness, no controls for register, nothing. Against this background, I found the results stunningly accurate.

3.5 Extending This to Structural Alternations

Given the above results, it is still very possible that lexically-defined contexts/prototypes correlate very strongly with lexical choices, but that does not also mean that lexically-defined contexts/prototypes do the same for structural choices. Again, expectations of good and bad outcomes are both defensible. On the one hand and from a construction grammar perspective, maybe especially from that of collostructional analysis – the extension of co-occurrence of lexical items (collocation) to co-occurrence of lexical items and grammatical constructions – one might again expect good results: Ever since the first distinctive collostructional alternation studies (Gries and Stefanowitsch, 2004), it has been shown time and again (mostly for verbs in constructions) that even functionally very similar constructions have apparently semantically/functionally-motivated preferences for certain words. More generally than these collostructional studies but coming from different angles, Levin (1991) or Hunston and Francis (2000) have also established strong and motivated correlations between verbs and the constructions they appear in.

On the other hand, we have the same valid reasons for expecting much worse results: The above kinds of studies were all syntactically very fine-grained, avoiding nearly all of the noise of a context window in favor of a precise and usually manual identification of the one relevant (often verb) slot of a construction. Also, it could be argued that verbs in particular, because of their rich and relational semantics are particularly good at distinguishing between (argument structure) constructions. Thus, the seemingly blunt approach of including all words in some context window again raises the specter of data too noisy to be useful for anything.

But there is another important issue that might further undermine any relevance of lexical context/prototypes as defined here. As discussed above, for many structural alternations, we already know a large number of predictors that already lead to often very good (regression) models and classification accuracies. Following Gries's (2018a) logic, we should only assume that a new predictor like these divergences is relevant if it either replaces what we already know or if it adds to what we already know, where *what we already know* could be expressed in a regression model involving predictors we know from previous work to be relevant. The next two sections do just that, explore two alternations to determine whether lexical context/prototypes replace or add predictive power to models of alternations that already involve several predictors *known* to be relevant.

4. Case Study 2: Preposition Stranding vs. Pied-piping

4.1 Introduction

This case study looks at preposition stranding in a by today's standards tiny data set from the BNC (originally explored in Gries, 2002); the alternants are shown in (8)b, with the stranded construction in (8)a and the pied-piped construction in (8)b.

(8) a. What shuttle] [$_{\text{bridging structure}}$ bay is the Vorlon ship] in?
b. In what shuttle bay is the Vorlon ship?

The present case study involves 299 cases (177 stranded, 122 pied-piped) and three predictors known to be correlated with preposition stranding:

- VerbType, the type of verb in the clause: *copula* (as in (8)) vs. *intransitive* vs. *transitive* vs. *(phrasal-)prepositional* verbs;
- PrepSem, the semantics of the preposition that is pied-piped or stranded: *spatial* (as in (8)) vs. *temporal* vs. *metaphorical* vs. *abstract*;
- a numeric predictor Length, which results from the merging of two highly-correlated values (r>0.92), namely the length of the bridging structure in words and its barrierhood (an index reflecting open-/closed-class words and frequency effects), with a principal components analysis into a factor score that covers 98.21% of the variance of the two original predictors.

Given that, unlike in the previous section, we now actually have (structural and semantic) predictors, the analysis here will use a slightly different route, which is discussed next.

4.2 Methods

The first parts will be the same as before, namely computing on the basis of the lexical contexts within the sentence the lexically-defined prototypes for the stranded and the pied-piped constructions and computing each of the 299 cases' divergences from each of the two prototypes. As a result, we know for each (stranded or pied-piped) construction how much it diverges from either prototype; that information will be kept in the variables DivFromPP and DivFromStrd.

However, since now we also have additional predictors to consider, one cannot just test the divergence vectors on their own with, say, spine plots or a monofactorial regression model (such as a generalized additive model (GAM) that allows for a curved effect of the divergences). One can *start* with that

(exploratorily), but then one needs to also show that the divergence vectors do something above and beyond the predictors we already know to be effective. Therefore, one analysis will involve only the divergence vectors, but a second one will involve the divergence vectors as well as the other predictors (VerbType, PrepSem, and Length) precisely to determine whether DivFromPP and DivFromStrd are still correlated with the constructional choices when VerbType, PrepSem, and Length are also available for the regression model. This will be done by doing two bidirectional stepwise model selection processes (using *AIC*):

- One involves using a regular generalized linear model (a binary logistic regression) starting with no predictors that is allowed to use VerbType, PrepSem, and Length and all their pairwise interactions to make the model as good as it can get;
- the other involves a generalized linear model starting either with DivFromPP and DivFromStrd that is then also allowed to use any predictors it wants to add (or drop: the divergence predictors can of course also be deleted if they do not make a worthwhile contribution to this model) or with an intercept-only starting model.

Both final models are then tested for collinearity and overdispersion and evaluated in terms of their fit, classification accuracy, etc. and, in particular, whether one is better than the other and, if so, how much.

4.3 Results

4.3.1 Evaluation of the Divergences

The evaluation of the classificatory power shows that DivFromPP and DivFromStrd are very strongly correlated with the constructional choices. The results of the first, more descriptive analysis is shown in the four panels of Figure 2. The top two panels show each divergence vector on the *x*-axis and the observed probabilities of the two constructions on the *y*-axes. Clearly, the greater the divergence from the pied-piping prototype, the less often that construction is used, and the same holds for the stranded construction.

The lower two panels show the results of two monofactorial GAMs in which the constructional choices were modeled using each just one divergence vector. The lower left panel shows the smoother returned by the GAM with DivFromPP as a predictor; this model's R^2 is 0.297 and its PRE-score is λ=0.361. In other words, just knowing the divergence of a concordance line from the lexical prototype of the pied-piped construction reduces one's error in guessing the constructional choices by >36% and the effect is the 'desired'

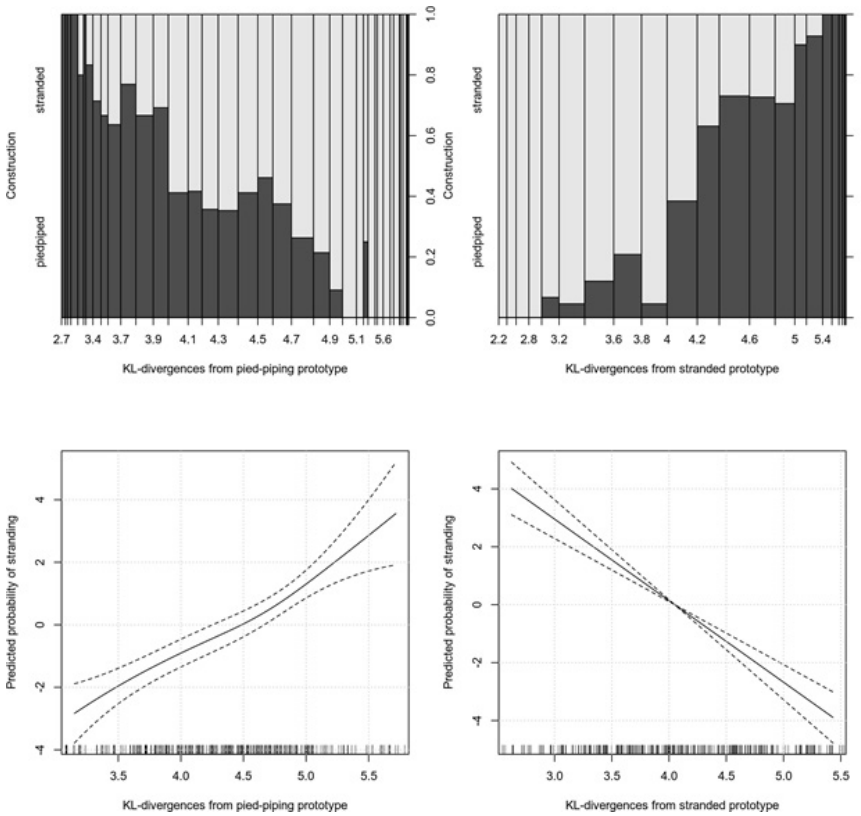

Figure 2: Spine plots and GAM results for DivFromPP (left) and DivFromStrd (right)

one: The higher the divergence from the pied-piped prototype, the more stranded constructions are used. The lower right panel has the corresponding results of a GAM with DivFromStrd as as predictor; this model's R^2 is 0.488 and its PRE-score is λ=0.574!

4.3.2 Evaluation of the Divergences Plus Other Predictors

The next analysis was concerned with the classificatory power of DivFromPP and DivFromStrd when the other predictors are considered at the same time. The final model of the first selection process *without* the divergences represents a very good fit (*LR*-statistic=185.02, *df*=4, p<0.0001, Nagelkerke R^2=0.622, C=0.912, no collinearity or overdispersion). Its precision and recall (for the stranded construction) are 0.862 and 0.881, which it achieves just with Length and VerbType: the longer and more complex the intervening material, the less likely stranding becomes, and stranding is most likely with (phrasal-)prepositional verbs and least likely with transitive verbs.

Amazingly enough, the final model of the second selection process now *with* the divergences represents an even better fit (*LR*-statistic=376.38, *df*=5, $p<0.0001$, Nagelkerke $R^2=0.966$, $C=0.998$, no collinearity or overdispersion); its precision and recall (for the stranded construction) are 0.983 and 0.989, which it achieves just with VerbType and both divergence predictors (which have the same effect as above). According to the relative likelihood test, the second model with the divergences is more than 10^{40} times as likely to be the better model than the first one and its classification accuracy is significantly better than that of the first one.

4.4 Interim Discussion

Again, we find the divergences are strongly correlated with the linguistic choices speakers make, here for a grammatical alternation. The divergence vectors already lead to good results and sizable R^2-values on their own (recall Figure 2), but when combined with other predictors, the model involving the divergences achieves nearly perfect classification accuracy, and it does so without random effects or curvature or any other more sophisticated techniques, making a strong case for its 'power'. However, before the divergences' classificatory power is overestimated on the basis of just this result, it needs to be stated that in this data set, the divergences are not really surprisingly fairly well predictive of the Length variable: If one forces Length into the model with the divergences, one runs into collinearity problems. In other words, in this data set, the divergences replace Length with something that then also has more classificatory power, yielding the extremely good model discussed above, but that also means one needs to pay attention to what variability in the data exactly the divergences are accounting for; the next case study will consider this additional piece of the puzzle.

5. Case Study 3: *of*- vs. *s*-genitives

5.1 Introduction

The second alternation case study is on the genitive alternation exemplified above in (2). The present case study involves 4,045 cases (3,052 *of*, 993 *s*) and, in Gries, Heller, and Funke (under revision), was analyzed using a conditional inference forest. The data are much more heterogeneous than the previously discussed data sets because that study was concerned with variation on two levels:

- On the level of variety contrasting British and Sri Lankan English. Previous studies of such variety differences have shown that predictors of alternations are differently strongly related to constructional

choices in different varieties on different points of emancipation from the historical source variety of British English; the data are based on the British and Sri Lankan components of the International Corpus of English.
- On the level of gender because (a) previous studies have shown that language change is often driven by female speakers and because (b) language external factors such as gender or register often mediate language-internal factors such as animacy or semantic relation.

The present data set contains annotation for the following well-known predictors of the genitive alternation:

- Modality, the mode of production: *speaking* vs. *writing*;
- Gender, the sex of the speaker: *female* vs. *male*;
- PorAnimacy, the animacy of the possessor: *animate* vs. *collective* vs. *inanimate* vs. *locative* vs. *temporal*;
- PorFinalSib, whether the possessor ends in a sibilant, which would make an *s*-genitive harder to articulate: *no* vs. *yes*;
- PorDef, the definiteness of the possessor: *indefinite* vs. *definite*;
- LengthDiff, the difference between the lengths of the possessor and the possessed (measured in characters and log-transformed to address skew);
- SemRel: the semantic relation expressed by the genitive: *prototypical* (part-whole, kinship, legal relations) vs. *non-prototypical* (other).

5.2 Methods

In this case, we have quite a few well-known predictors that are related to genitive choices; the analytical approach will be the same as in the previous section: first, an exploration of only the divergence predictors (computed as before) that involves spine plots and GAMs; second, an exploration where the divergences are competing with the other predictors for a slot in the final model. That second part will again involve bidirectional model selection processes using *AIC* as a criterion to pick the best generalized linear model.

5.3 Results

5.3.1 Evaluation of the divergences

As before, the evaluation of the classificatory power shows that DivFromOf and DivFromS are correlated with the constructional choices, as is shown in the four panels of Figure 3. The top two panels show each divergence vector on the *x*-axis and the observed probabilities of the two constructions on the

y-axes. As before, the greater the divergence from the *of*-prototype, the less often that construction is used, and the same holds for the *s*-genitive. The lower two panels show the results of two monofactorial GAMs in which the constructional choices were modeled using each just one divergence vector. The lower left panel shows the smoother return by the GAM with DivFromOf as as predictor; this model's R^2 is 0.18 and its PRE-score is λ=0.16, which is still in the right direction and highly significant, but weaker than in the case of preposition stranding. The lower right panel has the corresponding results of a GAM with DivFromS as a predictor; this model's R^2 is 0.021, but its PRE-score is λ=0 because it virtually always predicts *of*-genitives.

While these models' performances are worse than for preposition stranding, this may in part be due to the fact that the *of*-genitive is so much more frequent in the data than the *s*-genitive so that, if only one divergence vector is used as a predictor, then the dominance of the *of*-genitive is 'too much'. Before we bring in the other predictors, it is therefore useful to check how well just

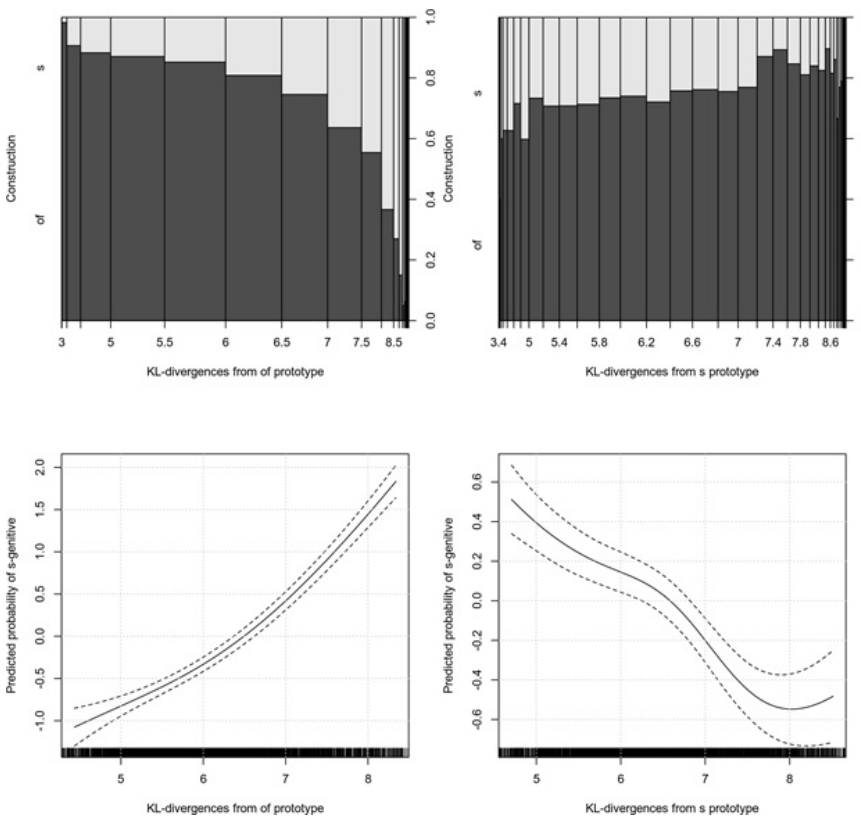

Figure 3: Spine plots and GAM results for DivFromOf (left) and DivFromS (right)

both divergences together predict genitive choice. I therefore fit two models: a GAM that featured smoothed versions of both divergences and their interaction as predictors of the genitive choices, and a more basic generalized linear model with just the two divergences and their interaction as predictors.

The results are very unambiguous: Both regression models reflect very strong correlations between the divergences and the genitives (R^2 of the GAM is 0.818, R^2 and C of the regular logistic regression are 0.869 and 0.988 respectively); the GAM leads to a huge proportional reduction of error of λ=0.81. This is because, as visual exploration of the predictions shows, the regular regression model's predictions look as if, anthropomorphizing a bit, the model looks at which divergence from which genitive is greater and then decides to predict the other one, and as was clear from the high R^2- and C-values, this 'strategy' makes for very good classifications.

5.3.2 Evaluation of the divergences plus other predictors

How do these numbers change when a range of powerful predictors is competing with the divergences? The first model selection process completes with a very significant final model (LR-statistic=1931.3, df=25, p<0.0001, Nagelkerke R^2=0.565, C=0.91, no collinearity or overdispersion); its precision and recall (for the *s*-genitive) are 0.719 and 0.669, which it achieves with a variety of predictors and interactions of predictors with PorAnimacy and VerbType The effects are largely along the lines of what one would expect: *s*-genitives are avoided more with possessor ending in sibilants and with inanimate possessors but more likely with animate possessors, short possessors, and prototypical semantic relations.

When the same model selection process is begun with either DivFromOf and DivFromS in the starting model (but eligible for deletion) or with an intercept-only starting model, the results change again markedly. The new final model includes the context-based divergence variables and their interaction, then what much previous work has shown to be the two strongest predictors of genitive choice, namely PorAnimacy and LengthDiff (whose effects are as expected from previous work), and PorFinalSib as well as Modality (also with effects supporting previous work). In spite of the reduction in the complexity of predictors (the only interaction in this model is the one of the divergences), this model fit is much better than that of the model without the divergences: LR-statistic=3836.9, df=10, p<0.0001, Nagelkerke R^2=0.912, C=0.994, no collinearity or overdispersion); its precision and recall (for the *s*-genitive) are 0.947 and 0.938; the relative likelihood of the model with the divergences is infinitely higher than that of the model without them.

5.4 Interim Discussion

This case study is more interesting than the previous one: The data set is much larger, much more diverse (especially in how it includes BrE and SrilE), and it involved many more and powerful predictors of the alternation on top of the divergences. The picture is similar in how it shows that even if many known powerful predictors are available, the divergences can trump some of them, simplify the model a bit, but still boost classification power to values that are rarely seen. Again, the divergences this time are not simply collinear with the remaining predictors, as can be seen from both variance inflation factors and the only moderate degrees with which the divergences correlate with the other predictors that were annotated. Thus, the divergences from the lexical prototype simplify the model by reducing interactions, but at the same time they are not just some straightforward transformation of the predictors that are usually considered to co-determine genitive choices – they do add information as well.

6. Case Study 4: The Functions of *Well*

6.1 Introduction

The last case study in this paper is somewhat different from the others: It is not about (predicting) a choice that a speaker makes but about (predicting) a function that a speaker puts a word to. The analysis in this section is based on an early part of the data from Rühlemann and Gries (under revision) and is concerned with an acoustic correlate of the function of the word *well*. The part of the data to be discussed here consists of 268 uses of *well* from nine-word turns from the conversational part of the spoken BNC, 221 of which are labeled *pragmatic* (those are cases of *well* functioning as markers of dispreference, quotes, restarts and others) and 47 of which are classed as *syntactic* (those are cases where *well* functions as an adverb, adjective, as an additive subjunct, or as a part of *as well*)

The point of Rühlemann and Gries is to address a gap in research when it comes to exploring *well*'s acoustic properties; they are trying to determine to what degree the duration of *well* (measured in ms) can help predict which of the two functions distinguished here the relevant *well* instantiates. In a first preliminary analysis of the data, they used Duration as a predictor and PosInTurn as a control variable – the latter because of some well-known correlations between especially pragmatic functions of *well* and their positions in turns; for instance, quote and restart markers are often turn-initial. An initial exploration using these two variables found that PosInTurn (as a binary factor with levels *initial* vs. *non-initial*, for data sparsity reasons) had a strong impact on classification accuracy, but it also interacted with Duration such that, for instance, intermediately long *well*s had a higher chance of being pragmatic even if they were not utterance-initial anymore.

However, in this kind of study, it does not even make much sense to *not* include some operationalization of lexical context. In the present case, context can be so important as to cancel out, or override, pretty much any other variable: If the word preceding *well* is *as*, but the word after *well* is not also *as*, then we have an instance of *as well* (i.e., here, the syntactic function) pretty much no matter what the duration of *well* is. Thus, if the goal is to show that Duration explains variability in the functions that *well* is used for, then we need to not just control for PosInTurn but also lexical context to make sure the variable Duration does not take credit for accounting for variability in the data that is actually perfectly accountable for by lexical context.

6.2 Methods

The computation of the prototypes was done the same as before, the only difference being that the collocate window is restricted here to L3 to R3 (to have one case study in which the context window size is varied). Then, the divergences vectors were computed also as before, leading to two variables called DivFromPra and DivFromSyn, which were explored using the same kind of regression modeling techniques as before.

6.3 Results

6.3.1 Evaluation of the Divergences

Even in this conceptually different case where the phenomenon is not an alternation as before, the overall results are similar, with the relevant visualization of spine plots and GAM results in the upper and lower row of Figure 4. The GAM predicting the function of *well* based on DivFromPra returns an R^2-value of 0.255 and the same as a PRE-score of λ=0.255; the corresponding results for the GAM with DivFromSyn as a predictor are even better: R^2=0.458 and λ=0.319. A GAM with both divergences and their interaction returns an R^2-value of 0.89 and a λ of 0.872.

6.3.2 Evaluation of the Divergences Plus Other Predictors

The model selection process for the regression without the divergences leads to the maximal model with PosInTurn, Duration, and their interaction in a way that makes a lot of sense: When PosInTurn is *initial*, Duration does not matter much and the model predicts pragmatic uses, but when PosInTurn is non-*initial*, then Duration does matter such that increasing durations increase the chances of syntactic *well*s. This model is very significant (*LR*-statistic=127.62, *df*=3, p<0.0001) and quite accurate (Nagelkerke R^2=0.626, *C*=0.941, no overdispersion and collinearity only for the interaction terms); its precision and recall (for the *syntactic well*) are 0.729 and 0.745.

When the same model selection process is done with DivFromPra and DivFromSyn, the new final model still contains PosInTurn*Duration, which also still has the same effect, but now the final model also contains both context-based divergence variables. How does that affect the model? As before, it is *much* better: *LR*-statistic=232.26, *df*=5, $p<0.0001$, Nagelkerke $R^2=0.958$, $C=0.997$, no overdispersion and collinearity for the interaction term); its precision and recall (for the syntactic *wells*) are 1 and 0.979; the relative likelihood of the model with the divergences is $>10^{21}$ higher than that of the model without them. Since here the smaller model is a sub-model of the larger one, we can make a LR comparison, which shows that the model with the divergences is significantly better than the one without (*LR*-statistic=104.65, *df*=2, $p<0.0001$), as is the larger model's classification accuracy: in fact, the model with the divergences classifies all cases but one correctly (accuracy=0.996)!

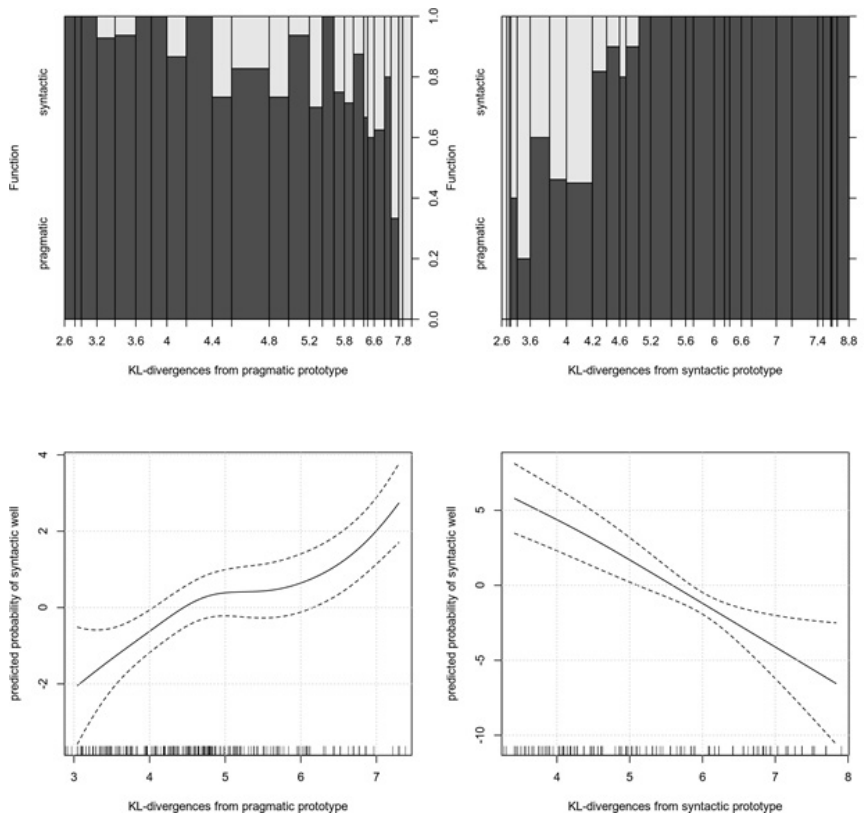

Figure 4: Spine plots and GAM results for DivFromPra (left) and DivFromSyn (right)

6.4 Interim Discussion

This case study is a bit different in how this was not an alternation study of the kind exemplified by the other case studies and, here, the inclusion of at least some sort of lexical-context predictor was maybe easier to justify *a priori*, given how lexical context is indubitably related to the function that *well* will have in a certain context. In spite of these differences, the empirical results are quite comparable, however. When the divergence predictors are permitted to be added to the first maximal model involving PosInTurn*Duration, both of them get added to the model and are providing a remarkable boost to the classification accuracy of the model. In other words, the divergence predictors add genuinely new information to the statistical model and that addition of lexical material is making the model classify nearly every function of *well* correctly.

7. Discussion and Concluding Remarks

7.1 Interim Summary

What have we seen? We have seen that recent work in computational psycholinguistics has proposed an operationalization of prototypes for morphological/inflectional classes (on the basis of relative frequencies of affixes in a morphological class) and that the divergence of a certain word of that class from the prototype of that class can be measured via the KL-divergence of the word's relative frequencies to those of the prototype. If we analogize from their suggestions – 'affix → collocate' and 'morphological class → synonym or alternant' – we see that, if we compute this not over affixes, but collocates, then that logic can be used to operationalize lexical-content prototypes for different members of a set of functionally similar words (near synonyms) or constructions (alternation studies). More interestingly, we have seen that the divergences of individual instances from the prototypes of their synonym/alternant lead to very high degrees of predictive power. In case study after case study, we find that

- the divergence vectors *on their own* exhibit monofactorial correlations that are comparable to many traditional predictors in alternation studies or in fact much higher;
- the divergence vectors *together* often exhibit extremely high correlations with the linguistic choices they are being used to predict, often rivaling R^2-values of the most advanced kinds of generalized additive mixed models of data sets having been built with possibly hundreds of hours of (semi-)manual annotation;

- the divergence vectors provide very substantial boosts to the classification accuracy of regression models using the traditional kinds of predictors;
- the results regarding traditional linguistic predictors remain interpretable: Sometimes, some traditional predictors do not make it into final models, sometimes they do, but the results never became erratic – if anything, the lexical-content variables usually added considerable amounts of discriminatory power to the analysis and sometimes focused the analysis on the most powerful predictors.

It is also worth emphasizing that these results are obtained

- for very different phenomena: lexical near synonymy, constructional near synonymy, and functions of a discourse marker;
- for data that other than the sparsity of their collocate frequencies (due to the usual Zipfian distributions of linguistic data) have very little in common, or are in fact actually quite different kinds of data: (a) small data sets of <300 data points involving ≈250 collocates (*well*) or ≈850 collocates (preposition stranding) and with only two or three other not super well-established independent variables; (b) intermediate data sets of ≈4,000 data points involving ≈20,000 collocates and seven well-established other independent variables; and (c) a large-ish data set of ≈17,000 data points, ≈30,000 collocates, and no predictors other than the collocates at all. Similarly, the results were as good as they were for all and completely unfiltered collocates, for collocates from very small context windows, and for lemmatized and POS-tagged collocates;
- with the simplest of corpus-linguistic and statistical tools, namely essentially just percentages of collocates of concordance lines that were grouped by the match (the synonym, the construction type, or the function) – the approach is not using dispersion statistics, no association measures or key words statistics, no significance tests, no vector space or word embeddings methods (such as word2vec or GloVe), …

Also, the approach will generalize robustly and well. We know this because the collocate prototypes already involve quite a lot of low percentages, and the collocate frequency vector of every individual instance is of course extremely sparse, but we still obtained the very high R^2-values, *C*-scores, precision, and recall scores. We also know this in practice because of how unfazed, so to speak, the approach was when we haphazardly lumped British English and Sri Lankan English genitives together in spite of how research into World English

in general and genitives in particular has uncovered how traditional predictors differ between varieties.

7.2 Implications and Where to Go From Here

What does this mean? Quite frankly, I am not sure. This is mainly because of the mismatch between the performance of the lexical-context predictors on the one hand, but their partial lack of interpretability on the other. 'Partial' lack because I would submit that their effect is interpretable in a sense in the near synonymy study – because there they are just a very convenient coarse-grained, but obviously highly predictive alternative to studies that have so far used manual and/or slot-based approaches involving distinctive collocates, etc. Also, their effect is interpretable in the *well* case study – because there the divergence vectors contain information on lexical context that will be correlated with *well*'s function as discussed above. In fact, it is possible to tease out from the D_{KL} computations which collocates contribute most to the D_{KL}-values by looking at a combination of the contributions to D_{KL} (i.e., the result of $p \times \log_2 {}^{p}/_{q}$) for each collocate i) and frequency. Fittingly, the most remarkable collocates for *brisk* when contrasted to *quick* (the most frequent adjective of the six) are *walk(ing)*, *wind*, *pace*, *business*, and *trade*, whereas the most remarkable collocates for *fast*, when contrasted to *quick*, are *rate*, *growth*, *speed*, *car*, *food*, *lane*, *reactor*, results that make a lot of sense intuitively.

The situation for the morphosyntactic alternations discussed above is slightly different (as it is for a third alternation, particle placement, which I did not discuss here for lack of space, but where the results are comparable to the alternations discussed here). I think it's fair to say that the results show that the method discussed here makes, minimally, for an extremely powerful control variable that should maybe be included in future studies if only to make sure that results for other predictors of interest are not anti-conservative. As mentioned above, there were some studies that involved lexical context – as when Szmrecsanyi (2006) measures lexical complexity of contexts with a type-token ratio – but that variable has nowhere near the discriminatory power that we have seen here. The idea that these divergence vectors might be extremely useful controls, therefore, does not seem too far-fetched at all.

Relatedly, it would be interesting to see to see whether the divergence vectors are correlated with what many current studies might incorporate as random effects such as varying intercepts for speakers, files, or conversations. For instance, to the degree that different conversations revolve around distinct topics, one might expect varying intercepts for conversations to be slightly correlated with the divergence vectors; however, this remains a topic for future research.

Much more speculatively and hesitantly and much less founded, maybe the above results can also be integrated into some theoretical perspective along the

lines of the discussion exemplar-based approaches and naive discriminative learning in Baayen *et al.* (2011) and/or Lester (2018). While both their foci are quite different, the results discussed here at least seem compatible with both and psycholinguistically more informed research than this paper can develop this further. The absence of well-founded theoretical implications notwithstanding, I hope this study has succeeded in documenting the extremely high degree of discriminatory power these divergences from lexical-context prototypes have – if that leads to studies with better statistical control of lexical context and maybe later also to better theoretical accounts of, maybe most ambitiously, the nature of the connections between different kinds of elements of the construction, then I will view this study as a success.

About the Author

Stefan Th. Gries is a Professor at the Department of Linguistics, University of California, Santa Barbara, CA, USA

Notes

1. In order to handle cases in which p_i is 0, one can either define the $\log_2$ of 0 as 0 or apply some sort of smoothing; I will not deal with these technicalities here.

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P. and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118 (3). 438–481. https://doi.org/10.1037/a0023851

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110-142.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology* 18 (3). 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Chen, P. (1986). Discourse and particle movement in English. *Studies in Language* 10 (1). 79–95. https://doi.org/10.1075/sl.10.1.05che

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1). 22–29.

Church, K. W., Gale, W., Hanks, P,, Hindle, D. and Moon, R. (1994). Lexical substitutability. In B. T. S. Atkins and A. Zampolli (Eds), *Computational Approaches to the Lexicon*, 153–177. Oxford: Oxford University Press.

Estival, D. (1985). Syntactic priming of the passive in English. *Text* 5 (1–2). 7–21. https://doi.org/10.1515/text.1.1985.5.1-2.7

Givón, T. (Ed.). (1983). *Topic Continuity in Discourse: A Quantitative Cross-language Study*. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/tsl.3

Gries, S. Th. (2001). A corpus-linguistic analysis of *-ic* and *-ical* adjectives. *ICAME Journal* 25. 65–108.

Gries, S. Th. (2003a). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London and New York: Continuum Press.

Gries, S. Th. (2003b). Testing the sub-test: a collocational-overlap analysis of English *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics* 8 (1). 31–61. https://doi.org/10.1075/ijcl.8.1.02gri

Gries, S. Th. (2018). Preposition stranding in English: Predicting speakers' behaviour. In V. Samiian (Ed.), *Proceedings of the Western Conference on Linguistics*. Vol. 12, 230–241. California State University, Fresno, CA.

Gries, S. Th. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34 (4). 365–399. https://doi.org/10.1007/s10936-005-6139-3

Gries, S. Th. (2010). Behavioral Profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5 (3). 323–346. https://doi.org/10.1075/ml.5.3.04gri

Gries, S. Th. (2018a). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1 (2). 276–308. https://doi.org/10.1075/jsls.00005.gri

Gries, S. Th. (2018b). Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics* 31 (1). 8–29. https://doi.org/10.1075/bjl.00001.gri

Gries, S. Th. and Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1). 97–129. https://doi.org/10.1075/ijcl.9.1.06gri

Gries, S. Th., Heller, B. and Funke, N. S. (under revision). The role of gender in postcolonial syntactic choice-making: Evidence from the genitive alternation in British and Sri Lankan English. In T. J. Bernaisch (Ed.), *Gender in World Englishes*. Cambridge: Cambridge University Press.

Harris, Z. S. (1954). Distributional structure. *Word* 10 (2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hunston, S. and Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/scl.4

Jaeger, T. F. and Snider, N. (2008). Implicit learning and syntactic persistence: Surprisal and cumulativity. In B. C. Love, K. McRae, and V. M. Sloutsky (Eds), *Proceedings of the Cognitive Science Society Conference*, 1061–1066.

Lester, N. A. (2018). *The syntactic bits of nouns: How prior syntactic distributions affect comprehension, production, and acquisition*. Unpublished Ph.D. dissertation, UC Santa Barbara.

Lester, N. A. (to appear). *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research*.

Levin, B. (1991). *English Verb Classes and Alternations: A preliminary Investigation*. Chicago, IL: University of Chicago Press.

Milin, P., Đurdević, D. F., del Prado Martín, F. M. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60 (1): 50–64. https://doi.org/10.1016/j.jml.2008.08.007

Pawley, A. and Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (Eds), *Language and communication*, 191–225. London: Longman.

Rohdenburg, G. (2003). Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In G. Rohdenburg and B. Mondorf (Eds), *Determinants of Grammatical Variation in English*, 2305-250. Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110900019.205

Rosenbach, A. (2002). *Genitive Variation in English: Conceptual Factors in Synchronic and Diachronic Studies*. Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110899818

Rühlemann, C. and Gries. S. Th. (to appear). How do speakers disambiguate multifunctional words? The case of *well*. *Functions of Language*.

Szmrecsanyi, B. (2006). *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis* Berlin and New York: Mouton de Gruyter. https://doi.org/10.1515/9783110197808

Wolk, C., Bresnan, J., Rosenbach, A., and Szmrecsanyi, B. (2013). Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* 30 (3). 382–419. https://doi.org/10.1075/dia.30.3.04wol

Wulff, S., Gries, S. Th. and Lester, N A. (2018). Optional *that* in complementation by German and Spanish learners. In A. Tyler, L. Huan, and H. Jan (Eds), *What is Applied Cognitive Linguistics? Answers from Current SLA Research*, 99–120. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110572186-004