EFL and/vs. ESL?
A multi-level regression modeling perspective on bridging the paradigm gap

Stefan Th. Gries & Sandra C. Deshors[*]
University of California, Santa Barbara & New Mexico State University

Abstract

The study of learner language and of indigenized varieties are growing areas of English-language corpus-linguistic research, which are shaped by two current trends: First, the recognition that more rigorous methodological approaches are urgently needed: with few exceptions, existing work is based on over-/under-use frequency counts that fail to unveil complex non-native linguistic patterns; second, the collective effort to bridge an existing "paradigm gap" (Sridhar & Sridhar 1986) between EFL and ESL research.

This paper contributes to these developments by offering a multifactorial analysis of seventeen lexical verbs in the dative alternation in speech and writing of German/French learners and Hong Kong/India/Singapore English speakers. We exemplify the advantages of hierarchical mixed-effects modeling, which allows us to control for speaker and verb-specific effects, but also for the hierarchical structure of the corpus data. Second, we address the theoretical question of whether EFL and ESL represent discrete English varieties or a continuum.

Key words: EFL, ESL, regression modeling, dative alternation

1 Introduction

1.1 The EFL-ESL paradigm gap: To be or not to be bridged?

The study of EFL (i.e. English as a foreign language, varieties of English spoken in countries such as France or Germany) and the study of indigenized English varieties (English as a second language, ESL, i.e., post-colonial English varieties spoken in countries like Singapore or Hong Kong) are two areas of corpus-linguistics that have developed rapidly over the past few years. Although both areas are concerned with modeling non-native English varieties, EFL and ESL analysts have adopted different foci. While learner corpus researchers mostly focus on structural and lexical differences between different EFL varieties as well as differences between EFL and ENL (English as a native language), ESL researchers mostly concentrate on identifying the linguistic patterns that characterize individual post-colonial English varieties and distinguish them from contemporary English or the English spoken at the time that the post-colonial variety established itself. The different contexts of acquisition and use of EFL and ESL have long influenced analysts to approach the two domains separately. This is despite Sridhar & Sridhar's (1986) call to bridge the 'paradigm gap' between the EFL and ESL research areas and to treat them in unified ways. Only recently corpus linguists started to address Sridhar and Sridhar's call by developing empirical methods to bridge the gap.

Mukherjee & Hundt's (2011) volume on *Exploring Second-Language Varieties of English and Learner Englishes* already presents the benefits of unified approaches to the paradigm gap to identify (dis)similarities of patterning across EFL and ESL. Hilbert (2011:142) notes, for

instance, that "within the field of research into L2 varieties of English, an integrated model is essential" (also see Bongartz & Buschfeld (2011) for a first attempt to integrate ESL and EFL). In addition, in the field of phraseology, integrating EFL and ESL helped Nesselhauf (2009) to identify similarities of the phraseology of institutionalized second language and foreign learner varieties that previously had gone almost unnoticed.

Despite the rapidly growing number of studies attempting to bridge the gap, the question of whether or not this gap should indeed be bridged remains to be empirically confirmed. In other words, it is necessary to establish whether EFL and ESL represent types of varieties that are similar enough in order to be contrasted reliably and meaningfully. This is an important point because at a theoretical level, combining EFL and ESL is not necessarily straightforward: The two varieties are distinct types of non-native English, and while ESL varieties are essentially institutionalized varieties (i.e., they have extended range of uses in the sociolinguistic context of a nation, an extended register/style range, a process of nativization, …), EFL varieties are primarily performance Englishes (i.e., they have no social status and they are used as a foreign language) (Kachru 1982). While studies such as Götz & Schilk (2011) have found this distinction between EFL and ESL to be linguistically reflected in corpus data, the corpus methodologies employed in such studies often exhibits limitations that prevent their authors from drawing theoretical conclusions on the (different) linguistic statuses of EFL and ESL. The relevant literature indicates that this type of issue is not unusual. In the next section, we identify a variety of specific limitations that characterize EFL and ESL research.

1.2 Existing attempts to bridge the paradigm gap

Corpus data are paramount to tease apart EFL and ESL varieties both at the descriptive and the theoretical levels:

> since both learner Englishes and second-language varieties are typically non-native forms of English that emerge in language contact situations and that are acquired (more or less) in institutionalized contexts, it is high time that they were described and compared *on an empirical basis* in order to draw conceptual and theoretical conclusions with regard to their form, function and acquisition (Hundt & Mukherjee 2011:2, our emphasis)

However, as mentioned above, existing corpus-based attempts to bridge the paradigm gap reveal a number of problematic issues. Those are mainly of two kinds: corpus-related and analytical. As for the first corpus-related issue, throughout the literature, there is a lack of a systematic distinction between the spoken and written language modes; a rare exception is Szmrecsanyi & Kortmann (2011), who include both spoken and written native English subcorpora to serve as reference data. Because "linguistic features from all levels – including lexical collocations, word frequencies, nominalizations, dependent clauses, and a full range of co-occurring features – have patterned differences across registers" (Biber et al. 2000:234), distinguishing between the two language modes is often essential. This assessment is echoed by McCarthy & Carter (2001:1): "Spoken grammars have uniquely special qualities that distinguish them from written ones, whenever we look in our corpus, at whatever level of grammatical category"). Thus, without a mode distinction, one cannot be sure that observed pattern differences across corpora are due to variation across varieties rather than registers. In the case of Hilbert (2011), it is almost impossible to know what the author's observed pattern differences

reflect since the author compares the spoken components of the Indian and Singapore subsections of the *International Corpus of English* (ICE) directly with the *Hamburg Corpus of Irish English* which is exclusively composed of written data.

Beyond mode, another potential problematic issue involves the lack of comparability between corpora at an even finer level of resolution, that is at the level of register. Götz & Schilk (2011) illustrate this issue clearly as they compare learner spoken data from the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) with broadcast discussions, interviews and unscripted speeches from the Indian subsection of ICE. While data sparsity issues may explain this decision, it still casts some doubt on the authors' results given the potential lack of comparability across the two corpora. Finally, some studies try to sample in such a way as to minimize the effect that corpous differences may have but do then fail to control for them statistically. For example, Gilquin & Granger (2001) hold the mode constant in their study of the uses of *into* and sample from the arguably related registers of essays and editorials, but they do not statistically control for any remaining potential differences of modes and genres (see below for how this can be done).

The above-mentioned limitations both culminate in the more general issue of corpus structure. Virtually none of the existing studies on learner or indigenized variety corpus data or properly account for the fact that corpus data come with a hierarchical structure, i.e. a structure involving multiple levels nested into each other. Specifically, in most corpora, speakers/writers are nested into files, which are nested into registers, which are nested into modes. For instance, a particular speaker is recorded, the recording is transcribed into one single file which represents one single register, which represents one single mode. Given that corpus design, however, analysts routinely jeopardize the validity of their results because they sometimes compare different corpora and/or different modes (speaking vs. writing) with each other, but they do so only *separately* (doing similar analyses to different (parts of) corpora) or *summarily* (by only discussing implications of different results). That is, a study that compares different corpora typically takes only that one level of variation into consideration instead of considering that one level of variation *at the same time as* a variety of other levels (e.g., CORPUS, MODE, REGISTER, SUBREGISTER, and SPEAKER). So more concretely, a study that compares speaking vs. writing (i.e., MODE) typically takes only that level of variation into consideration but does not consider that level at the same time as the other levels (i.e., the higher level of CORPUS and the lower levels of REGISTER, SUBREGISTER, and SPEAKER); similarly, a study that compares corpora (i.e., CORPUS) typically takes only that level of variation into consideration but does not consider that level at the same time as the other lower levels of MODE, REGISTER, SUBREGISTER, and SPEAKER. What needs to be done is exploring the variation on *all* the hierarchical levels resulting from the corpus design *at the same time* because such analyses can reveal that factors that seemed significant/insignificant in previous analyses may turn out to be insignificant/significant (cf. Gries, under revision for discussion/exemplification).

As for analytical limitations, much existing work is limited in two ways. First, many studies do not account for enough (or even any!) of the contextual information available in their corpus data. As we have shown in much more detail elsewhere (Gries & Deshors 2014), much research is still based on mere comparisons of frequencies of occurrences of a linguistic element *E* and immediate leaps towards claims of over-/underuses with little or no regard of the contextual conditions that facilitate/suppress the occurrence of *E*. For instance, if negation leads to a preference of *can* over *may* in native speech and if learners use *can* more than native speakers, then there are at least two possible explanations for this: either the learners overuse

can, or the learners overuse negation and then use *can* just like native speakers would (i.e., more often). It is probably fair to say that most learner/variety corpus research has so far adopted the first explanation without even considering the second. In addition, there is very little work that has taken lexical or speaker-specific variation into systematic consideration, i.e. variation that is peculiar to particular lexical items or particular speakers/writers.

The second analytical limitation is directly related to the first: Given the scarcity of contextual features included in analyses, existing studies are typically not multifactorial in nature and, thus, at a risk of (i) masking the real complexity of co-occurrence patterns in the data and (ii) therefore, making generalizations about the linguistic structure of non-native varieties (as in Nesselhauf 2009 and Biewer 2011) that may not be supported in more comprehensive studies. It is worth noting, however, that some studies recognize the need for contextual information and they compensate for it with qualitative observations, at least to some extent (e.g., Gilquin 2011, Hundt & Vogel 2011, or Laporte 2012). (We say "to some extent" because, while qualitative analysis and interpretation are necessary and can be useful, no analyst's mind is able to really uncover and realistically weigh the presence of, say, a dozen factors influencing a particular linguistic choice and their interactions.)

The above is not to say that no study addresses the various limitations we previously pointed out. One case in point is Smrecsanyi & Kortmann (2011), who bridge the paradigm gap by studying part-of-speech (POS) frequencies using a clustering technique to analyze and compare degrees of grammatical analyticity and syntheticity in five world Englishes, eleven learner Englishes, and across three standard British English registers (school essays, university essays and speech). Interestingly enough, the authors' results unveil strikingly *different* typological profiles of EFL and ESL. Thus, while their study is an exercise in bridging the gap between EFL and ESL (in that their analysis includes a wide range of EFL, ESL, and ENL data), they also show that bridging the gap may well yield results indicating that ESL and EFL speakers behave very differently from each other. Other interesting studies using multifactorial methods in the domain of learner corpus research (LCR) are Tono (2004) or Collentine & Asención-Delaney (2010).

Another research tradition with methodologically more advanced corpus-based studies involves alternations such as particle placement (cf. (1)), the genitive alternation (cf. (2)), or the much-studied dative alternation (cf. (3)). It is this body of work – specifically with regard to the dative alternation – that we now discuss in more detail.[1]

(1) a. John picked up the book
    b. John picked the book up
(2) a. the President's speech
    b. the speech of the President
(3) a. John gave Mary the book
    b. John gave the book to Mary

1.3 Corpus-based work on alternations

For more than a decade now, corpus linguists have been studying alternations of the above kinds in multifactorial ways. Outside of variationist sociolinguistics, the first corpus-based study of this kind is probably Leech, Francis, & Xu's (1994) study of the genitive alternation, but this approach only became more mainstream when larger number of predictors and different statistics were introduced in Gries (2000, 2002, 2003a, b) and then quickly adopted by others. Especially

the number of multifactorial studies of the dative alternation increased dramatically, with Bresnan et al. (2007) probably reflecting the current state of the art and confirming that the dative alternation is governed simultaneously by factors such as animacy, givenness, length, definiteness (of patients and recipients) as well as other factors.

Over time, this has also begun to influence both learner and variety corpus research. In learner corpus research, studies such as Gries & Wulff (2013), Gries & Adelman (2014), Deshors & Gries (to appear), Deshors (to appear a, b) are all multifactorial studies of alternative (lexical or grammatical) choices and all compare (in similar ways) the choices EFL and ENL speakers make and why. Similarly in variety research, studies like Bresnan & Hay (2008), Bresnan & Ford (2010), Bernaisch et al. (2014), Nam et al (2013), Schilk et al (2013), and Deshors (to appear c) all explore the dative alternation and have been moving the field along to its current relatively sophisticated state of the art. This desirable development notwithstanding, all of the above studies still exhibit one or more shortcomings of the kinds discussed in the previous section: most of these studies do not account for lexical/speaker-specific variation, do not take the hierarchical structure of the corpora into consideration, and – perhaps one of the most fundamental issues – do not make explicit comparisons of non-native and native speaker choices in precisely defined contexts.

This latter problem is of particular importance because while multifactorial regressions can shed light on how different factors affect linguistic choices differently in ENL and E[FS]L data, most of the above studies do not ask what is arguably one of the most meaningful questions when comparing non-native varieties, namely "in the situation that the E[FS]L speaker is in now (and that may not even be attested in the ENL data!), what would an ENL speaker do?" In this paper, we propose some solutions to the above problems. Specifically, we pursue three goals:

- a descriptive one, namely identifying the factors and their nature that make the dative-alternation choices of French and German learners of English as well as speakers of Hong Kong, Indian, and Singaporean English different from those of BrE speakers?
- a methodological one, namely demonstrating one way of how learner corpus studies need to take into consideration various patterns in the data (the hierarchical structure of corpus data and idiosyncratic effects) that no existing study has ever considered;
- a theoretical one, namely thereby beginning to address the question of how similar EFL and ESL patterning is and how much the paradigm gap can/should be bridged (when the most appropriate quantitative methods are used).

2 Data and methods

This section discusses how our data were extracted, annotated, and statistically analyzed.

2.1 Data

2.1.1 The corpus data

We extracted 1265 occurrences of ditransitive and prepositional dative constructions across five written and spoken corpora that were distributed as represented in Figure 1 and Table 1.

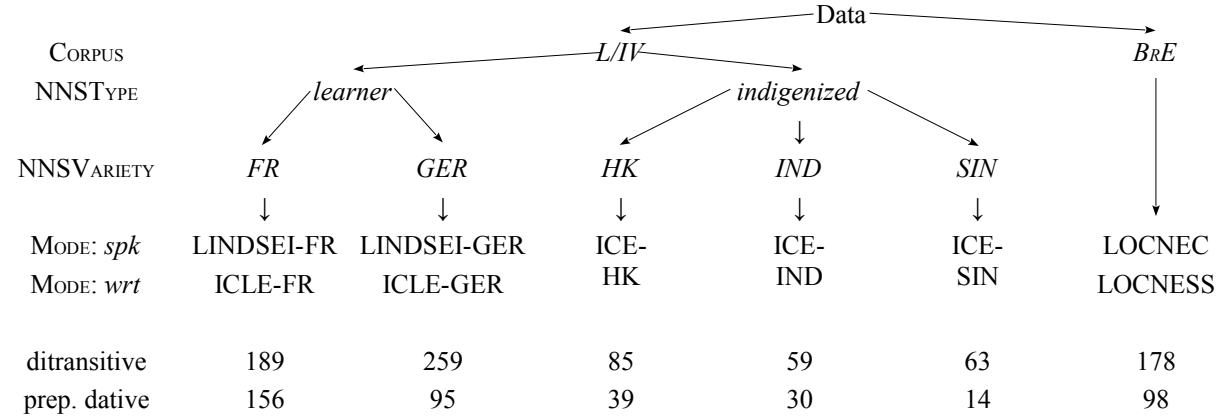


Figure 1: Composition of the corpus data set as determined by the CORPUS, TYPE, and MODE

Table 1: Abbreviations and references of the corpora used

Abbreviation	Full corpus name and reference
LINDSEI-FR, -GER	*Louvain International Database of Spoken English Interlanguage* (Gilquin et al 2010)
ICE-HK, -IND, -SIN	*International Corpus of English* (Greenbaum 1996)
ICLE-FR, -GER	*International Corpus of Learner English* (Granger et al 2009)
LOCNEC	*Louvain Corpus of Native English Conversation* (De Cock 2004)
LOCNESS	*Louvain Corpus of Native English Essays* (Granger et al 2009)

Our motivation behind this corpus sampling scheme was to minimize register differences between the corpora. For example, in order to ensure comparability across the ICLE and ICE corpora, we limited the ICE data to the *class lessons* subset of the spoken sub-corpus (files S1B-001 to S1B-020) and the *non-professional writing* subset (including student essays and examination scripts) of the written sub-corpus (files W1A-001 to W1A-020). Also, we sampled from both spoken and written corpus data to be able to control for any influence that the mode might have. With regards to the EFL data, we included the French and German subsections of ICLE and LINDSEI. Our main motivation here was to have one Germanic and one Romance native language represented in our corpus. Similarly, with regards to the ESL data, we wanted to include two native languages from different language families (i.e., Chinese for the Indo-European family and Hindi for the Sino-Tibetan family). Our native speaker data exclusively consist of British English.[2] Finally, with regards to the coding of the spoken data, contexts of utterance were checked rigorously to ensure that each annotated occurrence was uttered by a single speaker and that our coding would not suffer from corrections, false starts or any intervening material that conversational data can include.

As for the instances of the two constructions, we extracted all instances of the verbs listed in (4) from the corpora using the programming language R (R Core Team 2014). These verbs were chosen because, as Gries & Stefanowitsch (2004) showed, they prefer the ditransitive ((4)a), the prepositional dative ((4)c), or have no preference for either construction ((4)b) in ENL.

(4) a. *ask*, *give*, *offer*, *show*, *teach*, *tell*
b. *lend*, *owe*, *send*

c. *bring*, *hand*, *leave*, *pass*, *pay*, *play*, *sell*

After true ditransitives and prepositional datives were manually identified in the concordances, the resulting 1265 matches were then annotated as described below.

2.1.2 The annotation

We annotated our concordance lines for the following fixed-effect predictors (i.e., predictors whose levels in the sample cover and exhaust the levels this predictor would exhibit in the population because, say, there are no additional levels of VOICE that our current classification does not already cover):

- RECACCESS/PATACCESS: *given* vs. *new*, i.e. whether the referent of the recipient/patient was given (i.e., already mentioned in the preceding ten lines) or new;
- RECSEMANTICS/PATSEMANTICS: *abstract* vs. *concrete* vs. *human* vs. *informational*, i.e. what the referent of the recipient/patient referred to (examples of patient annotation include *give free rein to their imagination* vs. *giving bread and games to the people* vs. *give you a grandson* vs. *give us an answer*);
- RECANIMACY/PATANIMACY: *animate* vs. *inanimate*, i.e. whether the referent of the recipient/patient was animate (e.g. *John gave Mary a squirrel*) or not (e.g. *John gave Mary a letter*);
- RECPRONOUN/PATPRONOUN: *no* vs. *yes*, i.e. whether the recipient/patient was pronominal (e.g., *John gave it to her*) or not (e.g., *John gave the book to his father*);
- VOICE: *active* vs. *passive*, i.e. whether the clause with the ditransitive or prepositional dative was in active voice (e.g., *they gave the parliament too much power*) or not (e.g., *too much power was given to the parliament*);
- LENDIFF: the numeric difference of the length of the recipient minus the length of the patient (in words).
- MODE: *spoken* vs. *written*, i.e. what kind of file the concordance line is from.

Crucially, this study is among the first to also take the multi-level nature of the corpus data represented in Figure 1 into consideration. Therefore, every concordance line was also annotated with regard to a variety of other variables that will feature in the statistical analysis as random effects (i.e. predictors whose levels in the sample do not cover exhaust the levels this predictor would exhibit in the population because, say, future studies may involve lemmas or varieties we did not include):

- LEMMA/MATCH, where MATCH represents the actual verb form that was found in the corpus data (e.g., *given*), where LEMMA represents the lemma of that form (e.g. *give*), and where MATCH is nested into LEMMA since each verb form deterministically occurs with only one lemma;
- CORPUS: *BrE* vs. *L/IV*, i.e. whether the concordance line came from the British English data or the learner/indigenized variety data;
- for all concordance lines, we also identified the file name FILE (as a proxy for a specific speaker) and, for the L/IV data, we also annotated for TYPE/VARIETY/FILE, where FILE is nested into VARIETY, which is nested into TYPE as shown in Figure 1.

Finally, the dependent variable of this study is TRANSITIVITY: *ditransitive* vs. *prepositional dative*, i.e. whether the use of the verb constituted a ditransitive (e.g., *John gave* [$_{VP}$ [$_{NP\ Rec}$ *Mary*] [$_{NP\ Pat}$ *a book*]]) or a prepositional dative (e.g., *John gave* [$_{VP}$ [$_{NP\ Pat}$ *a book*] [$_{PP}$ *to* [$_{NP\ Rec}$ *Mary*]]]).

2.2 Statistical evaluation

So far, the best kind of existing multifactorial (regression) work in learner/variety corpus research is characterized by predicting a dependent variable – a lexical or constructional choice – on the basis of many predictors which, crucially, should be able to interact with a predictor called L1 (for learner corpus research) or SUBSTRATELANGUAGE (for variety research) because only by including this interaction can one determine whether the effect of a particular predictor is different for different speaker groups (cf. Gries & Deshors 2014). However, what this approach does *not* do is answer the above-formulated central question, "in the situation that the E[FS]L speaker is in now, what would an ENL speaker do?" In order to address that question as precisely as possible, Gries & Adelman (2014) and Gries & Deshors (2014) develop and exemplify an approach called MuPDAR (Multifactorial Prediction and Deviation Analysis with Regressions). For the present scenario, in which we study an alternation in native speakers of BrE as well as L/IV (learner/indigenized varieties), the MuPDAR approach can be explained as in Figure 2. This approach answers three questions:

- step 3 → "what are the factors that impact NS behavior?"
- steps 4-5 → "in the situation that the L/IVS is in, what would a NS do?"
- step 6-7 → "do the L/IVS do what the NS would have done, and if not, why?"

In the remainder of this section, we outline how we analyzed the annotated corpus data using the MuPDAR protocol. We proceed in three main steps: Section 2.2.1 discusses step 3 of the protocol, i.e. the regression that was fit on the BrE data; Section 2.2.2 then turns to steps 4-6, i.e. how the resulting regression model was applied to the L/IV data to generate predictions of which construction a NS of BrE would have chosen. Finally, Section 2.2.3 discusses step 7, i.e. the second regression in which we explore what determines whether the L/IVS made BrE-like choices or not. All statistical analyses were performed with R 3.0.2 (R Core Team 2013) and the packages effects 2.3-0 (Fox 2003) and lme4 1.0-6 (Bates et al. 2014); given that this is not a textbook, a certain degree of technicality is unavoidable and we refer the reader to Gries (2013) for a general introduction to multifactorial analysis techniques.

2.2.1 Regression R_1: exploring the choices made by the BrE NS

In a first series of steps, the data were explored to identify patterns that would pose problems to the subsequent regressions (such as data sparsity and collinearity). Therefore, several variables' coding was slightly changed by conflating levels based on their patterning with the dependent variable of R_1, TRANSITIVITY. For instance, we only distinguish the following levels of RECSEMANTICS: *human* vs. *non-human*, and only the following levels of PATSEMANTICS: *abstract/human* vs. *concrete* vs. *informational*.[3] Also, the variable PATANIMACY had to be discarded because of its near perfect correlation with PATSEMANTICS. Then, the data were split up by the variable CORPUS to retain, for now, only the BrE NS data, to which we fit R_1 as a hierarchical generalized linear mixed effects model as represented in (5):[4]

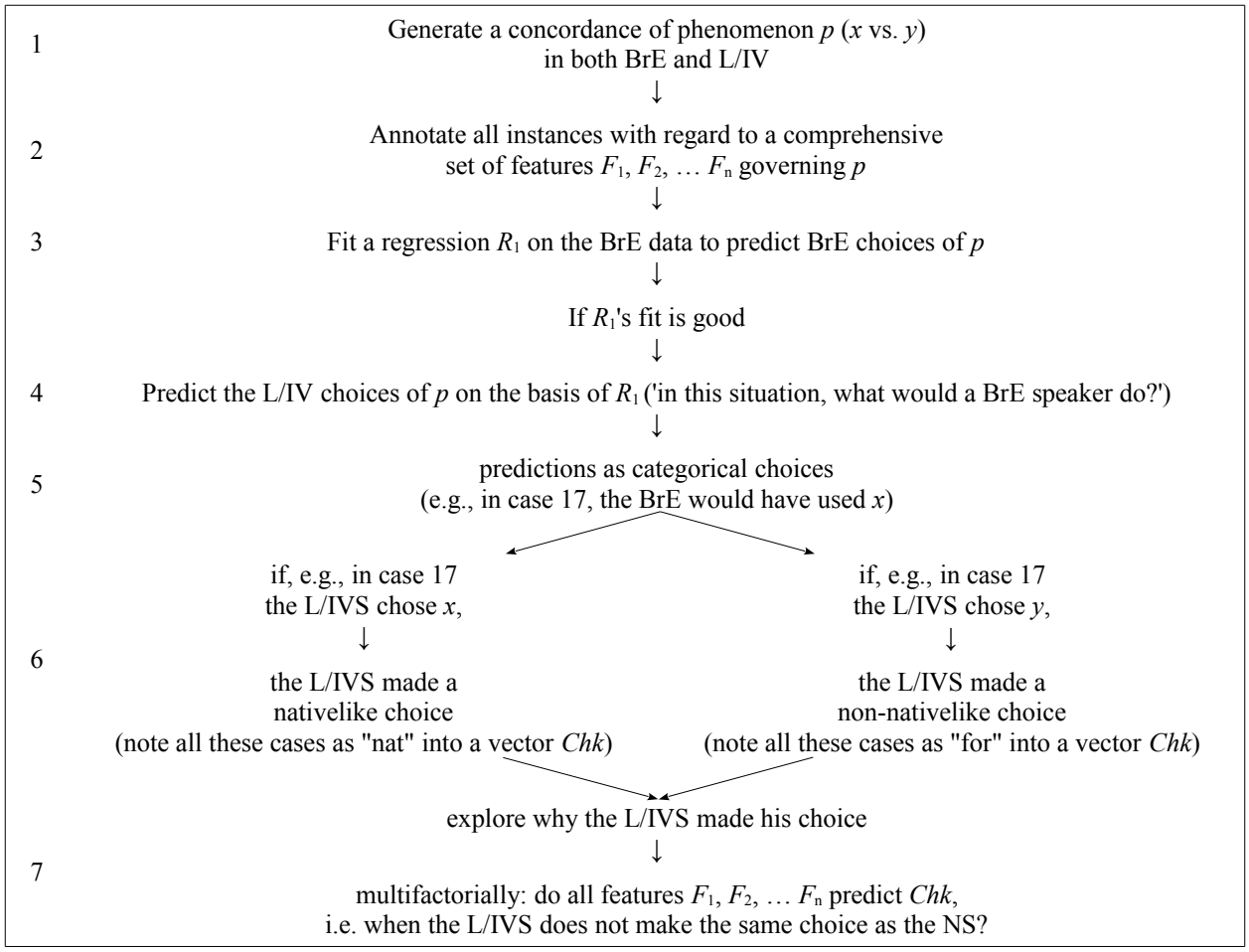


Figure 2: Flowchart of the MuPDAR approach applied to the present data

(5) TRANSITIVITY ~ RECACCESS + PATACCESS + RECSEMANTICS + PATSEMANTICS + RECANIMACY +
VOICE + LENDIFF + MODE +
(1|FILE) + (1|LEMMA/MATCH) (i.e. varying intercepts)[5]

Note in particular the last line, which allows for (i) file-specific idiosyncrasies (a heuristic to capture speaker-specific effects) and (ii) lexical idiosyncrasies. The latter are nested – a particular verb form is only attested with its lemma – such that there may be lexical effects on the level of the form (cf. Newman & Rice 2006) or on the level of the lemma (cf. Gries 2010) or on both. This is how R_1 takes some of the hierarchical structure of the data into consideration and note again that the crucial point is that our modeling process considers both levels of variability – LEMMA and MATCH – at the same time. Since, in this paper, we are not so much interested in the factors that govern NS behavior – cf. the huge amount of literature available on this topic – but rather in the predictions the model makes, we did not undertake a model selection process. Instead, we determined whether the above-defined model resulted in a good fit and a good classification accuracy to see whether proceeding with MuPDAR was feasible.

2.2.2 Applying R_1 to the L/IV data

The next step involved applying the equation of R_1 to the L/IV data,[6] and a *C*-value was computed to determine whether the regression equation based on the NS data can predict the L/IV choices well enough to proceed with the MuPDAR approach.[7]

2.2.3 Regression R_2: exploring the choices of the L/IV data

For each of the L/IV data points, we compared whether the L/IV speaker made the constructional choice that a BrE speaker would have made. The results of these comparisons were stored in a variable NATIVELIKE: *false* (the L/IV speaker did not make the choice predicted for the BrE NS) vs. *true* (the L/IV speaker made the same choice as that predicted for the BrE NS). This variable was then the dependent variable in R_2, whose initial model is represented in (6):[8]

(6) NATIVELIKE ~ RECACCESS + PATACCESS + RECSEMANTICS +
VOICE + LENDIFF + MODE + TRANSITIVITY +
(1|LEMMA/MATCH) + (1|TYPE/VARIETY/FILE) (i.e. varying intercepts)

Again, it is important to note the random-effects structure: The model again allows for idiosyncratic preferences of verb forms and lemmas – the former nested into the latter – but it also explores three levels of hierarchical structure for the non-ENL data: files (i.e. speakers) nested into the five varieties nested into the two corpus types (EFL vs. ESL). Unlike virtually all regressions in learner/variety corpus research, this kind of model can determine whether any of these levels has an effect – what we are of course particularly interested in this bridging-the-gap study is whether there are effects on the level of TYPE because those would imply that EFL and ESL speakers differ.

To arrive at a final model for R_2, we explored at each step how much the addition of an additional predictor (including all possible two-way interactions) or deletion of a predictor would improve the model.[9] For the final model – i.e. a model which could not be improved by adding to, or subtracting from it – we computed overall model summary statistics (R^2s and classification statistics) and represented the effects of all significant highest-level predictors as well as all varying intercepts.[10]

3 Results

3.1 Results of R_1

Even though R_1 was a relatively simple model (in the sense that no interactions between predictors were included) the fit is very good. Specifically, the classification accuracy is 91.7%, which is highly significantly better than both always choosing the more frequent ditransitive or choosing constructions randomly (both $p_{\text{binomial}}<10^{-25}$). Even more remarkably, the *C*-value for this regression is 0.973, i.e. very close to the theoretical maximum of 1. Lastly, the two R^2s for this model, a marginal one for only the fixed effects and the conditional one including both fixed and random effects (cf. Nakagawa & Schielzeth 2013), are likewise very high at 0.792 and 0.9 respectively. Since there was also no significant overdispersion, it was safe to proceed with the following MuPDAR steps.

3.2 Results of applying R_1 *to the L/IV data*

Given the excellent fit of R_1 to the BrE data, we applied its regression equation to the L/IV data,

which resulted in a very encouraging good fit: The *C*-value quantifying the classification accuracy best is an (again) excellent 0.925. In addition, for each L/IV data point, we computed a variable called DEVIATION, whose value quantified if/how much the L/IV speaker was off:

- if the L/IV speaker made the choice predicted for a BrE speaker, the value of DEVIATION was set to 0;
- if the L/IV speaker did not make the choice predicted for a BrE speaker, the value of DEVIATION was set to 0.5 minus the predicted probability of the prepositional dative that was returned by R_1.

This means, if DEVIATION is greater than 0, the L/IV speaker chose a prepositional dative although a BrE speaker wouldn't have, and if DEVIATION is smaller than 0, the L/IV speaker chose a ditransitive although a BrE speaker wouldn't have.

3.3 Results of R_2

The model selection process described in Section 2.2 resulted in the deletion of the insignificant predictor RECSEMANTICS and the addition of four significant interactions: LENGTHDIFF × TRANSITIVITY, LENGTHDIFF × MODE, PATACCESS × TRANSITIVITY, and RECACCESS × MODE. The corresponding final model represents again a very good fit to the data: the classification accuracy is 95.3% (this is highly significantly better than both baselines; both $p_{\text{binomial}}<10^{-22}$), the *C*-value is a remarkable 0.98, and the two R^2s for this model are $R^2_{\text{marginal}}=0.76$ and $R^2_{\text{conditional}}=0.89$, and there was no significant overdispersion or collinearity (all *VIF*s<3.5). Table 2 represents the coefficients of R_2.

Coefficient tables such as Table 2 are usually very hard to interpret, which is why we discuss and visualize all effects separately. While this results section is rather detailed, it is necessary to realize how its very high degree of precision compares favorably to the current state of the art in much of LCR, which does not go beyond simple cross-tabulation.

Table 2: Results of R_2 (predicted level of NATIVELIKE: *yes*)

Fixed effects (intercept = 1.781)	*coefficient*	*standard error*	*z*	*p*
$\text{RECACCESS}_{\text{given} \rightarrow \text{new}}$	0.36	0.403	0.895	0.371
$\text{PATACCESS}_{\text{new} \rightarrow \text{given}}$	-4.77	0.719	-6.634	<<<0.0001
$\text{VOICE}_{\text{active} \rightarrow \text{passive}}$	-1.23	0.543	-2.267	0.023
LENGTHDIFF	-1.09	0.231	-4.726	<0.001
$\text{MODE}_{\text{written} \rightarrow \text{spoken}}$	0.8	0.602	1.329	0.184
$\text{TRANSITIVITY}_{\text{ditr} \rightarrow \text{prepdative}}$	-0.513	0.416	-1.233	0.218
$\text{LENGTHDIFF} \times \text{TRANSITIVITY}_{\text{ditr} \rightarrow \text{prepdative}}$	2.48	0.309	8.037	<<<0.0001
$\text{LENGTHDIFF} \times \text{MODE}_{\text{written} \rightarrow \text{spoken}}$	-1.09	0.297	-3.659	<0.001
$\text{PATACCESS}_{\text{new} \rightarrow \text{given}} \times \text{TRANSITIVITY}_{\text{ditr} \rightarrow \text{prepdative}}$	10.82	1.384	7.817	<<<0.0001
$\text{RECACCESS}_{\text{given} \rightarrow \text{new}} \times \text{MODE}_{\text{written} \rightarrow \text{spoken}}$	-2.79	0.771	-3.62	<0.001

Random effects	*sd*		
TYPE/VARIETY/FILE	1.71	LEMMA/MATCH	0
TYPE/VARIETY	0.17	LEMMA	0.76
TYPE	0.41		

3.3.1 Significant fixed effects of R_2

Figure 3 represents the effect of VOICE in R_2 with a cumulative distribution plot (cf. Gries 2013:114, 175-177).

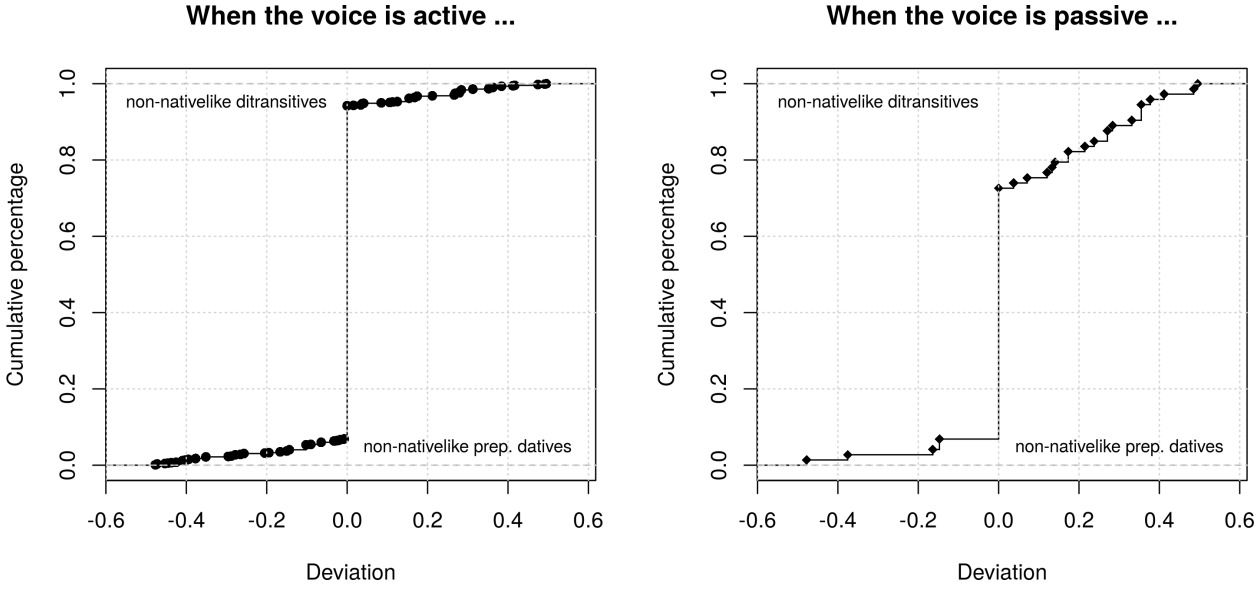


Figure 3: The effect of VOICE on NATIVELIKE and DEVIATION in R_2

On the *x*-axis, we show DEVIATION, on the *y*-axis, we show the cumulative percentage of

the values of DEVIATION, and the lines show the cumulative distribution functions for the two levels of VOICE (in the two panels). This plot shows two important aspects of the data: First, the L/IV speakers have more difficulties with making nativelike choices with passives than with actives, which is indicated by the fact that there are more data points with DEVIATION=0 for actives (the part of the line at x=0 in the left panel is much longer than the part of the line at the same x-axis value in the right panel). Second, when the I/LV speakers make non-nativelike choices, they do so by overusing passive prepositional datives (the line for passives in the right panel moves off to the right at y=0.726, indicating that more than 25% of the passives are non-nativelike prepositional dative passives).

Figure 4 visualizes the first significant interaction, LENGTHDIFF × TRANSITIVITY, with a scatterplot for each level of TRANSITIVITY. The x-axis represents LENGTHDIFF, while the y-axis represents DEVIATION (0-values of DEVIATION are jittered vertically). Each point in each panel represents a L/IV choice, with the two large x's representing the bivariate means and the two lines summarizing the point clouds. The plot shows that

- L/IV speakers are more likely to make nativelike choices when LENGTHDIFF differs more strongly from 0 (cf. how once LENGTHDIFF>5 or LENGTHDIFF<-5, deviation values are either zero (and jittered) or extremely small);
- the nativelike choices show that the L/IV speakers have mastered how this alternation is affected by short-before-long: when the recipient is longer than the patient, they usually correctly choose the prepositional dative (cf. the right half of the right panel), and when the recipient is considerably shorter than the patient, they usually correctly choose the ditransitive (cf. the left half of the left panel);
- L/IV speakers are more likely to make non-nativelike choices with prepositional datives (cf. the larger number of points with DEVIATION≠0 and the corresponding curve in the right panel).

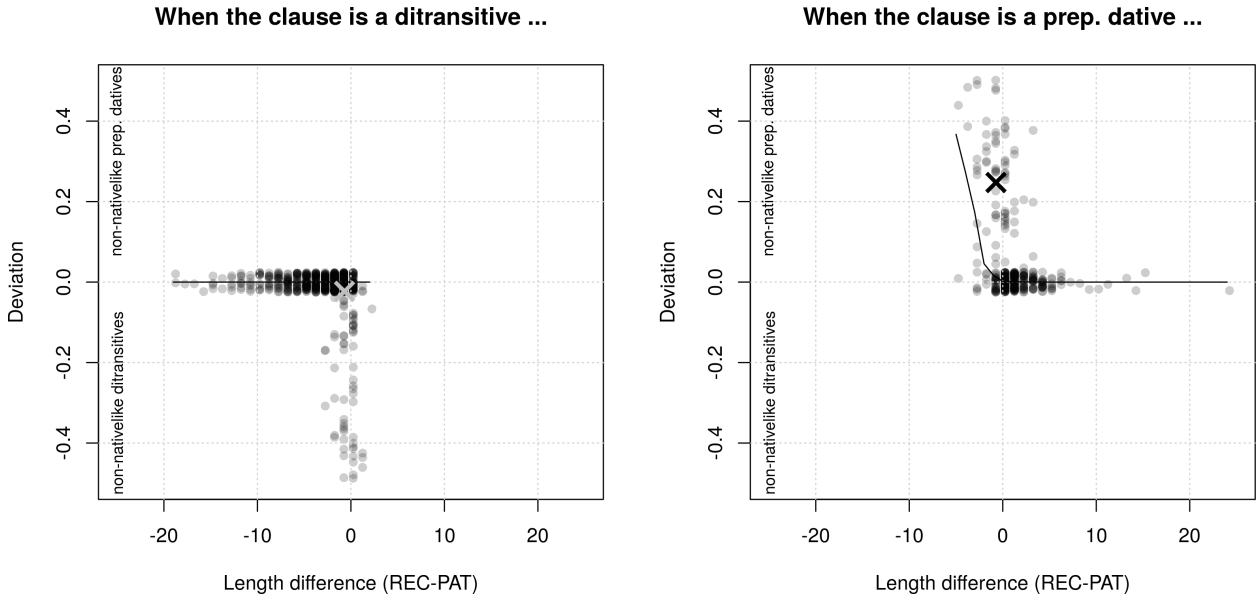


Figure 4: The effect of LENGTHDIFF × TRANSITIVITY on NATIVELIKE in R_2

In other words, L/IV speakers struggle with the middle ground, with cases where

LengthDiff does not provide them with good guidance – i.e., (more) extreme values – which construction to choose, which is reminiscent of Gries & Adelman (2014), who also found that intermediate degrees of the givenness of a referent – referents that are not completely given/topical or completely new – lead to least nativelike choices (by learners of Japanese).

Figure 5 is an analogous representation of LengthDiff × Mode, which shows that

- generally and somewhat unsurprisingly, the written data are characterized by a wider spread of LengthDiff than the spoken data (cf. the wider horizontal range of points in the right vs. the left panel);
- as before in Figure 4, L/IV speakers are more likely to make nativelike choices when LengthDiff differs more strongly from 0 and, thus, provides a good cue as to the more nativelike constructional choice;
- L/IV speakers make more non-nativelike prepositional dative choices in speaking and particularly more non-nativelike ditransitive choices in writing.

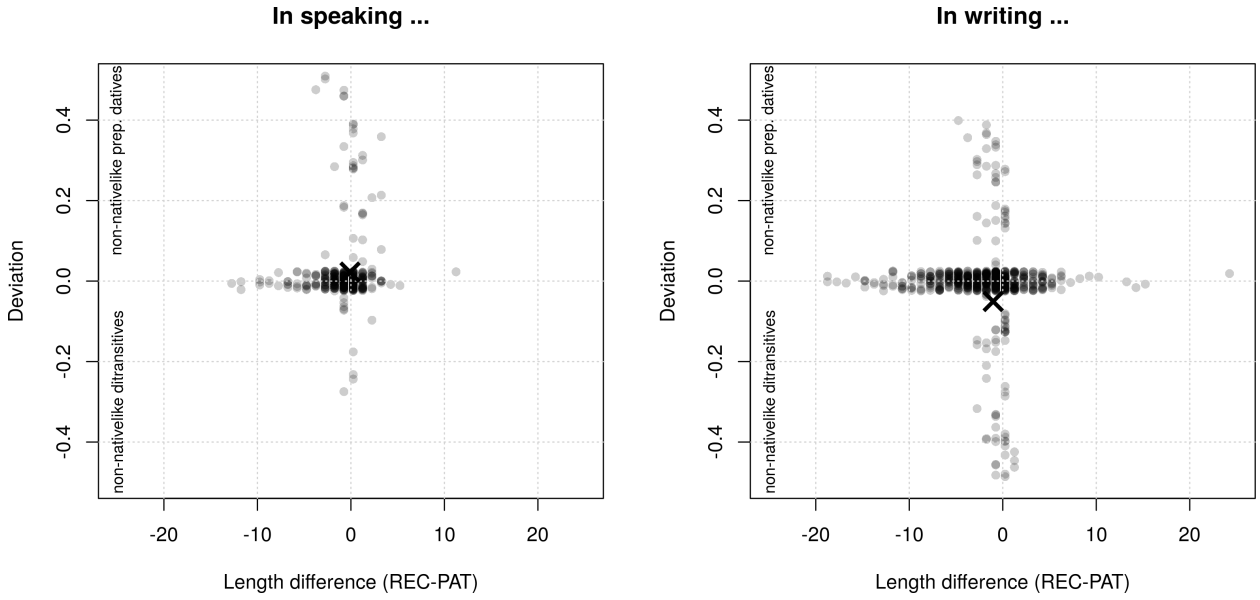


Figure 5: The effect of LengthDiff × Mode on Nativelike in R_2

Figure 6 visualizes the interaction PatAccess × Transitivity in two panels:

- just as the previous interactions showed that the L/IV speakers have mastered short-before-long, Figure 6 reveals they have also mastered the correlated tendency given-before-new: when the patient is new, L/IV speakers typically choose ditransitives in a nativelike way (cf. the long red vertical line at x=0), and when the patient is given, they also typically choose prepositional datives in a nativelike way (cf. the long blue vertical line at x=0).
- however, of the cases where the L/IV speakers do *not* make nativelike choices – i.e. when they choose ditransitives with given patients or prepositional datives with new patients – the results shows that the more frequent non-nativelike choice is that of ditransitives with given patients (the red line in the right panel reveals that ≈70% of these choices are not like those of BrE speakers). In other words, the L/IV speakers do not rely enough on, or

*under*estimate, the strength of the cue 'given patient' for the outcome/constructional choice 'prepositional dative'.

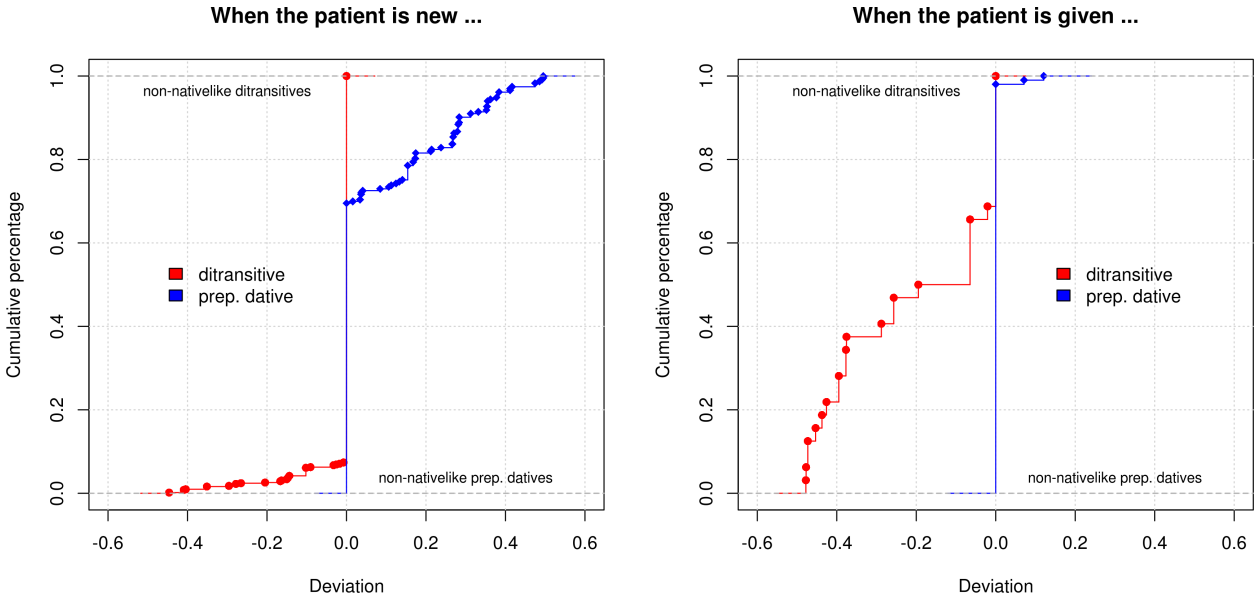


Figure 6: The effect of PATACCESS × TRANSITIVITY on NATIVELIKE in R_2

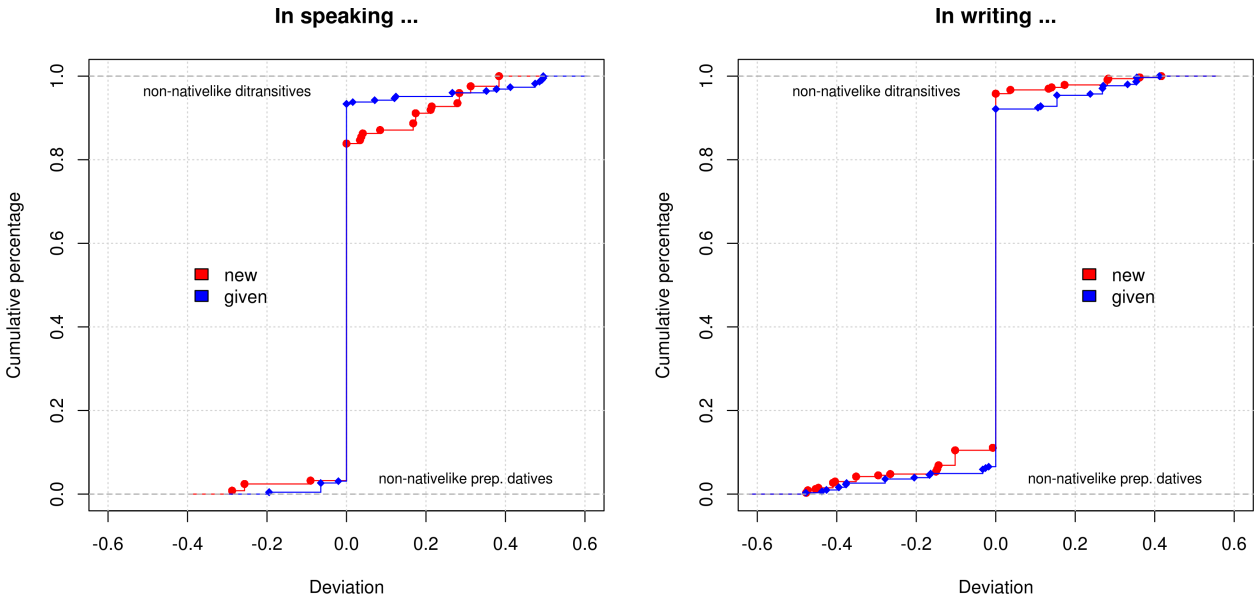


Figure 7: The effect of RECACCESS × MODE on NATIVELIKE in R_2

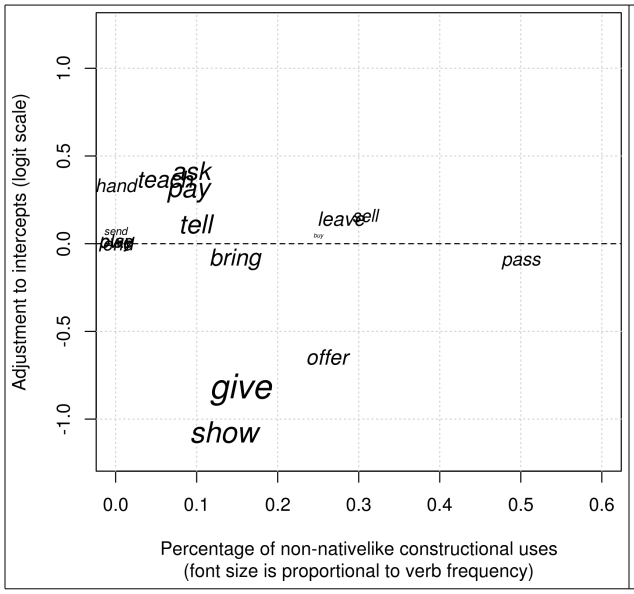
The final fixed effect is shown in Figure 7, the interaction RECACCESS × MODE. In general, both NS and L/IV speakers adhere to given-before-new and prefer prepositional datives with new recipients and ditransitives with given recipients. However, the L/IV speakers do not adhere to this pattern equally in both modes. Specifically, in writing, RECACCESS makes little difference for whether L/IV speakers make the BrE speakers' choices (cf. the close proximity of the red and blue lines in the right panel). However, in speaking, L/IV speakers make many more non-

nativelike choices of prepositional datives and these are largely with new recipients. Interestingly, new recipients normally lead to prepositional datives so the data show that the L/IV speakers overuse prepositional datives with new recipients. Thus, unlike in the previous interactions, here, the L/IV speakers *over*estimate the strength of the cue 'new recipient' to the outcome 'prepositional dative'.

3.3.2 Random effects of R_2

In this section, we now turn to the random-effects structure of R_2; the discussion of these is less in depth since the main point of including the quite complex random-effects structure of R_2 is to demonstrate how future work in LCR would be well advised to explore fixed effects – the predictors that are usually the target of a study – more reliably by, so to speak, partialing out random effects – the variables that usually contribute noise to be filtered out. Thus, this section is largely descriptive.

We begin with potential lexical effects. The first relevant observation is that the hierarchical effect of LEMMA/MATCH is non-existent: in this case, distinguishing the verb forms nested into the lemmas does not result in more predictive power (a result largely in line with Gries 2010). However, there is an effect on the level of LEMMA only but it is essential to note that this is a phenomenon-specific finding: in the very next study of another linguistic phenomenon, the results may be the opposite and lemmas may be unimportant whereas forms/matches are important – this is precisely why this kind of effect must be included in future studies. The effect of LEMMA is summarized in Figure 8: the verb lemmas are plotted into a coordinate system of their percentage of non-nativelike uses (on the *x*-axis) and the lemmas' varying intercepts (on the *y*-axis), which indicate how much a particular verb lemma's patterning differs from that of all. The font size represents the verb frequency. In addition, the right panel shows for each lemma how much in % its uses by the L/IV speakers were those predicted for the BrE speakers.



LEMMA	nativelikeness	LEMMA	nativelikeness
pass	0.5	*sell*	0.692
leave	0.722	*offer*	0.739
buy	0.75	*give*	0.846
bring	0.853	*show*	0.867
tell	0.902	*ask*	0.907
pay	0.91	*teach*	0.939
hand	1	*lend*	1
owe	1	*play*	1
send	1		

Figure 8: The effect of LEMMA on NATIVELIKE in R_2

Since R_2 is a regression not on the constructional choices per se but on the match of the L/IV speakers to the BrE speakers, the varying intercepts cannot be interpreted as revealing

something about the constructional preferences of the verbs (recall note 5 above). However, it is interesting to note that the three verb lemmas with the largest adjustments to the overall regression intercept – *show*, *give*, and *offer* – are all verbs that are very strongly attracted to the ditransitive in BrE as a whole. In fact, Gries & Stefanowitsch (2004:106) show that these three verbs are the verbs that rank 1st, 3rd, and 4th in their preference to the ditransitive. On the whole, there is also a pattern that it is the higher-frequency verbs that the L/IV speakers use more in more nativelike ways: the verbs that are less used in a nativelike fashion are typically less frequent; cf. all verbs to the right of x=0.2. However, although this random effect adds little in terms of interpretation, it clearly shows that verbs differ hugely in terms of much L/IV speakers make targetlike choices, which in turn means that LCR should take lexically-specific variability (more) into account.

Figure 9 represents the random effect with the potentially highest impact on R_2's classifications. For each of the 496 files, a vertical line indicates the specific intercept adjustment that FILE required. Obviously, many of these adjustments are much higher than those for most verbs (cf. Figure 8), which means including FILE as a random effect is important: the fixed-effects results are more precise (cf. also Section 4.2) and it makes sure the assumptions of regression models are not violated because speaker-specific variability is statistically controlled, another important item on LCR's to-do (more often) list.

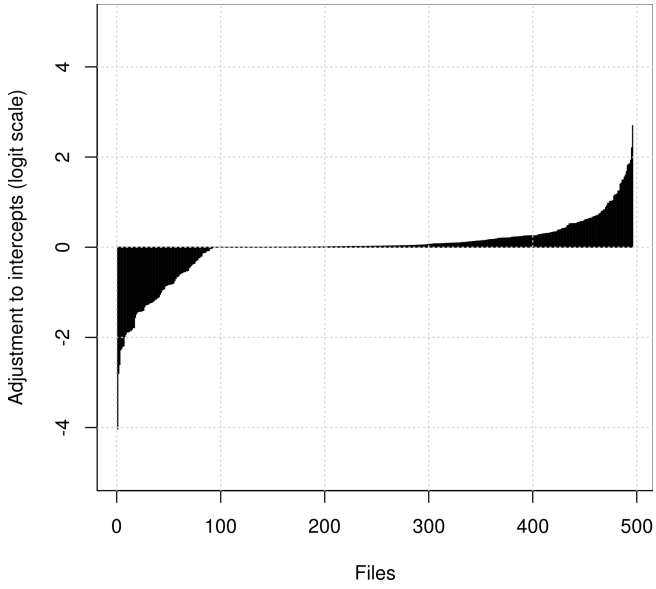


Figure 9: The effect of TYPE/VARIETY/FILES on NATIVELIKE in R_2

The theoretically interesting random effect, however, is Figure 10, which visualizes the effect TYPE/VARIETY. Two important observations: First, for TYPE: compared to the overall baseline, the estimates for the learner data have to be adjusted *up*wards (by 0.12) whereas the estimates for the indigenized varieties data have to be adjusted *down*wards (by -0.36). This reflects that the learners had a higher probability to make NS-like choices than the indigenized-variety speakers. Second, in addition to TYPE, there are also adjustments for VARIETY that have to be added to those for TYPE. Crucially, these are quite small compared to those for TYPE, which shows that the TYPE distinction (*learner* vs. *indigenized*) is more influential than the VARIETY distinction or, from the opposite perspective, that the between-TYPE differences are larger than

the within-TYPE differences. Those findings confirm that, qualitatively, the two kinds of NNS should be treated differently.

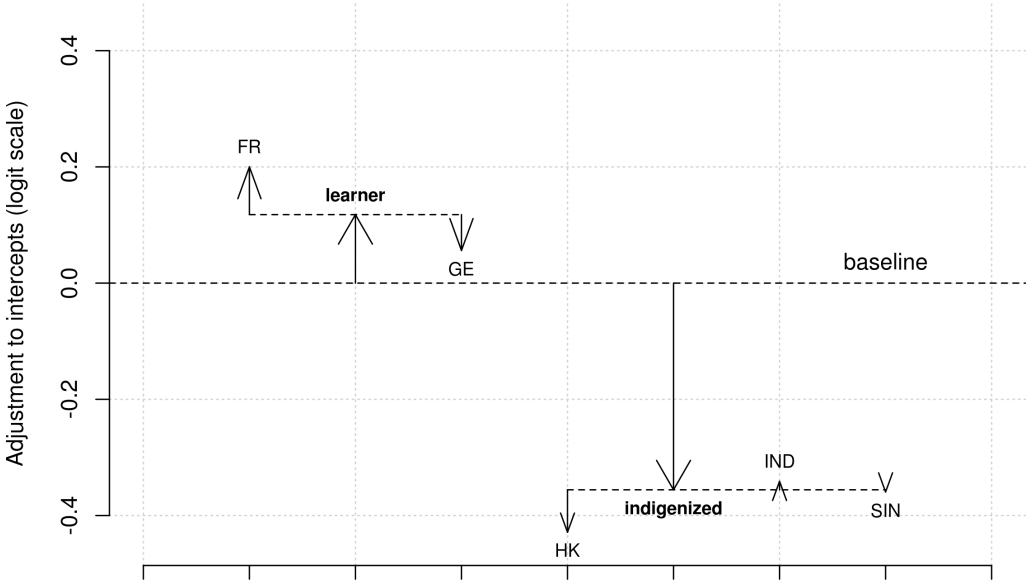


Figure 10: The effect of TYPE/VARIETY on NATIVELIKE in R_2

In the following section, we will discuss the results and conclude.

4 Discussion and concluding remarks

In this section, we discuss our results as they relate to our above three objectives.

4.1 Objective 1: description

Our first goal was to describe how the dative alternation differs between ENL, ESL, and EFL speakers. After the necessarily detailed results section above, it is instructive to zoom out a bit to see the bigger picture.

On the whole, the L/IV speakers make quite similar choices to the BrE speakers. However, they are less nativelike with passives, which is understandable given the structural complexities involved and given the fact that passives are so much less frequent in their input. L/IV speakers are also less nativelike in cases where predictors do not provide reliable cues to the choice of construction, as when patient and recipient are about equally long. It is worth pointing out that the non-nativelike choices were mostly prepositional datives.

The two effects involving accessibility were interesting in that they showcased how L/IV speakers can underestimate the relevance of a cue ('given patient') or overestimate it ('new recipient'). In the cases of the learners in particular, these effects clearly reflect a system in the making that has not yet 'understood' the how much information patient accessibility provides for the constructional choices (when given patients lead to non-native ditransitives) or how much information recipient accessibility provides for the constructional choices (when new recipients lead to prepositional datives too often even when other cues would make native speakers to choose a ditransitive). Thus the present results provide an interesting high-resolution snapshot of

language systems largely, but not (yet) completely, in sync with the native BrE system in a way reminiscent of Ellis (2006:1) statement that "[l]anguage learners are intuitive statisticians, weighing the likelihood of interpretations and predicting which constructions are likely in the current context."

One interesting commonality of most of these reflexes of non-nativelikeness is that they involve the prepositional dative more than the ditransitive. As we discuss below in Sections 4.3 and 4.4, our theoretical focus here is more on the EFL and/vs. ESL distinction, but obviously relating the L/IV speaker choices to their respective L1 preferences would be a natural next step. Thus, while more remains to be done, we submit that the present approach has a lot to offer at a level of resolution that much previous work in learner/variety corpus research has not provided precisely *because* it is methodologically advanced; in the next section we will provide further evidence why this is indispensable.

4.2 Objective 2: methodological implications

Our second goal was to demonstrate how learner and variety corpus research needs to be more careful and take into consideration various patterns in the data that are routinely ignored. On the one hand, such patterns can be the simple distinction of modes, which we have shown above to be important for the present question. Much more important, however, is the hierarchical structure of corpus data as well as idiosyncratic effects of speakers, lexical items, etc. These effects can be crossed (as when, here, the varieties contain most of the verbs studied here) or nested (as when varieties are nested into types). As another example consider Gries & Bernaisch (under revision), who work with data in which newspapers are nested into varieties, which a sound study would take into consideration. Other fields have taken this step already but most corpus studies still have not. What happens if one does not take these kinds of effects into consideration at the same time? The answer is straightforward: all statistics that ignore these kinds of structure will violate the standard assumption of non-longitudinal statistics, the independence of data points. That in turn means that the results for all predictors may turn out to deviate considerably from what the better analysis will return. In fact, this point can be made much more illustratively: Figure 11 shows how all model coefficients change in % if one does not include any random effects at all: one obtains regression results that are off by on average 40% / maximally 75%! Thus, while the kind of modeling advertised here involves a range of complexities, no serious quantitative study can afford to ignore corpus structure as well as lexical/speaker idiosyncrasies that may lead to such grossly incorrect estimates.

4.3 Objective 3: theoretical implications

Our final goal was to address the question of if and how the paradigm gap can/should be bridged. Strictly speaking, this question can be seen as involving two components: (i) should EFL and ESL data be studied together comparatively?, and (ii) are EFL and ESL qualitatively of the same kind (and maybe just on different points of some continuum) or are both qualitatively discrete varieties in their own right? As for (i), we believe the answer is yes, if only because such studies are required to answer question (ii). However, as Hundt & Mukherjee (2011:213) stress, "the degree to which a clear distinction between types of Englishes and individual varieties is possible may depend on the descriptive approach" and we hope to have shown here how important so far rarely used advanced methods are for such advanced questions and complex data.

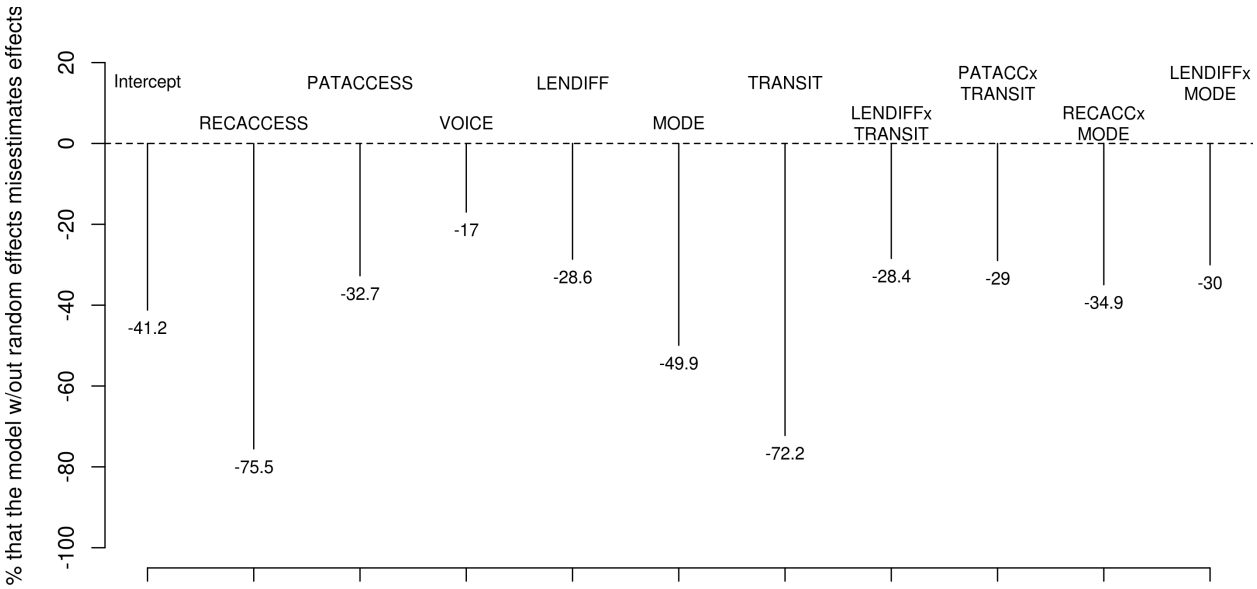


Figure 11: The degree to which a model without any random effects goes wrong

As for (ii), although this question has been discussed much in the very recent past (e.g. the contributions in Mukherjee & Hundt 2011), there is currently no consensus on the status of the two varieties in relation to one another and whether EFL and ESL varieties exhibit the same structural characteristics. One view is that "the distinction between EFL and ESL should be viewed as a continuum," as held by Gilquin & Granger (2011:56) or Nesselhauf (2009). Another view is that EFL and ESL are neither qualitatively different nor located on a continuum: Deshors (to appear c) uses a cluster-analytic approach and finds that individual ESL and EFL varieties are intermingled rather than grouped together according to TYPE and positioned distinctively closer or further away from the native variety. Finally, there is the view of Szmrecsanyi & Kortmann (2011:182), who argue for a clear-cut dichotomy between the two English varieties: "the two variety groups are both fairly discrete and internally coherent […] which confirms the need for drawing a distinction between *English as a Foreign Language* and *English as a Second Language* varieties on purely structural grounds."

Obviously, we are not in a position to provide the ultimate answer. However, our study is the first to simultaneously filter out multiple different sources of noise in the data, and its results, however preliminary they may seem, point more to EFL and ESL as discrete types of varieties (based on the varying intercepts for TYPE/VARIETY) just like Szmrecsanyi & Kortmann (2011). Thus, while the paradigm gap should be bridged in the sense of analyzing EFL and ESL data together, we should also be open to the possibility that analyzing EFL and ESL together may provide more support for their differences than their similarities.

4.4 Where to go from here

In our study, we were mostly concerned with the theoretical question of ESL vs. EFL and, thus, the role of TYPE/VARIETY. However, much more is needed: Obviously, we need a methodologically sophisticated studies on a wider range of phenomena – maybe EFL and ESL are more similar on some levels of linguistic analysis than others (cf. Mukherjee & Gries 2009 and Gries & Mukherjee 2010 for diverging findings even in the small area of lexicogrammar). However, we also need more information on how individual predictors' effects differ across TYPE

and VARIETY. In our study, the random effects were only modifying the intercept, i.e. the baseline probabilities of the constructions or of whether speakers make the BrE choice. This is because the size of our data set did not allow for a maximal random-effects structure of the type recommended by Barr et al. (2013) – with such a larger data set, one could then also explore whether individual predictors' effects varied across TYPE and/or across VARIETY etc. However, with the data sets currently available, this is quite difficult because such more fine-grained studies would result in small ratios of data points divided by estimated parameters and adjustments. Future work on the basis of larger data sets in which this exploration is possible can then serve to pinpoint in more detail how different learner/variety speakers differ and to what degree this may result from their L1s, and if this study will help to inspire more advanced work on these issues, then it has done most of its job.

References

Achard, Michel. 2004. Grammatical instruction in the natural approach: a cognitive grammar view. In Michel Achard & Susanne Niemeier (eds.), *Cognitive Linguistics, Second Language Acquisition and Foreign Language Teaching*, 165-194. Berlin: Mouton de Gruyter.

Barr, Dale J., Roger Levy, Christoph Scheepers, & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255-278.

Bernaisch, Tobias, Stefan Th. Gries, & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1). 7-31

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. <http://CRAN.R-project.org/package=lme4>.

Biber, Douglas, Susan Conrad, & Randi Reppen. 2000. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biewer, Carolin. 2011. Modal auxiliaries in second language varieties: A learner's perspective. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 7-33. Amsterdam & Philadelphia: John Benjamins.

Bongartz, Christiane M. & Sarah Buschfeld. 2011. English in Cyprus: Second language variety or learner English? In Joybrato Mukherjee & Marianne Hundt, eds. *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 35-54. Amsterdam & Philadelphia: John Benjamins.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer, & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69-94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, Joan & Marilyn Ford. 2010. Predicting Syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 186-213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2). 245-259.

Collins, Peter. 1995. The indirect object construction in English: An informational approach. *Linguistics* 33(1). 35-49.
Collentine, Joseph & Yuly Asención-Delaney. 2010. A corpus-based analysis of the discourse functions of *ser/estar* + adjective in three levels of Spanish as FL learners. *Language Learning* 60(2). 409-445.
De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series* 2: 225-246.
De Cuypere, Ludovic & Saartje Verbeke 2013. A corpus-based analysis of the dative alternation in Indian English. *World Englishes* 32(2). 169-184.
Deshors, Sandra C. to appear a. Towards an identification of prototypical non-native modal constructions in EFL: A corpus-based approach. *Corpus Linguistics and Linguistic Theory*.
Deshors, Sandra. C. to appear b. A multifactorial approach to linguistic structure in L2 spoken and written registers. *Corpus Linguistics and Linguistic Theory*.
Deshors, Sandra C. to appear c. A case for a unified treatment of EFL and ESL: A multifactorial approach. *English World-Wide*.
Deshors, Sandra C. & Stefan Th. Gries. 2014. A case for the multifactorial assessment of learner language: the uses of *may* and *can* in French-English interlanguage. In Dylan Glynn & Justyna Robinson (eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy*, 179-204. Amsterdam & Philadelphia: John Benjamins.
Divjak, Dagmar & Stefan Th. Gries. 2009. Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In Katarzyna Dziwirek & Barbara Lewandowska-Tomaszczyk, eds. *Studies in cognitive corpus linguistics*, 273-296. Frankfurt am Main: Peter Lang.
Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1-24.
Fox, John. 2003. Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15). 1-27. <http://www.jstatsoft.org/v08/i15/>.
Gilquin, Gaëtanelle, Sylvie De Cock & Sylviane Granger. 2010. The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
Gilquin, Gaëtanelle & Sylviane Granger. 2011. From EFL to ESL: Evidence from the *International Corpus of Learner English*. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 55-78. Amsterdam & Philadelphia: John Benjamins.
Götz, Sandra & Marco Schilk. 2011. Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 79-100. Amsterdam & Philadelphia: John Benjamin.
Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, & Magali Paquot. 2009. *International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses Universitaires de Louvain.
Green, Georgia M. 1974. *Semantic and syntactic irregularity*. Bloomington, IN: Indiana University Press.
Greenbaum, Sidney, ed. 1996. *Comparing English Worldwide: The International Corpus of*

English. Sidney Greenbaum. Oxford: Clarendon Press.
Gries, Stefan Th. 2000. Towards Multifactorial Analyses of Syntactic Variation: The Case of Particle Placement. Ph.D. dissertation, University of Hamburg.
Gries, Stefan Th. 2002. Evidence in Linguistics: three approaches to genitives in English. In Ruth M. Brend, William J. Sullivan, & Arle R. Lommel (eds.), *LACUS Forum XXVIII: what constitutes evidence in linguistics?*, 17-31. Fullerton, CA: LACUS.
Gries, Stefan Th. 2003a. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London, New York: Continuum.
Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1-27.
Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: a practical introduction.* London & New York: Routledge, Taylor & Francis Group.
Gries, Stefan Th. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In Mario Brdar, Stefan Th. Gries, & Milena Žic Fuchs (eds.), *Cognitive linguistics: convergence and expansion*, 237-256. Amsterdam & Philadelphia: John Benjamins.
Gries, Stefan Th. 2013. *Statistics for linguistics with R: a practical introduction*. 2nd rev. & ext. ed. Berlin & New York: De Gruyter Mouton.
Gries, Stefan Th. under revision. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models.
Gries, Stefan Th. & Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. *Yearbook of Corpus Linguistics and Pragmatics*, 35-54. Berlin & New York: Springer.
Gries, Stefan Th. & Tobias Bernaisch. under revision. Exploring epicenters empirically: Focus on South Asian Englishes.
Gries, Stefan Th. & Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*. 9(1). 109-136.
Gries, Stefan Th. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: an ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520-548.
Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1). 97-129.
Gries, Stefan Th. & Stefanie Wulff. 2013. The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics* 18(3). 327-356.
Harrell, Frank E. Jr. 2001. *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. Berlin & New York: Springer.
Hilbert, Michaela. 2011. Interrogative inversion as a learner phenomenon in English contact varieties: A case of Angloverbals?. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 125-143. Amsterdam & Philadelphia: John Benjamins.
Hundt, Marianne & Joybrato Mukherjee. 2011. Introduction: Bridging a paradigm gap. In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 1-7. Amsterdam &

Philadelphia: John Benjamins.
Hundt, Marianne & Katrin Vogel. 2011. Overuse of the progressive in ESL and learner Englishes – fact or fiction? In Joybrato Mukherjee & Marianne Hundt (eds.), *Exploring second-language varieties of English and learner Englishes: Bridging the paradigm gap*, 145-164. Amsterdam & Philadelphia: John Benjamins.
Jarvis, Scott & Scott A. Crossley (eds.). 2012. *Approaching language transfer through text classification explorations in the detection-based approach*. Multilingual Matters.
Kachru, Braj. 1982. Models for non-native Englishes. In Braj. B. Kachru (ed.), *The other tongue: English across cultures*, 48-74. Urbana & Chicago: University of Illinois Press.
Laporte, Samantha. 2012. Mind the Gap! Bridge between World Englishes and Learner Englishes in the making. *English Text Construction* 5(2). 265-292
Leech, Geoffrey, Brian Francis, & Xunfeng Xu. 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In Catherine Fuchs & Bernard Victorri (eds.), *Continuity in linguistic semantics*, 57-76. Amsterdam & Philadelphia: John Benjamins.
McCarthy, Michael & Ronald Carter. 2001. Ten criteria for a spoken grammar. In Eli Hinkle & Sandra Fotos (eds.), *New perspectives on grammar teaching in second language classrooms*, 51-75. Mahwah, NJ: Lawrence Erlbaum Associates.
Mukherjee, Joybrato & Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes: verb-construction associations in the International Corpus of English. *English World-Wide* 30(1). 27-51.
Mukherjee, Joybrato & Marianne Hundt (eds.). 2011. *Exploring second language varieties of English and learner Englishes: Bridging the paradigm gap*. Amsterdam & Philadelphia: John Benjamins.
Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133-142.
Nam, Christopher, Sach Mukherjee, Marco Schilk, & Joybrato Mukherjee. 2013. Statistical analysis of varieties of English. *Journal of the Royal Statistical Society* 176(3). 777-793.
Nesselhauf, Nadja. 2009. Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide* 30(1). 1-26.
Newman, John & Sally Rice. 2006. Transitivity schemas of English EAT and DRINK in the BNC. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 225-260. Berlin & New York: Mouton de Gruyter.
R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org/>.
Ransom, Elizabeth. 1979. Definiteness and animacy constraints on passives and double object constructions in English. *Glossa* 13(2). 215-240.
Schilk, Marco, Joybrato Mukherjee, Christopher Nam, & Sach Mukherjee. 2013. Complementation of ditransitive verbs in South Asian Englishes: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 9(2). 187-225.
Sridhar, Kamal K. & S. N. Sridhar 1986. "Bridging the paradigm gap: Second language acquistion theory and indigenized varieties of English". *World Englishes* 5(1). 3-14.
Szmrecsanyi, Benedikt & Bernd Kortmann. 2011. Typological profiling: Learner Englishes versus L2 varieties of English. In Joybrato Mukherjee & Marianne Hundt, eds. *Exploring*

second-language varieties of English and learner Englishes: Bridging the paradigm gap, 167-207. Amsterdam & Philadelphia: John Benjamins.

Tono, Yukio. 2004. Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In Guy Aston, Silvia Bernardini, & Dominic Stewart (eds.), *Corpora and language learners*, 45-66. Amsterdam & Philadelphia: John Benjamins.

Wulff, Stefanie & Stefan Th. Gries. to appear. Prenominal adjective order preferences in Chinese and German L2 English: a multifactorial corpus study. *Linguistic Approaches to Bilingualism*.

* The order of authors is arbitrary.

1 We are disregarding here the large body of multifactorial work done by Crossley, Jarvis, and collaborators (cf. in particular Jarvis & Crossley 2012) because much of that work focuses on detecting the L1 of a writer rather than, as here, understanding any one particular lexical or grammatical choice in detail.

2 Some readers may question the choice of the LOCNESS/LOCNEC corpora for the native data as opposed to ICE-GB. Our main motivations here are that given our goal to make all subcorpora as comparable as possible, (i) the EFL data set is approximately 2.5 times larger than the ESL data set (EFL=699 occurrences vs. ESL = 290 occurrences), (ii) only the *class lessons* and *non-professional writing* ICE-GB files would have been utilized, that is 40 files (or 80 000 words) against a total of 254 files across LOCNEC and LOCNESS (approximately 200 000 words), and therefore (iii) the LOCNESS/LOCNEC corpora provide a data set directly comparable with a larger portion of the non-native data.

3 While this grouping of variable levels may seem somewhat arbitrary, it is the one that is supported most strongly by the data: Likelihood ratio tests reveal that abstract and human patients did not differ from each other significantly (p=0.809) in terms of their patterning with TRANSITIVITY.

4 We could not include random slopes for all predictors etc. (as recommended by Barr et al. 2013) because of the small sample size.

5 Given the fairly small size of the data set and the already complicated nature of the statistical analysis, we are restricting our random-effects structure to the simplest possible case, namely varying intercepts. In a regression equation predicting a numeric response y on the basis of a numeric predictor x, the intercept represents the predicted value of y when x=0. By analogy, varying intercepts for files in R_1 represent a kind of baseline of the data in each file – do the data in one file exhibit an overall tendency to use more ditransitives or more prepositional datives? By the same token, varying intercepts for files in R_2 represent a kind of baseline of the data in each file – do the data in one file exhibit an overall tendency to make more or fewer nativelike choices?

6 Crucially and as in Gries & Adelman (2014), since R_1 includes random effects, those were not included in the application of R_1 to the L/IV data – only the coefficients of the fixed effects were included.

7 C-values range from 0.5 to 1 and the higher the value, the better a regression model is at classifying or predicting the dependent variable; C-values ≥0.8 are generally considered good (Harrell 2001:248).

8 The variables RECANIMACY and PATSEMANTICS were not included in R_2 because the former was very highly correlated with RECSEMANTICS and because the latter increased all confidence intervals to include the whole range from 0 to 1.

9 We used likelihood ratio tests and *AIC* for these comparisons, as is common practice.

10 It is instructive to briefly explain how MuPDAR differs from an approach in which just one regression is fit on all the data, i.e. NS and NNS at the same time (as in Gries & Wulff 2013). The results of both approaches can be similar, but the MuPDAR approach is more focused. For instance, the MuPDAR approach could return a result in which the effect of some predictor X in an NNS variety is considered statistically significantly different from the NS even if (i) the direction of effect of X and (ii) all linguistic choices following from it are identical for both NS and NNS. This can happen, for instance, if a variable such as LENGTHDIFF has a very strong effect in NS (e.g., a positive slope) and a significantly weaker but still positive slope in NNS. Since the MuPDAR approach compares NS-based predictions with NNS actual choices and focuses on the cases where NNS make non-NS-like choices, it is better at avoiding results that do not have consequences for actual speaker choices.