Particle placement alternation in EFL learner vs. L1 speech: assessing the similarity of probabilistic grammars

Magali Paquot,[a] Jason Grafmiller[b] & Benedikt Szmrecsanyi[c]

[a] FNRS - Université catholique de Louvain, [b]University of Birmingham, [c]KU Leuven

Abstract

The main objective of this study is to investigate the (linguistic) variables that may influence learners of English as a Foreign Language (EFL) in their selection of continuous vs. discontinuous particle verb constructions in the *Louvain International Database of Spoken English Interlanguage*. The study is driven by two research questions: (1) What factors influence EFL learners' particle placement alternation in speech?; (2) How do EFL learners' particle placement preferences compare with those of users of first language varieties, and more particularly British English? Results show that for EFL learners with Germanic L1s, the grammar overlaps with the native grammar both in terms of its overall complexity and in the relationship(s) between those factors. By contrast, for EFL learners with non-Germanic L1s, the grammar is notable for its simplicity and for its heavy reliance on semantics.

1. Introduction

While English generally exhibits a relatively rigid word order, there are a number of positional alternations where equivalent grammatical constructions are possible without any change in meaning, including particle placement alternation (for transitive phrasal verbs), as exemplified in (1).

(1) a. Verb-object-particle (continuous) order: *John picked up the book.*

b. Verb-particle-object (split) order: *John picked the book up.*

Such phenomena of syntactic variation have attracted considerable interest in different areas in linguistics (e.g. historical linguistics, sociolinguistics) and many studies have sought to identify and describe the variables that determine native speakers' choice governing the alternations (e.g. Gries, 2003; Bresnan et al., 2007; Grafmiller, 2014; Rosenbach, 2014). For example, studies have shown that native speakers of English are more likely to use phrasal verbs in the continuous order (see 1.a) if the resulting meaning of the verb + particle construction is idiomatic (e.g. *carry out, make up*) or if the direct object is long or complex. Crucially, such constraints are probabilistic rather than categorical, i.e. native speakers' choices for a construction are better described as (un)likely rather than as obligatory or impossible in a specific context.

It is precisely because of this that the particle placement alternation is an interesting phenomenon to study from the perspective of the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators (Bresnan 2007; Bresnan & Ford 2010). This is a research program that marshals essentially variationist analysis methods to investigate syntactic variation phenomena in naturalistic corpus data with optional experimental back-up. The aim is to gauge the extent and nature of grammatical knowledge from how language-internal constraints regulate language variation. Three assumptions underlie work in the probabilistic grammar framework:

1. Grammatical variation is regulated by multiple and sometimes conflicting probabilistic constraints, which can influence linguistic choice-making in more or less subtle ways.
2. Speakers have powerful predictive capacities, hence grammatical knowledge must have a probabilistic component.
3. This probabilistic component is derived in large part from language experience, and so is subtly, but fluidly (re)constructed throughout speakers' lives.

How do we know that speaker have powerful predictive capacities? Corpus-based probabilistic grammar analysis can be optionally supplemented by experimental methodologies. Bresnan (2007:76–84), for example, used a scalar rating task based on corpus materials (transcriptions of spoken dialogue passages) as stimuli to model subjects' responses regarding the naturalness of dative variants (*John sent the President a letter* versus *John sent a letter to the President*) in context. Subsequently, these responses were compared to the predictions of the dative alternation regression model reported in Bresnan et al. (2007). The experiment showed that the likelihood of finding a particular linguistic variant (in this case, ditransitive or prepositional dative variants) in a particular context in a corpus corresponds to the intuitions that speakers have about the acceptability of these variants. Therefore, speakers' implicit knowledge about language must be to some extent probabilistic in nature.

Now, until very recently the bulk of the probabilistic grammar literature – including variationist (socio)linguistic work – has restricted attention to alternation phenomena in native and indigenized L2 (or: ESL) varieties of English, as in an ongoing project based at KU Leuven, which investigates the plasticity and malleability of probabilistic knowledge of English grammar by language users with diverse linguistic, regional, and cultural backgrounds (see Szmrecsanyi et al., 2016 and https://researchportal.be/en/project/exploring-probabilistic-grammars-varieties-english-around-world). Grafmiller & Szmrecsanyi (to appear), for example, explored particle placement across nine varieties of English around the world, utilizing data from the *International Corpus of English* and the *Global Corpus of Web-based English* and reported a high degree of uniformity among the factors influencing particle placement in native varieties (e.g. British and Canadian English), while English as a second language varieties (ESL – English varieties that develop in contexts where English is used for intranational purposes; e.g. Indian and Singaporean English) exhibit a high degree of dissimilarity with the native varieties, and considerable intra-group diversity. All in all, studies carried out in the context of the abovementioned project at KU Leuven have shown that while English varieties share a core probabilistic grammar, grammatical variation is also subject to indigenization "at various degrees of subtlety, depending on the abstractness and the lexical embedding of the syntactic pattern involved" (Szmrecsanyi et al., 2016:111). Grafmiller & Szmrecsanyi (to appear) attribute what they refer to as 'probabilistic indigenization' among ESL varieties to "the complex interaction between the simplification processes due to second language acquisition and the unique cultural, political and sociolinguistic ecologies of the individual regions" (p.43) and call for more comparative studies of native and second/foreign language learners at varying levels of proficiency.

As to learner language, while there has been some research on alternation phenomena such as the genitive alternation (Gries & Wulff, 2013), the dative alternation (Gries & Deshors, 2015) and optional complementizer *that* (Wulff et al, 2014) in learner English, very little research so far has focused on the (linguistic) variables that may influence learners of English as a Foreign Language (EFL) in their selection of continuous vs. discontinuous particle verb constructions. EFL learners' use of phrasal verbs has been the target of many studies in second language acquisition and learner corpus research (e.g. Chen, 2013; Dagut and Laufer 1985; Gilquin, 2015; Hulstijn and Marchena, 1989; Laufer and Eliasson 1993; Liao and Fukuya 2004; Sung, 2017) but these studies have typically investigated underuse or avoidance of phrasal verbs and the impact of variables such as the mother tongue background and L2 proficiency on these phenomena. Gilquin (2015), however, compared the use of continuous vs. split transitive phrasal verbs by native speakers of English and French EFL learners and reported that learners rely on a lower percentage of verb + direct object + particle constructions. Similar results were reported by Sung (2017) for Korean EFL learners.

The main objective of this exploratory study is therefore to fill a gap in the literature by investigating particle placement alternation in EFL language. We decided to focus on EFL learner spoken language in this first study because, given that speech puts greater cognitive strain on production than writing does, and given that production systems are already heavily burdened in L2 speech (Linck et al., 2014), we might expect to find larger differences between processing-related factors in L1 and L2 users when we look at spoken as opposed to written data. The study is driven by the following research questions:

1. What factors influence EFL learners' particle placement alternation in speech?
2. How do EFL learners' particle placement preferences compare with those of users of a first language variety?

By answering these two questions, the study also aims to start exploring whether EFL learners share a core probabilistic grammar with users of first and second language varieties of English (Szmrecsanyi et al., 2016).

2. Data and methodology

The study focuses on learner speech (Section 2.1) and largely replicates the methods used in Szmrecsanyi et al. (2016) to identify interchangeable transitive phrasal verbs (Section 2.2.), code particle placement variants in EFL learner speech (Section 2.3) and analyze data statistically (Section 2.4).

2.1. Data

The learner data come from the *Louvain International Database of Spoken English Interlanguage*, i.e. a learner corpus made up of interviews in English with university students from several mother tongue backgrounds and whose English proficiency ranges from intermediate to advanced (LINDSEI; Gilquin et al. 2010). Each interview follows the same structure: it starts with a discussion on a set topic, then moves on to a free discussion and finishes with a picture description. As shown in Table 1, the seven sub-corpora under study represent English as a Foreign Language as spoken by a variety of learner populations with Indo-European languages as first languages.

As the use of phrasal verbs is very sensitive to text types (e.g. Dempsey et al., 2007), the *Louvain Corpus of Native English Conversation* (LOCNEC; De Cock 2004), i.e. a corpus of interviews with native speakers of British English collected under the same circumstances as the LINDSEI, is used as a comparable L1 corpus. Our selection of British English as the first language variety against which to compare learner speech is thus largely driven by the availability of a fully comparable corpus.

Table 1: Spoken samples

	Interviews	**Words (learners only)**
LINDSEI-DU	50	83,134
LINDSEI-FR	50	94,941
LINDSEI-GE	50	89,384
LINDSEI-GR	50	78,243
LINDSEI-IT	50	61,271
LINDSEI-SP	50	67,642
LINDSEI-SW	50	75,202
LOCNEC	50	125,069

2.2. Data identification

In line with Szmrecsanyi et al. (2016), the particle verb dataset consisted of all interchangeable transitive phrasal verb tokens that contained one of the ten most frequent particles: *around, away, back, down, in, off, out, over, on,* and *up* (Gries 2003: 67-68). Unlike in Szmrecsanyi et al. (2016), however, identification of particle placement variants was done fully manually with the help of WordSmith Tools 6.0 for two main reasons:

(1) the LINDSEI components are much smaller than the corpora used in Szmrecsanyi et al. (2016);
(2) relying on part-of-speech tagging to retrieve verb + particle sequences in learner speech would most probably have lowered our recall rate, something that was deemed unacceptable as it has repeatedly been reported in previous literature that phrasal verbs are not very frequent in learner language (e.g. Gilquin, 2015).

Obviously, this meant extensive manual weeding-out (from 18,108 to 470 lines of concordances) to exclude all instances where the ten particles were not used as part of phrasal verbs but as prepositions, adverbs, etc. In addition, only those particle placement instances where the competing variant could have been used were retained for further analysis and tokens that did not include genuine interchangeable uses – passives sentences, sentences with extracted direct objects, modified particles (e.g. *send the ball right back*), names, titles and other fixed phrases (e.g. *Take Me Out of the Ball Game*), etc. – were also removed.

This selection and filtering process resulted in a dataset of 470 interchangeable particle verb tokens (Table 2).

Table 2: Distribution of transitive particle verb variants across the learner corpora and LOCNEC

	DU	FR	GE	GR	IT	SP	SW	LOCNEC
V-Prt-OBJ	30	13	34	20	11	6	30	45
V-OBJ-Prt	39	24	33	15	3	8	42	117
Total	**69**	**37**	**67**	**35**	**14**	**14**	**72**	**162**

A closer look at the dataset showed that native speakers and EFL learners used the personal pronoun *it* as object exclusively in the verb – object – particle construction. As there was no alternation, tokens with *it* as direct object were deleted from the final dataset too (cf. Szmrecsanyi et al., 2016; see Table 3).

Table 3: Distribution of transitive particle verb variants across the learner corpora and LOCNEC (lexical direct objects only)

	DU	FR	GE	GR	IT	SP	SW	LOCNEC
V-Prt-OBJ	30	12	31	19	11	6	28	42
V-OBJ-Prt	10	9	10	4	0	0	16	56
Total	**40**	**21**	**41**	**23**	**11**	**6**	**44**	**98**

As the resulting dataset sometimes included too few cases per L1s (and even zero instances of V-OBJ-Prt constructions, cf. Table 3), the 284 cases were regrouped by L1 families.

Table 4: Distribution of transitive particle verb variants (lexical direct objects only) per L1 families

	British English (LOCNEC)	Germanic languages	Non-Germanic languages
V-Prt-OBJ	42 (42.9%)	89 (71.2%)	48 (78.7%)
V-OBJ-Prt	56 (57.1%)	36 (28.8%)	13 (21.3%)
Total	**98**	**125**	**61**

2.3. Predictor variables

Once the interchangeable transitive verbs were identified, they were annotated with a set of predictors that have been shown to have an effect on native speakers' choices governing particle placement alternations (e.g. Gries, 2003; Capelle, 2009; Grafmiller, 2015). These predictors largely relate to the discourse accessibility and length (or 'heaviness') of the direct object and the semantic properties of the verb:

- DIROBJWORDLENGTH : length of the direct object in number of words;
- DIROBJTYPE : the syntactic category of the direct object's head (see Table 5 for a list of categories);
- DIROBJTHEMATICITY: the extent to which a word represents the, or one of the, topics or 'themes' of a text; operationalized as the relative frequency of the head noun in the text in which it occurs;
- DIROBJDEFINITENESS : whether the direct object is 'definite' (proper nouns, NPs with a definite determiner, definite pronouns, s-genitive NPs, superlatives, temporal expressions) or 'not' (NPs with an indefinite determiner, indefinite pronouns, bare plural NPs, numbers that are not years or monetary amounts, gerunds not headed by definite determiners, any determinerless noun ending in *– tion*, *-ment*, *- sion*, *- ology*, or *–ism*) following the criteria of Garretson et al. (2004);
- DIROBJANIMACY: the direct object may be coded as 'human and animal', 'collective', 'inanimate', 'locative' or 'temporal';
- DIROBJGIVENESS: the direct object was coded as 'given' if its head noun (lemma) was mentioned in the 100 words prior to the actual occurrence (including the interviewer's turn), and as 'new' otherwise;
- DIROBJCOMPLEXITY : whether the direct object includes any kind of complement and/or postmodification (see Table 6 for a list of categories);
- DIROBJCONCRETENESS : whether the direct object is "visible and physically manipulable or not" (Gries, 2003: 71);
- PPADJUNCTS: the presence of a prepositional phrase (PP) following the target VP;
- VERBSEMANTICS: initially coded as 'literal', 'metaphorical' and 'idiomatic' (cf. Gries, 2003: 72) but results of an inter-rater reliability test indicated poor agreement (see below) so we recoded this variable as 'literal' vs. 'non-literal'.

Each particle placement alternation was also coded for:

- VARIETY: DU, FR, GE, GR, IT, SP, SW and GB (British English as represented in LOCNEC);
- FAMILY: Germanic L1 (including DU, GE, SW), NonGermanic L1 (including FR, GR, IT, SP), GB (see Section 2.2);

- FILEID: the number of the file in which the token was found (In LINDSEI and LOCNEC, a file corresponds to an interview with one learner or native speaker respectively);
- RESP: the response variable, i.e. 'continuous' vs. 'discontinuous' verb-particle placement.

For more information about the variables and how they were coded, see Grafmiller (2015) and Grafmiller et al. (2016).

Table 5: Types of heads (based on Grafmiller et al, 2016: 9)

Code	**Category**	**Examples**
'nc'	Common noun	*birds, the market, wisdom, this year*
'np'	Proper noun	*President Kennedy, Japan, the United Nations*
'pprn'	Personal pronoun	*me, theirs, yourself*
'iprn'	Impersonal pronoun	*everyone, something, whoever*
'dm'	Demonstrative	*This, that, these, those*
'ng'	Gerund	*give up drinking*

Table 6: Complexity of the direct object (Grafmiller et al, 2016: 13)

Code	Category	Comments	Examples
'co'	Coordinated NP	Noun phrases involving multiple heads joined with *and, or, but, though*, or any other conjunction	*the onions and the potatoes, Accounting or Economics, silt and floodwaters*
'cp'	Sentential complement	Complement clauses of nouns that take sentential complements; can have overt or null relativizer	*rumors that Obama was not born in the US*
'gn'	Genitive	NP with either an *s*- or *of* genitive	*my father's gun, the cause of all women*
'pp'	Prepositional phrase	Any PP that is unambiguously modifying the constituent NP (and not some larger constituent, e.g. the VP); this includes non-genitive *of*-PPs	*the lies about Obama, research on these writers, that line of work, his example of the Temperance Society*
'rc'	Finite relative clause	Finite, restrictive relative clauses. These can have overt or null relative pronouns	*the guy that caused the accident, the toys you thought were our favorites*
's'	Simple	Any pronoun or NP with [(Det) (A) N] structure	*subscriptions, today, any old rubbish, her head, its previous accomplishments, it, anyone else*
'vp'	Reduced relative clause	Reduced relatives headed by either present or past participles	*the one sitting on the log, the package damaged by the carrier, the point you made about a possible glut of graduates*

As the dataset was annotated by one of the authors, a MA student doing an internship at the Centre for English Corpus Linguistics and a student worker, an inter-rater reliability test was conducted. Results were good, ranging from 93.3% to 100% agreement, except for VERBSEMANTICS (56.7%). It proved very difficult to distinguish between metaphorical and idiomatic verb senses and we therefore opted for recoding this variable into a binary variable ('literal' vs. 'non-literal', see above).

2.4. Statistical analysis of data

Like in Szmrecsanyi et al. (2016), the effects of the different variables described in Section 2.3 were investigated with conditional inference trees and random forests. Conditional inference trees provide visual representations of (potentially) complex interactions that are relatively easy to interpret, yet challenging to model with other techniques. Random forests extend this method by creating hundreds or thousands of trees grown from random subsamples of both the data and predictors. Due to the large scale random sampling process, random forest models tend to be quite accurate. And crucially for the present study, conditional inference trees and random forests are more appropriate than regression models for 'small n large p' situations, that is, situations in which we have relatively few observations but potentially many predictor variables; they are also robust to problems of predictor correlation (cf. also Tagliamonte & Baayen 2012).

We used R (R Core Team, 2017) and the *party* and *partykit* packages (Hothorn et al. 2006; Strobl et al. 2009) to model the data using conditional inference trees and random forests.

3. Results

With very small datasets there is always a concern of overfitting, i.e. obtaining results too finely tailored to our specific dataset, therefore we limit our models to the five predictors that showed the strongest correlation with particle placement in the full dataset in preliminary tests. These are DIROBJWORDLENGTH, DIROBJDEFINITENESS, DIROBJTYPE, DIROBJCOMPLEXITY, VERBSEMANTICS, and obviously, FAMILY, as we are particularly interested in the effect of the language backgrounds.

The classification accuracy of the conditional inference tree, 70.7% (N = 284), is significantly better than the baseline accuracy of 63.0% we would obtain by simply always choosing the more common variant ($p_{binom} < 0.001$). We also calculate the concordance statistic C, which represents the probability that the model will rank any randomly chosen observation of a continuous token as more likely than any randomly chosen observation of a discontinuous token. This is a measure of accuracy independent of the baseline distribution in the dataset, and ranges from 0.5 (random chance) to 1 (perfect prediction). Values above 0.8 reflect a model with relatively good explanatory power. For the present tree we obtained a C statistic of 0.65, which indicates that the accuracy of our tree model over the entire dataset is quite low. Nonetheless, because the conditional inference tree method works by recursively partitioning the data into smaller and smaller subsets, meaningful patterns can be identified within different subregions of the data. We discuss some of these patterns below.

The resulting tree diagram is shown in Figure 1. Each node in the tree represents a split in the data into two subsets corresponding to the labelled predictor and its values shown on the connecting lines. At each node the model considers all possible ways of dividing the data according to all predictors, and chooses the predictor and values that provides the strongest correlation with the outcome (at the customary significance level $p < 0.05$) and splits the data accordingly. The topmost node thus represents the most predictive (binary) split in the dataset overall, and the lower nodes represent the most predictive splits within the ensuing sub-regions of the data. The terminal nodes, or "leaves", of the tree provide barplots of the observed proportions of the V-OBJ-Prt and V-Prt-OBJ variants, along with the total number of tokens observed in the corresponding subsets of the data (see also Szmrecsanyi et al. 2016:117).

The first thing to note is that not all predictors are present in Figure 1: DIROBJDEFINITENESS, DIROBJTYPE, and DIROBJCOMPLEXITY are absent, suggesting that these factors play a minor role in predicting the choice of variant in the model. As for the significant predictors, at Node 1, we see that the most significant predictor is FAMILY: there is a major (and most relevant given the topic of this paper) split into GB (i.e. British English) and non-native varieties of English (NONGERMANIC,

GERMANIC). Moving down from Node 1, the second predictor is DIROBJWORDLENGTH in the two sides of the tree, with a major split into relatively short (three words or fewer) and relatively long (more than three words) direct objects. The three main differences between British English and non-native varieties of English are that:

1) The V-OBJ-Prt construction is more frequent in British English than in learner varieties;
2) The V-OBJ-Prt construction is extremely rare in EFL varieties with direct objects of more than 3 words while it occurs in c. 30% of the cases in British English;
3) No further predictor seems to have a significant effect on particle placement alternation in British English while SEMANTICS plays a role in EFL varieties (Node 3, right side of the tree) in interaction with DIROBJWORDLENGTH. In learner language, interchangeable transitive phrasal verbs are likely to appear in a V-OBJ-Prt construction in between 40% and 50% of the cases. This probability drops dramatically (c. 10%) for non-compositional phrasal verbs used with a direct object of length > 1.

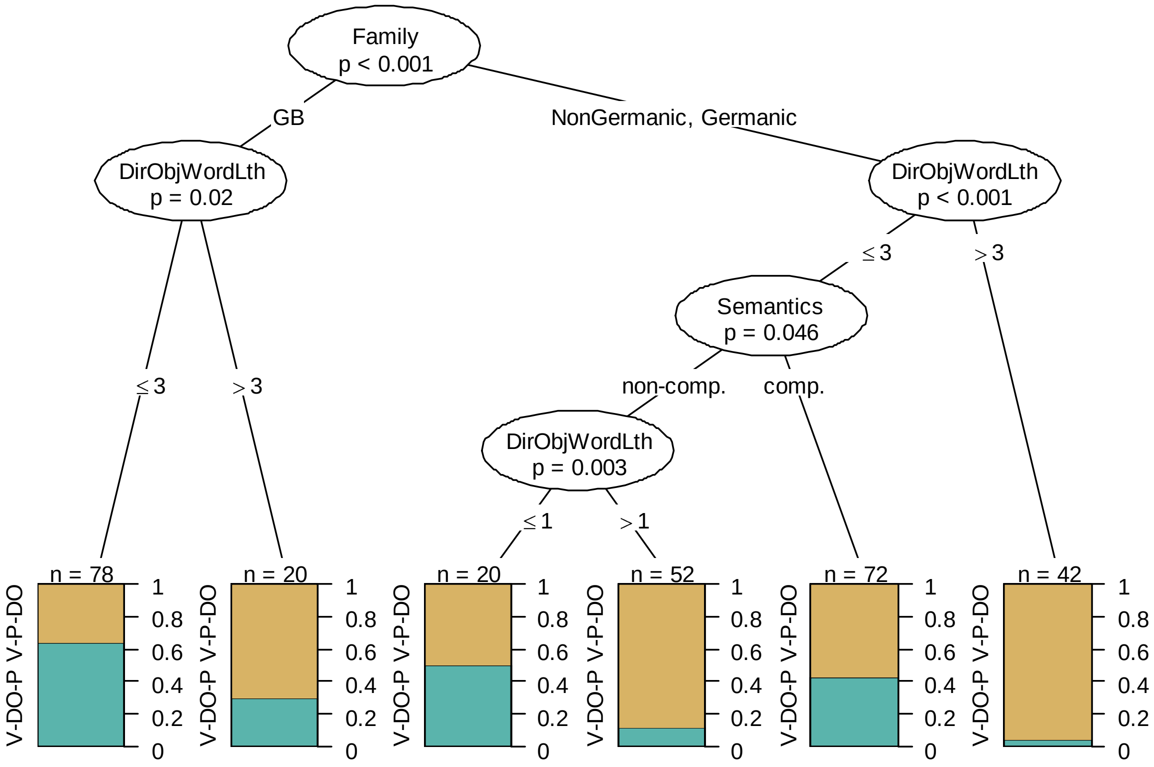

Figure 1: Conditional inference tree for particle placement

We now turn to a random forest analysis of the dataset. The random forest analysis performs significantly better than the single conditional inference tree, with a C statistic of 0.82, and a predictive accuracy of 73.9%, compared to the single conditional inference tree accuracy of 70.7% (pbinom < .001). In this instance we find that the random forest model is a substantial improvement over the single conditional inference tree model.

The explanatory power of individual predictors can be assessed by comparing the decrease in overall accuracy of the model, measured as the difference in the concordance statistic C, when each predictor's values are randomly permuted. The greater the decrease, the more important the predictor. The relative ranking obtained from the random forest analysis is shown in Figure 2. Here we see that more than any other predictor, FAMILY makes the largest contribution overall, followed by DIROBJWORDLENGTH, SEMANTICS and DIROBJDEFINITENESS. Neither DIROBJTYPE and

DIROBJCOMPLEXITY play much of a role however. This ranking is largely consistent with the results above.

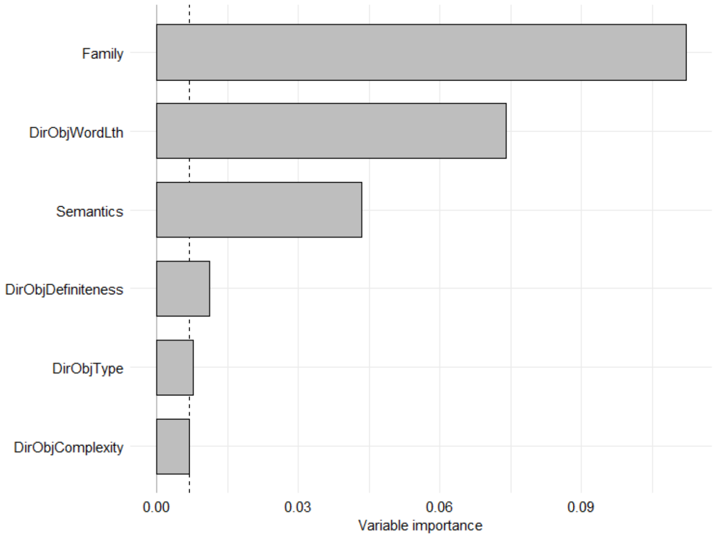

Figure 2: Predictor importance ranking for random forest analysis of particle placement. Values on x-axis reflect decrease in model fit (C) when predictor values are randomly permuted.

There is only one notable exception: DIROBJDEFINITENESS ranks somewhat higher than we expected in the random forest model, given that we found no splits at all in the single conditional inference tree model. This finding illustrates one potential danger of relying on a single conditional inference tree, which is that the position of a predictor in a single tree is not always a reliable indicator of its relative importance overall (see Szmrecsanyi et al, 2016). A single tree only displays the most predictive binary split within a (sub)region of the data, and the precise character of each subregion will depend on the splits that have been made above it. The usefulness of conditional inference trees lies in their ability to illustrate potential interactions among predictors, rather than their ability to assess predictors' overall importance. However, because the variable importance rankings in the random forest are based upon the conditional permutation of predictor variables over many trees, we can be confident that the rankings in Figure 2 reflect our best model of the relative explanatory power of each of our predictors. The superiority of the random forest method is attested by its substantial improvement in classification accuracy (C = 0.81) over the single inference tree method (C = 0.65).

The directions of the variable effects predicted by the random forest, as shown in Figure 3, are largely as we expect. The discontinuous variant is more likely when the direct object is short, structurally simple, and either an indefinite pronoun or a definite gerund or demonstrative. Semantically compositional verbs also favor the discontinuous variant. Finally, we find substantially higher probability of the discontinuous variant in the GB dataset than the other two, but no meaningful difference between the Germanic and Non-Germanic families.

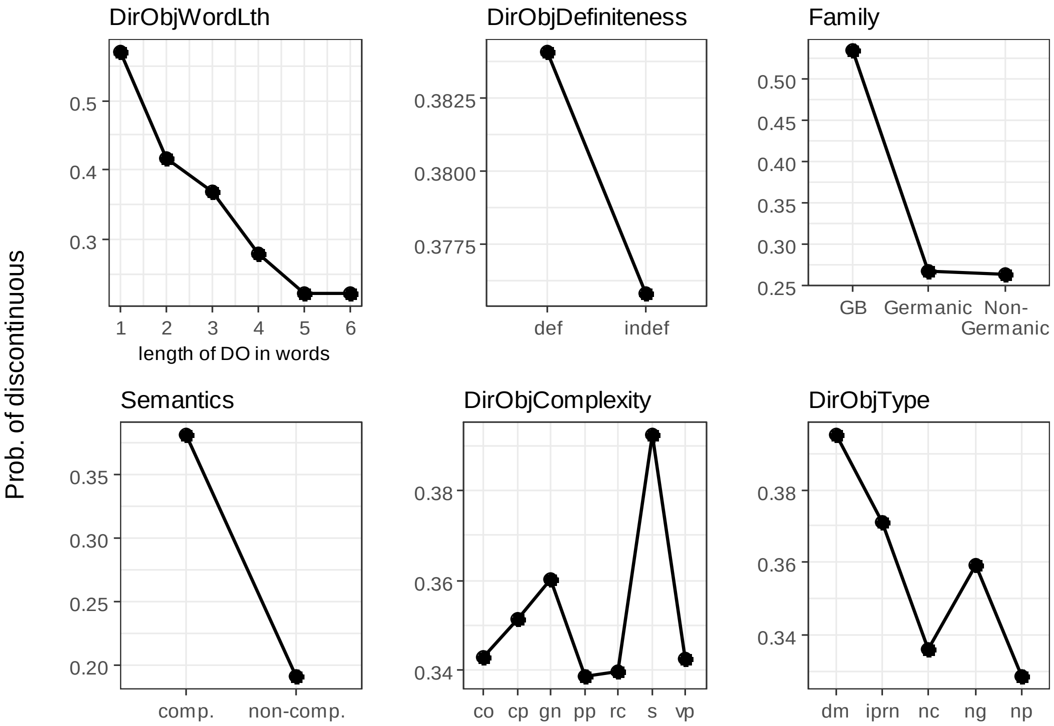

Figure 3: Predicted probabilities of the discontinuous variant derived from the random forest model.

In order to explore cross-family differences in predictor importance more closely, we next divide our dataset by FAMILY and compute a separate random forest analysis in each of the three resulting datasets. We use this approach, rather than a single model with interactions, for two reasons. First, interpretation of interaction effects in random forests is not as straightforward as in other methods such as regression (Wright et al. 2016). Second, logistic regression modeling is much less reliable for small datasets such as ours (Harrell 2015: 233). If we look at the permutation variable importance ranking across these datasets (Figure 4), the results provide a slightly more complex picture where:

1) More predictors as described in the literature have an effect on particle placement alternation in the native British dataset;
2) The EFL learners with Germanic languages resemble the British speakers to some extent; DIROBJWORDLENGTH exerts a considerable influence in both datasets, while VERBSEMANTICS and DIROBJDEFINITENESS also play a role but to varying degrees.
3) Particle placement alternation by EFL learners with French, Italian, Spanish and Greek L1 background is surprising in the sense that direct object length does not seem to be an important predictor while semantics really stands out as the most important predictor in this subset. Given the very small number of discontinuous tokens in this dataset, however, it remains to be seen to what extent our findings are representative.

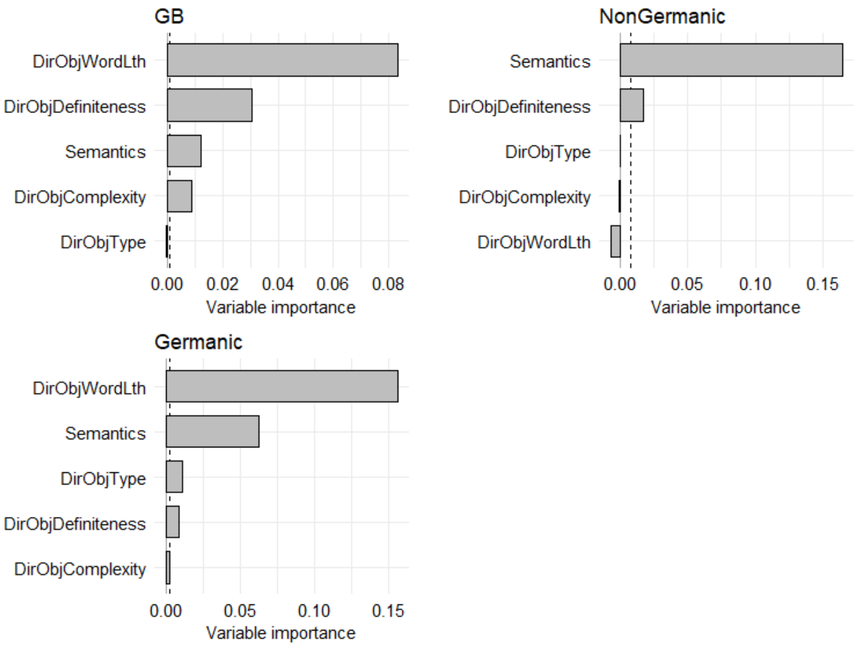

Figure 4: Predictor importance ranking by FAMILY

4. Discussion

The present study was constrained by the limited number of data available to us. This meant that our analysis of the factors influencing particle placement was limited to a smaller set of predictors than in other recent studies (Gries 2003; Szmrecsanyi et al, 2016; Grafmiller & Szmrecsanyi to appear). Nonetheless, using methods well-suited to the problem at hand, our analysis revealed a number of new insights regarding the relationship between the phrasal verb grammars of L1 and EFL language users.

Most importantly given our research questions, the particle placement alternation exhibits strong variety effects – there is a first split between British English and non-native varieties of English in the conditional inference tree, and FAMILY is ranked as the most important predictor of particle placement choice by the full dataset random forest analysis. This is mainly due to the fact that EFL speakers tend to use the V-OBJ-Prt construction far less frequently overall than L1 English speakers (see Table 4). As discussed by Gilquin (2015), there are a number of reasons why EFL learners may favour the V-Prt-OBJ construction: the V-Prt-OBJ "corresponds to the basic transitive scenario of an agent acting upon a patient, it requires less processing effort because the particle is presented immediately after the verb, it is less marked (subject to fewer restrictions) and more salient (this is the most frequently elicited word order in experimental designs)" (Gilquin, 2015: 64-65; cf. also Gries, 2003: 141–142). Gilquin (2015) further notes that learners often study phrasal verbs in decontextualized lists of vocabulary, which is likely to strengthen the sense of unity that is created in the mind of EFL learners between the verb and the particle.

Although not statistically significant, there is also a difference in the proportion of V-OBJ-Part variants between learner populations with Germanic vs. non-Germanic language background: EFL learners with Dutch, German or Swedish as L1 use more V-OBJ-Part constructions than EFL learners with French, Italian, Greek and Spanish as L1.[1] One obvious explanation for this difference lies in the fact that the latter L1s do not have English-like phrasal verbs. At this stage, however, and this is a major limitation of our study, it is impossible to conclude whether the difference is due to crosslinguistic influence or is the result of inter-group differences in L2 proficiency. Gilquin et al. (2010) submitted a random sample of five interview extracts from each of the L1 subcorpora in LINDSEI to a professional rater who was asked to rate them on the basis of the Common European

[1] The difference is not statistically significant ($X^2 = 1.18$; p = .28) but this is more probably due to the size of the dataset than to the lack of difference.

Framework of Reference for Languages (CEF) descriptors for speaking (Council of Europe 2001). While a majority of the samples from the Dutch, German and Swedish components of LINDSEI were rated at the C (i.e. ‘advanced’) level of the Common European Framework of References for Languages (CEFR; Council of Europe, 2001), samples from LINDSEI-FR, LINDSEI-GR, LINDSEI-IT and LINDSEI-SP mainly scored at the B (i.e. ‘intermediate’) level (cf. Gilquin et al, 2010: 7).

The second predictor in both the conditional inference tree and random forest, i.e. length of the direct object, significantly influences the choice of the particle verb variant in the dataset as a whole. As shown in Figure 5, the effect of length is an incremental one: with each additional word in length, the discontinuous order (V-OBJ-Prt) becomes significantly less likely. However, Figure 5 also shows that the influence of length is dramatically reduced for the EFL speakers of non-Germanic L1s. Speakers of Germanic L1s do show an effect of length, though the effect is sharper and less gradual than it is for L1 English speakers.

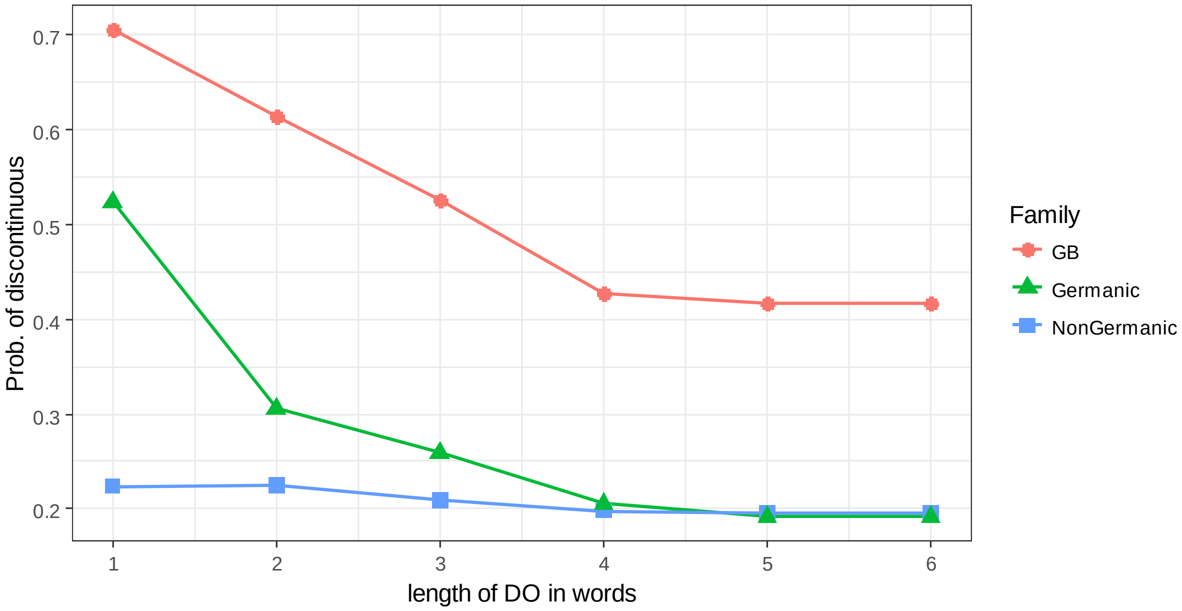

Figure 5: Effect of DIROBJWORDLENGTH in the individual FAMILY random forest models

The significant impact of SEMANTICS on the use of transitive particle verbs by EFL learners with non-Germanic L1 backgrounds may be explained as follows. A sizeable proportion of the very few transitive particle verbs that allow for particle placement alternation and do not have pronominal direct object in LINDSEI-FR (52%), LINDSEI-GR (82.6%), LINDSEI-IT (72%) and LINDSEI-SP (83.3%) are non-compositional. A large proportion of these non-compositional cases are continuous (from 63% in LINDSEI-FR to 84% in LINDSEI-GR and 100% in LINDSEI-IT and LINDSEI-SP). Examples include *carry out some research* (LINDSEI-IT), *found out the truth* (LINDSEI-IT), *gives up his work* (LINDSEI-SP), *make up my mind* (LINDSEI-FR), and *make up stories* (LINDSEI-GR). It is not clear why those non-compositional phrasal verbs would be so frequent in the non-Germanic L1 language varieties. Gilquin (2015) already made the point that learners often have to learn phrasal verbs in decontextualized lists of vocabulary, which is likely to create a strong link between the verb and its particle (see above). Arguably, these lists do not feature the most literal or directional phrasal verbs but tend to focus on the more salient metaphorical and idiomatic structures. It could also be hypothesized that the way in which EFL learners from different origins get exposure to different kinds of phrasal verbs, e.g. through media or education, might help explain the differences among EFL learners we observe here. People in Italy, Spain, Greece, etc. probably have much less ambient exposure to English (of all kinds) than people in Scandinavia, Holland, Flanders and Germany. For speakers who get practice with phrasal verbs in mainly educational settings, such limited exposure may affect their ability to master use of different verbs and the alternation itself, which is predominantly found in informal language (Dempsey et al. 2007; Grafmiller & Szmrecsanyi to appear).

All in all, the patterns that characterize the variation grammar in EFL learners' use of transitive particle verbs can be summarized as follows. For EFL learners with Germanic L1s, the grammar overlaps with the native grammar as found in LOCNEC both in terms of its overall complexity (the number of contributing factors) and in the relationship(s) between those factors. By contrast, for EFL learners with non-Germanic L1s (and lower levels of proficiency), the grammar is notable for its simplicity and for its heavy reliance on semantics. Broadly speaking, these findings parallel those of recent studies comparing variation in probabilistic grammars between ESL varieties at varying stages of nativization, which found that less developed varieties tend to exhibit simpler grammars dominated by a small set of factors (Grafmiller & Szmrecsanyi to appear; Heller et al. 2017; Röthlisberger et al. 2017; Szmrecsanyi et al. 2016). The extent to which these cross-varietal patterns in ESL and EFL are driven by differences in proficiency, L1/substrate influences, or simple exposure to English are an active area of research in probabilistic grammar studies.

5. Conclusion

The study presented here represents the first step in our collaborative work to explore whether EFL learners share a core probabilistic grammar with users of first and second language varieties of English (Szmrecsanyi et al., 2016). Although we investigated particle placement alternation in seven LINDSEI components and manually scanned 18,108 lines of concordance, the final dataset was rather small and all conclusions are therefore necessarily tentative. The study needs replication in a larger spoken learner corpus such as the *Trinity Lancaster Learner Corpus*, which also has the additional advantage of controlling for proficiency (Gablasova et al, forthcoming). To investigate the effect of mode, it should also be replicated on the basis of a large corpus of learner writing such as the *International Corpus of Learner English* (Granger et al, 2009). In the future, we also want to test a semi-automatic extraction procedure to retrieve transitive phrasal verbs and investigate the impact of other factors such as the frequency of the phrasal verbs, the association between a verb and a particle and the association between a verb and the V-OBJ-Part or V-Part-OBJ variants (see Gries & Stefanowitsch, 2004; Gilquin, 2015; Deshors, 2016) on learner spoken language. We are particularly interested in testing whether the prediction according to which higher phrasal verb frequency is correlated with (higher percentages of) V-OBJ-Part (Gries, 2011) also holds in learner language. We also intend to investigate other alternation phenomena such as the genitive alternation or the dative alternation to better understand to what extent factors that have been shown to influence native speakers' choice between two variants also play a role in EFL learners' choices governing the alternations.

Acknowledgements

We thank Marie Gabrys and Sarah Pesenti, two MA students at the Université catholique de Louvain, for their help with the data coding.

References

Bresnan, J. (2007). Is syntactic knowledge probabilistic ? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (eds). *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton de Gruyter, 75-96.

Bresnan, J. & M. Ford (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. Language 86(1): 168–213.

Bresnan, J., A. Cueni, T. Nikitina, & B. Harald (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds). *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Capelle, B. (2009). Contextual cues for particle placement. In Bergs, A. & Diewald G. (eds). *Contexts and Constructions*. Amsterdam: John Benjamins, 145-192.

Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics* 18(3): 418-442.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Dagut M. & B. Laufer (1985) Avoidance of phrasal verbs – A case for contrastive analysis. *Studies in Second Language Acquisition* 7:73-79.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures* (BELL), New Series, 2, 225–246.

Dempsey, K. B., McCarthy, P. M., & McNamara, D. S. (2007). Using phrasal verbs as an index to distinguish text genres. In *Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference* (pp. 217–222). Menlo Park, CA: AAAI Press.

Deshors, S. (2016). A co-varying collexeme analysis of verb-particle combinations in EFL and their semantic associations. *International Journal of Learner Corpus Research* 2(1): 1-30.

Gablasova, D., & Brezina, V. (2015). Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In J. Romero-Trillo (Ed.). *Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies* (pp. 117-136). Berlin: Springer International Publishing.

Gablasova, D., Brezina, V. & McEnery, T. (forthcoming). The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*.

Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11(1): 51-88.

Gilquin, G., De Cock, S. & S. Granger (2010). *Louvain International Database of Spoken English. Interlanguage*. Louvain-la-Neuve, Belgium: Presses. Universitaires de Louvain.

Grafmiller, J. (2014). Variation in English Genitives across Modality and Genres. *English Language and Linguistics* 18 (03): 471–96. doi:10.1017/S1360674314000136.

Grafmiller, J. (2015). Guidelines for selection and annotation of interchangeable particle verbs. Internal report (last modified, 11 October 2015).

Grafmiller, J., M. Röthlisberger, B. Heller & B. Szmrecsanyi (2016). Annotation of common features for the genitive, dative, and particle placement alternations. Internal report (last modified, 11 March 2016).

Grafmiller, J. and B. Szmrecsanyi (To appear). Mapping out particle placement in Englishes around the world. A case study in comparative sociolinguistic analysis. *Language Variation and Change.*

Granger S, Dagneaux E, Meunier F and Paquot M (2009). *The International corpus of learner English. Handbook and CD-ROM* (Version 2) Louvain-la-Neuve: Presses Universitaires de Louvain.

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

Gries, S. & S. Deshors (2015). EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research* 1(1): 130-159.

Gries, S. & A. Stefanowitsch (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97-129.

Gries, S. Th. & S. 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, F. E. (2015). *Regression Modeling Strategies*. 2nd ed. New York: Springer.

Heller, B., Szmrecsanyi, B., & Grafmiller, J. (2017). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics*, 45(1), 3–27. https://doi.org/10.1177/0075424216685405

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. https://doi.org/10.1198/106186006X133933

Hulstijn J. and E. Marchena (1989) Avoidance: Grammatical or semantic causes. *Studies in Second Language Acquisition* 11: 242-55

Laufer B. and S. Eliasson (1993) What Causes Avoidance in L2 Learning: L1 L2 difference, L1/L2 Similarity, or L2 Complexity? *Studies in Second Language Acquisition* 15(1): 35-48.

Liao Y. and Y.J. Fuyuka (2004) Avoidance of phrasal verbs: the case of Chinese learners of English. *Language Learning* 54(2):193-226.

Linck, J. A., P. Osthus, J.T. Koeth & M. F. Bunting. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review 21(4). 861–883*. doi:10.3758/s13423-013-0565-2.

Rosenbach, A. (2014). English genitive variation – the state of the art. *English Language and Linguistics* 18 (02): 215–62. doi:10.1017/S1360674314000021.

Röthlisberger, M., Grafmiller, J., & Szmrecsanyi, B. (2017). Cognitive indigenization effects in the English dative alternation. Cognitive Linguistics, 28(4). https://doi.org/10.1515/cog-2016-0051

Strobl, C., J. Malley, and G. Tutz (2009) An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods 14(4): 323–348.

Sung, M.-C. (2017). Underuse of English verb-particle constructions in an L2 learner corpus: Focus on structural patterns and one-word preference. *Corpus Linguistics and Linguistic Theory*. Advanced access. https://doi.org/10.1515/cllt-2017-0002

Szmrecsanyi, B., J. Grafmiller, B. Heller & M. Röthlisberger (2016). Around the world in three alternations: modeling syntactic variation in varieties of English. *English World-Wide*, 37(2): 109-137.

Tagliamonte, S. & H. Baayen (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2): 135-178.

Waibel, B. (2007). *Phrasal verbs in learner English: A corpus-based study of German and Italian students*. Doctoral dissertation. The Albert-Ludwigs-University, Freiburg.

Wright, M. N., A. Ziegler & I. R. König (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17(1). https://doi.org/10.1186/s12859-016-0995-8.

Wulff, S., Lester, N. & M. T. Martinez-Garcia. (2014). That-variation in German and Spanish L2 English. *Language and Cognition* 6(02). 271–299.