MuPDAR for corpus-based learner and variety studies

Two (more) suggestions for improvement

Stefan Th. Gries
University of California, Santa Barbara / Justus Liebig University Giessen

Corpus-based studies of learner language and (especially) English varieties have become more quantitative in nature and increasingly use regression-based methods and classifiers such as classification trees, random forests, etc. One recent development that is becoming more widely used is the MuPDAR (Multifactorial Prediction and Deviation Analysis using Regressions) approach of Gries and Deshors (2014) and Gries and Adelman (2014). This approach attempts to improve on traditional regression- or tree-based approaches by, firstly, training a model/classifier on the reference speakers (often native speakers in learner corpus studies or British English speakers in variety studies), then, secondly, using this model/classifier to predict what such a reference speaker would produce in the situation the target speaker is in (often non-native speakers or indigenized-variety speakers). The third step then consists of determining whether the target speakers made a canonical choice or not and explore that variability with a second regression model or classifier. The present paper is a follow-up to Gries and Deshors's (2020) and offers additional answers to a variety of questions that readers and audiences to MuPDAR presentations have been raising for a few years. First, I show how MuPDAR can be extended straightforwardly to alternations that involve more than the typically used binary choices; I do so in a way that also addresses another potential challenge and exemplify this with a case study from varieties research. Second, I outline a casewise-similarity approach towards predicting what reference speakers would do that avoids frequent regression modeling problems and exemplify, as well as compare, it to competing alternatives with a case study from learner corpus research.

Keywords: corpus-based alternation research, learner corpus research, variety research, MuPDAR, predictive modeling

https://doi.org/10.1075/scl.105.09gri

1. Introduction

1.1 General introduction

Corpus-based studies of learner language and (especially) English varieties have become more quantitative in nature and many studies in particular on morpho-syntactic alternation studies are increasingly using regression-based methods and other classifiers (e.g. classification trees or random forests). A common denominator of many such studies is that a binary morphosyntactic, or constructional choice (*constructional* being used in the Construction Grammar sense of the term) is studied in terms of two things:

– which predictors affect the constructional choice? (Often, a substantial number of predictors is taken into account.)
– what differences, if any, are there between native and (different kinds of) non-native speakers (in learner corpus studies) or between speakers of a native/inner-circle variety and (different kinds of) other variety speakers (in varieties research)?

Critically, in order to answer the second question, the regression model or, more generally, the classifier must allow for each predictor to interact with a predictor that encodes the status of the speaker: native speaker (NS) vs. non-native speaker (NNS, i.e. of often several different L1s) for learner corpus research and NS vs. 'other-variety' speaker (VS of often several different varieties). Such an approach – typically using generalized linear regression modeling or tree-based approaches and involving speaker status as one of often many predictors – has been used in quite a few studies, typically on binary alternation choices such as the genitive alternation (*of* vs. *s*), the dative alternation (ditransitive vs. prepositional datives), particle placement (V Part DO vs, V DO Part), prenominal adjective order (Adj_1 Adj_2 N vs. Adj_2 Adj_1 N), *that*-complementation (present vs. absent), and many others:

(1) a. the access code of Commander Sinclair
    b. Commander Sinclair's access code

(2) a. Mr Garibaldi gave his deputy the access code.
    b. Mr Garibaldi gave the access code to his deputy.

(3) a. Mr Garibaldi gave back the key card.
    b. Mr Garibaldi gave the key card back.

(4) a. the nice shiny key card
    b. the shiny nice key card

(5) a. The Vorlons didn't care that the Shadows were attacking.
    b. The Vorlons didn't care the Shadows were attacking.

In spite of the fact that these studies are a huge improvement over decades of monofactorial chi-squared tests, depending on one's perspective, Gries and Deshors (2014) and Gries and Adelman (2014) developed an alternative approach that has a slightly different focus. Their multi-step approach is called MuPDAR(F) (Multifactorial Prediction and Deviation Analysis using Regression/Random Forests), and it involves four steps:

i. one trains a first straightforward/regular model (R1, regression 1) or other classifier (e.g. RF1, random forest 1) on what I will here call the reference speakers (RS: in learner corpus studies those are the native speakers; in variety studies, those are often speakers of the historical source variety of British English);
ii. if this first model/classifier is good (enough), one uses it to predict what such a RS would produce in the situation the target speaker is in (TS: in learner corpus studies those are the learners; in variety studies, these are indigenized-variety speakers); that means, for every TS choice, one then also knows what a RS would have done in that context;
iii. one determines whether the TS's actual choice (as manifested in the data) is the same choice as the one the RS is predicted to have made); for instance, if a learner produced an *s*-genitive in the situation in which a native is predicted (from R1) to also have predicted one, then the learner made 'the canonical/nativelike choice' but if a learner produced an *of*-genitive in that same situation, they made a non-nativelike choice. Thus, the result of this step iii. is usually one vector/column with the predicted probability of the choices the RS would have made and another vector/column with the values TRUE/FALSE or nativelike/non-nativelike;
iv. one explores the degree to which the TS made canonical choices with a second regression model (R2) or classifier (e.g. RF2) that involves only the TS part of the data. That second model/classifier can have one of two kinds of responses: On the one hand, it can be the binary variable that states whether TS made the choice that a RS would have made in their place or not, i.e. the second vector/column mentioned in iii. On the other hand, sometimes a more precise version is used, one in which the second classifier's response variable is a numeric variable that quantifies how much 'off' a TS choice is, which is computed from the probability of what a RS is predicted to have chosen, i.e. the first vector/column mentioned in iii. This is more precise because it provides a finer resolution into the (non-)canonicalness of the TS choice than a mere binary classification. Crucially, this second model/classifier has as predictors the same linguistic and/or contextual predictors that the first classifier used, but also a predictor typically called L1 or COUNTRY/VARIETY that can interact with all others to see, for instance, whether learners from different L1s make non-nativelike choices in the same way or differently from each other.

The point of this approach, compared to a 'traditional' one-regression-only kind of approach is to explore in particular those kinds of differences between RS and TS that lead to different choices. A 'normal' regression approach might return significant differences between levels of a predictor even if they do not actually lead to different categorical decisions. However, by virtue of step iii. and iv, MuPDAR(F) by contrast focuses on those cases where the actual categorical choices are different; in addition, since it can include the actual linguistic choice of the TS as a predictor, the construction-specific effects of predictors can be highlighted better than in the traditional approach. (None of this is to imply that the traditional approach doesn't have its merits!)

MuPDAR(F) has led to many interesting results, including Gries and Deshors (2014) on *may* vs. *can*, Gries and Adelman (2014) on subject realization vs. omission in learner Japanese, Wulff and Gries (2015) on prenominal adjective order in English, Deshors and Gries (2016) and Kolbe-Hanna and Baldus (2018) on *ing* vs. *to*-complements, Heller et al. (2017) on the genitive alternation in British vs. Singaporean English, Wulff and Gries (2019) on particle placement, Wulff and Gries (2020) on genitives in learner data, Kruger and De Sutter (2018) and Lester (2019) on *that*-omission, Schweinberger (2020) on adjective amplification, Werner et al. (2020) on present perfect vs. simple past choices, etc.

While these applications have led to instructive results, the way MuPDAR(F) has most often been used can still be improved and especially conference presentations of the method have triggered several questions/concerns. In a first methodological follow-up paper, Gries and Deshors (2020) made two suggestions. First, they addressed the frequent question about how to deal with cases in which both of the alternants are in fact acceptable choices and whether MuPDAR(F) in particular, but actually any classifier-based approach in general, might not be overzealous in its labeling of certain constructional choices as non-canonical (much like none of the examples above in (1) are (5) is obviously unacceptable). Gries and Deshors (2020) propose to recognize a middle-ground, i.e. to improve step iii of MuPDAR above by permitting it to return a prediction that in effect says 'in this particular speech situation that the TS is in, a RS would find both alternants acceptable', a categorical version that could consist of treating predicted probabilities between 0.4 and 0.6 as 'not unambiguously favoring one choice'. Instead of using such a predicted probability cut-off, however, they suggest another possible way to quantify what counts as middle ground, namely 1 logloss unit (see Section 2.3.2 for concrete examples of how logloss is computed) around the predicted probability cut-off point of 0.5, which sets the stage Section 2.3 below). They then proceed to (i) exemplify that approach by applying it (with random forests) to the dative alternation and (ii) discuss qualitatively how the middle-ground examples differ from the others.

Gries and Deshors's (2020) second suggestion to improve alternation studies is more general and programmatic in nature and, while applicable to MuPDAR(F), not exclusively focused on it. This second proposal is essentially an argument in favor of studying something that hardly any previous work has paid any attention to so far: the most spectacularly misclassified/mispredicted cases, i.e. the cases where, say, a learner makes a *very* surprising choice. For instance, in a learner corpus study on the genitive alternation (see (1)) a NNS's choice of an *of*-genitive for this situation/context would be quite surprising:

- a short given/accessible/inferable human (and thus animate and concrete) possessor;
- a long new concrete inanimate object as the possessum;
- an obvious straightforward POSSESSION relationship between the two.

This would be a completely obvious case for a NS to avoid an *of*-genitive (?*the newly-acquired office furniture of my dad*) – they would use an *s*-genitive (*his newly-acquired office furniture*). If, however, the NNS nonetheless chooses an *of*-genitive, then this case, and others like it, should be explored (post hoc) qualitatively, but that is something that hardly ever happens, and Gries and Deshors (2020) provide a quick exploration of the 100 worst-predicted cases of the dative alternation by German and Chinese learners (in ICLE and LINDSEI).

1.2 Motivation of the present paper

In the present paper, I want to discuss two other questions that have sometimes been raised with regard to the MuPDAR(F) approach. The first question is fairly straightforward, but one that has not been discussed yet in any published or presented work: How can MuPDAR(F) be done when the alternation is not one between two choices (e.g. like all the examples in (1) to (5)) but between three or more choices? Examples for the latter might include

- future choice: *will* vs. *shall* vs. *going to* vs. present tense with future temporal adverbials;
- voice: active vs. passive with *by* vs. passive without *by*;
- deontic modality as in *The Drazi must/have to/need to stop fighting each other*; etc.

The second question to be discussed in this paper is more fundamental and is mostly concerned with steps i and ii from above. At this point, the first model/classifier has always been a regression model or a tree- or forest-based approach.

However, the distributional characteristics of corpus data can make both these approaches quite difficult: First, we often have really not that many data points so that, once random effects and/or interactions between predictors are considered, cell frequencies are often quite small, which is problematic for both regressions and tree-based approaches; also and somewhat ironically, tree-based approaches as they are usually employed can fail to detect interactions in the data (Boulesteix et al. 2015, Gries 2020), a potential problem that will also be addressed below. Second, many predictors are very Zipfian-distributed, meaning that sometimes very few levels of categorical predictors or very small value ranges of numeric predictors account for most of the data, leaving very little data from which to make robust estimates for much of the rest of the predictor(s); that, too, can be highly problematic for the models'/classifiers' robustness and predictive power. Thus, a question that has often been posed is how especially the first two steps of MuPDAR can be applied when the data are 'problematic' in such ways.

In what follows, I will discuss suggestions on how to deal with both of these questions. Section 2 will propose two straightforward ways to extend the so far usually binary-alternation applications to a case with a four-level constructional choice in the domain of variety research. Section 3, on the other hand, will develop a protocol to predict what a reference speaker would do even if the data are such that regression- or tree-based approaches might encounter difficulties in the domain of learner corpus research. Section 4 will then conclude.

2. Case study 1: The dative and voice alternation across varieties

2.1 Introduction

In this section, I will discuss two ways in which MuPDAR can be made to work with multinomial alternations (i.e. alternations involving more than two alternants); the first proposal is an extremely simple extension way that is basically a straightforward extension of the binary logistic regression application; the second one is much better and more precise in how it takes all predicted probabilities of each corpus example into consideration. To exemplify these two approaches, I will use the data from Bernaisch et al. (2014), who are concerned with cross-varietal differences and similarities in South Asian Englishes and British English for the four verb-complementation patterns with the verb *give* that arise from the combination of the dative alternation (ditransitive vs. prepositional dative as in (2)) and the voice alternation (active vs. passive), as exemplified in (6).

(6) a. Morden gave Londo a present.
b. Morden gave a present to Londo.
c. Londo was given a present by Morden.
d. A present was given to Londo by Morden.

Their data come from the South Asian Varieties of English (SAVE) corpus, which contains national newspapers featuring Bangladeshi, Indian, Maldivian, Nepali, Pakistani, and Sri Lankan English; in addition and to represent British English as the 'reference variety', they also use the news section from the British National Corpus.

Given the high frequency of *give*, Bernaisch et al. (2014) drew random samples from the SAVE corpus and the BNC and then focused on the above-mentioned four complementation patterns. Ultimately, 1871 instances of these constructions were annotated with regard to the following variables:

- TRANSITIVITY: the construction in which *give* is used *ditransitive active, ditransitive passive, prepositional dative active*, or *prepositional dative passive*;
- COUNTRY: *GB* vs. *Ban, Ind, Mal, Nep, Pak*, or *SL*;
- RECLENGTH and PATLENGTH: the lengths of the recipient and the patient in words;
- RECANIMACY and PATANIMACY: *animate* vs. *inanimate*, i.e. the animacy of the recipient and/or the patient;
- RECACCESSIBILITY and PATACCESSIBILITY: *given* vs. *new*, specifically the accessibility of the recipient and/or the patient operationalized as to whether the recipient/patient was mentioned in the preceding ten lines;
- RECPRONOMINALITY and PATPRONOMINALITY: *pronoun* vs. *np*, i.e. whether the recipient/patient was a pronoun or a lexical NP;
- PATSEMANTICS: *abstract* (as in *give him a hard time*), *concrete* (as in *give him a book*), or *informational* (as in *give him a warning*).

Bernaisch et al.'s (2014) analysis of these data involves a conditional inference tree and a random forest, but in this section, I will use it to exemplify two ways of using MuPDARF with a constructional choice involving more than two levels. For the analysis here, the variables RECLENGTH and PATLENGTH were logged (to the base of 2), plus I added a variable LENGTHDIFFERENCELOG (RECLENGTHLOG minus PATLENGTHLOG). In addition, PATANIMACY and PATPRONOMINALITY were not considered here any further because of their extreme imbalance (in each, one level accounted for > 99% of all cases).

Finally, another important aspect of the present analysis is concerned with the predictors (of especially the second model/classifier). Many studies do not

explore interactions between predictors much, neither in a regression context nor in tree-based approaches. This may in part be due to the fact that the inclusion of interactions usually comes with increasing demands on sample size, but also due to the fact that researchers believe that even the standard application of tree-based methods handles interactions well (enough). However, as Gries (2020) has discussed in detail, that is not necessarily the case: Random forests, for instance, can be quite bad at informing the researcher of interaction effects because the variable importance statistics they return for individual predictors do not unambiguously reveal whether a high variable importance of a predictor *x* is due to *x*'s marginal effect (i.e. Wright et al. 2016: 7; Gries 2020: 636–637). Therefore, an additional novelty aspect of the present study is to exemplify briefly how to deal with this. In the present case, I suspected that certain predictors might interact with each other: RECACCESSIBILITY and PATACCESSIBILITY (because it is conceivable that the effect of PATACCESSIBILITY is weakened depending on what RECACCESSIBILITY does) as well as RECPRONOMINALITY and PATSEMANTICS (because it is conceivable that the effect of PATSEMANTICS is weakened depending on what RECPRONOMINALITY does), which is why I followed the logic of Gries (2020, Section 3.1) and created interaction predictors for these two combinations of two predictors each, which were then also entered into the analysis (see Deshors & Gries 2020; Deshors 2020, to appear, for uses of this method).

2.2 MuPDAR: Steps i and ii

To implement the first two steps of the MuPDAR protocol, the data were split up into the data for the GB speakers (i.e. the RS) vs. all others (i.e. the TS). Then, in step i, a random forest was fit on the GB data part only, which included all of the above predictors. This random forest, RF1, achieved an OOB prediction accuracy of 66.78%, which, according to an exact binomial test, is significantly better than the no-information rate of 61.99% of always picking ditransitive actives ($p = 0.012$). In step ii, RF1 was used to generate RS-based predictions for all the TS data in the South Asian Englishes part of the data. It turns out that the TS choices are the same as the predictions of the RS-trained random forest RF1 in 55.29% of all cases, again significantly more often than the no-information rate ($p_{\text{binomial}} < 10^{-13}$).

It is step iii that is now potentially trickier in this situation than in MuPDAR approaches with a binary linguistic choice, which will be discussed in what follows.

2.3 MuPDAR: Step iii for a multinomial context

2.3.1 *The simple version*

The simplest way to do step iii, to compare the TS choices to the RS-based predictions, is to do this in a binary fashion; this is in fact what has been done in most MuPDAR applications and it works here as well. One just generates a variable that states for each TS choice whether it is the RS-based prediction or not and then this variable can be the response variable for the second model/classifier. If one does that here, the second classifier, another random forest, achieves an OOB prediction accuracy of 67.26%, which is significantly better than the no-information rate ($p_{\text{binomial}} < 10^{-22}$).

While appealing in its simplicity, this strategy is perhaps also too simple because it loses a lot of information. If a RS is predicted to use a ditransitive active with a predicted probability of 0.93 but a TS does not use such a ditransitive active, this is (likely) a fairly obvious non-canonical choice. If, on the other hand, a RS is predicted to use a ditransitive active with a predicted probability of 0.43 and a prepositional dative active with a predicted probability of 0.41 and a TS produces the prepositional dative active, then a RS might find that constructional choice perfectly fine and acceptable as well, but the simple version discussed here would label it – overzealously, see Gries and Deshors's (2020) middle ground argument – as non-canonical. For this reason, even the first MuPDAR publications already proposed the more fine-grained alternative briefly discussed above, which will now be extended as well as improved, in the next section.

2.3.2 *The better version*

As mentioned above, a more precise version of measuring how much a TS choice may deviate from a RS choice has also been used. This first extension of MuPDAR was already discussed in the first such study (Gries & Deshors 2014: 128); it still focuses on the cases where the choices of TS differ from those predicted for the RS, but it also considers the severity of any deviations:

- if a TS made the choice a RS is predicted to make, then a so-called DEVIATION score is set to 0 (because there is no deviation between the RS prediction and the TS choice);
- if a TS is predicted to make choice *a* with a predicted probability $pp = 0.51$, but a RS uses choice *b*, then this is considered non-canonical but with a very small DEVIATION score of just $pp - 0.5 = 0.01$ (the subtrahend is 0.5 because when two choices are available, then a classifier predicts the one whose $pp \geq 0.5$ (with the other choice having a predicted probability of $1 - pp = 0.49$)); however,

– if a TS is predicted to make choice *a* with a predicted probability *pp* = 0.91 but a RS uses choice *b*, this is considered non-canonical with a now much larger DEVIATION score of *pp* −0.5 = 0.41.

As pointed out in Gries and Deshors (2020: 73), the DEVIATION score is, in some sense, a signed, but less punitive, 'version' of the classification measure of logloss (see below for examples), which (just like the Brier score) penalizes false classifications/predictions but penalizes false classifications/predictions that were made confidently/boldly more; in other words, the worst predictions are confident/bold predictions that turn out to be wrong. This approach is definitely much more fine-grained than the binary approach and has been used successfully in, for instance, Gries and Deshors (2014) or Lester (2019). However, there is also an at least potential shortcoming of this approach. Consider what happens if the TS make 'canonical' choices in the vast majority of cases, e.g. 80% or 85% or even more. With data like these, something statistically undesirable can happen: Most of the DEVIATION scores will then be 0 and only a potentially tiny number of cases will differ from 0. Once the second model/classifier is fit with DEVIATION as the response variable, then, especially if that is a regression model, that may lead to difficulties with model fitting, at least when the usual straightforward linear (mixed-effects) models are used. Thus, it would be nice to be able to improve this, and improve it we can in a way already alluded to in Gries and Deshors (2020), namely by using the above-mentioned measure of logloss, specifically, the multi-class version of logloss, i.e. one that is designed to handle response variables with more than two levels.

Let us first consider how multi-class logloss is computed. Logloss is a single value that quantifies how good a model/classifier is at making classifications/predictions for a complete data set, and that single value is the average of a 'contribution to logloss' for every single case in the data set. That contribution to logloss of a case in turn is the negative log of the model/classifier's predicted probability for the level that was observed in that case, i.e. that the model/classifier should have predicted. Consider scenario 1 shown in Table 1: The response variable has four levels (*a, b, c*, and *d*) and their predicted probabilities from some classifier are 0.97, 0.01, 0.01, and 0.01 respectively. If the speaker's choice was *a* (i.e. that's what the model should have predicted), then the classifier did a really good job here because it did indeed assign a huge predicted probability to outcome *a* and really low predicted probabilities to all other outcomes; the contribution to logloss would correspondingly be the bold values in Table 1, i.e. a value close to 0, namely 0.0305.

However, if the speaker's choice was *b* as in Table 2, then the classifier did a really bad job here because it assigned a huge predicted probability to a wrong outcome (*a*) and a really low one to the one it should have predicted (*b*); the contribution

Table 1. Scenario 1 (the speaker chose and the correct prediction would be *a*), contribution to logloss in bold type

	Response: *level a*	**Response: level *b***	**Response: level *c***	**Response: level *d***
predicted probability	0.97	0.01	0.01	0.01
-log of pred. prob	**0.0305**	4.6052	4.6052	4.6052

to logloss would correspondingly be the bold value in Table 2, i.e. a value much greater than 0, namely 4.6052. In other words, the more the contribution to logloss value for a prediction deviates from 0, the more, in this case, non-canonical the TS choice (relative to the RS prediction).

Table 2. Scenario 1 (the speaker chose and the correct prediction would be *b*), contribution to logloss in bold face

	Response: *level a*	**Response: level *b***	**Response: level *c***	**Response: level *d***
predicted probability	0.97	0.01	0.01	0.01
-log of pred. prob	0.0305	**4.6052**	4.6052	4.6052

This also helps avoid the potential problem of the DEVIATION score that, if most predictions are correct (e.g. 80%), then 80% of the DEVIATION scores will be 0, which can pose problems to regressions: the logloss values will be much more diverse/variable because there is no binary 'as-predicted-vs-not' classification and many different constellations of predicted probabilities for the four response levels give rise to many different values, which is good in how it makes for a better response variable to be modeled in step iv. In this particular case, for instance, the 1579 NNS cases lead to 727 different logloss values. How are these multi-class logloss values used then? That is the topic of the next section.

2.4 MuPDAR: Step iv for a multinomial context

Just like the DEVIATION scores from some previous analyses, the multi-class logloss values computed in the previous step are used as the response variable in the second model/classifier of the MuPDAR(F) protocol: the higher the logloss value, the more the NNS made a 'non-canonical' choice and the second model/classifier will determine which linguistic/contextual predictors are most associated with high logloss values or, in other words, with what seems to be difficult for NNS. As for predictors, I used all predictors used in RF1 above (including the two interaction predictors) but, since we now have data for multiple varieties/

countries, I again followed Gries (to appear) and also created interaction predictors of COUNTRY (i.e. variety) with every categorical linguistic/contextual predictor[1] – the point is to see whether some of these predictors have different effects in different countries. Then RF2 was trained on the logloss values from step iii to obtain a forest of, now, regression trees. Which variables are driving this forest? Figure 1 represents permutation-based variable importance values (predictors involving COUNTRY are highlighted in grey, the error bars represent one standard error of the importance scores). Interestingly, most of the (interaction) predictors involving COUNTRY do not seem to be noteworthy because they end up having negative importance scores.

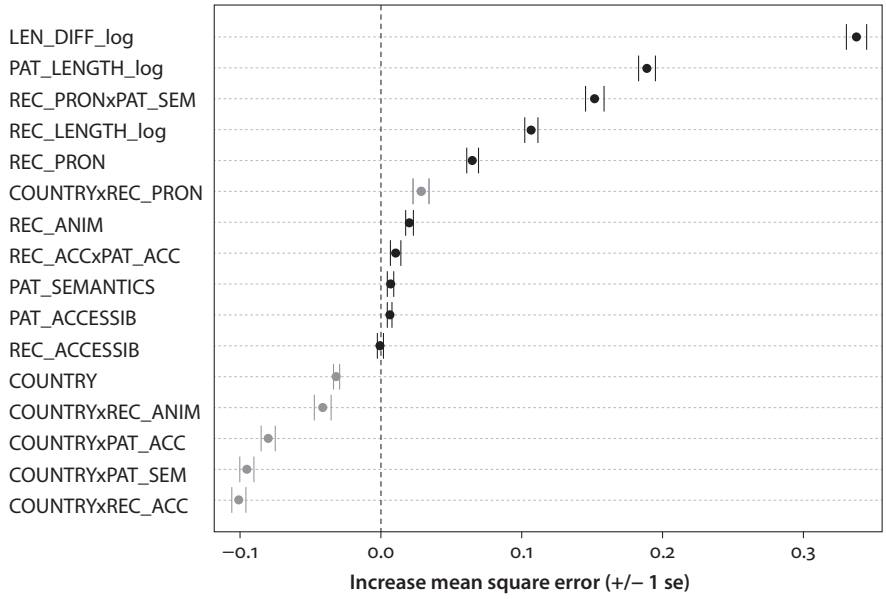

Figure 1. Variable importance scores of RF2 (modeling logloss)

1. I am not including interaction predictors here that combine COUNTRY and the numeric length predictors. While it is possible to do so, it requires a few non-trivial statistical considerations that go beyond what can be reasonably discussed here. Also, the reported RF2 does not include the construction chosen by the TS; including TRANSITIVITY as a predictor, which means one also needs to include all its interactions with all other predictors, just makes everything more complex without, in this more programmatic paper, adding to the explanatory value of a section whose main point is the 'multinomiality' of the example. In actual research examples, including the TS' linguistic choice and all its interaction can in fact be useful.

These results are already interesting, but also require more scrutiny than one might guess at first. The first interesting aspect is the central role played by the length predictors, but it is of course not enough to look at variable importance plots – we need to see not just the effect sizes, but also the effect directions, which we can do with partial dependence plots or with descriptive summaries. Consider Figure 2 for a descriptive summary of the effects of PATLENGTH and RECLENGTH; the lengths are on the *x*- and *y*-axes respectively, the plotted numbers and their darkness represent the average logloss value (higher/darker numbers meaning higher logloss/non-canonicalness) and the physical sizes of the numbers represent the numbers of cases with such lengths (and the findings discussed here represent those that would also emerge in more complex ways from partial dependence plots).

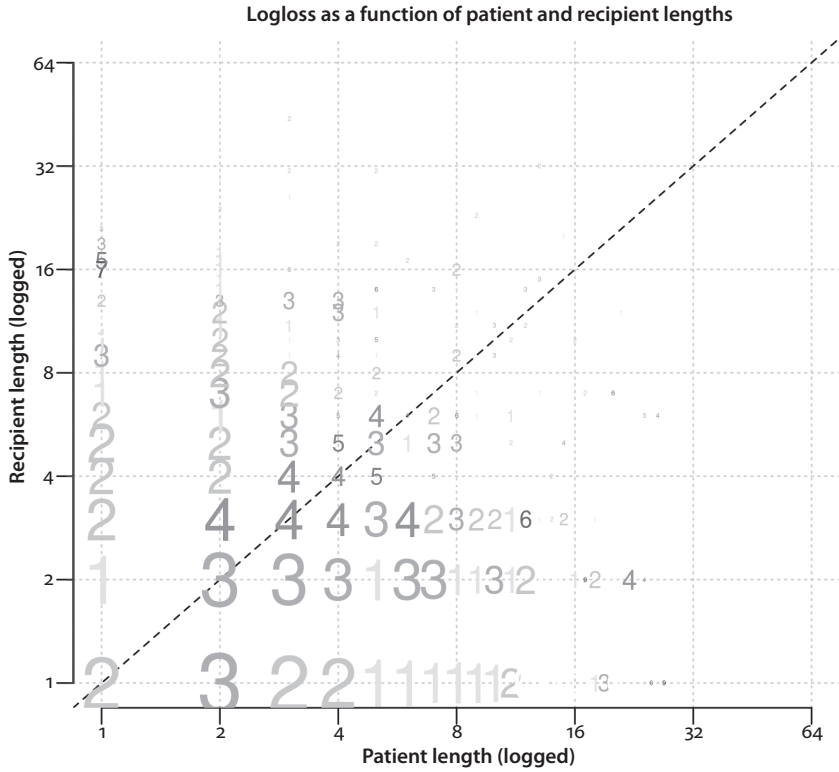

Figure 2. Observed logloss means for each combination of RECLENGTH and PATLENGTH (here unlogged for ease of comprehension)

First, we can see that most patients and recipients are short – no surprise there. Second and more interestingly, we see that TS clearly make least RS-like choices with patients and recipients that are 2 to 4 words long (even though such patients and recipients are quite frequent, as indicated by the physically large numbers in these ranges). The least canonical choices are observed when, in spite of the short and fairly even lengths, the NNS go with the one of the rarer choices (like the passives). However, post-hoc exploration also indicates that the NNS are also responding to the overall short-before-long principle like the RS.

What about other predictors? Space precludes an exhaustive discussion of all findings in this programmatic context, but since random forest studies hardly ever discuss interactions, it is instructive to briefly demonstrate how to work with the interaction predictors. Let us begin with partial dependence plots for the main effects and the interaction of RECPRONOMINALITY and PATSEMANTICS as shown in Figure 3. These plots are variants of the regular plots returned by randomForest::partialplot customized such that:

– the *y*-axis values and the darkness of the bars indicate the effect strength (i.e. the degree to which the relevant represented predictor increases the predicted logloss values; thus, the heights of the bars are similar to what in a regression model would be the coefficient of a predictor);
– the bar width reflects the frequency of the relevant level(s) in the data (wider bars indicate more frequent levels).

Recall from Figure 1 that this interaction predictor of RECPRONOMINALITY:PATSEMANTICS came with a quite high variable importance, which is what studies usually base their interpretation on. However, Figure 3 indicates that that high importance value is misleading because the lower panel (for the interaction predictor) looks a lot like an additive combination of the two upper panels (the marginal/main effects): The upper panels show that NP recipients are correlated with higher loglosses than pronominal ones and abstract patients are correlated with lower loglosses than concrete and informational ones; the lower panel shows that the interaction predictor merely adds those two effects up in a predictable, additive way, which would correspond to a non-significant interaction in a regression model. For instance, in the lower panel the highest logloss values are associated with concrete NPs, which is what an additive prediction would predict from the upper two main-effect panels. Thus, while the variable importance plot suggests that RECPRONOMINALITY:PATSEMANTICS is important, it is in fact only important because of the two main effects it contains, not because the predictors actually interact. This exemplifies how, counter to what many believe, a high variable importance

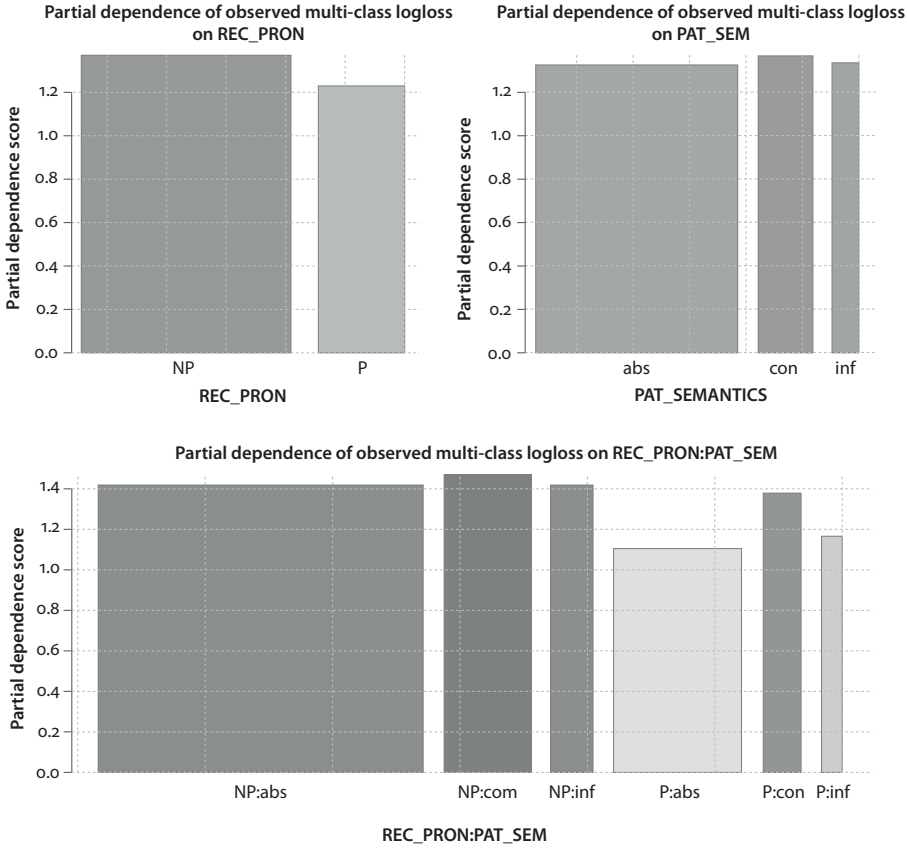

Figure 3. Partial dependence plots for RECPRONOMINALITY, PATSEMANTICS, and their interaction predictor

value of a predictor in a random forest does *not* indicate that that predictor is an important (main) effect per se – a variable importance value of a predictor means that a predictor is important somehow, e.g. as a main effect or because of its interaction with another predictor (see Gries 2020, 2021: Section 7.3).

As it turns out, in this data set there seem to be hardly any relevant interactions – at least among the predictors used here – one at least slightly noteworthy hint of an interaction arises with COUNTRY:PATACCESSIBILITY, which is shown in Figure 4.

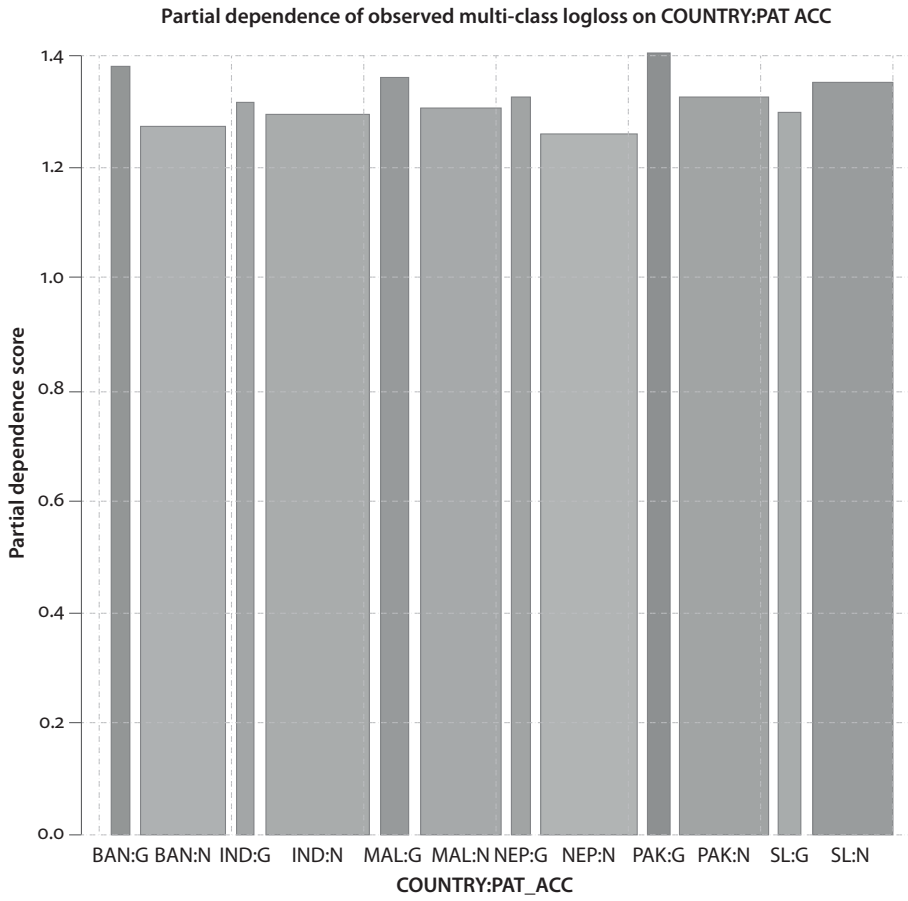

Figure 4. Partial dependence plots for COUNTRY:PATACCESSIBILITY, and their interaction predictor

We can see that in nearly every country loglosses are a bit higher with given than with new patients (supporting the weak main effect of PATACCESSIBILITY) – the only minor exception to this trend are the Sri Lankan English speakers that make slightly more non-canonical choices when the patient is new (see right).

2.5 Interim conclusion

While the current discussion was shorter than it would be in a less programmatic paper, I hope to have shown a few things. First, the MuPDAR(F) approach is indeed straightforwardly extensible to linguistic choices involving more than two alternants and the way that is done best – using (contributions to) multi-class

logloss – also addresses a (potential and purely statistical) shortcoming of previous studies' reliance on DEVIATION as computed from predicted probabilities).

Second, given how random forest results are not really interpretable with regard to whether predictors interact or not, we need to consider including interaction predictors for all combinations of predictors we suspect to interact (just like we would in a regression-based approach – but for random forests this is hardly ever done).

Finally, even when interaction predictors are included, one cannot take their importance values at face value – one needs to compare the interaction predictors to their constituent predictors using either descriptive statistics or partial dependence comparisons. Only then do we gain clarity about whether a variable importance score reflects a main effect, an interaction, or both, which is particularly important for all interaction predictors with L1 (in learner corpus work) or COUNTRY/VARIETY (in variety studies) – if one does not do a some such post hoc exploration of the predictors with high variable importance values, it is extremely likely that one will misinterpret an interaction effect in one's data as a main effect and, thus, overgeneralize.

3. Case study 2: The dative alternation by learners

3.1 Introduction

In this section, I will discuss a possibility to generate predictions for MuPDAR(F)'s step i when the data are not well-suited for the most widely-used analytical approaches/classifiers because of their distributional characteristics, which could include violating assumptions of the methods intended to be used, data sparsity/rarely populated cells, generally small sample sizes, etc. In such situations, the (current) default of especially regression-based, but also tree-based, approaches can be hard to implement properly even for expert users. The point of this section is to propose, exemplify, and evaluate an alternative approach that makes essentially no assumption in terms of the distributions of (combinations of) predictors, or that of the response, the residuals, or similar parameters. I will use as a test case the extremely well-studied example of the dative alternation as exemplified again here in (7).

(7) a. Mr Garibaldi gave his deputy the access code.
    b. Mr Garibaldi gave the access code to his deputy.

Specifically, the data set studied is that analyzed in Gries and Deshors (2020), who searched four corpora for the ten verb lemmas listed in (8) (because Gries and Stefanowitsch (2004) found the verbs in (8a) to prefer the ditransitive, those in (8b) the prepositional dative with *to*, and those in (8c) to have no strong preference for either construction).

(8) a. *give, tell, show, ask*
b. *bring, sell, pass*
c. *send, lend, write*

The corpora Gries and Deshors (2020) searched cover NS and NNS speaking and writing: the Louvain Corpus of Native English Essays (LOCNESS) and the Louvain Corpus of Native English Conversation (LOCNEC; representing NS data) as well as the International Corpus of Learner English (ICLE) and the Louvain International Database of Spoken English Interlanguage (LINDSEI; representing NNS data). As far as NNS are concerned, to arrive at manageable sample sizes, they restricted their attention to learner data representing speakers with Chinese, Germanic (German and Swedish) and Romance (French, Italian, and Spanish) L1s. The results of the concordance were then read to retrieve all and only all instances of the constructional alternation in question; the composition of the resulting sample is represented in Table 3.

Table 3. Sample composition

CX	English	Chinese	Germanic		Romance			
			German	Swedish	French	Italian	Spanish	Totals
Ditransitive	293	205	226	177	194	167	184	1446
Prep. dative	113	221	79	62	113	117	83	788
Totals	406	426	305	239	307	284	239	2234

These instances were then annotated for a small subset of all predictors that are known to be correlated with the dative alternation:

- RECANIMACY and PATANIMACY: the (degree of) animacy of the recipient and the patient: *humanimate* vs. *animate* vs. *inanimate*; the level of *animate* was applied to NPs such as *society, families, the public, the next generation, our country*, etc., but no particular theoretical commitment is implied here;
- RECLENGTH and PATLENGTH: the lengths of the recipient/patient in characters logged to the base of 2;
- LENGTHDIFF: the length difference between logged RECLENGTH and PATLENGTH;
- L1FAMILY, based on the L1s of the NNS as discussed above: *Chinese* vs. *Germanic* vs. *Romance*;
- VERBLEMMA and VERBMATCH for what, in a mixed-effects regression model, would be sources of random-effects variation, especially given the large frequency of *give* in our sample.

3.2 The proposed classifier

As argued in, for instance, Gries (2018), MuPDAR(F) is essentially a missing-data-imputation method: It is used, to use a learner corpus example, to

> generate a native-speaker prediction for every data point in the learner data; […]: we have annotations for each learner choice and we have the learner choices, but what we need is missing, namely for every learner choice what a native speaker would have done and instead of asking several native-speaker annotators to provide that information, we 'ask' a statistical model to provide it instead
>
> (Gries 2018: 299–300)

Dealing with missing data can be done in various ways, as discussed in literature on data mining (e.g. Torgo 2011). The simplest approach that can be applied to rare cases of individual missing data points (other than removing the data) is to replace the missing data point with a measure of central tendency for that column, that speaker, …; this measure could be the mode or an average (Torgo 2011: Section 2.5.2) – clearly, this would not be useful here. Another option is essentially the one that MuPDAR(F) has been using: using correlational structure in the data to make the best 'guesstimate' of what the missing value(s) – the RS choices for the TS data – would be (Torgo 2011: Section 2.5.3), However, as argued above, this can be quite tricky for various distributional reasons. There is another option, however, which is captured well by the title of Section 2.5.4 in Torgo (2011: 60): "filling in the unknown values by exploring similarities between cases", which will be introduced now on the basis of this data, which is a good test case for the classifier to be proposed below given that the sample size, while certainly not bad, is also pushing the limits of minimal sample size recommendations (especially once interactions of predictors within the L1s are included).

This casewise-similarity approach was, here, based on a dissimilarity, or distance, metric that permits the user to compare usage events to each other. The dissimilarity measure computed was the Gower metric, a measure that can compare cases characterized by numeric, ordinal, and categorical predictors (Gower 1971; Podani 1999). Specifically, this measure's distance value for two cases *x* and *y* is based on,

- for numeric variables, the difference between values normalized by their range;
- for ordinal variables, the difference between ranks normalized by their range;
- for categorical variables, whether the value of *x* is equal to that of *y* or not.

As an example for numeric values, consider three people *x, y*, and *z* who are 180, 181, and 179 cm tall respectively and weigh 80, 75, and 80 kg respectively. Then, Gower's dissimilarities between *x* and *y, x* and *z*, and *y* and *z* are 0.75, 0.25, and 1 respectively. The value for *x* and *z* is lowest (i.e. their similarity is highest) because

their height is only 1 cm apart and the weight is the same whereas the value for *y* and *z* is highest (i.e. their similarity is lowest) because their height difference is 2 cm and their weight difference is 5 kg, both of which are the range values of *all* heights and weights.

As an example for ordinal and categorical values, consider three NPs *x, y*, and *z* which have human, animate, and inanimate referents respectively (on a 4-point animacy hierarchy that also includes the level 'abstract') and are all definite (on a 2-point definiteness scale). If we treat animacy as a categorical variable (i.e. as unordered, as virtually all studies using regression models and tree-based approaches have done), then *x, y*, and *z* all score a Gower's distance value of 0.5: the average of 1 from the categorical animacy differences and 0 from the lack of differences regarding the binary definiteness variable. However, if we – more realistically – treat the animacy hierarchy as ordinal, then Gower's distance 'recognizes' that *x* is closer to *y* (distance = $^1/_6$) than to *z* (distance = $^1/_3$) because while both are definite, human (*x*) is closer to animate (*y*) than to inanimate (*z*). Crucially and as this case exemplifies, the Gower metric can return a single value quantifying how different two cases are even if each case contains multiple mixed – numeric, ordinal, and categorical – values.

How is this applied here? In a way reminiscent of leave-one-out validation, a short R script computed for every single case of the TS/NNS data how dissimilar (in terms of the Gower metric) it is to each of all the actual dative alternation choices by the RS/NS; the computation of the Gower metric was based on the predictors RECANIMACY and PATANIMACY, RECLENGTH and PATLENGTH and LENGTHDIFF, and VERBLEMMA as well as VERBMATCH. That means, for each of the 1828 NNS cases, I obtained 406 distance values, namely values that indicate how dissimilar each NNS case is to each of the 406 NS cases.

The next step consisted of computing the mean distance of each NNS case to all 293 NS ditransitives and to all 113 NS prepositional datives. For example, the first NNS case's mean Gower distance to the 293 NS ditransitives is 0.2347 while the same case's mean Gower distance to the 113 NS prepositional datives is 0.3275 (rounded to four decimals). This means that the situation that the NNS is in right now, as he has to make a choice for a construction, is more similar to the situations in which NS chose ditransitives than to the situations in which NS chose prepositional datives; in fact, according to a *t*-test, this difference is significant ($p_{\text{corrected for 1828 tests}} < 10^{-6}$) meaning the present approach predicts that the NNS in case 1 should choose a ditransitive (which the NNS did in fact choose), and analogous computations were made for all cases.

In a final step, I followed Gries and Deshors (2020) and allowed for an 'either' prediction: If there was no significant difference between a NNS case's similarities

to all NS ditransitives and the same NNS case's similarities to all NS prepositional datives, the present approach returns 'either (is fine)' as a prediction. Both the binary and the ternary predictions were then evaluated in terms of accuracy.

3.3 Results

3.3.1 *Prediction accuracies without and with 'either' cases*

The first relevant result is the confusion matrix relating (i) traditional binary predictions of the two constructions based on the casewise-similarity predictions just discussed to (ii) the actual observations, which is shown in Table 4: The prediction accuracy is high (at $^{(1103+482)}/_{1828} = 0.8671$), which is also significantly higher than the no-information-rate baseline of $^{1153}/_{1828} = 0.6307$ ($p_{\text{exact binomial test}} < 10^{-112}$).

Table 4. Confusion matrix for binary ditr.-vs.-prep.-dative predictions (italicized = correct predictions)

Similarity-based preds.	Pred.: ditransitive	Pred.: prep. dative	Totals
obs.: ditransitive	*1103*	50	1153
obs.: prep. dative	193	*482*	675
Totals	1296	532	1828

Table 5. Confusion matrix for ternary ditr.-vs.-either-vs.-prep.-dative predictions (italicized = 'correct' predictions)

Similarity-based predictions	Pred.: ditransitive	Pred.: either	Pred.: prep. dative	Totals
obs.: ditransitive	*995*	*154*	4	1153
obs.: prep. dative	53	*397*	*225*	675
Totals	1048	551	229	1828

What about the confusion matrix that includes the 'either' predictions? Consider Table 5. As is obvious, there are quite a few cases (551) that this approach considers 'either' cases, i.e. cases where a NS should be accepting of either construction. Since these 'either' cases are the ones that boost the approach's accuracy score (to 0.9688), it is important to establish that they are in fact realistic 'either' cases and that this is not just a statistical trick to boost accuracy. Therefore, in (9), I list a variety of examples that, according to the present approach should be acceptable both with the construction they were produced with and the alternative option; arguably, that is indeed the case.

(9) a. Technology has given a great number of advantages to man's life
b. this love gives a certain unity to the book
c. he gives more importance to quantity than to quality
d. Tony, making a joke, give some misdirections to Marlow and Hastings
e. This intervention of women in the stage gave the theatre a more sexual sense of acting
f. when she shows the results to her friends
g. she shows the picture to her friends

The distribution of cases is then also represented in Figure 5: the *x*-axis represents the mean distances of corpus examples from RS/NS ditransitives whereas the *y*-axis represents the mean distances of corpus examples from RS/NS prepositional datives. This in turn means that points close(r) to the upper left corner are cases predicted to be ditransitives while points close(r) to the lower right corner are cases predicted to be prepositional datives. The point characters – *d* or *p* – indicate the construction the TS/NNS chose while the colors indicate the prediction for each TS/NNS utterance (as per the legend in the top left corner); that means that red *p*s and blue *d*s are cases where the TS/NNS made non-canonical choices.

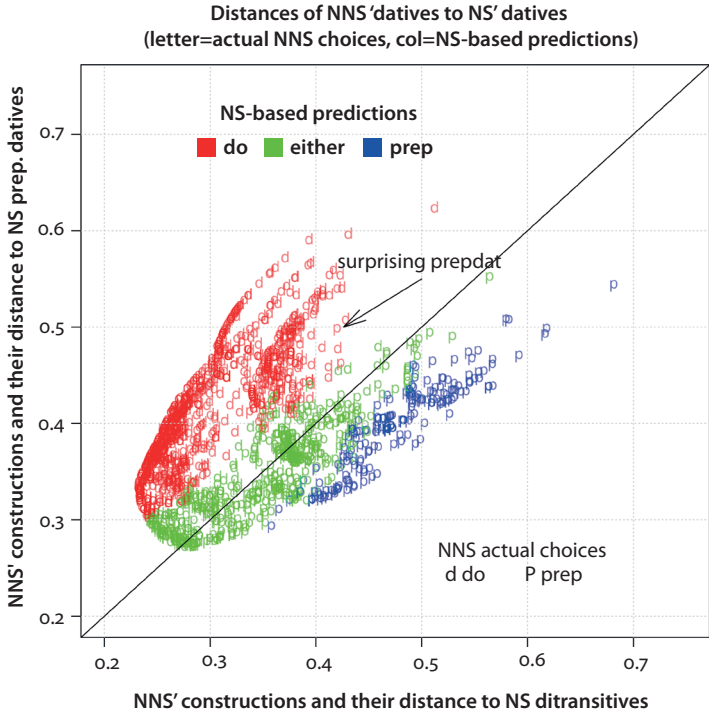

Figure 5. The distribution of NNS constructions based on their median similarities to NS ditransitives (*x*-axis) and NS prepositional datives (*y*-axis)

The results reflect, or are compatible with, the logic of this approach in several ways: (i) Obviously, the (green) 'either' predictions are mostly around the main diagonal where NNS utterances are similarly different from both NS ditransitives and NS prepositional datives; after all, this is a design feature of the approach. (ii) As we move away from the main diagonal, (red) ditransitives and (blue) prepositional datives are predicted more. In particular the blue predictions are nearly all correct (as we expected to see from Table 5) whereas among the (many more) red predictions, one particularly unexpected prepositional dative is highlighted in the plot.

3.3.2 *Comparison and validation*

Before concluding this section, it is prudent to compare this approach to alternatives that a reader might think of at this point. First, how does this prediction-based-on-similarities approach, essentially a *k*-nearest neighbor approach with $k = n_{\text{RS cases}}$) compare against the currently 'conventional' choice of researchers who feel that they or their data cannot handle regressions, i.e. random forests. If one trains a random forest on the RS/NS data and makes corresponding (binary) predictions for the TS/NNS data, one obtains the confusion matrix in Table 6.

Table 6. Confusion matrix from the similarity-based approach for binary ditr.-vs.-prep.-dative predictions (italicized = correct predictions)

Random forest predictions	Pred.: ditransitive	Pred.: prep. dative	Totals
obs.: ditransitive	*1071*	82	1153
obs.: prep. dative	163	*512*	675
Totals	1234	594	1828

This random forest result, too, has a high prediction accuracy (at ${}^{(1071+512)}/_{1828} = 0.866$), but while this is significantly higher than the no-information-rate baseline of ${}^{1153}/_{1828} = 0.6307$ ($p_{\text{exact binomial test}} < 10^{-111}$), it is actually a tiny bit less successful than that of the similarity-based approach mentioned above (which was 0.8671). In other words, the above similarity-based approach outlined above (non-significantly) outperforms random forests here while still being conceptually simpler: The similarity-based approach requires fewer steps, no hyperparameter-tuning at all (at least as introduced so far), no random sampling of cases or predictors, no amalgamation of potentially thousands of forests, yet it still produces prediction, not classification, accuracies right away even without cross-validation.

Second, some readers may also note some similarity to the notion of memory-based learning (see, e.g. Daelemans & van den Bosch 2005) and the corresponding TiMBL implementation (see Daelemans et al. 2018, <https://languagemachines.github.io/timbl/>). If that tool (Version 6.4.8-1) is applied to the current RS/NS data to make predictions for the current TS/NNS data (with all default settings),

we obtain the confusion matrix shown in Table 7, which comes with a prediction accuracy of 0.7861, i.e. (significantly, $p_{\text{exact binomial test}} < 10^{-139}$) lower than both the random forest and the similarity-based variant proposed here.

Table 7. Confusion matrix from TiMBL for binary ditr.-vs.-prep.-dative predictions (italicized = correct predictions)

TiMBLpredictions	Pred.: ditransitive	Pred.: prep. dative	Totals
obs.: ditransitive	*1050*	103	1153
obs.: prep. dative	288	*387*	675
Totals	1338	490	1828

The fact that this result is significantly worse than both others must *not* be understood as a verdict on TiMBL in general, which has many different metrics and parameters other than the default ones used here. The strikingly lower accuracy in this particular case is in part due to the fact that TiMBL's basic similarity metric does not distinguish categorical and ordinal variable (although changing my algorithm to this as well reduces accuracy only by 1.25%), but other differences between basic TiMBL and the above that are probably more relevant are that (i) basic TiMBL uses only the one nearest neighbor (rather than a mean of all) (ii) and how those are weighted in TiMBL's default (rather than them not being weighted).

However, the main point for the present discussion is a that similarity-based approach like the one outlined above, in spite of its relative simplicity relative to random forests and TiMBL, can generate very good prediction accuracies without any distributional assumptions and even for data that would pose problems for other models/classifiers; also, note that, in the spirit of Section 2, this approach can obviously handle linguistic choices with more than two options as well.

4. Concluding remarks

As mentioned at the beginning, the main point of this paper is programmatic; it is a follow-up to Gries and Deshors (2020) in how it, too, proposes answers to questions that readers and listeners of MuPDAR(F) papers and presentations and reviewers have been asking. Accordingly, Section 2 outlined a way in which the approach can be extended/improved in a way that not only handles alternation phenomena with more than two alternants straightforwardly but does so in a way – multi-class log loss – that is more precise than a binary canonicalness variable would be while also addressing distributional problems that could at least potentially arise from the DEVIATION scores used in previous work. The approach was shown to have,

in this example at least, excellent variance explanation. In addition, Section 2 also exemplified the use and interpretation of interaction predictors in random forests.

Section 3, on the other hand, proposed and evaluated a very different way of generating the what-would-the-reference-speaker-do predictions for data sets for which other models/classifiers, in particular regression models, might be less appropriate – a casewise-similarity approach that is a specific variant of a $k = n$ nearest-neighbor approach whose main metric permits the user to not only make predictions for otherwise problematic data sets, but also does so by doing something most analyses to data are not doing, namely respecting the nature of ordinal predictors (such as animacy or complexity scales, proficiency bins, etc. etc.). The approach's predictive accuracy outperformed a random forest (ever so slightly) and the default settings of TiMBL (more significantly), making this appear to be a nice alternative on the grounds of statistical performance alone. However, it is worth pointing also to the theoretical advantages of an approach such as memory-based learning. At a time when much work in cognitive and psycholinguistics is

- embracing usage- and exemplar-based models according to which categorization and processing are related to the similarity/distance of exemplars in a multidimensional exemplar space;
- discussing the degree of cognitive realism of our statistical tools (Baayen & Ramscar 2015; Divjak et al. 2016; Milin et al. 2016; Klavan & Divjak 2016),

an approach like this should be of interest also to those corpus-linguistic research areas that have not yet taken too much notice of this kind of work, and I think it is fair to say that both learner corpus research and varieties research are among those. I therefore hope, minimally, that this paper provided the kinds of improvements to MuPDAR(F) that scholars in those two areas have been asking about and, ideally that the casewise-similarity approach and its relation to memory-based learning can inspire new ideas and developments in those two fields.

References

Baayen, R. Harald & Ramscar, Michael. 2015. Abstraction, storage, and native discriminative learning. In *Handbook of Cognitive Linguistics*, Ewa Dąbrowska & Dagmar S. Divjak (eds), 100–120. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110292022-006

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

Boulesteix, Anne-Laure, Janitza, Silke, Hapfelmeier, Alexander, Van Steen, Kristel & Strobl, Carolin. 2015. Letter to the editor: On the term 'interaction' and related phrases in the literature on Random Forests. *Briefings in Bioinformatics* 16(2): 338–345. https://doi.org/10.1093/bib/bbu012

Daelemans, Walter & van den Bosch, Antal. 2005. *Memory-Based Language Processing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486579

Daelemans, Walter, Zavrel, Jakub, van der Sloot, Ko, & van den Bosch, Antal. 2018. TiMBL: Tilburg Memory-Based Learner. Version 6.4 Reference Guide. *ILK Technical Report – ILK 11–01*. <https://github.com/LanguageMachines/timbl/raw/master/docs/Timbl_6.4_Manual.pdf> (4 April 2022).

Deshors, Sandra C. 2020. English as a Lingua Franca: A random forests approach to particle placement in multi-participant interactions. *International Journal of Applied Linguistics* 30(2): 214–231. https://doi.org/10.1111/ijal.12275

Deshors, Sandra C. To appear. Contextualizing past tenses in L2: Combined effects and interactions in the present perfect vs. simple past alternation. *Applied Linguistics*. https://doi.org/10.1093/applin/amaa017

Deshors, Sandra C. & Gries, Stefan Th. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis to ing vs. to complements. *International Journal of Corpus Linguistics* 21(2): 192–218.

Deshors, Sandra C. & Gries, Stefan Th. 2020. Mandative subjunctive vs. *should* in world Englishes: A new take on an old alternation. *Corpora* 15(2): 213–241. https://doi.org/10.3366/cor.2020.0195

Divjak, Dagmar S., Arppe, Antti & Dąbrowska, Ewa. 2016. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1): 1–34. https://doi.org/10.1515/cog-2015-0101

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4): 857–871. https://doi.org/10.2307/2528823

Gries, Stefan Th. 2020. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory* 16(3): 517–647. https://doi.org/10.1515/cllt-2018-0078

Gries, Stefan Th. 2021. *Statistics For Linguistics with R*, 3rd rev. and ext. edn. Berlin: De Gruyter. https://doi.org/10.1515/9783110718256

Gries, Stefan Th. & Adelman, Allison S. 2014. Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*, Jesús Romero-Trillo (ed.), 35–54. Cham: Springer. https://doi.org/10.1007/978-3-319-06007-1_3

Gries, Stefan Th. & Deshors, Sandra C. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1): 109–136. https://doi.org/10.3366/cor.2014.0053

Gries, Stefan Th. & Deshors, Sandra C. 2020. There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal* 44: 69–96. https://doi.org/10.2478/icame-2020-0003

Heller, Benedikt, Bernaisch, Tobias, & Gries, Stefan Th. 2017. Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME Journal* 41: 111–144. https://doi.org/10.1515/icame-2017-0005

Klavan, Jane & Divjak, Dagmar S. 2016. The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica* 50(2): 355–384. https://doi.org/10.1515/flin-2016-0014

Kolbe-Hanna, Daniela & Baldus, Lina. 2018. The choice between *-ing* and *to* complement clauses in English as first, second and foreign language. Paper presented at ICAME 39, University of Tampere.

Kruger, Haidee & De Sutter, Gert. 2018. Alternation in contact and non-contact varieties: Reconceptualising *that*-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior* 1(2): 251–290. https://doi.org/10.1075/tcb.00011.kru

Lester, Nicholas A. 2019. *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research* 5(1): 1–32. https://doi.org/10.1075/ijlcr.17013.les

Milin, Petar, Divjak, Dagmar S., Dimitrijević, Strahinja & Baayen, R. Harald. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4): 507–526. https://doi.org/10.1515/cog-2016-0055

Podani, János. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon* 48: 331–340. https://doi.org/10.2307/1224438

Schweinberger, Martin. 2020. A corpus-based analysis of differences in the use of very for adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research* 6(2): 163–192. https://doi.org/10.1075/ijlcr.20011.sch

Torgo, Luis. 2011. *Data Mining with R: Learning with Case Studies*. Boca Raton FL: Chapman & Hall/CRC. https://doi.org/10.1201/9780429292859

Werner, Valentin, Fuchs, Robert, & Götz, Sandra. 2020. L1 influence vs. universal mechanisms: An SLA-driven corpus study on temporal expression. In *Learner Corpora and Second Language Acquisition Research*, Bert Le Bruyn & Magali Paquot (eds), 39–66. Cambridge: CUP. https://doi.org/10.1017/9781108674577.004

Wright, Marvin N., Ziegler, Andreas, & König, Inke R. 2016. Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17(145). https://doi.org/10.1186/s12859-016-0995-8

Wulff, Stefanie & Gries, Stefan Th. 2015. Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism* 5(1): 122–150. https://doi.org/10.1075/lab.5.1.05wul

Wulff, Stefanie & Gries, Stefan Th. 2019. Particle placement in learner English: Measuring effects of context, first language, and individual variation. *Language Learning* 69(4): 873–910. https://doi.org/10.1111/lang.12354

Wulff, Stefanie & Gries, Stefan Th. 2020. Explaining individual variation in learner corpus research: Some methodological suggestions. In *Learner Corpora and Second Language Acquisition Research*, Bert Le Bruyn & Magali Paquot (eds), 191–213. Cambridge: CUP.