sciendo

ICAME Journal, Volume 44, 2020, DOI: 10.2478/icame-2020-0003

There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies

Stefan Th. Gries
University of California, Santa Barbara and Justus Liebig University Giessen
Sandra C. Deshors, Michigan State University

Abstract

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 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 on the reference speakers (often native speakers (NS) in learner corpus studies or British English speakers in variety studies), then, secondly, using this model to predict what such a reference speaker would produce in the situation the target speaker is in (often non-native speakers (NNS) or indigenized-variety speakers). Crucially, 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.

Both regression-based modeling in general and MuPDAR in particular have led to many interesting results, but we want to propose two changes in perspective on the results they produce. First, we want to focus attention on the middle ground of the prediction space, i.e. the predictions of a regression/classifier that, essentially, are made non-confidently and translate into a statement such as 'in this context, both/all alternants would be fine'. Second, we want to make a plug for a greater attention to misclassifications/-predictions and propose a method to identify those as well as discuss what we can learn from studying them. We exemplify our two suggestions based on a brief case study, namely the dative alternation in native and learner corpus data.

1 Introduction

1.1 General introduction

Corpus-based studies of learner language and English varieties especially have become more quantitative in nature over the last ten years and are increasingly using regression-based methods and classifiers. Many of these studies model a certain linguistic choice (often a binary variable such as ordering$_1$ vs. ordering$_2$, or construction$_1$ vs. construction$_2$) on the basis of a variety of linguistic and/or contextual predictors X_1, X_2, …, X_n such that

– in learner corpus research, they also involve an additional predictor **L1** (i.e., learners' native language) that can interact with $X_{1\text{-}n}$;

– in corpus-based variety research, they also involve an additional predictor VARIETY (e.g. British English vs. American English vs. Indian English vs. Hong Kong English and/or many others) that can interact with $X_{1\text{-}n}$.

This approach allows one to see whether X_1, X_2, …, X_n behave differently depending on L1 or VARIETY (see Gries and Deshors 2014: Section 3), which is often the main question of such studies (even if this is often not discussed using the above kind of statistical-interaction terminology).

As mentioned above, such studies are becoming increasingly widespread; examples include but are not limited to Deshors (2014, 2018), Wulff et al. (2014), Heller et al. (2017), Szmrecsanyi et al. (2017). In spite of the fact that these studies are a huge improvement over decades of monofactorial chi-squared or loglikelihood ratio 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.

Consider as an example a study of the genitive alternation, i.e. the choice of *of*- vs. *s*-genitives, in, say a learner corpus research study. Such a study is very likely to include as a predictor the variable POSSESSORANIMACY, which has been shown to be highly correlated with genitive choice (animate possessors preferring *s*-genitives and abstract possessors preferring *of*-genitives). In, say, a binary logistic regression (fixed- or mixed-effects), there might be an interaction (i.e., a combined effect) between POSSESSORANIMACY and VARIETY, such that the model predicts probabilities of *s*-genitives 0.85 and 0.75 in native speaker data and learner data respectively. This 0.1 difference in predicted probabilities might be significant but it also means that the model actually makes the same categorical prediction for the relevant cases: The predicted probabilities of *s*-genitives (0.85 and 0.75) are higher than the corresponding ones of *of*-genitives (1-0.85=0.15 and 1-0.75=0.25 respectively) so in both cases the model will predict an *s*-geni-

tive. Depending on one's question, the 0.1 difference may still be theoretically or conceptually interesting, but it might also not be, given that the model's predictions for both cases are the same.

The recent methodological development offering a different perspective on this issue is the so-called MuPDAR approach, a multi-step technique that has mostly involved regression modeling (as per its name), but recently also random forests. Conceptually, the MuPDAR approach is based on missing-data imputation, i.e., the kinds of methods used in other disciplines to guess what a missing data point would have been, had it been provided (in the experiment or the survey). The missing data that are imputed in MuPDAR are,

– in learner corpus research, native speaker (NS) judgments of non-native speaker (NNS) language use;

– in corpus-based variety research, reference-variety speaker judgments of, say, outer-circle variety speaker use.

That is, and to stick with the example of the genitive alternation, the MuPDAR approach consists of applying a first regression (or some (machine-learning) classifier) to only the reference speakers, i.e. the native speakers or the reference-variety speakers (e.g. British English speakers). If that first model/classifier works well enough, then it is used to impute for each non-native speaker's or each variety speaker's actual genitive in the data what a native speaker or a reference-variety speaker would have said in the exact same linguistic context. Thus, if a NNS used an *s*-genitive in a corpus example and a NS speaker is predicted to have used one as well, then the MuPDAR approach would consider the NNS choice 'nativelike'. If, on the other hand, a NNS used an *s*-genitive in a corpus example but a NS speaker is predicted to have used an *of*-genitive instead, then the MuPDAR approach would consider the NNS choice 'non-nativelike'.

What is done with all these 'nativelike' vs. 'non-nativelike' decisions/labels? In MuPDAR kind of analyses, these labels often become the dependent/response variable of a second regression/random forest to determine under which conditions NNS/variety speakers make choices that deviate from NS/reference-variety speaker choices. In other words, all the predictors that were used in the first regression/random forest (and sometimes more, such as the L1 of the NNS) are now used to predict nativelike vs. non-nativelike choices, which is just technical-sounding language for 'what NNS still have difficulties with' (given that, in conditions *A*, *B*, *C*, …, they are still making non-nativelike choices). Now it is true that a traditional one-regression kind of approach would also help

explore, say, NS vs. NNS differences, but one advantage of the MuPDAR approach is precisely that its design focuses specifically on the cases where NS and NNS actually make different choices, and not on the cases where a regression coefficient (i.e. the coefficient that indicates how much a variable (level or increase) changes the predicted outcome) may be significant even if it leads to the same predictions for NS and NNS alike.[1]

MuPDAR 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 Japanese, Wulff and Gries (2015) on prenominal adjective order in English, Deshors and Gries (2016) and Kolbe-Hanna and Baldus on *ing* vs. *to*-complements, Heller, Bernaisch and Gries (2017) on the genitive alternation in British vs. Singaporean English, Wulff and Gries (2019) on particle placement, Wulff and Gries (to appear) on genitives in learner data, Kruger and De Sutter (2018) and Lester (2019) on *that*-omission, Werner, Fuchs, and Götz (to appear) on present perfect vs. simple past choices, etc. However, in this paper we want to discuss two areas in which we see room for improvement of both MuPDAR in particular and general regression/classifier approaches in corpus linguistics in general.

1.2 Room for improvement 1: The middle ground

Ever since MuPDAR was developed, we were aware of one potential shortcoming, namely that even though constructional choices by the NNS are categorized as nativelike or not, the approach has no mechanism to state 'in this context, a NS would be fine with either constructional choice' – there is not just one correct choice, and this is a question MuPDAR practitioners have encountered sometimes at conferences. Specifically, the so far most frequent implementation of MuPDAR at least proceeds such that

– if a NS is predicted to use an *of*-genitive with 51 per cent and if a NNS uses an *s*-genitive, the NNS choice is considered non-nativelike, just like
– if a NS is predicted to use an *of*-genitive with 91 per cent and if a NNS uses an *s*-genitive, which would also be considered a non-nativelike choice.

A first extension of MuPDAR already discussed in the first MuPDAR publication (Gries and Deshors 2014:128) is already able to handle this situation better because, while it still focuses on the cases where the choices of the NNS/variety speakers differ from those predicted for the NS/reference variety speakers, it *also* considers the severity of the deviations between NNS and NS choices:

– if a NNS made the choice a NNS is predicted to make, then a so-called DEVIATION score is set to 0;

– if a NS is predicted to use an *of*-genitive with 51 per cent and if a NNS uses an *s*-genitive, this is considered non-nativelike, but with a very small DEVIATION score of just 0.51-0.5=0.01; however,

– if a NS is predicted to use an *of*-genitive with 91 per cent and if a NNS uses an *s*-genitive, this is non-nativelike with a much larger DEVIATION score of 0.91-0.5=0.41.

In other words, these two cases, where a NNS made non-nativelike choices, are still identified as such, but this approach also quantified the degree to which they are non-nativelike (0.01 vs. 0.41); readers with some expertise in machine-learning methods will recognize that the DEVIATION score is, in some sense, a signed, but less punitive, 'version' of the classification measure of log loss, which is computed as, in pseudocode, `if prediction is correct, -log (predicted probability); if it is not, -log(1-predicted probability);` in other words, log loss (just like the Brier score) penalizes confident false classifications/predictions, but penalizes false classifications/predictions that are made confidently/boldly more; in other words, in terms of log loss, the worst predictions are confident/bold predictions that turn out to be wrong.

Arguably, this is not just an issue for MuPDAR but at least in part also one of traditional one-step regression analyses of such kinds of alternation data. Consider a situation in which a NNS has a slight preference for, and thus would produce, an *s*-genitive (i.e. imagine the predicted probability of an *s*-genitive is 0.55) but where a NS has a slight preference for, and thus would produce an *of*-genitive (i.e. imagine the predicted probability of an *s*-genitive is 0.45). This is how such a case would be considered in the three different approaches outlined so far:

– in a traditional one-step regression analysis, this case might be considered interesting because it would increment the count of misclassified cases: For this case, the model would, from what NS are doing, expect an *of*-genitive, but the NNS picked an *s*-genitive, making this an instance counting against precision and recall (and contribute a value of 0.597837 to the overall computation of log loss);

– in the more common MuPDAR analysis, this case would be interesting because it would be a case classified as non-nativelike and, thus, become a case of interest for the second regression model/classifier;

– in the more precise MuPDAR approach using DEVIATION (or log loss), this case would also be interesting because it would be a case classified as non-nativelike, but less so, because the DEVIATION (or log loss) score is relatively low/close to 0.

What all of these approaches, by now common methods in corpus linguistics, in a sense, are not concerning themselves with, however, is considering the possibility that, in certain contexts, really both choices – *of*- and *s*-genitive – are just about equally acceptable or at least so acceptable that a NS would not think twice even if he saw the linguistic choice he might not have made himself. That is, traditional regression methods might too eagerly label a certain choice as 'misclassified by the regression model' whereas MuPDAR might too eagerly label a certain choice as 'not what a reference speaker would have done'.

Taking a step back for a moment, this situation might in part be a result of applying the kinds of methods we are using to the kinds of data we are studying. The situation we have described above, a case where a NS might use construction *X* but would also be perfectly fine with hearing someone else use construction *Y* is a situation that the methods we are using in corpus linguistics – binary logistic/multinomial regression, classification/conditional inference trees, random forests, support vector machines, neural nets, etc. – are not usually used for: When these methods are getting a binary response variable, their point usually is not to return a shrug 'either one's fine I guess' kind of response. And in the other disciplines, from which we borrow these methods, such an 'either/both levels of the dependent variable' kind of response from a classifier is not really an option, in fact completely undesirable (because of the costs that it introduces):

– a credit card transaction is either fraudulent or it is not – a credit card company is not interested in classifying a particular transaction as 'well, could be fraudulent, could be fine' because from that classification no advice follows, namely whether to call the credit card holder to alert him and/or verify the transaction or whether to not intervene lest the customer gets annoyed at too many false alarms;

– person recognition (e.g. using various sensors on a smartphone for unlocking it) is not designed to return as a result 'maybe that's my owner, I better turn on, maybe not, I better stay off' – in fact, if such a result was obtained from, say, an iris or fingerprint scanner, the classifier would probably equate that uncertainty with a negative recognition event lest the device be too accommodating when presented with irises or fingerprints it was actu-

ally not trained on; same with cancer screening, HIV tests, etc. where an 'uncertain' response by a classifier would do little more but impose costs (e.g., for a second test to get a more certain response).

Thus, the contexts from which the kinds of models and classifiers we are currently using a lot are actually not exactly structurally identical to those we are facing. Yes, we, too, have two or sometimes multiple options that we hope a regression/classifier can distinguish well, but unlike in the above examples, our choices are 'less drastically incompatible', so to speak:

– a prediction of 'either one' in the fraudulent credit card transaction scenario makes no sense and given the cost it introduces – e.g. human attention being required to make a disambiguating decision – 'either one' *should* be counted as a misclassified case;

– a prediction of 'either one' in the person recognition scenario makes no sense and given the cost that it introduces – the smartphone's OS prompting the user to enter a pin, which takes time and causes annoyance – 'either one' *should* be counted as a misclassified case.

By contrast, a prediction of 'either genitive would be fine' is not an a priori nonsensical response and if really either genitive was possible, our approaches should be able to say so because that means an 'either' decision by the classifier would *not* be a misclassified case. In fact, this latter case would be just like a human reader or learner corpus error annotator who would accept both *the speech of the President was well received* and *the President's speech was well received*; somewhat funnily, whenever linguists provide minimal pairs as examples to exemplify the alternation they are studying, they – as will we below – usually provide alternants that are acceptable and differ only in the constructional choice. Plus, no one has any problem recognizing this 'either is fine' kind of situations in other contexts, for instance in the lexical domain, as when no reader would rigorously demand 'you can pick only one!' with *the economy experienced* fast *growth* as opposed to *the economy experienced* rapid *growth*. Thus, the middle ground/'either' scenario is known to everyone, just not studied much in its own right. Given the ubiquity of such classification studies but also situations where both/all choices of an alternation are acceptable, we submit that this is an important issue to explore: Our analyses and understanding of such phenomena does not necessarily benefit from our ignoring the middle ground by counting them as 'misclassified' (in the traditional regression) or as a 'deviation from some reference' (in MuPDAR).

1.3 Room for improvement 2: Spectacularly misclassified/mispredicted cases

The second issue we want to discuss is a simpler and more general one. It is concerned with the fact that, in most alternation studies of either the traditional kind or the MuPDAR kind, there is too little exploration of the cases that are misclassified or mispredicted. The typical Results and Discussion sections of studies present one (final) model or classifier, discuss its accuracy and maybe *C*-score, sometimes precision and recall scores, and the significant effects or important variables of the (final) model are discussed – ideally in terms of the predicted values resulting from each predictor of the model while controlling for everything else.

None of this is problematic, but we submit that what would also be interesting but is often not discussed, are the cases where the regression model/classifier made wrong classifications/predictions and maybe particularly so where these classifications are most spectacularly erroneous, i.e. and again using the above example, where the DEVIATION scores are far from 0 because

– a regression model/classifier returns a predicted probability of an *s*-genitive of, say, 0.85 or even much higher, but the actual speaker choice was an *of*-genitive;

– a regression model/classifier returns a predicted probability of an *s*-genitive of, say, 0.15 or even much lower, but the actual speaker choice was still an *s*-genitive.

(While the above bullet points used predicted probabilities / DEVIATION scores, one could of course also look at the log loss scores of these examples and consider every case with a positive log loss score 1.897 (corresponding to wrong predictions with predicted probabilities of p=0.85) to be spectacularly erroneous classification/prediction.) We feel these cases should receive much more attention than they have in the past (including much of our own previous work!) because what better indicator for 'future work' is there than the cases where a model/classifier is convinced a speaker will use *X* only to see the speaker use *Y*? The cases where we are most wrong are those that are most in need of explanation: They are the ones where, e.g.,

– the learner makes the most surprising linguistic choices (from the NS perspective), which might point to where intervention can lead to learning/corrections most efficiently;

– the variety speaker makes the most surprising linguistic choices (from the speaker-of-another-variety perspective), which might point to precisely the

situations where an indigenized variety resists the tendencies of a historical source variety or creates patterns at odds with what a historical source variety is doing;

– the analyst is pointed to missing predictors most clearly: If many such spectacularly erroneous predictions share a feature or a combination of features we are not covering with our current predictors, we probably should include it by adding a new predictor or a new interaction.

So far, it seems as if misclassified cases were never ignored – because they are outside of the main diagonal of a confusion matrix – but this means the misclassified cases are only considered *quantitatively* – what is their impact on accuracy, precision, recall, etc.? – but we are hard-pressed to come up with a study that has a dedicated section to discussing (the most egregiously) misclassified cases *qualitatively*, and we think the field is missing out on the most obvious pointers for future research.

1.4 Overview of the present paper

In this paper, we are exploring first steps towards addressing the above two points. Specifically and with regard to Section 1.2, we are proposing to tweak MuPDAR so that it can return a prediction of 'either', indicating that the NS would accept either construction. That is and as discussed above, we are imputing a human grader or annotator who reads every instance of an alternation and, for each case, makes one of three determinations:

– "perfectly acceptable to me" as in 'yes, I'd say it like that, too';

– "acceptable" as in 'I wouldn't necessarily say that but it's ok (I wouldn't blink if I heard that said to me)';

– "unacceptable, one cannot say that here".

Thus, the application of this revised middle-ground MuPDAR to the above example – NS preference for *s*-genitive = 0.45 but the NNS chose an *of*-genitive – might lead to a classification of this case as an 'either genitive', which would then also mean that the NNS's choice would not be considered this 'non-native-like' (or that a traditional regression approach would not consider this 'misclassified').We then also provide a few first pointers towards how such cases can be explored in more detail.

With regard to Section 1.3, we are exemplifying how one might go about identifying the most egregious misclassified/-predicted cases; conveniently, we

think, the way in which these might be explored can be informed by the same strategies we use to explore the middle-ground cases.

While we think the suggestions to be made below make an important contribution to the field, this paper is largely programmatic. In Section 2, we will begin by discussing the data set we are using to exemplify both the above suggestions and then turn to exploring a revised MuPDAR approach and our current suggestion(s) regarding how to go about identifying the cases that are now to be members of the 'either' category. In Section 3, we will then discuss the exploration of misclassified/mispredicted cases. Finally, Section 4 will discuss some implications of the proposed new analyses and outline what we think are the most obvious future steps.

2 Towards middle ground/either–or predictions

2.1 Data and annotation

Our first case study is concerned with the dative alternation as exemplified in (1), an extremely well-studied alternation that is, therefore, a good test case:

(1) a. Mr Garibaldi gave his deputy the access code.

b. Mr Garibaldi gave the access code to his deputy.

Specifically, to obtain a decent representation of the constructions, we searched through four corpora for the ten verb lemmas listed in (2); this is because Gries and Stefanowitsch (2004) found the verbs in (2) a to prefer the ditransitive, those in (2)b the prepositional dative with *to*, and those in (2)c to have no strong preference for either construction.

(2) a. *give*, *tell*, *show*, *ask*

b. *bring*, *sell*, *pass*

c. *send*, *lend*, *write*

The corpora we searched were intended to cover NS and NNS speech and writing; correspondingly, we searched 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, we restricted our 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 1:

Table 1: Sample composition

			Germanic		Romance			
CX	**English**	**Chinese**	**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 we imply no particular theoretical commitment here;

– RecLength and PatLength: the lengths of the recipient and patient in characters logged to the base of 2;

– LengthDiff: the length difference between RecLength and PatLength: as an ordinal variable with 5 levels that essentially mean 'Rec>>Pat', 'Rec>Pat', 'RecPat', 'Rec<Pat', and 'Rec<<Pat'; this variable was develop to allow for the inclusion of interaction predictors (see below);

– L1Family, based on the L1s of the NNS as discussed above: *Chinese* vs. *Germanic* vs. *Romance*;[2]

– VerbLemma and VerbMatch for what, in a mixed-effects regression model, would be source of random-effects variation, especially given the large frequency of *give* in our sample;

– finally and following Gries's (to appear) recommendation, we also included interaction predictors namely RecAni:PatAni as well as RecAni:LenDiff and PatAni:LenDiff.

2.2 Statistical evaluation

We then proceeded with a random forest on the 406 NS data (with *ntree*=2000 and *mtry*=4) to see whether this forest achieves a satisfactory prediction accuracy to be able to continue with the MuPDAR approach; this is necessary because only if the first model/classifier is good enough does it make sense to use it to impute NS/reference speaker judgments for the NS/variety speakers.

Our application of the random forest with interactions to the NS data led to a good OOB prediction accuracy of 83 per cent and a *C*-score of 0.873, which we deemed appropriate to proceed. We then applied the random forest trained on the NS data to the NNS data to impute the choices NS are most likely to have made in those cases. We did that first in the traditional way, i.e. with a predicted-probability cut-off point of 0.5 as just about everyone has been doing, which led to the usual kind of confusion matrix shown in Table 2:

Table 2: Confusion matrix for the traditional cut-off at 0.5 (bold = accurate/nativelike)

	RF prediction: ditransitive	**RF prediction: prep. dative**	**Totals**
NNS choice: ditransitive	**1059**	94	1153
NNS choice: prep. dative	219	**456**	675
Totals	1278	550	1828

To address the 'both constructions are equally acceptable' situation, we also adopted a second, different approach. Specifically, we retrieved from the random forest object all 2000 prediction votes of each tree of the forest for each of the 1828 NNS choices. For instance, for a certain NNS genitive choice, 1300 of the trees in the random forest might predict an *s*-genitive whereas the remaining 700 trees would then predict an *of*-genitive.

But how is the middle ground defined? This is a question we will return to in more detail below, but in the absence of any criteria for this, we decided to use as a cut-off point the predicted probability that corresponds to one unit of log loss, i.e. to classifications/predictions that, if wrong, lead to log loss values of 1 (according to the formula provided above). This logic and threshold value of 1 log loss unit amounts to the following decision tree:

– if the predicted probability of an *s*-genitive was 0.632121, i.e. if 1265 or more of the trees in the forest predicted an *s*-genitive, then the random forest confidently predicts "*s*-genitive" (-log(1-0.632121) is a tiny bit >1);

– if the predicted probability of an *s*-genitive was 0.36788, i.e. if 735 or fewer of the trees in the forest predicted an *s*-genitive, then the random forest predicts confidently "*of*-genitive";

– if the predicted probability of an *s*-genitive was within the interval [0.36788, 0.63212], then the random forest does not confidently predict one of the two constructions but is interpreted to say "either genitive is fine".

Informally speaking, this logic can be paraphrased as "if 2000 trees in our forest can't make a clear-cut recommendation, both constructions are probably ok", with 'clear-cut' being defined as <1 log loss.

The next obvious step consisted of determining how much the existence of the new 'either' prediction changes the picture in terms of (i) NNS' accuracy and (ii) which instances are now predicted differently. The former can be done by simply cross-tabulating the NNS choices with both the two traditional MuPDAR predictions, which leads to Table 3:

Table 3: Confusion matrix for the new proposal (bold = accurate/nativelike)

	RF prediction: ditransitive	**RF prediction: either**	**RF prediction: prep. dative**	**Totals**
NNS choice: ditransitive	**1024**	**61**	68	1153
NNS choice: prep. dative	204	**138**	**333**	675
Totals	1228	199	401	1828

While the accuracy attained by considering the 'either' cases 'accurate/nativelike' is only higher by 85.12 per cent - 82.88 per cent = 2.24 per cent, according to exact binomial tests, that is in fact a significant increase (*p*s in both directions < 0.0054).[3] However, it is more instructive to determine which instances' classifications now changed in an interesting way – from 'non-nativelike' in the traditional approach to 'nativelike' (because of 'either') in the new approach. For this, consider Table 4:

Table 4: Both confusion matrices combined (bold = where the traditional and the new approach are making the same predictions, italic = 'improved' predictions)

NNS choices	Traditional binary prediction	Ternary pred.: ditransitive	Ternary pred.: either	Ternary pred.: prep. dative	Totals
ditransitive	ditransitive	**1024**	35	0	1059
	prep. dative	0	*26*	**68**	94
prep. dative	ditransitive	**204**	*15*	0	219
	prep. dative	0	123	**333**	456
	Totals	1228	199	401	1828

That is, there are 41 cases of NNS choices that the traditional approach would have labeled as non-nativelike, but that the new ternary approach labels 'native-like' because that approach recognizes that their predicted constructional probabilities are so close to 0.5 that either construction would in fact be acceptable. Which of course raises the obvious question(s): What do these NNS choices look like and are they really cases where both constructions would be acceptable?

2.3 Specific results

2.3.1 A brief qualitative look at a few selected examples

To get a first qualitative impression, consider (3) to (5) for a few examples from these 16 (slightly edited for presentation, but not in pertinent ways), and try to guess which are the actual NNS choices …

(3) a. It is manifest that use of credit cards can bring students many benefits
b. It is manifest that use of credit cards can bring many benefits to students

(4) a. But what gives the novel a certain unity, a certain harmony, is Mrs. Ramsay
b. But what gives a certain unity, a certain harmony, to the novel, is Mrs. Ramsay
c. But what gives a certain unity to the novel, a certain harmony, is Mrs. Ramsay

(5) a. This secret story that brings the pieces of the play together also gives the play its happy ending
b. This secret story that brings the pieces of the play together also gives a happy ending to the play

While we did not provide any context that might sway a reader towards one or the other construction (given the impact that discourse accessibility of the patient and the recipient and focus considerations have on the dative alternation), we submit that none of the examples in (3) to (5) should be considered clearly unacceptable (the actual NNS choices are all the (a) choices), which means that a more nuanced 'either' prediction is more useful here than a categorical prediction that would label some of the examples in (3) to (5) unacceptable.

2.3.2 A quantitative look at all examples

In order to get a clearer picture of how these 41 cases differ from the data as a whole, we compared their characteristics against those of all cases; given the smallness of the number of changed cases, we only did this monofactorially (i.e. we only explored each predictor on its own rather than all of them in concert): For the categorical predictors L1Family, PatAnimacy, RecAnimacy, VerbLemma, and LengthDiff, we computed multinomial tests for goodness-of-fit comparing the observed percentages of these predictors' levels in the 41 changed cases to those expected from the overall frequencies of these levels in all of the data;[4] for the numeric predictors PatLength and RecLength, we computed Kolmogorov-Smirnov tests for goodness-of-fit comparing the length distributions of the 41 cases to those of all 1828.

For L1Family and PatAnimacy, the 41 cases did not differ significantly from the data as a whole (p_{mnt}=0.316 and p_{mnt}=1 respectively). For VerbLemma, RecAnimacy, PatLength, RecLength, and LengthDiff there were indeed significant differences between the 41 cases with 'either' predictions and the data as a whole, which are shown in figures below. Figure 1 shows the results for VerbLemma and, as in all following plots, the overzealously labeled 'non-canonical' NNS uses are represented with black 1s, all NNS data are represented with black 2s, and all NS data are represented with grey 3s.

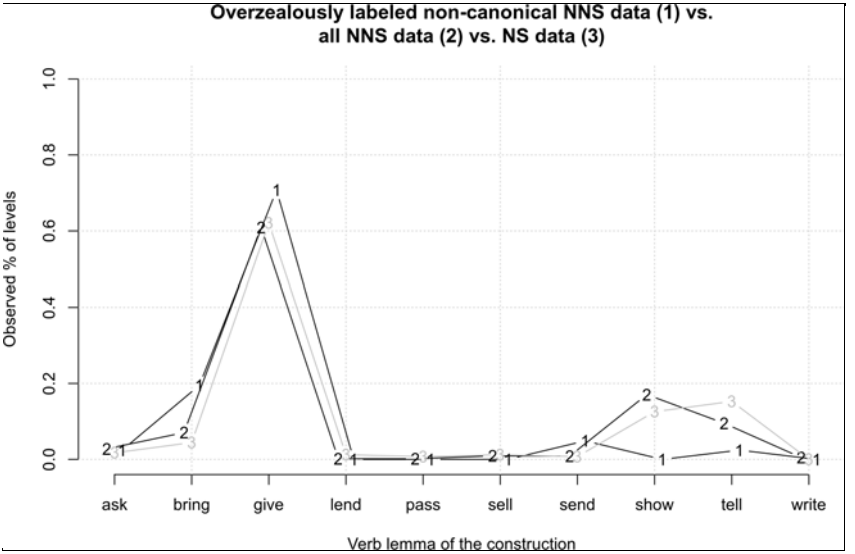


Figure 1: How the 41 re-classified cases (black 1s) differ from other data with regard to VERBLEMMA

It is clear that the NNS' overall frequencies of verb uses are very comparable to those of the NS (see the close fit of the lines of the 2s and the 3s). However and more importantly, it also shows that there is a difference between all NNS uses and the constructions predicted only non-confidently by the random forest. i.e. the ones we submit are cases where in fact either construction would be acceptable; that difference is weakly significant in spite of the small number of tokens ($p_{\text{mnt-sim}}$=0.0215). The verb lemmas *bring* and *give* are responsible for most of the non-confident predictions that would have been considered non-nativelike overzealously whereas *show* and *tell* are underrepresented, so to speak, among the 'either' cases. That of course also means that *show* and *tell* are particularly represented in cases with confident predictions (right or wrong ones). Given the small number of cases, it is hard to speculate on why these four verbs are noteworthy in this way, but a look at the actual contexts suggests that this might have to do with the (kinds and numbers of) senses the verbs have in these cases. For instance, relatively few of the instances of *give* actually involve the prototypical transfer scenario that is commonly so strongly associated with *give* – instead,

many examples with *give* involve expressions such as *giving* [humans] *a certain dimension*, *a new future*, *a bad name*, …, and the situation is similar with *bring*. In other words, these are cases that do not combine all of the usual indicators pointing to a ditransitive construction, which is what makes them more likely to be classified as non-canonical overzealously.

As for RecAnimacy, Figure 2 again shows NNS' recipient animacies are distributed much like those of the NS, but that the non-confident predictions of the cases previously considered non-nativelike are mostly non-human recipients (p_{mnt}<0.0001).

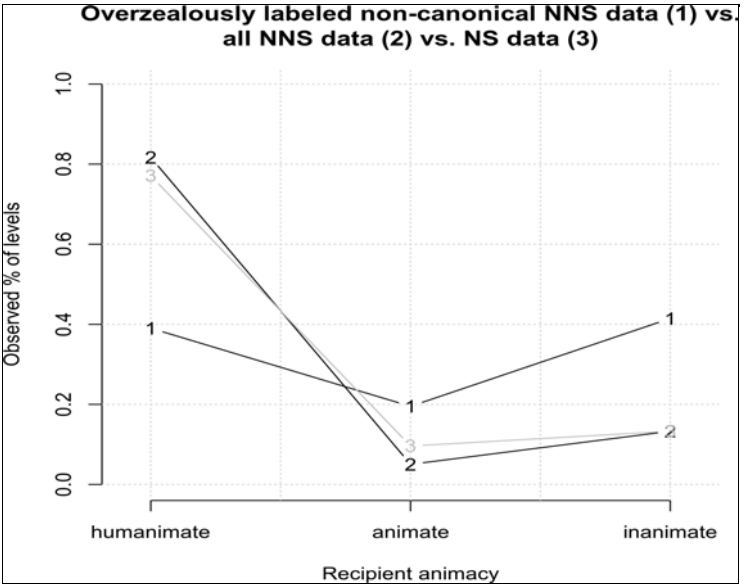
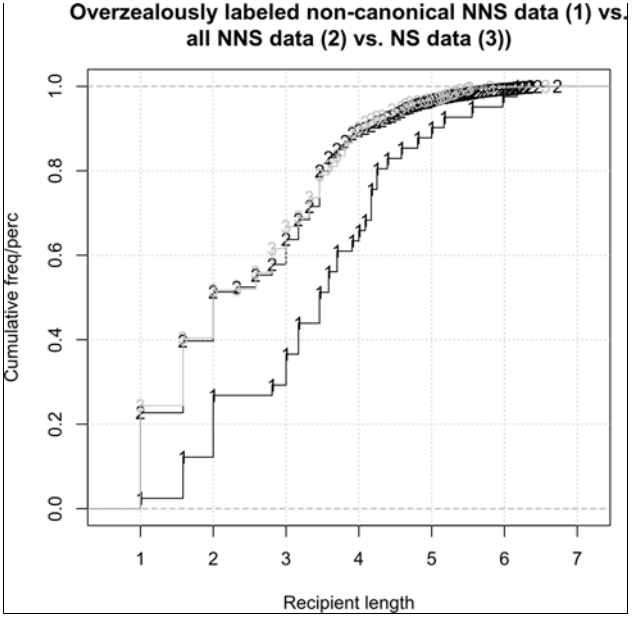


Figure 2: How the 41 re-classified cases (black 1s) differ from other data with regard to RecAnimacy

For the numeric variables, we are showing results using ecdf plots. Such (empirical cumulative distribution function) plots show the numeric variable being studied on the *x*-axis and use a line representing cumulative percentage distribution as *y*-axis values. That means, each point at some *x*-*y*-coordinate answers the following question: "How much in percent of the data (= the *y*-axis value) is smaller than or equal to this value on the *x*-axis?" For instance, the second 1

from the left in the upper panel indicates that 12.2 per cent of all overzealously labeled non-canonical recipients have a logged length of 3 (antilogged 1.58) or less.

With regard to the (logged) recipient and patient lengths, the results in Figure 3 show that their lengths in the NNS data as a whole are very similar to those of the NS, but that the non-confident/'either' predictions indicated by the black 1s involve significantly more longer recipients ($p_{KS\text{-}test}$<0.003) and significantly more shorter patients ($p_{KS\text{-}test}$<0.002).



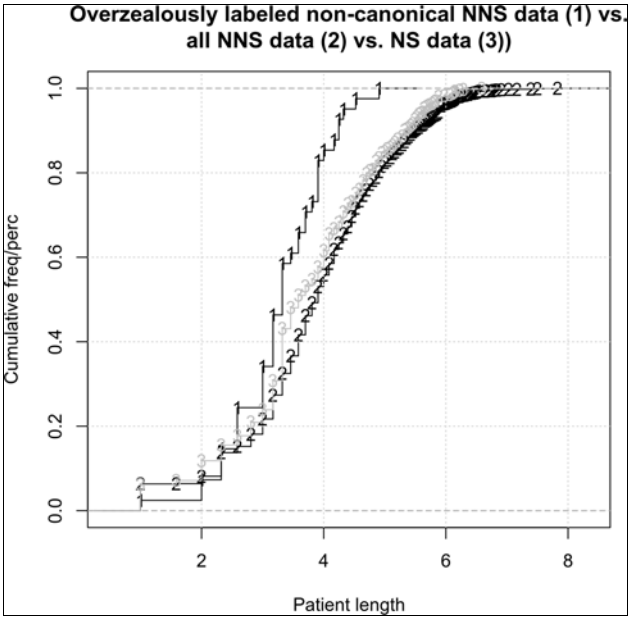


Figure 3: How the 41 re-classified cases (black 1s) differ from other data with regard to RecLength *(upper) and* PatLength *(lower)*

Finally, for LengthDiff, there is again the reassurance that the length differences in NNS and NS data are overall the same and reflect a situation where recipients are on average 2 units shorter than patients. However, we also find that the non-confident predictions are significantly different ($p_{KS\text{-}test}<0.0001$): there are hardly any cases with no length difference between patient and recipient (not the short horizontal red stretch around $x=0$) – instead, the non-confident predictions involve cases where there is a little bit of a length difference, i.e. when the patient is either 2-4 characters longer or shorter than the recipient; see Figure 4.

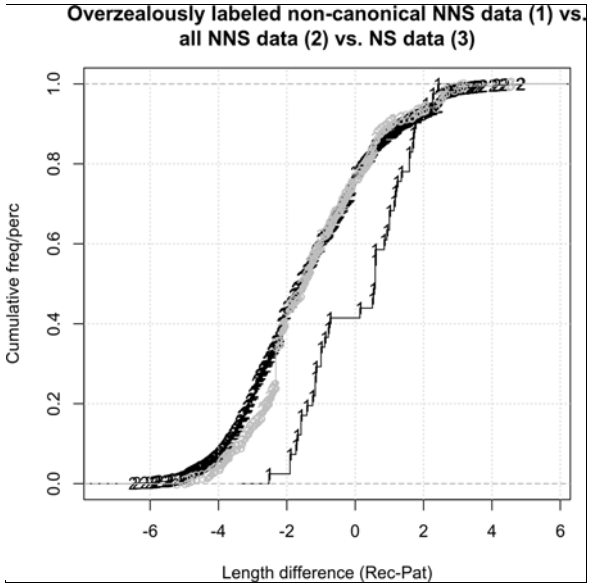


Figure 4: How the 41 re-classified cases (black 1s) differ from other data with regard to LengthDiff

2.4 Interim discussion

We think these results are both reassuring and interesting. They are reassuring because they show clearly that the overall distributions of the predictors we annotated in the NS and the NNS data are very similar, as indicated by the uniformly close fits of the 2- and 3- lines, which renders the data nicely comparable. We cannot help but pointing out that this is not trivial: As far as we can tell, most learner corpus studies or corpus-based variety studies do not actually show that the predictor distributions are similar across the different L1s/varieties they are studying – while that similarity is not a necessary precondition for different kinds of modeling, it is relevant information about one's data. For instance and in general at least, it stands to reason that studies that compare learners from different L1s with each other would benefit from the degree of comparability that would result from the annotated predictors being very similarly distributed in each learner variety much as experimentation usually involves trying to control circumstances – i.e. predictors and controls – as much as possible so that whatever changes are observed in the dependent variable are truly interesting differ-

ences. Inversely, would not any learner corpus researcher comparing learners from L1 *a* and L1 *b* become concerned if already the data from learner groups *a* and *b* are completely different already even before any modeling of, say, a constructional choice is done?

The results are also interesting because they show that the non-confidently predicted examples, the ones that we propose should be considered 'acceptable' even in the not-predicted construction, exhibit a significant pattern on their own that sets them apart from the more confidently predicted clear-cut cases. Even this finding on its own is already interesting because it means that the not-so-clear-cut cases are indeed different from the clear-cut ones, which in turn means that they should maybe not simply be lumped together with the clear-cut cases; that in turn is empirical support for our theoretical argument above that, in the kind of corpus research we are concerned with here, recognizing the middle-ground cases and treating them separately from the clear-cut cases is justifiable and instructive.

3 Towards the worst classifications/predictions

Now what about the most egregious misclassifications/predictions? To first determine the set of instances to look at, we cross-tabulated the (floors of the) log loss values with whether the prediction of the random forest is correct (TRUE) or not (FALSE);[5] the result is shown in Table 5:

Table 5: Log loss floors against prediction correctness

Log loss floors	0	1	2	3	4	5	7	34	Totals
FALSE	41	102	41	3	2	24	82	18	313
TRUE	1515	0	0	0	0	0	0	0	1515
Totals	1556	102	41	3	2	24	82	18	1828

Since there is no real precedent for this and 82+18 is such a nice round number, we decided to explore the top 100 most spectacularly mispredicted cases. Conveniently, we can study these 100 badly mispredicted cases in exactly the same way we have before studied the 41 middle-ground cases – the logic is the same: we want to see how a certain group of cases differs from the larger group of all NNS cases that they come from. Thus, we again used multinomial and Kolomogorov-Smirnov tests for goodness-of-fit, but add L1Family and the chosen construction CX as variables. To save space and given the extremeness of at

least some results to be discussed presently, we do not provide a lengthier discussion and plots as before but just summarize the main findings in a listwise fashion.

The 100 worst predictions as defined here differ from all NNS data; *all 100* involve

- NNS using a prepositional dative when the random forest was *very* certain a NS would have used a ditransitive (for CX, $p_{mnt}<10^{-40}$);
- humanimate recipients (for RecAnimacy, $p_{mnt}<0.001$);
- inanimate patients (for PatAnimacy, $p_{mnt}=1$).

In addition, the worst predictions involve

- nearly always (99 cases) the verb *give*, plus 1 case of *show* (for VerbLemma, $p_{mnt\text{-}sim}=0.0003$), which means that none of the verbs other than *give* is overrepresented in the worst predictions;
- mostly the Chinese learners, who are strongly overrepresented among the worst-predicted cases, whereas the learners from Romance and Germanic L1s are strongly underrepresented (for L1Family, $p_{mnt}<10^{-18}$);
- with regard to the lengths, recipients and patients that are nearly exactly equally long, or have recipients that are 1-2 characters shorter than the patients (for LengthDiff, $p_{KS\text{-}test}<10^{-10}$). This is interesting when compared to the discussion of LengthDiff in Figure 4 regarding the non-confident/middle ground predictions: There we saw that "the non-confident predictions involve cases where there is a little bit of a length difference, i.e. when the patient is either 2–4 characters longer or shorter than the recipient", and here we see that the worst predictions involve cases where the recipient and patient are about equally long; and of course the best predictions – confident and correct ones – are those with big differences between recipient and patient length, because then end weight/short-before-long makes a strong case for one construction.

In other words, and a bit simplistically because this is coming from a monofactorial exploration only, the cases where the NS-based random forest fails most at are cases where Chinese learners are using *give* in prepositional datives with inanimate patients that are as long as the humanimate recipients they are used with. In a more comprehensively-annotated data set, this observation would be

the starting point for a qualitative analysis of why these cases prove to be so problematic for the model/forest: can the unexpected choices be explained by

– additional predictors (e.g., syntactic priming, information status, or phonological effects)?
– transfer effects: if one L1-specific group of learners is responsible for most unexpected choices, does this hint at transfer effects?
– reference to specific speakers?
– prompt effects: are the learners re-using a structure or even an explicit formulation they saw in a prompt or heard from a NS interview?
– fixed expression or idiomatic status of the constructions in question? etc.

In sum, just like researchers using regression-based methods explore plots of residuals for model diagnostics (see Fox and Weisberg 2019: Section 8.1-8.2) and just like researchers attempting to establish from the data what might be prototypical instances of the alternants in question (see Gries 2003a, b for the first discussions along these lines), we are proposing to try to establish from the data the 'prototypical instances' of the worst predictions – the most confident yet wrong ones – and use them to zoom in on where the model needs improvements to minimize error (variance/deviance), and where better to start where the current analysis/model fails most spectacularly.

4 Discussion and concluding remarks

4.1 Summary

While the current discussion is largely programmatic and, in this particular data set, the difference between the traditional binary classification and the newly-developed ternary one is not huge, we do believe this is food for thought nonetheless. This is because the current approach does something that, as far as we can recall, we have not seen in corpus-linguistic alternation studies of all kinds, namely focus the analyst's attention on the cases alluded to in the title of the paper. The literature so far, especially the literature on native speaker data, has mostly been concerned with the main diagonal of a 2×2 confusion matrix (e.g., Table 2:), i.e. where the correctly-predicted cases in an either-this-construction-or-the-other scenario are located. Our first suggestion was inspired by (i) our realization that our middle ground cases are more important than they are in many other classificatory applications and (ii) by (correct) perception of conference audiences that MuPDAR was maybe too 'divisive' or 'harsh' around the

cut-off point; we therefore suggested to take more seriously the notion that there will be many cases where both/all constructions of an alternation are (roughly) equally acceptable, which amounts to extending the regular confusion matrix to a 2×3 confusion matrix such as Table 3.

Our second suggestion in turn was inspired by the sensation that the other cells of a confusion matrix are for the most part very much underutilized although the misclassifications really are among the best pointers to future research (especially if one focuses not just on incorrect predictions, but the worst ones).

With regard to both these issues, while we have concentrated on MuPDAR here, both of these points also have implications for 'traditional' one-step regression/classifier approaches. First, the failure to at least consider the middle ground more may make analysts end up with accuracy, precision, and recall statistics that, while statistically accurate, are conceptually a bit too stringent/stifling. Second and less controversially, of course a closer analysis of the worst predictions of a single one-step regression/classifier analysis can be beneficial along the same lines as discussed here for MuPDAR.

4.2 Where to go from here

Both suggestions here relied on an underutilized statistic well-known in machine-learning circles and competitions, log loss. While its application for identifying the worst predictions is probably fairly uncontroversial, its use for identifying the middle ground is admittedly less so and we feel that the most pressing issue for subsequent studies along these lines would be to develop 'the best way' of determining the size of the middle ground. Above, we used one log loss unit, which translates into the interval of [0.36788, 0.63212] for predicted construction probabilities, i.e. an interval of 0.2642 around the usual cut-off point of 0.5. While we think this is a good-sounding size – not too narrow, not too wide – obviously, this is not the only choice one could make.

In our first exploration of this, we actually relied on a confidence-interval kind of approach. That is, we collected for each of the 1828 NNS cases the 2000 votes generated by a forest with *ntree*=2000 and then determined, using an exact binomial test, for each prediction whether the 95 per cent confidence interval of the proportion of *s*-genitive predictions, once adjusted for 1828 comparisons, includes 0.5. Put differently, the middle ground was defined as the 95 per cent confidence interval around 0.5 of 2000 votes adjusted for 1828 tests. However, after some thought, that approach seemed problematic: First, because this interval, which amounts to [0.461, 0.539],[6] just seems unrealistically narrow. Second, because the size of this interval is correlated with the *ntree*, i.e. the number

of trees the forest contains, which is undesirable because it means that the desire to get more robust/representative forests leads to increase *ntree*, which will make the middle ground even narrower than this.

Another obvious alternative would be to just pick a middle ground percentage range, just as [0.4, 0.6] or [0.333, 0.667], etc. One possibility (mentioned to us by Martin Hilpert) would of course be experimental validation, i.e. to determine which percentage interval is most compatible with independently-collected judgment data from human annotators. (Plus we have other ideas but those are not ready for prime time yet.) However, we wish to state emphatically that the absence of an obvious interval size is a problem to be solved, yes, but it is not something that brings down the whole idea. This is because (i) there *are* ways in which the size of this interval can be determined (e.g. the experimental validation) and (ii) let's face it, even something as ubiquitous as *p*-values are ultimately based on the arbitrary selection of a cut-off point (at, typically, 0.05) so it's not like empirical sciences are never relying, at least for some time, on values that are not (yet) empirically firmly grounded. Thus, while we do not have a ready-made solution yet for this one aspect of our first proposal, we hope to at least stimulate some new thoughts and work going 'beyond the main diagonal'.

Notes

1. Obviously, the quality and reliability of the results of any such study involving comparisons between native and non-native speakers or speakers of different varieties will be correlated with how similar/comparable the corpora used are; note that this is the case regardless of whether such studies involve mere over- and underuse frequencies, traditional regression or classifier methods, or MuPDAR. Our stance with regard to that issue is no different from that of any other learner corpus study or varieties study: Since the quality of the results will be correlated with the degree of homogeneity of the corpora in question, which is why we chose LOCNESS and LOCNEC for the NS data as opposed to, say, the BNC or ICE-GB, but other than that there is really little else that has been or can be done and most learner corpus or variety studies make similar choices.
2. Of course, Chinese is not a language family, we are using L1Family as a convenient cover term without any theoretical significance for the Germanic and Romance L1s and prefer to call the Chinese data "Chinese" rather than, for our data, "other".

3. These are computed in R with
   `sum(dbinom(0:1515, 1828, 1556/1828))`
   `and sum(dbinom(1556:1828, 1828, 1515/1828)).`
4. For VerbLemma, we had to compute a Monte Carlo simulation version (3,000,000 simulations) since different R functions for an exact test crashed under a RAM demand in excess of 30GB (given that 2,505,433,700 possible events would have to be evaluated).
5. The notion of 'floors' is defined as in R: the floor of a number *x* is the largest integers not greater than *x*. That is,
   `floor(c(-1.5, -0.5, 0.5, 1.5, 1.75)) in R will return this vector c(-2, -1, 0, 1, 1).,`
   which makes this a very practical way to reduce a large set of many minimally different real numbers to a much smaller, more manageable set.
6. This is computed with
   `binom.test(1000, 2000, 0.5, conf.level=1-(0.95/1828)).`

References

Deshors, Sandra C. 2014. A case for a unified treatment of EFL and ESL: A multifactorial approach. *English World-Wide* 35(3): 279–307.

Deshors, Sandra C. 2018. Simple Past meets Present Perfect meets Passé Composé: A semantic exploration of the Present Perfect in French-English interlanguage. *International Journal of Learner Corpus Research* 4(1): 23–53.

Deshors, Sandra C. and Stefan Th. Gries. 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.

Fox, John and Sanford Weisberg. 2019. *An R companion to applied regression*. 3rd ed. Los Angeles, London, etc.: Sage.

Gries, Stefan Th. 2003a. *Multifactorial analysis in corpus linguistics: A study of Particle Placement*. London and New York: Continuum Press.

Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1: 1–27.

Gries, Stefan Th. to appear. On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory.*

Gries, Stefan Th. and Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new para-

digm for learner corpus research. In Jesús Romero-Trillo (ed.). *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms*, 35–54. Cham: Springer.

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

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

Heller, Benedikt, Tobias Bernaisch and Stefan Th. Gries. 2017. Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME Journal* 41: 111–144.

Heller, Benedikt, Benedikt Szmrecsanyi and Jason Grafmiller. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45 (1): 3–27.

Kolbe-Hanna, Daniela and Lina Baldus. 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 and Gert De Sutter. 2018. Alternation in contact and non-contact varieties: Reconceptualising *that*-omission in translated and non-translated English using the MuPDAR approach. T*ranslation, Cognition and Behavior* 1(2): 251–290.

Lester, Nicholas A. 2019. *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research* 5(1): 1–32.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte and Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa* 2(1): 1–27.

Werner, Valentin, Robert Fuchs and Sandra Götz. To appear. L1 influence vs. universal mechanisms: An SLA-driven corpus study on temporal expression. In Bert Le Bruyn and Magali Paquot (eds.). *Learner corpora and second language acquisition research*. Cambridge: Cambridge University Press.

Wulff, Stefanie, Nicholas Lester and Maria T. Martinez-Garcia. 2014. *That*-variation in German and Spanish L2 English. *Language and Cognition* 6: 271–299.

Wulff, Stefanie and Stefan Th. Gries. 2015. Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism* 5(1): 122–150.

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

Wulff, Stefanie and Stefan Th. Gries. To appear. Explaining individual variation in learner corpus research: Some methodological suggestions. In Bert Le Bruyn and Magali Paquot (eds.).di *Learner corpora and second language acquisition research*. Cambridge: Cambridge University Press.