Prenominal adjective order preferences in Chinese and German L2 English

A multifactorial corpus study

Stefanie Wulff and Stefan Th. Gries
University of Florida / University of California at Santa Barbara

This study presents a contrastive analysis of 3624 instances of prenominal adjective order retrieved from the Chinese and German sections of the *International Corpus of Learner English* and the *International Corpus of English*. The data was annotated for nine determinants of adjective order, including semantic, frequency-related, and articulatory features. Applying a two-step regression procedure called MuPDAR (*Multifactorial Prediction and Deviation Analysis Using Regressions*), the present study finds that overall, the intermediate-advanced level learners are well-aligned with native speakers' preferences. However, we also see that while the German learners seem generally better aligned with regard to frequency-related factors, the Chinese learners behave more target-like with regard to the effect of adjective gradability, and they seem more sensitive to segmental alternation constraints. In discussing these and other results, the study hopes to illustrate how corpus-based methods can make a valuable contribution to contemporary SLA research, specifically with regard to multifactorially determined phenomena such as adjective order.

Keywords: Corpus linguistics, adjective order, regression analysis, gradability, length, frequency, segment alternation, rhythmic alternation, Chinese learners, German learners

1. Introduction

As corpus linguistics is increasingly recognized as a useful addition to the linguist's methods tool box, the number of studies in second language acquisition (SLA) research that adopt a corpus-linguistic perspective is on the rise as well. Corpus data are being used to inform the development of experimental stimuli (the MRC Psycholinguistics Database and the English Lexicon Project are widely used, for

Linguistic Approaches to Bilingualism 5:1 (2015), **122–150**. DOI 10.1075/lab.5.1.05wul
ISSN 1879–9264 / E-ISSN 1879–9272

instance); corpus analyses are presented alongside experimental case studies to gain complementary insights into second language production, development, and the role of second language input (Gries & Wulff, 2005, 2009; Gilquin, 2007; Siyanova & Schmitt, 2008); and a growing number of published studies is entirely corpus-based (Tono, 2004; Paquot & Granger, 2012; Gries & Wulff, 2013).

As with any other method, the researcher's biggest challenge is to employ that method which best fits the research question at hand. This intricate relationship between theory and method has always been a primary concern in SLA research, and this awareness is reflected increasingly so in the steady increase of publications devoted to methodological issues in SLA research (Plonsky, 2013, 2014 provides excellent overviews). More specifically, all contemporary theoretical approaches to SLA, in spite of differences regarding the specific nature of SLA with regard to questions such as how much of language is innate, to what extent it is modular, or how exactly a learner's first and second language interact, agree on at least one fundamental point: SLA is a complex process, shaped simultaneously by various factors, both linguistic and extra-linguistic. This understanding has non-trivial consequences for any empirical study in SLA, regardless of how it is framed theoretically: the methods we choose need to match the assumed underlying complexity of the process we are trying to describe, model, and/or predict. As far as corpus-linguistic analyses of learner language are concerned, we believe that the great potential of corpus linguistics, especially with regard to its capacity to model complex phenomena, has yet to be realized in the majority of corpus-based SLA research.

In the present paper, we aim to demonstrate how one of the most pertinent questions in SLA research — where and to what extent learners deviate from target-like behavior — can be addressed using a new corpus-linguistic approach called *Multifactorial Prediction and Deviation Analysis Using Regressions* (MuPDAR; Gries & Adelman, 2014; Gries & Deshors, 2014). We hope to illustrate how the MuPDAR approach goes beyond most of current learner corpus research as it (i) does justice to the assumed complexity of the target phenomenon (here: adjective order) by including all predictors that have been argued to impact the phenomenon, (ii) aids in establishing a native speaker standard of comparison for the specific phenomenon under investigation, (iii) points to the specific predictors on which learners deviate from native-like behavior, i.e. points to the nature of learner errors, and (iv) furthermore quantifies the degree to which learners deviate from the native norm, i.e. informs the researcher about the severity of the non-targetlike choice, or error.

The structure of this paper is as follows: In Section 2, we discuss the phenomenon of prenominal adjective order, which we study here; Section 3 discusses the operationalization of our variables as well as the retrieval of our data and their

annotation; Section 4 provides a detailed overview of all results; Section 5 discusses these results and concludes with an outlook on the possibilities of further exploration that the MuPDAR approach affords.

2. The target alternation: prenominal adjective order (AO)

In order to illustrate how the MuPDAR approach works, we here present a case study of prenominal adjective order (AO). Native English speakers generally prefer (1a) to (1b):

(1) a. the big red squirrel
    b. the red big squirrel

Previous studies suggest that at least one factor at work here is the semantic class the two adjectives *big* and *red* belong to: *big* denotes the size of the referent encoded by the head noun while *red* is a color term; size adjectives are placed before color adjectives. While the preference for (1a) may be accounted for in this straightforward manner, previous research also suggests, however, that other factors such as the length, frequency, and other qualities of the adjective determine speakers' choices. In accordance with this view, Wulff (2003) presented an overview of previous research on AO, as well as the results of a multifactorial analysis based on data from the British National Corpus (BNC). This analysis included the factors listed in Table 1 below. We briefly introduce each factor in turn here;

Table 1. Overview of predictors studied in Wulff (2003) and additional predictors

Variable/predictor	Adjective$_1$	Adjective$_2$
SemClose	semantically versatile/flexible	semantically specialized
IndComp (Gradability)	more gradable	less gradable
NomChar	less nouny / more adjectival	more nouny / less adjectival
AffLoad	more positively-loaded	less positively-loaded
Length	short(er)	long(er)
Freq	more frequent	less frequent
NounSpecFreq	low(er)	high(er)
Variable/predictor	Predicted order	
SegAlt	adjectives will be ordered so as to optimize ideal syllable structure (CV)	
RhythAlt	adjectives will be ordered so as to optimize rhythmic alternation (su)	

in Section 3.2, we present the rationale for the corpus-based operationalizations adopted for each factor.

Semantic Closeness (SemClose) expresses the idea that adjectives that denote qualities that are less inherent to the referent of the head noun in question will precede adjectives that denote more inherent qualities (Whorf, 1945; Biber et al., 1999, p. 599). This line of reasoning is more generally reflected in Behaghel's Law that things that are conceptually similar are likely to be expressed in close proximity linguistically. To give one (fictitious) example, *shiny stainless steel* should be preferred over *stainless shiny steel* since *stainless* is a property more inherent to *steel* (and few other things) whereas *shiny* is a quality that can describe a wide range of things; in other words, *shiny* is more versatile (because it is semantically less close to the things it describes) than *stainless* (because it is semantically close to the things it describes).

Independence from Comparison, or Gradability, (IndComp) predicts that more gradable adjectives will precede less gradable adjectives (Martin, 1969; Posner, 1986). According to this factor, *heavy red box* will be preferred over *red heavy box* because *heavy* is a quality that is more dependent on comparison than *red*: in order to decide whether a box is heavy, some (if only implicit) comparison to other boxes has to be made — in contrast, the color of an object appears to be a quality that can be attributed to the object with less or no comparison to other objects.

A third semantic factor is the Nominal Character (NomChar; Posner, 1986, p. 13) of the adjectives in question: as Biber et al. (1999, p. 599) put it, there is "an overall tendency for the most nounlike modifiers to occur closest to the head noun." To give an example, *woolen white hat* is preferred over *white woolen hat* because *woolen* is much more adjectival in nature (it ends with a typical adjectival suffix) compared to *white* (which carries no adjectival morphology and may either be an adjective or noun depending on the context).

Some previous studies of native speakers' ordering preferences furthermore suggest an impact of the adjectives' affective load (AffLoad; Richards, 1977). Motivated by Boucher & Osgood's (1969) "Polyanna Hypothesis," which states that humans have a universal tendency to report "first the good news, then the bad news," this factor has been argued to render a sequence like *powerful dangerous medication* preferable to *dangerous powerful medication* since *powerful* denotes a positive quality while *dangerous* denotes a negative quality.

Psycholinguistic studies of ordering phenomena have furthermore provided compelling evidence that frequency (Freq) plays a major role in the sequencing of entities, including the ordering of adjectives (Bock, 1982; Lapata et al., 1999). In accordance with general claims regarding the nature of frequency effects on ordering phenomena, the more frequent adjective is expected to precede the less frequent adjective because of its higher resting activation level in the mental lexicon,

which entails that it is likely readied for production sooner than a less frequent adjective. Given the well-known inverse correlation of the length of words and their frequencies, this effect also entails a correlation of AO with length constraints (Length) such that shorter words should precede longer words because they are readied for production faster.

Lockhart and Martin (1969) found a more specific frequency effect: they demonstrated that the adjectives that are remembered most easily upon the occurrence of a noun tend to be those that tend to stand closest to these nouns in actual language use more frequently. The ease with which an adjective can be elicited through the presentation of a noun can be understood as the conditional probability of the adjective-noun sequence: the adjective with the higher Noun-Specific Frequency (NounSpecFreq) should occur closer to the head noun than the adjective with the lower noun-specific frequency. For a noun like *story*, for example, *true* has a higher conditional probability (at least in the BNC) to precede *story* than *fantastic*, which should render *fantastic true story* preferable over *true fantastic story*.

Next to these factors captured in Wulff (2003), we here also included two factors related to articulatory constraints since our research into other alternation phenomena suggests that they may have some impact on ordering preferences in general: Segment Alternation (SegAlt) and Rhythmic Alternation (RhythAlt). Segment Alternation (also referred to as "ideal syllable structure") has been suggested to impact adjective ordering such that the preferred ordering will be the one with the more ideal syllable structure in the sense of strict consonant-vowel alternation (Venneman, 1988, p. 13–29; Schlüter, 2003: Section 3.1). Accordingly, *lovely bright eyes* should be preferred over *bright lovely eyes* because the latter exhibits two clashes of two consonants and then two vowels at the word boundaries, while the former exhibits the supposedly desirable consonant-vowel alternation at the word boundaries.

The second phonological factor included here is Rhythmic Alternation, which denotes the postulated universal tendency for sequences to be preferred when they display a strict alternation of stressed and unstressed syllables (Couper-Kuhlen, 1986, p. 60; Shih et al., to appear; Gries & Wulff 2013 provide evidence that rhythmic alternation impacts the genitive alternation in English). That is, when considering this factor in isolation, *Chinese traditional band* will be preferred over *traditional Chinese band* because the latter contains a sequence of three unstressed syllables.

While AO has been researched quite intensively in native English speakers, the present study is one of the first to examine English L2 learners' AO preferences, especially from a multifactorial perspective.

3. Data and methods

3.1 Data retrieval

We retrieved exhaustive samples of adjective-adjective-noun sequences from the spoken section (10 million words) of the *British National Corpus* (BNC) to represent what is arguably the spoken equivalent of the target language of native English speakers, the German section of the *International Corpus of Learner English* (G-ICLE; ~250,000 words) to represent intermediate-advanced level German learners of English as a second language, and the Chinese section of ICLE (C-ICLE; ~500,000 words) to represent intermediate-advanced Chinese learners of English as a second language. The ICLE corpora contain academic student writing. 3015 attestations were obtained from the BNC, 323 from G-ICLE, and 286 from C-ICLE for a total sample size of 3624 attestations. More specifically, the BNC was queried first by taking advantage of the fact that this corpus is annotated for parts of speech, which allowed us to run an R script that browsed the corpus for all occurrences of adjective-adjective-noun sequences. The resulting list of candidate phrases was then manually checked for true hits (that is, to exclude false hits based on tagging errors introduced by CLAWS, the POS-tagger used for the BNC). The final sample of attestations thus yielded then served as the basis for queries on the learner corpora of all adjectives that occurred at least once in an adjective-adjective-noun sequence in the BNC. The two sets of candidate hits were then also manually inspected for true hits.

3.2 Data annotation

Semantic Closeness was operationalized by identifying for each adjective that occurred at least once in our data sample how many different word types it precedes, normalized by its token frequency in the BNC. For instance, adjectives with low values in SemClose — i.e. with many different word types following them — were *social* (0.0487), *difficult* (0.0692), *good* (0.0805), and *important* (0.0878) — whereas adjectives with high values in SemClose — i.e. with much fewer different word types following them — were *fanatical* (0.6824), *traditionalist* (0.8209), *existentialist* (0.875), and *monkish* (0.9231). Based on these values for each adjective, we then computed a SemCloseDiff value for each Adj-Adj-N triple by subtracting the value for each adjective$_1$ from the value of the adjective$_2$ it occurred with. The expectation from the literature was that AO choices would be correlated with higher values of SemCloseDiff, because higher values of SemCloseDiff would result from adjective$_1$ being quite flexible and adjective$_2$ being less flexible in terms of the number of words it precedes.

To measure each adjective's Independence from Comparison, we identified for each adjective how often it was used in an analytic and/or synthetic comparative or superlative, and divided that number by the adjective's overall token frequency in the BNC. To give a few examples, adjectives with low values in IndComp — i.e. with many cases of comparatives and superlatives — were *sophisticated* (0.7019), *versatile* (0.7547), *prestigious* (0.7719), or *influential* (0.8003) — whereas adjectives with high values in IndComp — i.e. with few cases of comparatives and superlatives — were *factual* (0.9869), *religious* (0.9947), *pancreatic* (1), or *Gaelic* (1). As for Semantic Closeness, we combined the adjectives' IndComp values into an IndCompDiff value for each Adj-Adj-N triple. The IndComp values were again computed as the difference of IndCompDiff_2 minus IndCompDiff_1. According to the literature, AO choices should be correlated with higher IndCompDiff values because higher values result from adjective_1 being less gradable than adjective_2.

In order to measure each adjective's Nominal Character, we looked up how often each adjective was tagged as an adjective and how often it was tagged as a noun and then computed the percentage of noun tags out of both noun and adjective tags. This led to $\text{NomChar}_{1/2}$ values of, for instance, *famous*, *inefficient*, and *ugly* (0), *private* (0.025), *academic* (0.078), *ideal* (0.189), *public* (0.406), *light* (0.645), and *material* (0.857). From this, for each Adj-Adj-N triple, we computed a NomCharDiff value, namely the difference of NomChar_2 minus NomChar_1. The expectation from the literature is that AO choices should be correlated with higher NomCharDiff values because higher values express that adjective_1 is less nouny than adjective_2.

Each adjective's Affective Load was coded on a simple ternary scale: −1 for negatively-loaded adjectives (e.g., *bloody*, *messy*, *bad*, *awful*, *terrible*, *dangerous*, *poor*, …), +1 for positively-loaded adjectives (e.g., *great*, *solid*, *conscious*, *brave*, *natural*, *productive*, *sunny*, …), and 0 for the vast majority of everything else. The decision to classify an adjective as either positively or negatively loaded was guided by consulting the adjective's entry in the *Collins Cobuild* dictionary and by having the data coded by both authors independently, who were in full agreement on all cases. We then computed an AffLoadDiff value, namely the difference of AffLoad_1 minus AffLoad_2 so that positive values indicate that the actually chosen order is more compatible with the hypothesis that positive adjectives precede negative ones.

Segment Alternation was coded by considering two transition points in each Adj-Adj-N triple: the one between the final segment of adjective_1 and the first segment of adjective_2, and the one between the final segment of adjective_2 and the first segment of the noun. Both transition points were coded for whether it involved a strict consonant-vowel (CV) alternation (in which case it was coded as 0), a sequence of two different consonants (C_1C_2) or vowels (V_1V_2) (which were

assigned a value of 1), or a sequence of identical consonants (C_1C_1) or vowels (V_1V_1) (which were assigned a value of 2). For each observed attestation in our data set, we then added up the values arising from this coding of the two transition points, captured that value in a variable called SegAltObs, and then added up the values that would have arisen from the reverse ordering of the adjectives and stored that in a variable SegAltAlt. The variable SegAltDiff was then computed as SegAltAlt minus SegAltObs so that

- positive values indicate that the actually chosen order of adjectives is more compatible with segment alternation than the theoretically possible reverse order (e.g., *important*$_{-CV-}$*early*$_{-VC-}$*training* is better than *early*$_{-V1=V2-}$*important*$_{-C1=C2-}$*training*);
- negative values indicate that the actually chosen order of adjectives is less compatible with segment alternation than the theoretically possible reverse order (e.g., *political*$_{-CC-}$*key*$_{-VV-}$*issues* is worse than *key*$_{-VC-}$*political*$_{-CV-}$*issues*);
- values of 0 indicate that both are equally (in)compatible with segment alternation (e.g., *happy*$_{-VC-}$*suburban*$_{-CC-}$*families* and *suburban*$_{-CC-}$*happy*$_{-VC-}$*families*).

Rhythmic Alternation was coded as the sequence of stressed (*s*) and unstressed (*u*) syllables of the actually observed ordering (as in *happy suburban families* → *suusususuu*), which was stored in a variable called RhythAltObs; the sequence of stressed (*s*) and unstressed (*u*) syllables of the theoretically possible alternate ordering (*suburban happy families* → *usususuu*) was stored in a variable called RhythAltAlt. For both these variables, the sequences of *s*'s and *u*'s was converted into a number between 0 and 1 that quantifies the degree to which the sequence exhibited stress clashes (i.e. sequences of 2+ *s*'s) or stress lapses (i.e. sequences of 3+ *u*'s), which were then normalized against the length of the sequence. For example,

- a sequence such as *ssssuuuuuu*, which violates rhythmic alternation maximally, scored 1;
- a sequence such as *susuususu*, which adheres to rhythmic alternation perfectly, scored 0;
- a sequence such as *uususs*, which somewhat violates rhythmic alternation, scored 0.4.

Finally, we then computed a RhythAltDiff value, namely the difference of RhythAltAlt minus RhythAltObs so that

- positive values indicate that the actually chosen order of adjectives is more compatible with rhythmic alternation than the theoretically possible reverse order (e.g., *sheer rational expectations* is better than *rational sheer expectations*);

– negative values indicate that the actually chosen order of adjectives is less compatible with segment alternation than the theoretically possible reverse order (e.g., *warm sensuous particulars* is worse than *sensuous warm particulars*);
– values of 0 indicate that both are equally (in)compatible with segment alternation (e.g., *warm weak tea* and *weak warm tea*).

The length of each adjective was comparatively simple to measure. Since all measures of word/constituent length — phonemes, characters, morphemes, syllables, words, … — are extremely highly correlated with each other (with *r*-values often exceeding 0.9), we adopted the simple strategy of counting each adjective's length in characters. From those, we computed for each adjective$_1$-adjective$_2$ pair a LengthDiff value, namely Length$_2$ minus Length$_1$; the expectation from the literature is that AO choices should be correlated with higher values of LengthDiff, because higher values of LengthDiff result from adjective$_1$ being shorter than adjective$_2$ as in, for instance, *new technological* or *long strenuous*.

Each adjective's frequency was determined by determining how often the corresponding form was tagged as an adjective in the BNC. From this, we then computed a FreqDiff value for each Adj-Adj-N triple, namely the difference of *ln*(FreqDiff$_1$) minus *ln*(FreqDiff$_2$). The expectation from the literature is that AO choices should be correlated with higher values of FreqDiff because higher values of FreqDiff express adjective$_1$ being more frequent than adjective$_2$ as in, for instance, *good safe* or *long strenuous*.

The adjectives' Noun-Specific Frequency was measured as the relative frequency with which the adjective preceded the noun of the triple. In other words, we determined the transitional probability *p*(noun|adjective) by dividing the frequency of each adjective-noun sequence by the token frequency of the adjective in the BNC. For instance, adjective-noun pairs with low values in NounSpecFreq — i.e. pairs where the adjective was not followed by their particular noun frequently — were *local bullfrogs* (< 0.0001), *huge fridge* (0.0001), or *solid hedge* (0.0003) — whereas adjective-noun pairs with high values in NounSpecFreq — i.e. pairs where the adjective was frequently followed by their particular noun — were *venereal disease* (0.5526), *focal point* (0.6127), or *stainless steel* (0.9283). From this, for each Adj-Adj-N triple, we then computed a NounSpecFreqDiff value, namely the difference of NounSpecFreq$_2$ minus NounSpecFreq$_1$. The expectation from the literature is that AO choices should be correlated with higher values of NounSpecFreqDiff because higher values of NounSpecFreqDiff result from the adjective$_1$ being less associated with the noun than adjective$_2$.

Finally, each instance was annotated for the predictor Language, which simply captured the L1 of the speaker, i.e. English for the NS and Chinese or German for the NNS.

As for the dependent variable of the first regression we ran (R_1; see Section 3.3 for details on the methodological approach), AO poses a difficulty that other common alternation studies do not have. In the many studies on syntactic alternations such as the dative alternation, particle placement, the genitive alternation etc. there are two (or more) alternatives all of which are realized in the data set. For instance, depending on givenness, length, animacy, and other considerations, speakers will sometimes use *give* in a ditransitive, sometimes in a prepositional dative. For AO, the situation is different because there is only one attested order. Thus, in some sense, there can be no variation in the corpus data — the first adjective is the first one, the second is the second. This raises the question of what to use as a dependent variable: unlike with other alternations, it cannot be a variable ORDER because that would invariably be *first-second*. To address this issue, we, first, created a copy of the data set that reversed the order of the two adjectives (and thus their annotations and the resulting difference computations) and, second, created a variable POSITION that marked whether a case was from the original data (i.e., an actually-attested ordering and, thus, the ordering of the two adjectives that the speaker preferred) or whether a case was from the reversed copy (i.e., a theoretically possible alternate ordering that the speaker did not in fact choose). Thus, all independent variables/predictors as described above were then used to try to predict whether they allow to predict what the speakers did (POSITION: *chosen*) vs. what they didn't (POSITION: *alternate*). This logic is visualized in Table 2 using only one hypothetical Adj-Adj-N triple ($x_{\text{Adj}}\ y_{\text{Adj}}\ z_{\text{Noun}}$) and two predictors (NOUNSPECFREQDIFF and INDCOMPDIFF).

Table 2. Raw data design for setting up the dependent variable POSITION and the independent variables for R_1

Adj1	Adj2	N	Position (NS)	NounSpec Freq1	NounSpec Freq2	NounSpec FreqDiff	Ind Comp1	Ind Comp2	IndComp Diff
x	*y*	*z*	chosen	0.2	0.5	0.3	0.1	0.3	0.2
y	*x*	*z*	alternate	0.5	0.2	−0.3	0.3	0.1	−0.2

As for the dependent variable of the second regression we ran (R_2), we will explain its operationalization in Section 4.3 below for expository reasons.

3.3 Statistical data analysis: MuPDAR

3.1.1 *Introduction*

The majority of learner corpus studies to date present simple frequency analyses of learner errors and/or analyses of overuse and underuse; a good share of

these studies are entirely descriptive in the sense that they present these frequency counts but cannot really offer deeper exploration of the causes that drove learners' non-target-like behavior. This is for several reasons: Firstly, frequency counts (in whatever form: as raw frequencies, percentages, or the like) do not imply any explanation of *why* learners commit errors, overuse, or underuse a specific target structure. Secondly, considering that native speaker research suggests that speakers' choices are always determined by not one, but a variety of factors — as we have seen in Section 2 for AO, for example — even the fewer studies that frame frequency counts by one variable still paint an incomplete picture. Thirdly, and maybe most crucially for those studies that seek to compare native speaker (NS) and non-native speaker (NNS) production, comparing frequency counts across different groups of NSs and/or NNSs does not license any conclusions regarding the (degree of) native-like behavior on the part of the learners.

To give just one (fictitious) example here of how simple contrastive frequency counts can lead to misleading interpretations, imagine that you examined NSs and NNSs' choices of active and passive voice in English, and that you found that the learners used passive voice significantly less often than the NSs did. One might jump to the intuitively reasonable conclusion that the NNSs underuse passive voice because it is a more complex structure that presents more difficulty for the learners than active voice. However, we can only draw that conclusion if we can be sure that we held all variables regarding NSs' choice between active and passive voice constant. We know NSs choose either voice depending on a variety of factors, including the definiteness of the subject (definite subjects invite passivization more than indefinite ones); the definiteness of the object (indefinite objects invite passivization more than definite ones); whether the verb is dynamic or stative (stative verbs dislike the passive); not to mention factors such as register (passives are comparatively rare in oral language) and genre (passives are much more prominent in academic writing than in personal correspondence, for example). What might appear as NNSs overusing active voice might be masking their non-target-like behavior on any of these variables (or any of their interactions, for that matter): For instance, NNSs might in fact overuse indefinite subjects compared to NSs, which triggers a higher share of active voice sentences, but in all other regards use active and passive voice just like NSs do. That is to say, their behavior with regard to the target structure might in fact be native-like and the true source of their non-native-like behavior may reside somewhere else.

A methodologically more adequate approach that addresses these issues is a regression analysis that includes (ideally all) the predictors assumed to impact the choice of the target structure, importantly also featuring the speakers' L1 (NS or NNS) as one independent variable that is allowed to interact with any of the other predictors included in the analysis (Gries & Deshors, 2014). Such an analysis

allows us to address the question: "How do native speakers and (different groups of) non-native speakers differ from each other?" One thing needs to be borne in mind when interpreting the results of such a regression analysis, however: the regression coefficients it provides for each predictor or interaction of predictors are based on all the data in the sample, that is, both the NS and the NNS data. Thus, in the absence of some very careful setting of *a priori* orthogonal contrasts — something we have never seen reported in any methods section of a corpus-based [F/S]LA paper — the results of regression approaches do not necessarily answer the questions of how the NNSs differ from the NSs and how different NNS groups differ from each other, because the regression coefficients that reflect the effects of predictors are then based on a combination of both NS and NNS data.

The MuPDAR approach presents a way to solve these issues by breaking down the analysis into a 3-step, 2-regression procedure. While these three steps build on each other, it is worth pointing out that depending on the research question, each of the three steps obtains results that one may choose to examine depending on the focus of the research question. The three steps are as follows:

- First, a regression R_1 is fitted on the NS data only to determine whether the included predictors in fact predict AO. In essence, this first step addresses the question: "What are the factors that impact native speakers' behavior?" The final model yielded from R_1 can be used to describe the NS system (as in Gries & Adelman 2014) and/or to make NS predictions.
- In a second step, the regression equation of R_1 is applied to the NNS data (be that one group of learners or several groups of NNSs with different L1 backgrounds). This yields, for every attestation in the NNS sample, a prediction of which variant of the target structure the NS would have picked in the same contextual configuration. The second step therefore addresses the question: "What would a native speaker have done in the same context?" (cf. Gries & Deshors 2014 for further explanation of the different ways in which these predictions can be interpreted, depending again on the specific focus of the research question).
- In a third step, a second regression is run, this time over the learner data, addressing the question: "Do the learners do what the native speaker would do in their place, and where they do not, why?"

In the following, we outline how these three steps were carried out in the present paper.

3.1.2 *Regression R_1: NS choices*

We first fit a binary logistic regression (Gries, 2013: Section 5.3) on the NS data that included

- Position: *chosen* vs. *alternate* as the dependent variable;
- RhythAltDiff + SegAltDiff + LengthDiff + FreqDiff + NomCharDiff + SemCloseDiff + IndCompDiff + NSpecFreqDiff + AffLoadDiff as main-effect predictors;
- all two- and three-way interactions of these main effects as interaction predictors.

Since, in this paper, we will actually not explore this regression model and its coefficients in detail — this regression is only run to obtain coefficients for predictions for the NNS data — no model selection process was undertaken. Rather, we determined whether the maximal model resulted in a good fit and a good classification accuracy to see whether proceeding with MuPDAR was feasible.

3.1.3 *Applying R_1 to the NNS data*

The second step of this MuPDAR analysis involves the application of the regression equation of R_1 to the NNS data. In essence, this amounts to answering the question "Would a NS have chosen/avoided this particular ordering which the NNS chose or avoided?" The fit of the NS model to the NNS data is then also quantified with a classification accuracy and a *C*-value in order to determine how well the NNS choices/preferences can be predicted from the NS at all.

3.1.4 *Regression R_2: NNS choices*

Given the results from step 2 as discussed in Section 3.1.1, we first determined for each NNS preference/choice whether it was the same as that predicted by R_1 for a NS and then compared all 1473 observed NNS preferences/choices and predicted NS preferences/choices in a pairwise fashion. The result of all these pairwise comparisons was stored in variable Correct: *true* (the NNS made the choice predicted for a NS) vs. *false* (the NNS did not make the choice predicted for a NS), which was added to the data frame. Of that data frame, we then took all the rows that contained NNS data with Position: *chosen* (i.e., the rows that reflect what the NNS actually wrote). This is because given the above-discussed symmetry of the data, the data points with Position: *alternate* are just the mirror image of those with Position: *chosen* and we did not want to inflate the sample size (and thus risk anti-conservative *p*-values) by including all cases with Position: *alternate*. Finally, we applied a binary logistic regression to the data with Correct as the dependent variable, this time with a model selection process because we are interested not only in overall predictive power but also in each the predictor's coefficient. This logic is visualized in Table 3.

Table 3. Raw data design for setting up the dependent variable Correct and the independent variables for R_2

Adj$_1$	Adj$_2$	N	Position (NNS)	Prediction (NS, from R_1)	Correct	Noun Spec Freq$_1$	Noun Spec Freq$_2$	Noun Spec Freq Diff
x	*y*	*z*	chosen	chosen	TRUE	0.2	0.5	0.3
a	*b*	*c*	chosen	alternate	FALSE	0.1	0.3	0.2

R_2

The overall direction of the model selection process was forward: We began with a model that included only Language as a predictor and then successively added the predictor to the model that most significantly improved the regression model while not unduly inflating the variance inflation factor of the model, where the candidate set of predictors contained all predictors and their interactions with Language. Once no more predictors could be added (with an exploratory threshold of $p_{\text{critical}} = 0.1$), we did a final test of whether, given everything currently in the model, any predictors would need to be deleted again (because of their now insignificant contribution to the model), and the final model was then evaluated with the usual summary statistics and plots.

4. Results

In this section, we summarize the results of R_1 (Section 4.1), the application of R_1 to the NNS data (Section 4.2), and the results of R_2 (Section 4.3).

4.1 Regression R_1: NS choices

The maximal model yielded for R_1 provided a good fit of the NSs' AO choices. The maximal model was highly significant (L.R. chi-squared = 2870.28, $df = 129$, $p < 0.0001$) and returned a good overall correlation (Nagelkerke's $R^2 = 0.47$). More importantly for our present purposes, however, is the classification accuracy of the model: Cross-tabulating the NS choices observed in the data and the NS choices predicted by the maximal regression model R_1 yields the classification matrix shown in Table 4.

Table 4. Classification accuracy of R_1 when applied to the NS data

	Position: *chosen*	Position: *alternate*	Totals
Predicted: *chosen*	2574	730	3304
Predicted: *alternate*	727	2571	3298
Totals	3301	3301	6602

In other words, R_1 classifies 77.93% of all choices correctly, which is highly significantly better than the chance accuracy of 50% ($p_{\text{exact binomial test}} < 10^{-100}$). More precisely, the *C*-value of this regression is 0.857, markedly exceeding the rule-of-thumb threshold of 0.8 (cf. Baayen 2008, p. 204).

4.2 Applying R_1 to the NNS data

Given the good fit of R_1 to the NS data, we applied the equation of R_1 to the NNS data, obtaining a NS preference/dispreference for every NNS choice. R_1 predicted the NNS choices relatively well, reflecting that, while the NNS choices on the whole were fairly compatible with those of the NSs, there were also some obvious differences worth exploring with R_2. Table 5 shows the classification accuracy that R_1 attained for the NNS data: the classification accuracy is 70.2%, i.e. a bit lower than what R_1 attained for the NS data.

Table 5. Classification accuracy of R_1 when applied to the NNS data

	Position: *chosen*	**Position:** *alternate*	**Totals**
Predicted: *chosen*	569	272	841
Predicted: *alternate*	167	465	632
Totals	736	737	1473

4.3 *Regression* R_2*: NNS choices*

Because of the risk of collinearity, automatic (stepwise) model selection was not feasible and we proceeded in a manual fashion. In this process, the variable FreqDiff was supposed to be added at step 4 of the model selection process, but its addition to the model increased the variance inflation factors of FreqDiff and SemCloseDiff to > 11 (because of the high intercorrelation between these two variables; $r > 0.8$). We therefore residualized SemCloseDiff out of FreqDiff and the new variable, FreqDiffWithoutSemCloseDiff, was then selected for inclusion. (This new variable's correlations with SemCloseDiff and FreqDiff were 0 and 0.43 respectively.) Once no more predictors could be added, a test for dropping predictors led to the final deletion of one effect. The final model of R^2 provided an exceptionally good fit (L.R. chi-squared = 500.86, $df = 12$, $p < 0.0001$) and a very high correlation (Nagelkerke $R^2 = 0.751$). Correspondingly, the model was able to classify very well where the NNSs' preferences/choices differ from those of the NSs, which is also reflected in the very good classification accuracy of 92.8% and a *C*-value of 0.965; cf. Table 6.

Table 6. Classification accuracy of R_2 when applied to CORRECT

	CORRECT: *true*	CORRECT: *false*	Totals
Predicted: *true*	554	38	592
Predicted: *alternate*	15	129	144
Totals	569	167	736

The highest-level predictors leading to this high accuracy are summarized in Table 7.

Table 7. Highest-level predictors in R_2, their significance tests, and the main predictor's main effect coefficient in R_1 (for comparison; previous literature on AO would lead to the expectation that all should be >0)

Highest-level predictor	Likelihood ratio test	*p*	Coef in R_1
SEMCLOSEDIFF	19.22 ($df = 1$)	< 0.0001	−7.72
LANGUAGE: FREQDIFFWITHOUTSEMCLOSEDIFF	47.75 ($df = 1$)	< 0.0001	0.67 (FREQDIFF)
LANGUAGE:INDCOMPDIFF	19.84 ($df = 1$)	< 0.0001	12.24
LANGUAGE:AFFLOADDIFF	10.94 ($df = 1$)	< 0.0001	−0.46
LANGUAGE:SEGALTDIFF	6.32 ($df = 1$)	0.012	0.47
LANGUAGE:NSPECFREQDIFF	3.51 ($df = 1$)	0.061	46.9

We now turn to the significant highest-level predictors to discuss their nature — direction and strength — in more detail. To present the results, we will use descriptive strip charts that have

- on the *x*-axis the significant predictor from the final model — a predictor that significantly interacts with LANGUAGE or the main effect of SEMCLOSEDIFF;
- on the *y*-axis the interaction of CORRECT:LANGUAGE: the first two rows of each plot show the cases where the NNS made a nativelike choice (first the Chinese, then the German NNSs); the bottom two rows show the cases where NNSs made non-nativelike choices (again, first the Chinese, then the German NSSs);
- the plotted unfilled squares represent NNS corpus examples that are plotted at their respective coordinates; the vertical bars represent the mean values of the plotted points within each of the four groups of CORRECT:LANGUAGE;
- the numbers plotted on the left and right margins of each plot represent the frequencies with which negative (left) and positive (values) of the variable on the *x*-axis are observed for the four groups of CORRECT:LANGUAGE; values of 0 (what we below refer to as *zero-condition*) are omitted from these counts because they reflect variable values that come with no preference for either ordering.

The first interaction with LANGUAGE was only created during the model selection process — FREQDIFFWITHOUTSEMCLOSEDIFF — to avoid high collinearity from the variables FREQDIFF and SEMCLOSEDIFF. R_1 showed that NS choices are on the whole compatible with the hypothesized effect of FREQDIFF or its residualized version of FREQDIFFWITHOUTSEMCLOSEDIFF: more frequent adjectives precede less frequent adjectives. Figure 1 shows that, on the whole, the NNS adjective orderings are also compatible with the hypothesized effect of FREQDIFFWITHOUTSEMCLOSEDIFF, but more precise exploration revealed that this is really only true for the German NNSs. This is because, of 169 + 154 = 323 nativelike Chinese NNS examples (cf. the top row), 169 (52.3%) actually exhibit a value of FREQDIFFWITHOUTSEMCLOSEDIFF that is <0 and, thus, dispreferred. By contrast, of the 79 + 167 = 246 nativelike German NNS examples (cf. the second row from the top), only 79 (32.1%) exhibit a dispreferred value of FREQDIFFWITHOUTSEMCLOSEDIFF < 0 — the majority of 167 (67.9%) examples exhibit the hypothesized positive value of FREQDIFFWITHOUTSEMCLOSEDIFF. This interpretation is supported by looking at the lower half of Figure 1: Of the 73 + 6 = 79

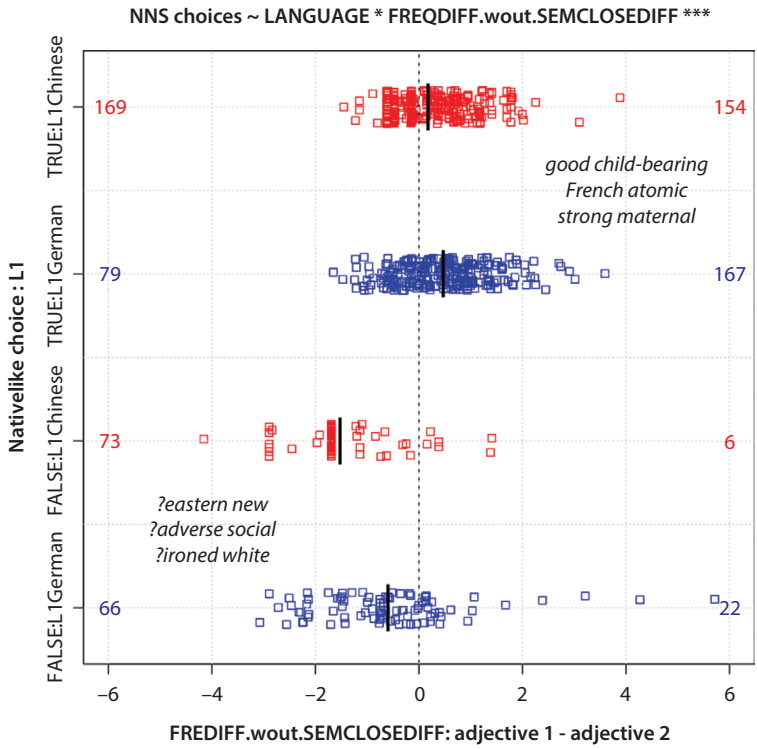

Figure 1. Strip chart representing the interaction LANGUAGE:FREQDIFFWITHOUTSEMCLOSEDIFF for the NNS corpus examples grouped by CORRECT:LANGUAGE

non-nativelike choices that the Chinese NNSs made altogether, 73 (92.4%) exhibit the dispreferred negative value of FreqDiffWithoutSemCloseDiff, compared to the still sizable but lesser 66 (75%) out of 66 + 22 = 88 examples of the German NNSs. Thus, the German NNS choices are more compatible with the effect of FreqDiffWithoutSemCloseDiff observed for NSs than the Chinese NNSs.

Consider now Figure 2 for the interaction of Language:NounSpecFreqDiff. R_1 showed that NounSpecFreqDiff has the hypothesized effect in the NS data: adjectives that are less connected to/frequent with the relevant noun tend to precede adjectives that are more connected to/frequent with the relevant noun. Figure 2 shows that the NNS adjective orderings are also compatible with this effect because the majority of points in the upper two rows of Figure 2 — where the NNSs made the same choice as predicted for the NSs (from R_1) — are on the right side of the plot, which is where NounSpecFreqDiff ≥ 0.

For instance, of the 76 + 165 = 241 nativelike Chinese NNS examples in non-zero conditions, 165 (68.5%) had the hypothesized value of NounSpecFreqDiff > 0; of the 53 + 132 = 185 nativelike German NNS examples, 132 (71.3%) had the

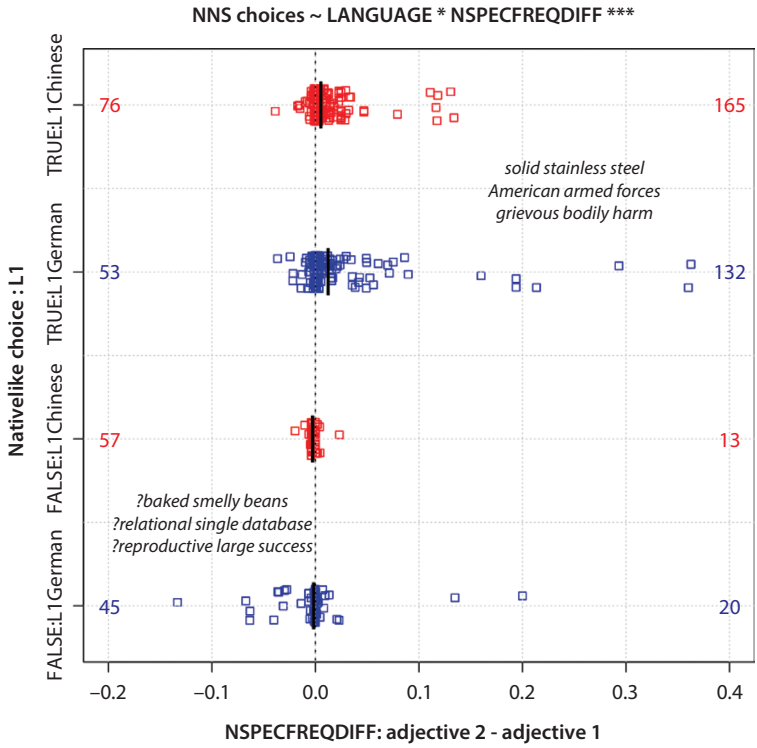

Figure 2. Strip chart representing the interaction Language:NounSpecFreqDiff for the NNS corpus examples grouped by Correct:Language

hypothesized value of NounSpecFreqDiff ≥ 0. While these percentages seem to indicate little difference between the two NNS groups, this impression changes when the non-nativelike choices of the NNSs in the bottom two rows are explored. Of the 57 + 13 = 70 non-nativelike choices that the Chinese NNSs made altogether, 57 (81.4%) exhibit dispreferred values of NounSpecFreqDiff < 0 whereas, of the 45 + 20 = 65 non-nativelike choices that the German NNSs made altogether, a lesser 45 (69.2%) exhibit dispreferred values of NounSpecFreqDiff < 0. Thus, the German NNS choices are more compatible with the effect of NounSpecFreqDiff observed for NSs than the Chinese NNSs.

Let us now turn to the interaction Language:IndCompDiff. R_1 showed that NS choices are on the whole compatible with the hypothesized effect of IndCompDiff: more gradable adjectives precede less gradable ones. Figure 3 reveals that the NNS adjective orderings are also compatible with the hypothesized effect of IndCompDiff. This is because the majority of points in the upper two rows of Figure 3 — where the NNSs made the same choice as predicted for the NSs (from R_1) — are on the right side of the plot, which is where IndCompDiff ≥ 0.

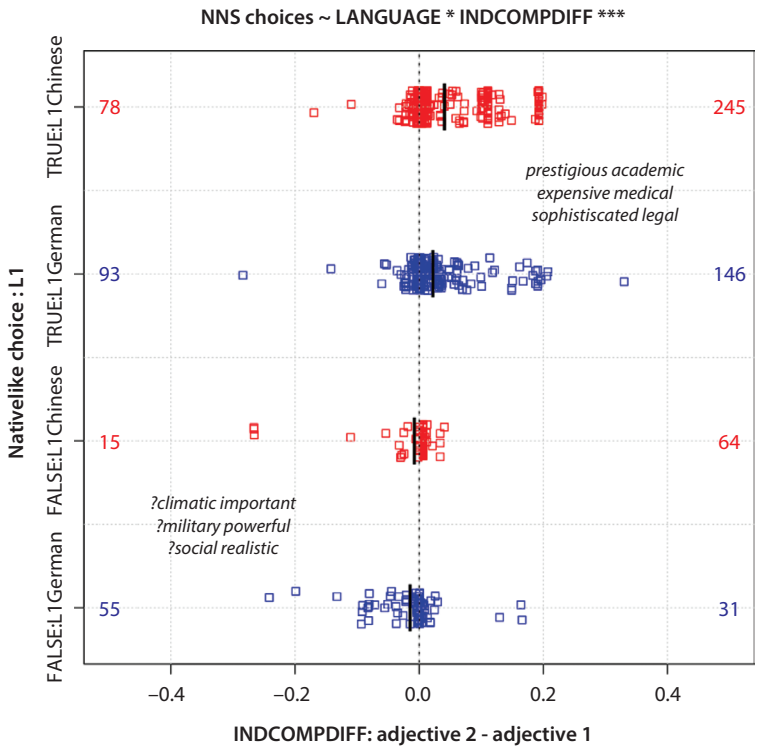

Figure 3. Strip chart representing the interaction Language:IndCompDiff for the NNS corpus examples grouped by Correct:Language

For instance, of the 78 + 245 = 323 nativelike Chinese NNS examples, 245 (75.8%) had the hypothesized value of IndCompDiff > 0; compared to that, of the 93 + 146 = 239 nativelike German NNS examples, only 146 (61.1%) had the hypothesized value of IndCompDiff > 0. Comparing these percentages shows that the choices of the Chinese NNSs are more compatible with the way NSs appear to react to IndCompDiff, an impression that is supported by the fact that, of the 86 non-nativelike choices that the German NNSs made altogether, 55 (64%) come with an IndCompDiff value that is, counter to expectation, <0, whereas such dispreferred values are only observed in 15 (19%) of the Chinese NNS data.

In Figure 4, we represent the interaction Language:SegAltDiff.

R_1 showed that NS choices are compatible with the hypothesized effect of SegAltDiff: adjective orderings that result in a more ideal syllable structure are preferred over those that result in a less ideal syllable structure. Figure 4 reveals that, on the whole, the NNS adjective orderings are also compatible with the expected effect of SegAltDiff, but more precise exploration shows that this is particularly true for the Chinese NNSs. This is because 87 of their 38 + 87 = 125

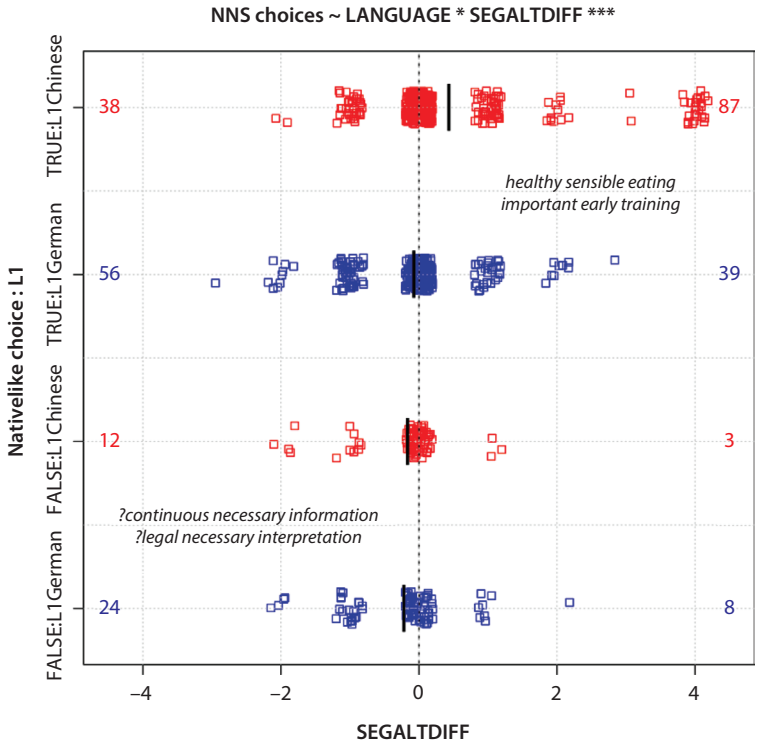

Figure 4. Strip chart representing the interaction Language:SegAltDiff for the NNS corpus examples grouped by Correct:Language (jittered along the *x*-axis)

nativelike choices in non-zero conditions (69.6%) exhibit the hypothesized (and, for the NS, observed) ≥ 0 values of SegAltDiff whereas only 39 of the 56 + 39 = 95 nativelike choices (41.1%) of the German NNSs do so, too. The findings regarding the non-nativelike utterances are less clear: the Chinese NNSs exhibit a slightly higher proportion of dispreferred negative values of SegAltDiff than the German NNSs (80% vs. 75%) but the mean SegAltDiff value for the German NNS is smaller. However, with regard to the non-nativelike utterances, the findings should be interpreted with some caution given the small number of cases. Still, as with IndCompDiff, the Chinese NNS choices are more compatible with NS preferences.

Figure 5 represents the interaction Language:AffLoadDiff.

This variable is interesting because R_1 showed that the NS choices are *not* compatible with the hypothesized effect of AffLoadDiff: Previous literature argued that AO should result in positive values of AffLoadDiff as operationalized here, but R_1 reveals that the NSs, if anything, exhibit the opposite tendency: there is a significant effect in R_1 showing that NSs place more negative adjectives before

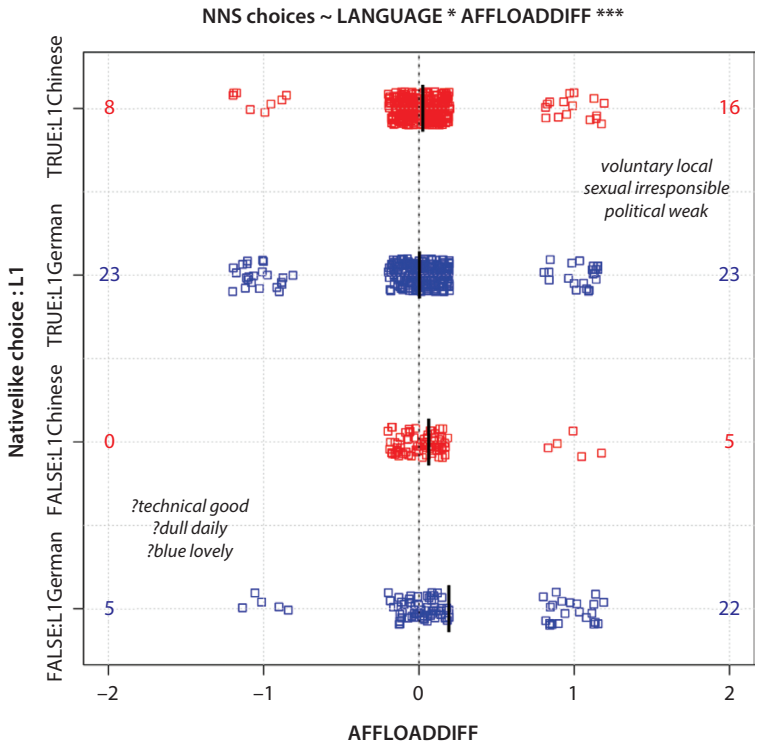

Figure 5. Strip chart representing the interaction Language:AffLoadDiff for the NNS corpus examples grouped by Correct:Language (jittered along the *x*-axis)

more positive ones. Interestingly, however, Figure 5 shows that the NNSs do not make choices that are compatible with the NS choices — as with all previous predictors — but they make choices that are compatible with the effect as postulated in the literature: unlike the NSs, the NNSs put the more positive adjective first, and this is particularly true of the Chinese NNSs: 16 of their 24 nativelike choices (66.7%) have the hypothesized positive values of AffLoadDiff, but all of their non-nativelike choices (5 out of 5) do as well. For the German NNSs, half of their nativelike choices have the hypothesized positive values of AffLoadDiff, and 22 of their 27 non-nativelike choices (81.5%) do as well. Thus, while neither NNS group follows NS preferences, the Chinese NNSs make more choices compatible with AffLoadDiff (72.4%) than the German NNSs (61.6%).

Let us conclude by briefly discussing the single main effect, i.e. a predictor whose effect did not differ significantly between the two levels of Language: SemCloseDiff; consider Figure 6. R_1 showed that the NS choices are *not* compatible with the hypothesized effect of SemCloseDiff: Previous literature argued that AO should result in positive values of SemCloseDiff as operationalized here,

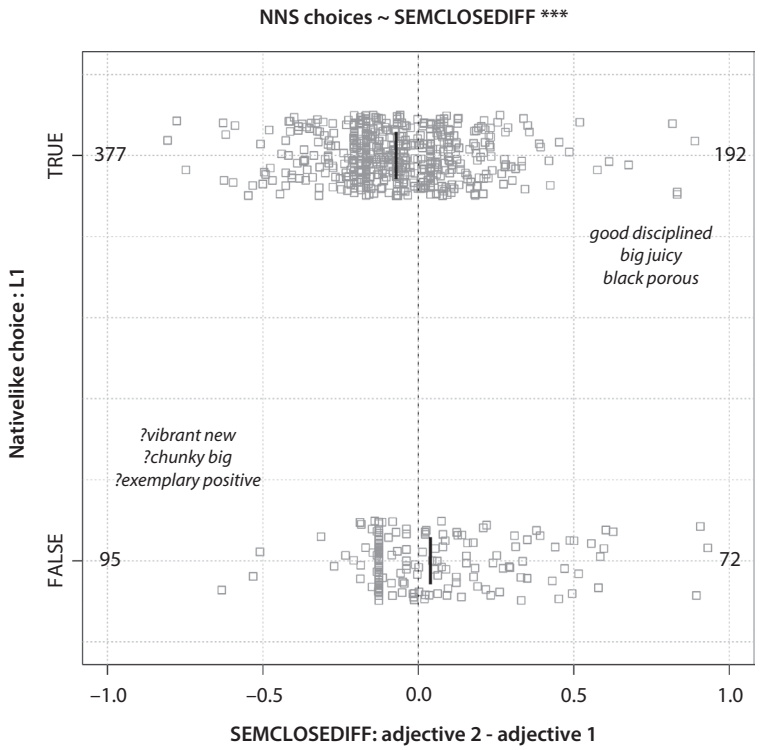

Figure 6. Strip chart representing the main effect SemCloseDiff for the NNS corpus data grouped by Correct

but R_1 reveals that the NSs exhibit, if anything, the opposite tendency. As for the NNSs, they again do not follow the NS patterning but instead the suggestions made in previous research: both NNS groups alike — which is why this predictor does not interact with LANGUAGE — put the more versatile adjective before the less flexible one more often (472 times) than not (264 times).

In the next section, we will discuss the above results from a broader perspective and conclude.

5. Concluding remarks

In summary, the MuPDAR analysis revealed that overall, the intermediate-advanced level Chinese and German learners captured in the ICLE corpora are quite well-aligned with native speakers' AO preferences. A closer look at the interactions featuring L1 as a predictor showed that the two learner groups deviate from the native speaker preferences in different ways. The German learners seem generally better aligned with regard to frequency-related predictors compared to their Chinese peers. Conversely, the Chinese learners have a head start with regard to the effect of gradability on ordering preferences, and they seem more sensitive to segmental alternation constraints. These findings are reminiscent of those we found for the genitive alternation in the same two learner groups (cf. Gries & Wulff 2013), so it appears that a systematic L1-specific pattern emerges here. As to the deeper causes for this pattern, we can only speculate. Maybe the German learners receive denser and high-quality input that allows their interlanguage system to be more adequately fine-tuned to these frequency effects than their Chinese peers do — a hypothesis that requires empirical validation, however. Finally, the observation that the German learners behave less native-like than their Chinese peers with regard to segment alternation invites speculation about the underlying cause being one of negative transfer from the L1: German arguably permits a higher number of (different and complex) consonant clusters than Chinese, which "does not have consonant clusters" (Li & Thompson, 1981, p. 3).

Two predictors, Affective Load and Semantic Closeness, pattern such that while the learners behaved in accord with previous literature on how these variables affect AO, the native speakers exhibited significant tendencies in the opposite direction: first, for Affective Load, the learners (and especially the Chinese learners) preferred positively loaded before negatively loaded adjectives (as hypothesized in the literature), while the native speakers exhibited a significant preference for negatively loaded adjectives to precede positively loaded ones. We have no straightforward explanation for this finding; bearing in mind that we here contrasted native speaker preferences in both oral and written language with learner

preferences in written language, we might be seeing a register effect here. Second, for Semantic Closeness, the native speakers also did not adhere to the directionality of the variables as postulated in previous studies: rather than preferring the semantically less versatile adjective to be closer to the head noun, they preferred it to be realized as adjective$_1$. The learners, in contrast, were in accord with the previous literature. Accounting for the fact that the native speakers do not behave as expected is challenging. However, our corpus-linguistic operationalization of Semantic Closeness cannot be responsible for that because the learners indeed display preferences that are captured quite well by our measure; also, Semantic Closeness appears to be relevant when it comes to positioning adjectives or relative clauses before nouns in Chinese (cf. Li & Thompson, 1981, p. 119).

All in all, we hope to have shown how the MuPDAR approach aids in uncovering some very interesting results: some point to differences between native and non-native speakers, some point to differences between different learner groups, and most give rise to new research questions at much higher levels of granularity and specificity than research based on over-/underuse counts can afford. Such follow-up studies may include further corpus analyses (of, say, corpus data from different registers, a wider range of proficiency bands, and/or different L1s) and/or experimental validation.

One interesting follow-up exploration, for example, is invited by exploring the results of R_1 to the NNS data. On the whole, the Chinese NNSs make significantly more nativelike choices than the German NNSs ($p_{\text{chi-squared test}} = 0.038$). While this result may suggest a higher degree of proficiency of the Chinese NNSs, there is an alternative hypothesis, one that would in fact be correlated with a lower degree of proficiency of the Chinese NNSs: Maybe the Chinese NNSs' language is more formulaic, in the sense that they (re-)use a smaller number of adjective, noun, and adjective-noun types in all their tokens than the Germans. Indeed that seems the case, as is indicated in Table 8: Even though the sample size is larger for the Chinese NNSs, each of the slots in an Adj-Adj-N triple — adjective$_1$, adjective$_2$, the N slot — and all combinations of adjectives and nouns are more diverse and less predictable in the German data.

In other words, the Chinese NNSs make more nativelike choices, but they do not employ the full range of nativelike choices — instead, they recycle a much

Table 8. Type/token frequencies/entropies for slots in Adj-Adj-N triples in the NNS data

Language (tokens)	Adj$_1$ (types/H)	Adj$_2$ (types/H)	Noun (types/H)	Adj N (types/H)
Chinese (402)	97 / 5.15	80 / 4.25	110 / 5.26	175 / 5.92
German (334)	216 / 7.41	230 / 7.51	243 / 7.66	319 / 8.29

smaller inventory of types. This is compatible with how four Chinese native speakers described their English learning experience in Chinese schools to us: to a considerable extent, classroom instruction involved rote-learning lists of adjective-noun collocations. Furthermore, collocational information appears to motivate not only their English ordering preferences: when we asked them to translate examples from English into Chinese and tell us which ordering they would prefer in Chinese, they would more often than not justify their preference with reference to collocations and fixed expressions in their L1 (Simpson (submitted) reports similar findings for Korean).

Additional areas of exploration involve the specific adjectives for which NNSs make non-nativelike choices. Space precludes an exploration of this topic, but it is easy to cross-tabulate the adjectives NNSs use with the variable CORRECT, i.e. whether they use them nativelike or not, which then allows for very specific post-hoc exploration. For example, one can see that every token of the Chinese speakers' use of *potential*, *Eastern*, *modern*, and *weak* is used in a non-nativelike way, and again many of these involve identical Adj-Adj-N triples. Yet another feature of applying R_1 to the NNS data is that one can also determine the degree to which a NNS choice is non-nativelike. This is because the NS prediction for the NNS data come in the form of predicted probabilities, which, as with nearly all regressions, are then dichotomized at the cut-off point of 0.5. That means, one can compare whether a NNS made a choice that, given the data, is far from clear-cut for a NS (because the predicted probability of the chosen order is around 0.5) or whether a NNS made a choice that is clearly dispreferred for a NS (because the predicted probability for the chosen order is less than, say, 0.2). The Chinese NNSs' uses of *potential* are only a few percent points off the NS prediction because the predicted probability for the NS to use that same ordering (*potential new customer*) is quite high (0.463). By contrast, the Chinese NNSs' uses of *Eastern* are quite off the mark because the predicted probability of the NSs to use that same ordering (*Eastern new territory*) is quite low (0.164). Thus, MuPDAR offers enormous potential for extremely fine-grained exploration of the data.

From a more global perspective, the above shows that the present study also provides yet another illustration of the fact that native speaker knowledge is probabilistic and multifactorial in nature, and that it can only be studied in precisely such ways. When seen in complementation of recent studies of different alternations such as the genitive alternation (Gries & Wulff, 2013), the choice of modal verbs (Gries & Deshors, 2014), or subject realization in Japanese (Gries & Adelman, 2014), to give but a few examples, the present study also underscores the point that NS knowledge underlying different phenomena is differently predictable. This implies that there is no 'native speaker norm' in the strict sense — rather, if comparisons with learner behavior are intended, what constitutes native-like

behavior has to be identified quantitatively and phenomenon-specifically. This is what the first regression in the MuPDAR approach serves to do.

In the same vein, the present study demonstrated NNS knowledge to be probabilistic and multifactorial in nature, and it is likewise differently predictable depending on the phenomenon under investigation. In order to uncover differences between NSs and NNSs, learner corpus research must answer the question "What would a native speaker do in a comparable situation?", where *comparable* must be described multifactorially — simply assuming that taking data from similar speech situations will license comparisons is insufficient. More specifically, the differences between NSs and NNSs, and/or different groups of NNSs, have to be established on a multifactorially-described, case-by-case basis rather than on the basis of decontextualized counts as they still prevail in contemporary learner corpus research. This is what the second regression of the MuPDAR approach serves to do.

It goes without saying that a learner corpus study is not the methodological answer to all questions in SLA research. Certain questions, including many regarding bilingual processing and mental representation, are more adequately addressed with experimental methods. However, we hope to have illustrated how a multifactorial corpus study is particularly useful in the context of variation-related questions such as accounting for learners' choices with regard to alternation phenomena. Furthermore, the results of the present corpus study may aid in the formulation of more pointed follow-up research questions that are most adequately addressed experimentally.

Last but not least, we would like to close with a brief note on the relationship between corpus linguistics as a method and its compatibility with different theoretical positions regarding L2 learning. The observant reader may have noticed that we do not attempt to account for the results we report here in terms of any one language learning theory. We deliberately do not engage in theoretical interpretations here because our main goal was to illustrate the potential of methodologically sophisticated learner corpus research. While we would personally lean towards a cognitive-functional account of language (learning), we would like to explicitly emphasize here that in our view, corpus linguistics is a method, not a theory, and as such equally compatible with *any* theoretical vantage point that allows for quantifiable operationalization of its hypotheses or predictions. We thus hope that this case study inspires researchers to include corpus linguistics in their methods tool box regardless of their theoretical standing — after all, any theory is only as good as the data (analysis) it rests on.

Acknowledgements

We would like to thank Hanlu Gong, Jing Li, Yuan Linfang, and Jingjing Zhao for their input regarding adjective ordering in Chinese.

References

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511801686

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bock, J.K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formation. *Psychological Review*, *89*, 1–47. DOI: 10.1037/0033–295X.89.1.1

Boucher, J., & Osgood, C.E. (1969). The Polyanna Hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*, 1–8. DOI: 10.1016/S0022–5371(69)80002–2

Couper-Kuhlen, E. (1986). *An introduction to English prosody*. London and Tübingen: Edward Arnold & Niemeyer.

Deshors, S.C. (2012). *A multifactorial study of the uses of may and can in French-English interlanguage*. Unpublished Ph. D. dissertation, University of Sussex.

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik*, *55*(3), 273–291. DOI: 10.1515/zaa.2007.55.3.273

Gries, St.Th. (2013). *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin and New York: De Gruyter Mouton.

Gries, St.Th., & Adelman. A.S. (2014). Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms* (pp. 35–54). Cham: Springer.

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

Gries, St.Th., & Wulff, S. (2005). Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, *3*, 182–200. DOI: 10.1075/arcl.3.10gri

Gries, St.Th., & Wulff, S. (2009). Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, *7*, 164–187.

Gries, St.Th., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, *18*(3), 327–356. DOI: 10.1075/ijcl.18.3.04gri

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. *Proceedings of the 9*th *Conference of the European Chapter of the Association for Computational Linguistics*, 30–36.

Li, C.N., & Thompson, S.A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley and Los Angeles: University of California Press.

Lockhart, R.S., & Martin, J.E. (1969). Adjective order and the recall of adjective-noun triples. *Journal of Verbal Learning and Verbal Behavior*, *8*, 272–275. DOI: 10.1016/S0022–5371(69)80075–7

Martin, J.E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, *8*, 697–704. DOI: 10.1016/S0022–5371(69)80032–0

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. DOI: 10.1017/S0267190512000098

Plonsky, L. (2013). Study quality in SLA: As assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. DOI: 10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*.

Posner, R. (1986). Iconicity in syntax. The natural order of attributes. In P. Bouissac, M. Herzfeld, & R. Posner (Eds.), *Iconicity. Essays on the nature of culture* (pp. 305–337). Tübingen: Stauffenburg.

Richards, M.M. (1977). Ordering preferences for congruent and incongruent English adjectives in attributive and predicative contexts. *Journal of Verbal Learning and Verbal Behavior*, *16*(4), 489–503. DOI: 10.1016/S0022–5371(77)80042-X

Schlüter, J. (2003). Phonological determinants of variation in English: Chomsky's worst possible case. In G. Rohdenburg, & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (pp. 69–118). Berlin and New York: Mouton de Gruyter.

Shih, S., Grafmiller, J., Futrell, R., & Bresnan, J. (to appear). Rhythm's role in genitive construction choice in English. In: R. Vogel & R. van de Vijver (Eds.), *Rhythm in phonetics, grammar and cognition*. Berlin: De Gruyter Mouton.

Simpson, H.E. (submitted). Structural, social, and cognitive factors driving adjective order in Korean: A multifactorial corpus analysis.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocations: A multi-study perspective. *The Canadian Modern Language Review/La revue canadienne des langues vivantes*, *64*(3), 429–458. DOI: 10.3138/cmlr.64.3.429

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernadini, & D. Stewart (Eds.)., *Corpora and language learners* (pp. 45–66). Amsterdam: John Benjamins. DOI: 10.1075/scl.17.05ton

Vennemann, T. (1988). *Preference laws for syllable structure and the explanation of sound change. With special reference to German, Germanic, Italian and Latin*. Berlin and New York: Mouton de Gruyter.

Whorf, B.L. (1945). Grammatical categories. *Language*, *21*, 1–11. DOI: 10.2307/410199

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, *8*(2), 245–282. DOI: 10.1075/ijcl.8.2.04wul

Authors' addresses

Stefanie Wulff, Ph.D. (corresponding author)
Linguistics Department
University of Florida
4131 Turlington, Box 115454
Gainesville, FL 32611–5454
Phone: 352-294-7455

swulff@ufl.edu

Stefan Th. Gries, Ph.D.
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106–3100

stgries@linguistics.ucsb.edu