Applied Psycholinguistics (2021), **42**, 279–299
doi:10.1017/S014271642000048X

CAMBRIDGE
UNIVERSITY PRESS

ORIGINAL ARTICLE

Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering

Stefan Th. Gries[1,*] and Stefanie Wulff[2]

[1]UC Santa Barbara/JLU Giessen and [2]University of Florida/UiT The Arctic University of Norway
*Corresponding author. Stefan Thomas Gries, Department of Linguistics, 3506 South Hall, University of California Santa Barbara, Santa Barbara, CA 93106-3100. E-mail: stgries@linguistics.ucsb.edu

(Received 10 January 2020; revised 18 August 2020; accepted 20 August 2020; first published online 04 December 2020)

ABSTRACT
This study examines the variable positioning of a finite adverbial subordinate clause and its main clause with the subordinate clause either preceding or following the main clause in native versus nonnative English. Specifically, we contrast causal, concessive, conditional, and temporal adverbial clauses produced by German and Chinese learners of English with those produced by native speakers. We examined 2,362 attestations from the Chinese and German subsections of the International Corpus of Learner English (Granger, Dagneaux, Meunier, & Paquot, 2009) and from the Louvain Corpus of Native English Essays (Granger, 1998). All instances were annotated for the ordering, the subordinate clause type, the lengths of the main and subordinate clauses, the first language of the speakers, the conjunction used, and the file it originated from (as a proxy for the speaker producing the sentence so as to be able to study individual and lexical variation). The results of a two-step regression modeling protocol suggest that learners behave most nativelike with causal clauses and struggle most with conditional and concessive clauses; in addition, learners make more non-nativelike choices when the main and subordinate clause are of about equal length.

Keywords: Chinese learners; English adverbial clauses; German learners; learner corpora; regression modeling

As the contributions to this Special Issue illustrate, we have made quite some headway since Dörnyei (2005, p. 7) offered the first formal definition of individual differences between learners as a "rather loose" concept that included "certain core variables and many optional ones." A few individual differences that have been receiving increasing attention in research are cognitive predispositions such as executive function and working memory (Wen, Borges Mota, & McNeill, 2015), declarative versus procedural memory (e.g., Hamrick, 2015; Morgan-Short, Faretta-Stutenberg, Brill-Schuetz, Carpenter, & Wong, 2014), aptitude (e.g., Granena, 2013, 2016), and resting qEEG activity (Prat, Yamasaki, Kluenda, & Stocco, 2016). Likewise, there is a solid number of studies examining differences in personality types, learning styles, and learning strategies (e.g., Grey, Williams, & Rebuschat, 2015). Also under scrutiny for their potential contributions to individual differences between learners are the learners' external circumstances and environment, including the amount of input they receive (e.g., Unsworth, 2016), the quality of the input (e.g., Rothman & Guijarro-Fuentes, 2010), and the specific learning context (e.g., Collentine & Freed, 2004; Grey et al., 2015). Several studies even consider the combined effects of internal and external factors (e.g., Chondrogianni & Marinis, 2011; Courtney, Graham, Tonkyn, & Marinis, 2017; Li, 2013; Sun, Streinkrauss, Tendeiro, & de Koot, 2016).

Corpus-based research appears well suited to contribute to this growing body of research, especially when it comes to investigating the other side of individual differences, so to speak, namely, the individual variation in performance that is clearly to be expected as a result of individual differences. The vast amount of production performance data a corpus makes available for a solid number of learners enables us to not only look at effects of individual differences in terms of global attainment outcomes as reflected in, say, overall accuracy or proficiency scores but also allows us (a) to look at very specific target structures and how they are used with variable accuracy by a learner, and (b) to look at the context conditions of each production. Especially the latter is a major advantage over language elicited in laboratory settings, where the diversity of contextual configurations is often kept low and/or constant and frequencies of exposure to settings are often artificially balanced. In other words, corpus data exhibit a much higher degree of ecological validity, and while they are admittedly noisier than experimental data, there are now more statistical approaches that allow us to handle such noise in meaningful ways.

However, the vast majority of individual differences and individual variation research to date is experimental, and only few studies are based on corpus data. As Kerz and Wiechmann (in press) state in their review of individual differences in corpus-based SLA research, "[a]lthough there is a tradition of studying the role of cognitive and affective IDs [individual differences] in SLA, their role has been a neglected area of research in L2 production studies"; notable exceptions are Dewaele and Furnham (2000), Kormos and Trebits (2012), and Möller (2017). We believe there are several reasons for this. For one, comprehension and processing-related questions are nearly impossible to investigate using corpus data, as most corpora provide production data only. Another reason likely is that even some corpus linguists treat corpora as massive pools of largely anonymized data points, and are not interested in relating data to the individual speakers who produced them. The biggest practical hurdle, however, seems to be that, to date, few learners of corpora are enriched with sufficient meta-data about the speakers they capture; even those that contain some meta-data only provide a small number of variables with speaker-specific information. Correspondingly, our main goal with this paper can only be to provide a mostly programmatic complement to the largely experimental studies that comprise this Special Issue.

The specific target structure we chose to examine here is the variable positioning of a finite adverbial subordinate clause and its main clause with the adverbial clause either preceding (SM) or following the main clause (MS). Specifically, we investigate four different types of finite adverbial clauses exemplified in (1)–(4): causal, concessive, conditional, and temporal adverbial clauses.

(1) a. Because he didn't each much, Max is hungry. [causal]
b. Max is hungry because he didn't eat much.
(2) a. Although Drew ate a lot, he is still hungry. [concessive]
b. Drew is still hungry although he ate a lot.
(3) a. If it doesn't snow, Fatih will go fishing. [conditional]
b. Fatih will go fishing if it doesn't snow.
(4) a. When Stefan visits Norway, he eats salmon. [temporal]
b. Stefan eats salmon when he visits Norway.

Regarding the factors that govern speakers' choice to either have the adverbial subordinate clause precede or follow their main clause, there is some research on first language (L1) speakers of English that we summarize in the next section, but no study to date (that we are aware of) has considered if and to what extent the same factors also play a role when second language (L2) speakers of English choose between the two possible orderings. The present study attempts to begin to close this gap by investigating data from Chinese and German L2 learners of English at the intermediate to advanced level of proficiency, and thus complements the authors' long-standing research agenda of examining alternation phenomena in learner language, including the genitive alternation (Gries & Wulff, 2013, Wulff & Gries, to appear), adjective order (Wulff & Gries, 2015), particle placement (Wulff & Gries, 2019), the double object alternation (Gries & Wulff, 2005, 2009), gerundial versus infinitival complementation (Martinez-Garcia & Wulff, 2012), and optional realization of *that*-complementizers (Wulff, Lester, & Martinez-Garcia, 2014; Wulff, Gries, & Lester, 2018); we chose Chinese and German learners mainly for comparability with our previous work that also often involved those two populations, and because the L1s are markedly different in their morphology and syntax.

This paper is structured as follows: in the following section, we provide a brief summary of previous research on adverbial clause ordering. We then describe how the corpus data was retrieved and annotated. Then we explain the statistical approach we employed called Multifactorial Prediction and Deviation Analysis Using Regression (MuPDAR; Gries & Deshors, 2014). We present our results before we close with a summary of our findings and a discussion of their implications.

Previous Research

Constituent order phenomena, of which the variable positioning of adverbial clauses is one example, have received considerable attention in previous research; for reasons of space, we here focus on more recent work (but see also Altenberg, 1984; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Diessel, 1996, 2001; Ford, 1993; Ford & Thompson, 1986; Quirk, Greenbaum, Leech, & Svartvik, 1985; Ramsay, 1987). Maybe the most comprehensive study to date is Diessel (2005), who presents results based on 2,034 attestations from conversation, fiction, and science writing. His main findings can be summarized as follows. The majority of attestations (1,252 out of 2,034) occurred sentence-finally. This distribution was reflected in all three genres, with the highest shares of the MS ordering in conversation (67.9%), followed by fiction (62.4%), followed in turn by science writing (56.3%). When adverbial clauses are in initial position, the adverbial clause tends to be shorter than the main clause (52.1% of the time), and only in 15.3% of all cases is the adverbial clause longer than the main clause; conversely, in cases where the adverbial clause is in final position, the adverbial clause is longer than the main clause in 36% of all cases, and shorter in 28% of all cases. Diessel interprets this distribution of lengths as evidence for a general observance of the short-before-long principle, which reflects speakers' general attempts to minimize cognitive effort in utterance planning. This default preference, however, can be overridden by discourse-pragmatic motivations. For example, the function of many adverbial clauses is to provide thematic background information to set the scene for upcoming information; this is particularly true for conditional clauses, and can likewise be true for temporal clauses when they describe a situation prior to the one encoded in the main clause. These functional considerations, then, can lead to speakers opting for an initial ordering of the adverbial clause, as the resulting ordering iconically mirrors the speakers' construal of events. Correspondingly, Diessel observes that conditional clauses are placed initially more often than temporal clauses, which in turn precede their main clause more often than causal clauses do (for a deeper analysis of the pragmatic functions of different types of adverbial clauses in initial and final position, see Diessel, 2008, 2013).

While adverbial clause ordering has received at least some attention in native speaker production data, there are only very few studies on how learners order adverbial clauses. To some extent, this reflects the comparatively lower number of studies on alternations in learner language more generally. Among the few exceptions is van Vuuren (2013), who examined L1 Dutch L2 English learners' use of adverbials, and found that they tend to overuse place and addition adverbials compared to native speakers, which is likely due to transfer of information-packaging preferences in Dutch. Van Vuuren and Laskin (2017) presented a more comprehensive study that examined learners' use of adverbials over time, with the interesting finding that while learners generally develop more nativelike usage patterns, their use of linking adverbials, specifically, becomes less nativelike with time (van Vuuren & Laskin speculate that this may be an instructional effect). Kerz (2013) closely examined L1 German L2 English learners' usage patterns of concessive adverbials; she found that while learners adopt the information-packaging constraints of concessives in a nativelike fashion, their constructional repertoire remains smaller than that of native speakers.

In the present study, we elaborate on previous research by (a) including not one, but two learner populations, namely Chinese and German learners; (b) expanding the scope to causal, concessive, conditional, and temporal adverbials; and most important in the context of this special issue, (c) presenting more complex statistical analysis that includes examination of individual variation.

Data Retrieval and Annotation

We retrieved exhaustive samples of finite adverbial clauses as in (1)–(4) from the International Corpus of Learner English (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009), specifically the Chinese section (C-ICLE; ~500,000 words) to represent Chinese intermediate to advanced learners of English as a L2, and the German section (G-ICLE; ~250,000 words) to represent German intermediate to advanced level learners of English as a L2. For a sample of native speaker data at least somewhat comparable to ICLE, we followed a fairly widespread practice and complemented the data from ICLE with attestations from the Louvain Corpus of Native English Essays (~300,000 words; Granger, 1998), which contains data from L1 speakers of English and who are responding to argumentative essay prompts similar to those that were used to solicit the data for ICLE. For data retrieval, we first ran concordance searches for the following adverbs (following Diessel, 2005): *after*, (*al*) *though*, (*as long*/*soon*) *as*, *because*, *before*, *now*/*so that*, *once*, *since*, *unless*, *until*, *when*, and *while*. We manually inspected the resulting 46,975 candidate hits to identify true hits of sentences with either ordering, thus, for instance, excluding adverbial clauses that interrupt the main clause as in (5), nonalternating adverbial clauses as in (6), nonfinite clauses as in (7), and subjectless and/or verbless adverbial clauses as in (8).

(5) Eloi went, while Jorge stayed in the lab, to get some food.
(6) The day when we have to make a decision is now.
(7) While speaking to Marit, Stefan took notes.
(8) When alive, spiders give Dave the heebie-jeebies.

From the resulting data sample with 9,218 true hits, we then identified speakers who contributed at least 8 attestations to the sample; speakers who contributed fewer than 8 attestations were excluded from the present study to strike a balance between this study's focus on individual differences and variation on the one hand (requiring a sufficient number of attestations per speaker) and a desire to retain as high an overall sample size as possible on the other hand. The final data sample thus included a total of 2,362 attestations, with 1,205 attestations by 68 L1 English speakers, 423 attestations by 35 L1 Chinese speakers, and 734 attestations by 53 L1 German speakers. Table 1 provides an overview. Only 1 out of all 156 speakers (1 of the 88 learners) contributed only examples of one semantic type; 93% of all learners contributed examples of three of the four semantic types.

Each attestation in the final data sample was then annotated for the following variables:

- ORDER: the ordering (*MS* vs. *SM*);
- SEMTYPE: the subordinate clause type (*causal*, *concessive*, *conditional*, or *temporal*);
- LENMC/LENSC: the numbers of words of the main and subordinate clause (including the subordinator), respectively.

Table 1. Overview of the sample composition

		SemType				
L1	Order	*Causal*	*Concessive*	*Conditional*	*Temporal*	Sum
CN	*MS*	134	4	86	38	262
	SM	36	21	77	27	161
EN	*MS*	388	39	98	183	708
	SM	59	53	215	170	497
GE	*MS*	209	18	91	147	465
	SM	52	23	88	106	269
Sum		878	158	655	671	2362

In addition, for each sentence we also included the variables L1 ([Native speaker] *NS/Eng* vs. [Nonnative speaker] *NNS/Chin* vs. *NNS/Germ*), Match (the conjunction used), and File (a proxy for the speaker producing the sentence).

In order to statistically control for any effects that speaker proficiency might have (the German learners' proficiency is considered to be higher than that of the Chinese learners) we created a variable to operationalize speaker proficiency to at least some degree. Specifically, Crossley, Salsbury, and McNamara (2011) found that lexical diversity is one of the strongest predictors of an individual's overall proficiency level, so we computed for each essay represented in our data 15 different lexical diversity scores, namely, all measures implemented in the R package::function quanteda::textstat_lexdiv (see Benoit et al., 2018). Then we performed a principal components analysis on the 15 measures, which indicated that the first principal component accounted for more than 77.6% of the variance of the 15 different measures (i.e., for more than 11.5 of the original 15 measures), so we used its principal component scores as a proxy of proficiency in a control variable LDComp1; reassuringly, the German learners do score significantly higher on this proficiency operationalization, as one would have expected from what is known about the corpus composition.

In the following section, we outline the statistical approach we used to analyze the data in more detail.

Statistical Analysis

As mentioned above, in this study we are using a relatively new approach called MuPDAR that has recently been developed for research on learner corpora or varieties of English, that is, research domains where it may be useful to consider the data as consisting of a "reference speaker" part and a "target speaker" part. In the present case of a learner corpus study, the "reference speakers" are the native speakers and the "target speakers" are the learners. Specifically, MuPDAR is similar in spirit to the technique of missing-data imputation and based on the following two questions:

1. Given the situation that the target speaker is in (as defined by linguistic, contextual, and maybe other variables), what would the reference speaker have done?
2. Did the target speaker make the same choice the reference speaker would have made and, if not, why not?

In order to address the above two questions, most learner corpus MuPDAR studies so far have proceeded as follows. First, one fits a regression (or some other classifier such as random forests) on reference speaker data to develop a hopefully good model of reference speaker behavior. Second, if that first model/classifier is sufficiently good, it is applied to the target speaker data to impute for every target speaker choice what a reference speaker would have done in the same situation. Third, one can then determine whether the target speaker made the same choice a reference speaker would have made, which can be quantified either in a binary (yes/no) format or numerically (how much does the target speaker's choice deviate from the reference speaker's imputed choice?). Fourth, a second model or classifier is fit to explore the (mis)match between the imputed target speakers' choices and the observed reference speakers' choices. In the context of a learner study, this amounts to determining which (combinations of) variables make learners make non-nativelike choices. Crucially, in both regressions or classifiers, one can use random effects (or their equivalents in other classifiers) to take individual speakers' or items' idiosyncrasies into consideration, as we will do below. In the next subsection, we discuss the application of this protocol to our data.

MuPDAR on Our Data

In order to make the data more amenable to regression modeling, we thoroughly explored and prepared them as follows. For the numeric predictors of LenMC and LenSC, for instance, we found them to be considerably right skewed, so we applied power transformations to them (using the optimal lambda and gamma values returned by the function car::powerTransform). We also computed a variable LenDiff, which is the difference LenMC minus LenSC, so the value of 0 means that both clauses are equally long. For the categorical predictors, no changes needed to be made.

For the random-effects variables, we conflated several levels of the variable Match in fairly straightforward ways: *till* and *until* were conflated, as were *while* and *whilst*, *cause* and *because*, *even if* and *if*, and a variety of different variants of *though* (*although*, *even though*, *though*, and *tho*). In addition, a variety of very low-frequency conjunctions including *insofar as*, *as soon as*, *so long as*, *so that*, *unless*, and *whenever* were grouped together as *other*; this conflation applied to a mere 54 out of 2,362 cases.

We then fit the first regression model of the MuPDAR protocol. In it, Order was the response variable, SemType and a polynomial (second degree) version of LenDiff were the fixed-effects predictors, and we had a maximal random-effects structure with varying intercepts and slopes of SemType and LenDiff for speakers (File) and varying intercepts and slopes of LenDiff for conjunctions (Match).

That model led to quite a good accuracy (80.8%) that is significantly better than baseline ($p_{\text{binomial test}} < 10^{-56}$), and an excellent *C* score of 0.9, which considerably exceeds the usual quality threshold of 0.8.

We then applied this model to the NNS data (using fixed effects and random effects for MATCH, but not for FILE, as the NNS speakers are different speakers) to predict for each NNS utterance that clause order a native speaker would have used in their place (given the predictors and MATCH). As these predictions were probabilities of SM, in a first step (used in all MuPDAR studies), if the predicted probability of SM was $\geq$.5, the prediction became SM; if the predicted probability of SM was <.5, the prediction became MS. A new second follow-up step will be discussed further below.

Compared to previous MuPDAR applications with intermediate to advanced learners, the predictions made by the model did not coincide quite as well with the actual NNS choices: whereas in previous MuPDAR applications, often >75% of the NNS choices were those predicted on the basis of the NS choices, here "only" 66.4% were. However, this is not a concern for three reasons: (a) it is precisely this kind of difference that MuPDAR was designed to explore, if all NNS choices were exactly like those of the NS, there would be no difference between the two speaker groups, the NNS would be perfectly nativelike; (b) most of the other alternations to which MuPDAR was applied were of a kind where, at least in some linguistic contexts, one of the options is strongly dispreferred or even ungrammatical (e.g., in the case of the genitive alternation, an *s*-genitive with a possessor containing a clause), whereas there is hardly any combination of factors that would categorically rule out an MS or an SM ordering; and (c) this alternation is much less understood than "the usual suspects" such as the dative alternation, particle placement, or the genitive alternation, for which there are hundreds of studies discussing literally dozens of predictors. Given that higher degree of flexibility coupled with our comparative lack of prior knowledge of the factors governing this alternation, a higher degree of mismatch between NS and NNS choices is only to be expected: there is less categorical patterning that learners can pick up on.

In order to quantify how much a NNS choice was nativelike, we then followed Gries and Deshors (2020) and computed the logloss statistic for all predictions but also, more important here, the contributions to logloss from each NNS choice. Logloss is a metric widely used in machine learning contexts to evaluate how well classifiers predict data that is based on two pieces of information:

- whether a classifier's prediction was correct: *yes* or *no*; and
- how confidently a classifier's prediction was made: very confidently (i.e., with predicted probabilities "far away" from the cutoff point of .5) or more tentatively (i.e., with predicted probabilities close to the cutoff point of .5).

The former is based on comparing the classifier's prediction to the observed reality; the latter is based on the predicted probability of the outcome. Specifically, logloss is computed as, in pseudocode, if prediction is correct, –log (predicted probability); if it is not correct, –log (1 – predicted probability). Thus, logloss (just like the Brier score) penalizes false classifications/predictions, and penalizes false classifications/predictions that are made confidently/boldly more. In other words, in terms of logloss, the worst predictions are confident/bold predictions that turn out to be wrong. The overall logloss statistic of a classifier is the average of the above pseudocode computations that are, for each case, what one might call the contribution to logloss from each case. For this analysis, we multiplied the contribution to logloss with −1 (making it positive) when the NNS chose SM (so that the direction of logloss is informative); this directional logloss then became the response variable of the second regression of MuPDAR.

In this second regression of the MuPDAR protocol, we fit a linear mixed-effects model with

- the logloss contribution of each case as the dependent/response variable;
- as fixed-effects predictors of interest L1, SemType, and LenDiff (as a polynomial to the 3rd degree to allow for some flexibility in its curvature) and all their interactions;
- as a fixed-effects control variable LDComp1 (as a polynomial to the 2nd degree to allow for some curvature);
- uncorrelated varying intercepts and slopes for File and for LenDiff; and
- varying intercepts for Match.

Results

We used a model selection process as proposed by Zuur, Ieno, Walker, and Saveliev (2009, Chapter 5), in which we first tried to identify the most useful random-effects structure (using REML estimation and LR tests), then tried to identify the optimal (in the Occam's razor sense) fixed-effects structure (using maximum likelihood estimation and LR tests with LDComp1 not being eligible for deletion), and finally refit the final model (using REML estimation again), which was then interpreted with effects plots (Fox & Weisberg, 2019). The following section will discuss the results of this process.

Overall model results and fixed effects

The above model selection process indicated that the random-effects structure could not be simplified without a significant loss of explanatory power. The fixed-effects structure, in contrast, was reduced by LR tests to a final model that, apart from the control variable LDComp1, contained the two interactions L1:LenDiff and SemType:LenDiff, with LenDiff remaining a polynomial to the third degree. This model is highly significant ($LR = 190.44$, $df = 21$, $p < 10^{-15}$), with a relative likelihood over a null model exceeding 10^{31}. The final model is also significant compared to a null model with just the control variable of LDComp1 ($LR = 182.12$, $df = 19$, $p < 10^{-15}$), with a relative likelihood over a null model exceeding 10^{30}. No overdispersion was observed, collinearity was moderate ($\kappa = 17.4$) but nearly exclusively restricted to the interaction of L1:LenDiff. However, the overall explanatory power of the model was not quite as good as hoped for: $R^2{}_{\text{marginal}} = .17$ and $R^2{}_{\text{conditional}} = .38$; the relevant numeric information regarding the final model is shown in Table 2 (subscripts *l*, *q*, and *c* refer to the linear, quadratic, and cubic parts of the polynomials).

Table 2. Overview of the final regression model

	CI_{lower}	*Estimate*	CI_{upper}	*SE*	$df_{Satterthwaite}$	*T*	*p*
(Intercept)	−0.361	−0.034	0.221	0.127	21.479	−0.263	.795
L1$_{chin\rightarrow germ}$	0.031	0.208	0.437	0.089	93.876	2.343	.021
SemType$_{caus/other}$	0.020	0.378	0.763	0.171	37.063	2.207	.034
SemType$_{temp/conx}$	−0.325	0.105	0.490	0.216	23.078	0.486	.632
SemType$_{conc/cond}$	−1.087	−0.532	−0.039	0.272	69.91	−1.958	.054
LENDIFF$_{l}$	−1.827	5.135	12.393	3.472	253.682	1.479	.14
LENDIFF$_{q}$	−7.004	0.110	5.775	3.081	1040.173	0.036	.972
LENDIFF$_{c}$	4.430	10.826	18.956	3.931	864.82	2.754	.006
LDComp1$_{l}$	−5.843	−2.659	0.281	1.394	84.339	−1.908	.06
LDComp1$_{q}$	0.009	2.406	4.590	1.054	85.11	2.284	.025
L1$_{chin\rightarrow germ}$:LenDiff$_{l}$	−20.703	−12.733	−4.043	3.648	159.651	−3.490	.002
L1$_{chin\rightarrow germ}$:LenDiff$_{q}$	2.906	8.716	14.283	2.923	1036.699	2.982	.003
L1$_{chin\rightarrow germ}$:LenDiff$_{c}$	−24.721	−15.145	−8.772	4.029	853.656	−3.759	<.001
SemType$_{caus/other}$:LenDiff$_{l}$	−14.533	−9.095	−4.669	2.398	1070.123	−3.793	<.001
SemType$_{temp/conx}$:LenDiff$_{l}$	−4.833	0.813	6.669	3.047	1031.73	0.267	.79
SemType$_{conc/cond}$:LenDiff$_{l}$	−11.242	−0.511	11.165	5.072	1050.924	−0.101	.92
SemType$_{caus/other}$:LenDiff$_{q}$	2.157	7.763	14.045	2.843	1103.474	2.730	.006
SemType$_{temp/conx}$:LenDiff$_{q}$	−8.588	−2.830	4.162	3.665	1011.238	−0.772	.44
SemType$_{conc/cond}$:LenDiff$_{q}$	−9.284	4.233	18.979	6.503	1085.563	0.651	.515
SemType$_{caus/other}$:LenDiff$_{c}$	−11.642	−7.395	−2.626	2.47	1079.76	−2.994	.003
SemType$_{temp/conx}$:LenDiff$_{c}$	−6.463	0.093	4.614	3.074	1088.77	0.030	.976
SemType$_{conc/cond}$:LenDiff$_{c}$	−14.074	−5.506	5.687	5.386	1079.795	−1.022	.307
Fixed-effects predictor	$LRT_{deletion}$	*npar*	*p*				
LdComp	7.887	2	.01938				
L1:LenDiff	32.656	3	<.0001				
SemType: LenDiff	53.084	9	<.0001				
Random-effects *SD*							
Intercepts for File	0.18531						
LenDiff slope for File	0.03438						
Intercepts for Match	0.33016						
Residual	0.77970						

We now first discuss the two significant fixed effects in general. Then we turn to comparing the performance of the speakers from the different L1s. Finally, we discuss the findings that involve speaker-specific results.

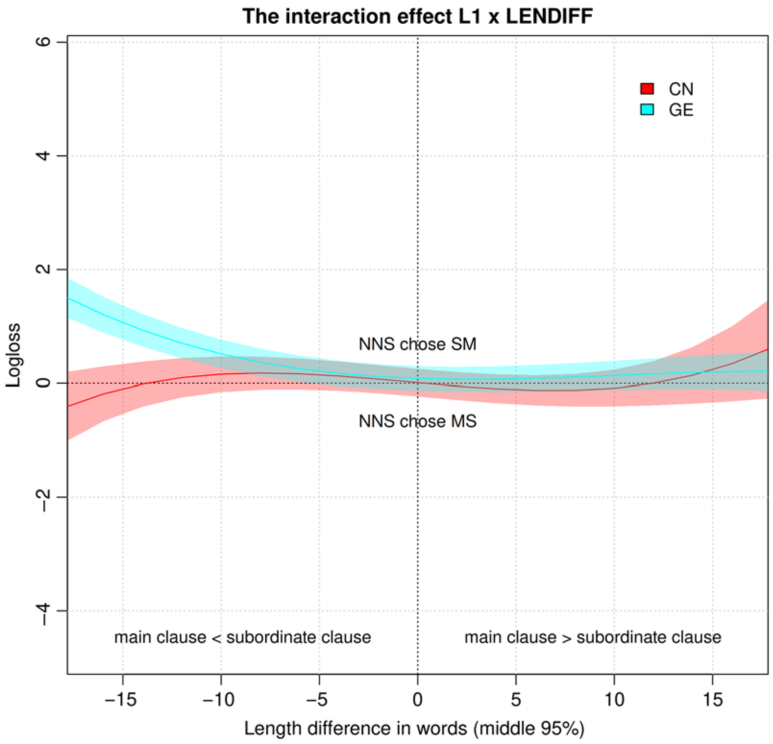

Figure 1. (color online) The effect of L1:LENDIFF on directional logloss.

The effect of L1:LENDIFF is shown in Figure 1. The *x*-axis represents the values of LENDIFF, the *y*-axis represents directional logloss, and the two curves represent the predictions for the Chinese and German learners with their 95% confidence bands.

The Chinese learners are predicted to behave very nativelike regardless of the length difference between the clauses: their confidence band always includes 0. The German learners, in contrast, behave differently depending on the length difference between the clauses:

- when the main clause is much shorter than the subordinate clause (on the left), they are predicted to exhibit high logloss values (i.e., use SM when they should have used MS);
- in all other cases, that is, when main clauses are approximately as long as or longer than the subordinate clauses, they are predicted to behave nativelike.

The effect of SEMTYPE:LENDIFF is shown in Figure 2. The overall coordinate system is the same as in Figure 1, but now the four lines represent the four clause types. The plot is harder to interpret (because of the overplotting), but the main conclusion is that the causal subordinate clauses are the ones that are associated with non-nativelike choices. More precisely, for just about all of the data, all predictions include the prediction of logloss = 0 in the confidence interval; however, with causal

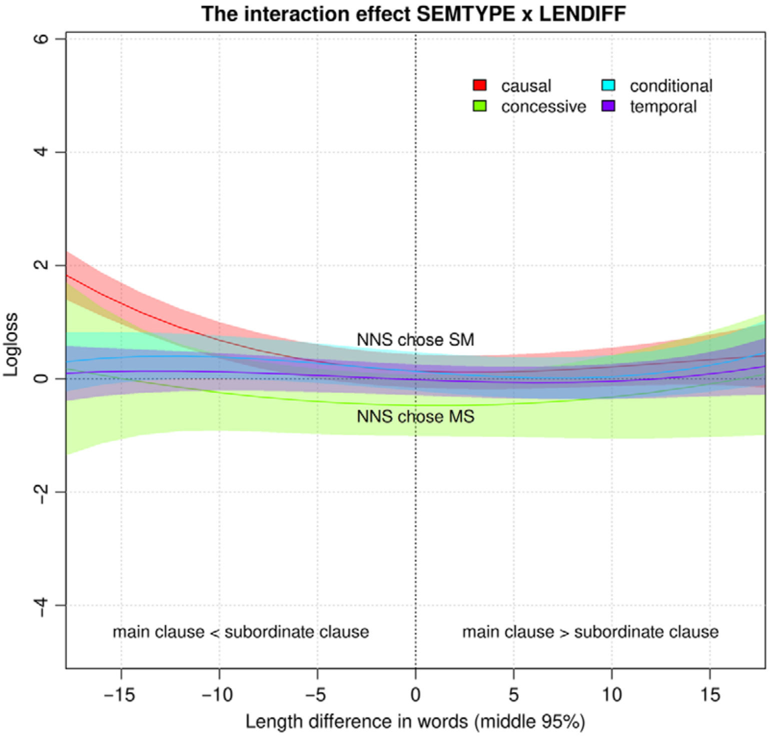

Figure 2. (color online) The effect of SEMTYPE:LENDIFF on directional logloss.

subordinate clauses that are considerably longer than their main clauses (i.e., the red curve on the left), the learners are predicted to choose SM even though they should position the longer subordinate clause after its main clause. This is unexpected because in the NS data, all four subordinate clause types, but especially the causal and the temporal clauses, exhibit notable (and expected) short-before-long effects. Accounting for this observation as transfer from the L1s seems rather unlikely: causal adverbials can occur both sentence-initially and sentence-finally in both Chinese and German, yet are reported to be predominantly placed clause-finally in both languages (Diessel & Hetterle, 2011). We can thus only interpret the learners' preference for sentence-initial placement as a genuine interlanguage phenomenon, which has also been reported in other studies (see, e.g., Paquot 2010, p. 177 for similar findings and discussion).

Comparisons of "nativelike choices" across L1

Nearly all previous MuPDAR studies considered the performance of speakers (from different L1 backgrounds) on the basis of, for instance, proportions of nativelike and non-nativelike NNS choices. The straightforward way to do this would be to cross-tabulate for each L1 how many NNS choices were identical to those predicted from the NS data, which could be represented in a table like Table 3; there is a bit of a

Table 3. Nativelikeness per L1 (based on binary NS prediction)

	Non-nativelike	Nativelike	Sum
L1: CN	154 (36.4%)	269 (63.6%)	423
L1: GE	235 (32%)	499 (68%)	734
Sum	389 (33.6%)	768 (66.4%)	1157

Table 4. Nativelikeness per L1 (based on ternary NS prediction)

	Non-nativelike	Nativelike	Sum
L1: CN	141 (33.3%)	282 (66.7%)	423
L1: GE	200 (27.2%)	534 (72.8%)	734
Sum	341 (29.5%)	826 (70.5%)	1157

difference such that the German learners' choices are nativelike slightly more often (odds ratio = 1.216).

However, as has occasionally been pointed out, proceeding this way might be overly stringent/harsh because, if a NS is predicted to use SM with a probability of .51 but the NNS used MS, then this is categorized as non-nativelike even though the NS was nearly as likely to use SM (.51) as MS (1 – .51 = .49). In addition, a binary classification of nativelike versus non-nativelike also means that such a NNS choice of MS would be classified just as non-nativelike as one where a NS was predicted to use SM with a probability of .99, even though in this latter case, the NNS choice seems much more egregiously non-native as the NS was extremely unlikely to produce MS (with a probability of 1 – .99 = .01).

In order to address this issue, Gries and Deshors (2020) proposed to recognize a middle-ground category, that is, cases where, while a NS obviously would make just one ordering choice, they would not blink if someone else made the other one. In other words, they argued in favor recognizing an "either" scenario, one where both ordering choices would seem perfectly alright to a NS. The question then of course becomes how to identify those cases. While Gries and Deshors (2020) used one log-loss unit as a threshold value for "either" cases, we will proceed differently here. When we applied the first regression model to the NNS data, we not only computed the predicted point probability of SM but also used bootstrapping to compute a 95% confidence interval for each prediction. We then considered a case an "either-ordering-would-be-acceptable" case if that confidence interval included the cutoff point of 0.5, which also means that when the NS prediction is "either," either NNS choice is considered nativelike. This changes Table 3 to Table 4, and the effect that the German learners are doing a bit better increases ever so slightly to an odds ratio of 1.335.

This kind of approach provides a slightly more realistic view by avoiding the "harshness" of the original binary categorization, but for individual variation, a look at the most fine-grained resolution, random-effects results from the final model is of course most instructive.

Speaker-specific results

The above results offer a general understanding of how the learners' ordering of clauses differs from that of the native speakers. However, it is also instructive to inspect the speaker-specificity of the results. Note that this is not just the case in the context of this special issue. Much like fixed-effects regression modeling might be subjected to model diagnostics involving influence measures (to determine which data points exert [unduly much?] influence on the overall outcome), random-effects results should be checked as part of model diagnostics (e.g., are they really approximately normally distributed as the vast majority of mixed-effects models in linguistics presupposes?). However, this often does not happen and, even more regrettably, random effects are often not explored/interpreted at all: researchers use them to get more robust/generalizable results, but do nothing else with them, neither visualization nor further exploration nor correlating them with predictors not included in the model (as in Miglio, Gries, Harris, Wheeler, & Santana-Paixão, 2013, where post hoc exploration of intercept adjustments revealed a significant correlation with geographical/dialectal speaker information).

Thus, given the amount of information that random effects can sometimes provide, they should always be explored a bit, not just in a context concerned specifically with individual variation. In this paper, we will look at speaker-specific results via the results from the random-effects structure for FILE. The simplest kind of representation that is at least sometimes provided is a dot chart of intercept or slope adjustments together with their 95% confidence intervals; this can at least be used to check for normality. Another possibility is a plot that we have not seen used anywhere, but which we feel is informative and shown in Figure 3.

The *x*- and *y*-axes represent the speaker-specific intercept adjustment (*x*-axis) and slope adjustment for LENDIFF (*y*-axis), and each red *C* and blue *G* represent a speaker's values. Any lines shown in Figure 3 show the confidence intervals of those adjustments that do *not* include 0; thus, there are no horizontal lines because the intercept adjustments of all speakers include 0, but the six vertical lines (for five German and one Chinese learners) indicate speakers whose adjustment to the slope of LENDIFF is so notable that its confidence interval does not include 0 anymore (GESA3007.txt [slope adjustment: 0.051] and GEAU4013.txt [0.048] in the top part of the plot, GESA5010.txt [–0.05], CNHK1369.txt [–0.084], GESA3003.txt [–0.073], and GEBA1031.txt [–0.084] in the lower part of the plot). This plot immediately indicates how many and which speakers differ significantly from the rest, in what way, and with how much uncertainty. While the ICLE does not offer much in terms of metadata, this is nonetheless instructive because of the six speakers with significant slope adjustments, the two speakers with positive adjustments are those that have spent time in an English-speaking country, whereas the four speakers with negative adjustments have not. While the limited number of significant adjustments and the lack of other instructive metadata render this observation merely speculative, it still serves as an example of what kind of post hoc exploration is feasible.

Another useful way to get an impression of speaker variability is by enriching the above kinds of effects plots in Figure 1 and Figure 2 with the corresponding regression lines for each speaker. Figure 4 is the result of enriching Figure 1 with the thin regression lines for all speakers on top of the confidence bands of the overall fixed effect as

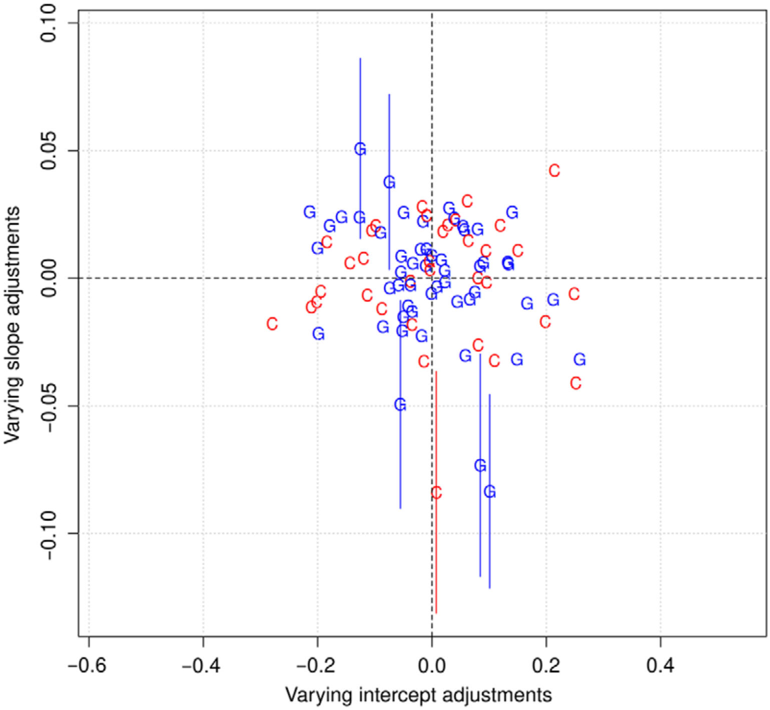

Figure 3. (color online) Intercept and slope adjustment for FILE.

shown in Figure 1. The six bold regression lines are for the speakers with significant slope adjustments.

Figure 4 shows us in a maybe more relatable fashion what the effect of the random effects for FILE are. For instance, the two highest blue lines (at $x = -17$ and $y > 2$) are speakers we above saw have notably different slopes for LENDIFF: the highest one is for GEBA1031.txt, the second highest one is for GESA3003.txt. Their lines "go down" more steeply because of the negative slope adjustment, reflecting that their reaction to LENDIFF is more incorrect especially with long subordinate clauses. A closer look at these two speakers shows that they have a few atypically long and thus convoluted (but not ungrammatical) productions. Examples (9) and (10) are of a long adverbial clause preceding a short main clause, which runs against the native speaker preference (at least based on LENDIFF alone).

(9) because political ideas can cause a lot of disagreement among different social groups and different mentalities, they separate people. [GEBA1031]

(10) the leader of the race, the spaniard miguel indurain, who is a member of the spanish "banesto" team, was attacked by the swiss alex zülle, who is a member of the "once" team, the spanish press accused the "once" team of treason. [GEBA1031]

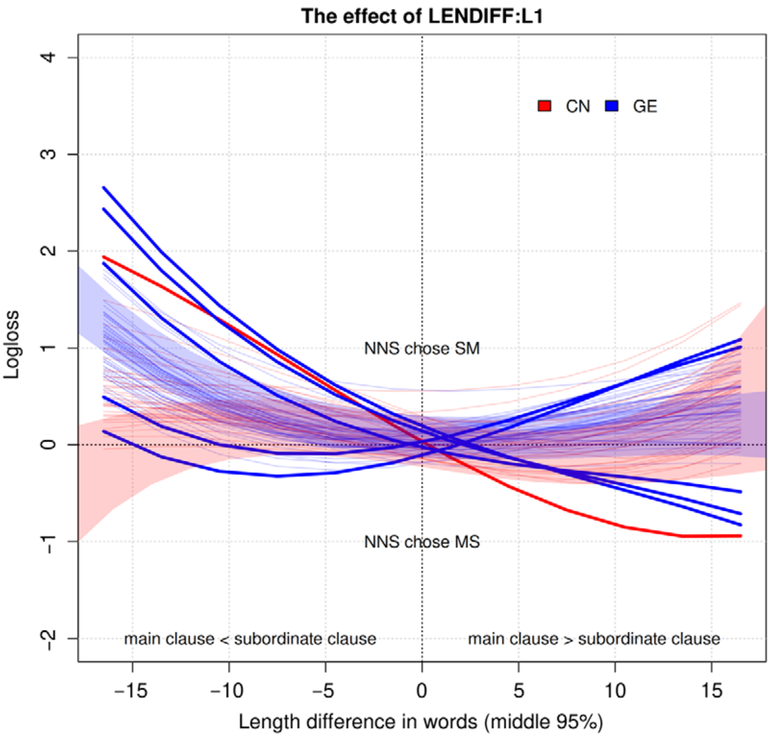

Figure 4. (color online) The effect of L1:LENDIFF on directional logloss.

At the same time, the two lowest bold blue lines are for GESA3007.txt and GEAU4013.txt. Their lines have positive adjustments, meaning the overall negative slope of LENDIFF is flattened for them, so to speak.

The one Chinese regression line is for CNHK1369.txt. On the basis of her metadata, the fact that her coefficients are so different is surprising. The vast majority of the Chinese learners in our data set are extremely homogeneous in terms of their metadata: most of them are female Cantonese-speaking women between 19 and 21 years with 13 years of English instruction in school, with as yet no years of English instruction at university and no other language spoken at home, and CNHK1369 is no different. Her proficiency score is also not worse than the average of the Chinese learners. That said, a closer look at her data reveals that she tends to make non-nativelike choices when the length difference between the main and subordinate clause is extreme. Example (11) is a long adverbial clause preceding a short main clause, so on the basis of LENDIFF, we would predict native speakers to prefer MS.

(11) as the report mentioned above said that breathing secondhand smoke can highly increase the risk of having lung cancers and diseases, it’s badly affected our health. [CNHK1369]

What is interesting about this particular example is that while it is technically an instance of SM, it is the long intervening relative clause that renders this example rather unique. The main clause (*it's badly affected our health*) is just a more compact summary of the content of the relative clause (*breathing secondhand smoke can highly increase the risk of having lung cancers and diseases*). One could argue that the non-nativelike nature of this production ultimately resides less in the ordering of adverbial and main clause, but in the interjection of the rather long relative clause.

The above discussion should serve to exemplify how the speaker-specific random-effects information from our modeling process can be used: to identify which speakers behave notably different, visualizing how their behavior is different compared to their learner group and all data, and exploring metadata or ultimately even the original corpus file for clues as to what might be responsible for the notable difference in behavior.

Discussion

While the above has exemplified one way of studying individual variation in corpus data, namely exploring speaker-specific predictions for relevant predictors, other avenues of research are theoretically possible, but often practically not easily available at this time; it is in this sense that the present study is programmatic in nature. Maybe the most dramatic gap, as we stated in our introduction, has to do with the lack of corpora that provide not only access to large amounts of production data by a sizable sample of language learners but also the pertinent individual difference measures. As a result of corpus compilation practices, this information about speakers cannot be added as predictors and/or controls in models/classifiers of corpus-based work. For instance, we usually do not have measures of learners' proficiency and even where, say, Common European Framework Reference for Languages levels of learners are provided, they are not precise or reliable enough to function as useful predictors, or their classificatory power is so small that they can often not even be predicted reliably from other text-specific indices. Approximating proficiency with text-based measures—for instance, lexical diversity, as in this study and in Wulff and Gries (to appear)—can be a useful heuristic, but is also just that, a heuristic. In addition, other kinds of metadata such as aptitude, motivation, working memory capacity, and so on, are usually not available.

Correspondingly, the present study faced the same problems: the amount of information on the individual speakers is relatively small/sparse and limited in terms of helping modeling power. As we mentioned above, the Chinese students that made it into our frequency-per-file-based sample were so homogeneous in terms of the metadata that were available that individual differences based on these metadata could not be expected.

These caveats and concerns notwithstanding, we hope the above shows at least programmatically that corpus data have the potential of offering insights about the role of individual variation in L2 acquisition research. If the relevant data sets are sufficiently large and comprehensively annotated, the right kind of statistical analysis can very well separate effects on the level of the overall sample from speaker-specific patterns, such as speakers whose behavior goes against well-established trends, speakers who react to certain (combinations of) characteristics in surprising ways, or theoretically, even speakers who react to certain lexical items in certain ways. However, that previous sentence already implies the biggest obstacles: the sizes of currently available corpora, the sampling schemes they are based on, and the degree of detail their metadata contain about speakers. While mixed-effects modeling and other statistical techniques can identify speaker-specific variability, the absence of rich metadata on speakers makes it quite hard for any analyst to determine which of the effects manifested in varying intercepts and/or slopes are truly individual effects or instead an effect of, say, working memory shared across many speakers.

In addition, the field needs to continue to develop the right ways of exploring these kinds of complex data sets. Too few learner corpus studies involve the right kind of statistical analyses, which can quickly become immensely complex. For many studies, some version of mixed-effects modeling might be required with the corresponding analysis of random-effects results, but other alternatives are conceivable. For example, individual variation can also be explored with a model-diagnostic tool that is often recommended even just for fixed-effects regression, namely, measures that quantify the influence that data points have on the (stability of the) regression coefficients. In fixed-effects regression, these are often computed by determining for every data point, how much the regression coefficients of the model change when it is deleted, and obviously, the more the regression results change as a result of omitting a certain data point, the more influential that data point is.

In mixed-effects models, the computation of such measures naturally takes into consideration the interdependence, or relatedness, of all the data points contributed by the same speaker. Thus, here influence measures require dropping each speaker, that is, level of FILE, from the analysis to see how the regression coefficients change in response to that omission from the data.

In addition and especially if we model observational data *and* include predictors quantifying individual cognitive factors, we need more modeling that can accommodate nonlinear effects as well as priming effects, adaptation, and so on; for very complex data sets such as those, the above approach of polynomial predictors might become too crude and generalized additive models may become necessary, which of course raises the bar considerably in terms of required expertise. However, given the combination of noisy and Zipfian-distributed observational data with many nonlinear predictors and controls, there are not many other options.

In summary, if we want to be able to see how individual differences are reflected in, and contribute to, individual variation in performance, the corpus developers among us have their work cut out for themselves. As we submit this paper, several research projects are under way that aim to deliver such data sets, including research by Gilquin and Laporte as well as Laporte and Gries, who are collecting non-native and native speaker essay-writing data, respectively, to be accompanied by measures of cognitive characteristics such as fluid intelligence, working memory span, inhibition, and different aspects of language aptitude. Another research project that promises to develop combined data sets of language production data and cognitive batteries is the Heritage-bilingual Linguistic Proficiency In their Native Grammar (HeLPING) project under the supervision of Jason Rothman. We look forward to the new avenues for analysis of both overall speaker behavior and individual variation that these and other future data sets will afford.

References

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica*, **38**, 20–69. doi: 10.1111/j.1467-9582.1984.tb00734.x

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, **3**, 774. doi: 10.21105/joss.00774

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

Chondrogianni, V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism*, **1**, 318–342. doi: 10.1075/lab.1.3.05cho

Collentine, J., & Freed, B. F. (2004). Learning context and its effects on second language acquisition. *Studies in Second Language Acquisition*, **26**, 153–171. doi: 10.1017/s0272263104262015

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, **29**, 243–263. doi: 10.1177/0265532211419331

Courtney, L., Graham, S., Tonkyn, A., & Marinis, T. (2017). Individual differences in early language learning: A study of English learners of French. *Applied Linguistics*, **6**, 824–847. doi: 10.1093/applin/amv071

Dewaele, J. M., & Furnham, A. (2000). Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365. doi: 10.1016/S0191-8869(99)00106-3

Diessel, H. (1996). Processing factors of pre- and postposed adverbial clauses. *Berkeley Linguistic Society*, **22**, 71–82. doi: 10.3765/bls.v22i1.1344

Diessel, H. (2001). The ordering distribution of main and adverbial clauses: A typological study. *Language*, **77**, 343–365. doi: 10.1353/lan.2001.0152

Diessel, H. (2005). Competing motivations for the ordering of main and adverbial clauses. *Linguistics*, **43**, 449–470. doi: 10.1515/ling.2005.43.3.449

Diessel, H. (2008). Iconicity of sequence. A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics*, **19**, 457–482. doi: 10.1515/cogl.2008.018

Diessel, H. (2013). Adverbial subordination. In S. Luraghi & C. Parodi (Eds.), *Bloomsbury companion to syntax* (pp. 341–354). London: Continuum.

Diessel, H., & Hetterle, K. (2011). Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. In P. Siemund (Ed.), *Linguistic universals and language variation* (pp. 21–52). Berlin: de Gruyter.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. New York: Routledge.

Ford, C. E. (1993). *Grammar in interaction. Adverbial clauses in American English conversations*. Cambridge: Cambridge University Press.

Ford, C. E., & Thompson, S. A. (1986). Conditionals in discourse: A text-based study from English. In E. C. Traugott, A. ter Meulen, J. Snitzer Reilly, & C. A. Ferguson (Eds.), *On conditionals* (pp. 353–378). Cambridge: Cambridge University Press.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). London: Sage.

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, **63**, 665–703. doi: 10.1111/lang.12018

Granena, G. (2016). Cognitive aptitudes for implicit and explicit learning and information-processing styles: An individual differences study. *Applied Psycholinguistics*, **37**, 577–600. doi: 10.1017/s0142716415000120

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London: Addison Wesley Longman.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International corpus of learner English* (Vol. **2**). Louvain-la-Neuve: Presses universitaires de Louvain.

Grey, S., Williams, J. N., & Rebuschat, P. (2015). Individual differences in incidental language learning: Phonological working memory, learning styles, and personality. *Learning and Individual Differences*, **38**, 44–53. doi: 10.1016/j.lindif.2015.01.019

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

Gries, St. Th., & Deshors, S. C. (2020). There's more to alternations than the main diagonal of a 2×2 confusion matrix: Improvements of MuPDAR and other classificatory alternation studies. *ICAME Journal*, **44**, 69–96.

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. doi: 10.1075/arcl.7.07gri

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**, 327–356. doi: 10.1075/ijcl.18.3.04gri

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, **44**, 9–15. doi: 10.1016/j.lindif.2015.10.003

Kerz, E. (2013). Concessive adverbial clauses in L2 academic writing. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 263–276). Louvain-la-Neuve: Presses universitaires de Louvain.

Kerz, E., & Wiechmann, D. (in press). Individual differences. In N. Tracy-Ventura & M. Paquot (Eds.), *Routledge handbook of SLA and corpora*. London: Routledge.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning*, **62**, 439–472.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal*, **97**, 634–654. doi: 10.1111/j.1540-4781.2013.12030.x

Martinez-Garcia, M. T., & Wulff, S. (2012). Not wrong, yet not quite right: Spanish ESL students' use of gerundial and infinitival complementation. *International Journal of Applied Linguistics*, **22**, 225–244. doi: 10.1111/j.1473-4192.2012.00310.x

Miglio, V. G., Gries, St. Th., Harris, M. J., Wheeler, E. M., & Santana-Paixão, T. (2013). Spanish *lo(s)-le(s)* clitic alternations in psych verbs: A multifactorial corpus-based analysis. In J. Cabrelli Amaro, G. Lord, A. de Prada Pérez, & J. E. Aaron (Eds.), *Selected proceedings of the 15th Hispanic linguistics symposium* (pp. 268–278). Somerville, MA: Cascadilla Press.

Möller, V. (2017). A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches. In P. de Haan, S. van Vuuren, & R. de Vries (Eds.), *Language, learners and levels: Progression and variation* (pp. 409–439). Louvain-la-Neuve: Presses universitaires de Louvain.

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, **17**, 56–72. doi: 10.1017/s1366728912000715

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. London: Continuum.

Prat, C. S., Yamasaki, B. L., Kluenda, R. A., & Stocco, A. (2016). Resting-state qEEG predicts rate of second language learning in adults. *Brain and Language, 157–158*, 44–50, doi: 10.1016/j.bandl.2016.04.007

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (Eds.) (1985). *A grammar of contemporary English*. London: Longman.

Ramsay, V. (1987). The functional distribution of preposed and postposed "if" and "when" clauses in written discourse. In R. Tomlin (Ed.), *Coherence and grounding in discourse* (pp. 383–408). Amsterdam: Benjamins.

Rothman, J., & Guijarro-Fuentes, P. (2010). Input quality matters: Some comments on input type and age-effects in adult SLA. *Applied Linguistics*, **31**, 301–306. doi: 10.1093/applin/amq004

Sun, H., Streinkrauss, R., Tendeiro, J., & de Boot, K. (2016). Individual differences in very young children's English acquisition in China: Internal and external factors. *Bilingualism: Language and Cognition*, **19**, 550–566. doi: 10.1017/s1366728915000243

Unsworth, S. (2016). Early child L2 acquisition: Age or input effects? Neither, or both? *Journal of Child Language*, **43**, 608–634. doi: 10.1017/s030500091500080x

van Vuuren, S. (2013). Information structural transfer in advanced Dutch EFL writing: A cross-linguistic longitudinal study. *Linguistics in the Netherlands*, **30**, 173–187. doi: 10.1075/avt.30.13van

van Vuuren, S., & Laskin, L. (2017). Dutch learner English in close-up: A Bayesian corpus analysis of pre-subject adverbials in advanced Dutch EFL writing. *International Journal of Learner Corpus Research*, **3**, 1–35. doi: 10.1075/ijlcr.3.1.01van

Wen, Z., Borges Mota, M., & McNeill, A. (Eds.). (2015). *Working memory in second language acquisition and processing*. Bristol: Multilingual Matters.

Wulff, S., & Gries, St. Th. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, **5**, 120–148. doi: 10.1075/lab.5.1.05wul

Wulff, S., & Gries, St. Th. (2019). Particle placement in learner language. *Language Learning*, **19**, 873–910. doi: 10.1111/lang.12354

Wulff, S., & Gries, St. Th. (to appear). Exploring individual variation in learner corpus research: Some methodological suggestions. In B. S. W. Le Bruyn & M. Paquot (Eds.), *Learner corpus research and second language acquisition*. Cambridge: Cambridge University Press.

Wulff, S., Gries, St. Th., & Lester, N. A. (2018). Optional *that* in complementation by German and Spanish learners. In A. Tyler, L. Huang, & H. Jan (Eds.), *What is applied cognitive linguistics? Answers from current SLA research* (pp. 99–120). New York: de Gruyter Mouton.

Wulff, S., Lester, N. A., & Martinez-Garcia, M. T. (2014). *That*-variation in German and Spanish L2 English. *Language and Cognition*, **6**, 271–299. doi: 10.1017/langcog.2014.5

Zuur, A. F., Ieno, E. N., Walker, N., & Saveliev, A. A. (2009). *Mixed effects models and extensions in ecology with R*. Berlin: Springer.

Cite this article: Gries, S. Th. and Wulff, S. (2021). Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied Psycholinguistics* **42**, 279–299. https://doi.org/10.1017/S014271642000048X