Subject Realization in Japanese Conversation by Native and Non-native Speakers: Exemplifying a New Paradigm for Learner Corpus Research

Stefan Th. Gries and Allison S. Adelman

Abstract In the field of Learner Corpus Research, Gries and Deshors (Corpora 9(1):109–136, 2014) developed a two-step regression procedure (MuPDAR) to determine how and why choices made by non-native speakers differ from those made by native speakers more comprehensively than traditional learner corpus research allows for. In this chapter, we will extend and test their proposal to determine whether it can also be applied to pragmatic and grammatical phenomena (subject realization/omission in Japanese), and whether it can help study categorical differences between learner and native-speaker choices; we do so by also showing that the more advanced method of mixed-effects modeling can be very fruitfully integrated into the proposed MuPDAR method. The results of our study show that Japanese native speakers' choices of subject realization are affected by discourse-functional factors such as givenness and contrast of referents and that, while learners are able to handle extreme values of givenness and marked cases of contrast, they still struggle (more) with intermediate degrees of givenness and unmarked/non-contrastive referents. We conclude by discussing the role of MuPDAR in Learner Corpus Research in general and its advantages over traditional corpus analysis in that field and error analysis in particular.

Keywords Learner corpora • Regression modeling • Subject realization • Japanese • Givenness and contrast

St.Th. Gries (✉) • A.S. Adelman
Department of Linguistics, University of California, Santa Barbara,
Santa Barbara, CA 93106-3100, USA
e-mail: stgries@linguistics.ucsb.edu

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, Yearbook of Corpus Linguistics and Pragmatics 2,
DOI 10.1007/978-3-319-06007-1_3, © Springer International Publishing Switzerland 2014

1 Introduction

1.1 The State of the Art in Learner Corpus Research

Given the increasing availability of learner corpora, learner corpus research (LCR) is a growing sub-field of corpus linguistics. Much of the work done in LCR is concerned with "bring[ing] out the words, phrases, grammatical items or syntactic structures that are either over- or underused by the learner" (Granger 2002: 132) and/or seeks to "uncover factors of 'foreign-soundingness'" (Granger 1996: 43), specifically "foreign-soundingness *even in the absence of downright errors*" (Granger 2004: 132, our emphasis).

The way much of the work in LCR proceeds can be summarized as follows:

> In either case the learner deviates in plus or minus from a certain statistical norm which characterizes native performance in a particular language. To ascertain such an error [though see below], one has to perform a *quantitative contrastive study of texts* written by native users of a particular language and by a non-native user of the same language and *compare the frequencies of use* of the investigated forms. (Krzeskowski (1990: 206), quoted from Granger 1996: 45, our emphases)

That is to say, one generates concordances of a phenomenon in question, determines the frequencies with which it is attested both in native language (NL)/native-speaker (NS) data and in non-native speaker (NNS) data, and compares them to determine whether, relative to the NS standard, the NNS over- or underuse the linguistic unit under consideration. Examples include

- Aijmer (2005), who explores the frequencies of use of modal verbs in NS English (in the LOCNESS corpus) and NNS English (in the Swedish component of the ICLE corpus) with multiple chi-squared tests;
- Altenberg (2005), who discusses frequencies/percentages of uses of English *make* and Swedish *göra* in four different constructional patterns and an 'other' category;
- Cosme (2008), who discusses (cross-linguistic) transfer-related issues based on the over-/underuses of adverbial and adnominal present/past participle clauses by French- and Dutch-speaking learners of English;
- Hundt and Vogel (2011), who explore the frequencies of progressives in data from corpora covering English as a NL, English as a second language, and English as a foreign language on the basis of likelihood-ratio tests;
- Hasselgård and Johansson's (2012) case study of the use of *quite* in the LOCNESS corpus and four components of the ICLE Corpus (Norway, Germany, France, and Spain) involving chi-squared tests comparing *quite*'s frequency (both on its own and with a colligation) from the ICLE components to the LOCNESS frequency;
- Neff van Aertselaer and Bunce (2012), who discuss the frequencies of reporting verbs in the Spanish component of the ICLE corpus and a small academic-writing corpus compiled from Spanish EFL students;
- Rogatcheva (2012), who compares the uses of present perfects by Bulgarian and German learners of English in the corresponding parts of ICLE; etc.

While the above kinds of studies appear to be what is currently the state of the art, this state of the art is severely lacking even if compared to two quite basic and very reasonable desiderata stated a long time ago. First, chi-squared tests of goodness-of-fit (of mere frequencies of occurrence) or of independence (of frequencies of co-occurrence) are certainly not the "massive statistical research" called for by Krzeszowski as early as 1990 (p. 212). Second, they are also not "comparing/contrasting what non-native and native speakers of a language do *in a comparable situation*" (Pery-Woodley 1990: 143, quoted from Granger 1996: 43, our emphasis). Both of these problems have a similar root, namely the fact that many studies reduce the context of a phenomenon under investigation to maximally one co-occurring factor/predictor, such as when Altenberg (2005) explores the use of *make* based on one predictor – patterns that *make* co-occurs with – or when Hasselgård and Johansson (2012) explore the use of *quite* based on one predictor – its colligation. However, this is neither comprehensive enough – surely the use of *make* or *quite* is co-determined by more than this one predictor – nor does a single predictor make the situations of use of *make* and *quite* comparable. As Gries and Deshors (2014) argue on the basis of the alternation of *may* vs. *can*,

> for example, the choice of the modal verbs *can* vs. *may* is determined by 15 or so different factors $F_{1–15}$ including syntactic characteristics of the clause and various morphological and semantic features of the subject [...], and maybe also by the circumstances of production, which we may call register. Thus, the traditional interpretation of "in a comparable situation" leads to the somewhat absurd assumption that we compare uses of NS and NNS that are completely different in terms of $F_{1–15}$ and only share the single factor that they were produced in an essay-writing situation in school.

Without wanting to be alarmist or polemic, it is not clear how the study of any phenomenon *P* that is determined by 15 or so different linguistic $F_{1–15}$ can be studied with over-/underuse counts at all. If a study on *P* bases a whole theory about how learners' use of *X* is affected by L1 influence/interference, teaching materials, etc. on just F_1 while completely ignoring $F_{2–15}$, how insightful can it be? Again, Gries and Deshors (2014) is instructive and merits a long-ish quote:

> From this perspective, it is obvious how lacking mere over-/underuse counts are: If a learner used *may* 10 % less often in a corpus file than a native speaker did, that discrepancy may be completely due to individual cases where closer inspection would reveal that, in many of these specific situations, a native speaker would also not have used *may*. Maybe the learners even wrote about the same topic as the native speaker but used more negated clauses than the native speaker. Negation is inversely correlated with the use of *may* so the fact that the learner used *may* 10 % less often than the native speaker says nothing about proficiency regarding *can*/*may* or over-/underuse as it is traditionally used – that 10 % difference is completely due to the learners' use of negations and, crucially, had the native speaker chosen negations as well, he would have exhibited the same perceived dispreference of *may*.

1.2 First Improvements

Given the above severe shortcomings of the state-of-the-art over-/underuse counts, what can been done to address this? So far, three main kinds of suggestions stand out.[1] One kind is exemplified by Tono (2004) or Collentine and Asención-Delaney (2010). The former studies verb subcategorization patterns by Japanese learners of English and is particularly instructive in how he takes interactions between predictors into consideration.

The latter explore the use of *ser/estar* + adjective using multifactorial regression modeling. Their work is highly interesting as it is one of the few published LCR studies that uses a regression-based approach and, thus, cover a large number of linguistic and contextual factors. Unfortunately, Collentine & Asención-Delaney's methodology has critical problems: one conceptual in nature, two statistical. The conceptual problem is that their study involves two regression models – one for *ser* + adjective and a separate one for *estar* + adjective – when what they should have done is one regression model for all the data including a predictor Verb: *ser* vs. *estar* that is allowed to interact with all others. This would have allowed them to see whether any effects differ significantly between the two verbs. As for the statistical problems, a somewhat subjective one is that best-subsets analyses are far from uncontroversial and have been surpassed by other methods (e.g., Lasso and Least Angle regressions). However, the authors do not provide enough information on how their statistical analysis proceeded, but typical implementations of this method neither include interactions between predictors in their computations nor allow for non-linear effects, which is problematic since we know from now two decades of research on lexical and syntactic alternation phenomena in linguistics that they usually involve interactions between predictors and sometimes also non-linear effects.[2]

The second kind of approach addresses several of Collentine & Asención-Delaney's problems and involves regression analyses of corpus data where

- the choice constituting phenomenon *X* to be studied is the dependent variable;
- many linguistic/contextual variables are the independent variables;
- an additional independent variable is the L1 of the speaker, which should minimally compare NS data to one NNL, but multiple NNLs would be better and more in line with, for example, Granger's (1996) Contrastive Interlanguage Analysis;
- the L1 variable can interact with all other predictors because only that will bring out whether any linguistic/contextual variable differs across the L1s.

This approach has been discussed in detail in Gries and Deshors (2014) and Gries and Wulff (2013) as well as several conference papers by the latter two authors.

The most fine-grained approach so far, however, is the so-called MuPDAR approach (Multifactorial Prediction and Deviation Analysis with Regressions) of

[1] We are disregarding here the large body of multifactorial work done by Crossley, Jarvis, and collaborators (cf. in particular Jarvis & Crossley 2012) because much of that work focuses on detecting the L1 of a writer rather than, as here, understanding any one particular lexical or grammatical choice in detail.

[2] An additional problem may involve the fact that the authors used a linear regression on data that might violate the assumptions of such regressions. However, we were unable to infer from the paper what the dependent variable was – possibly a frequency of *ser/estar* + adjective per file? – so the above has to remain speculation for now.

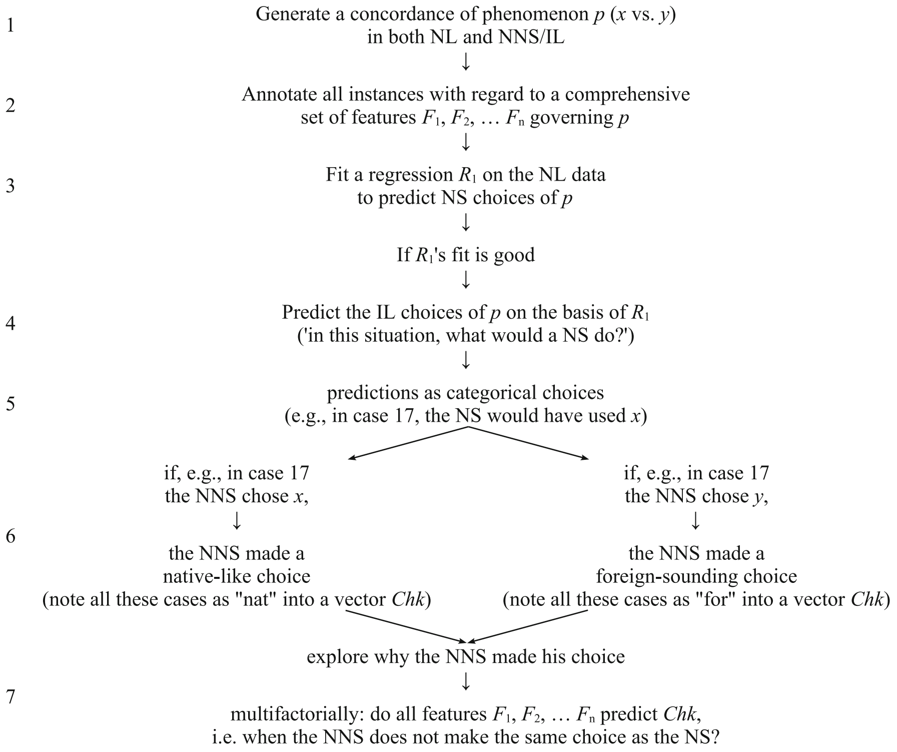
1
Generate a concordance of phenomenon p (x vs. y)
in both NL and NNS/IL
2
Annotate all instances with regard to a comprehensive
set of features F1, F2, … Fn governing p
3
Fit a regression R1 on the NL data
to predict NS choices of p
If R1's fit is good
4
Predict the IL choices of p on the basis of R1
('in this situation, what would a NS do?')
5
predictions as categorical choices
(e.g., in case 17, the NS would have used x)
6
if, e.g., in case 17
the NNS chose x,
the NNS made a
native-like choice
(note all these cases as "nat" into a vector Chk)
if, e.g., in case 17
the NNS chose y,
the NNS made a
foreign-sounding choice
(note all these cases as "for" into a vector Chk)
7
explore why the NNS made his choice
multifactorially: do all features F1, F2, … Fn predict Chk,
i.e. when the NNS does not make the same choice as the NS?

Fig. 1 Flowchart of the MuPDAR approach (Gries and Deshors 2014)

Gries and Deshors (2014). It involves a two-step regression procedure that offers an unprecedented level of precision in the analysis of learner language and is represented in a simplified version in Fig. 1.

First, one generates a concordance of phenomenon P and annotates it for an ideally large number of factors/predictors F_{1-n} that influence P. Then, P is modeled in a first regression R_1, but only on the basis of the NS data. If that regression model fits the data well, then its regression equation does a good job at quantifying each predictor's importance and predictive power for P and that means one can apply it to the NNS data. This is the first most essential step: What it does is answer for every NNS choice with regard to P the question "what would a native speaker have done?" These answers, i.e., the predicted NS choices, can then be compared to what the NNS did: either the NNS made the same choice as is predicted from the NS data, then he 'got it right', or the NNS made a choice that differs from what a NS would have done, in which case the NNS choice may not be prescriptively wrong, but at least not native-like. The final step then consists of a second regression R_2, in which one tries to identify which of the factors/predictors F_{1-n} result in the NNS making non-native-like choices. The results of this regression R_2 can then be interpreted in various ways; one of the most natural ways is that predictors that lead to higher NNS error rates can be considered 'difficult' for the learners.

1.3 Goals and Structure of the Present Chapter

Gries and Deshors (2014) exemplify the above approach with regard to modal choice by French learners of English. Their approach and results are quite promising but we want to explore two things they have not done. First, Gries and Deshors (2014) actually adopt a finer level of granularity than shown above: Rather than just considering categorically whether a NNS speaker makes a NS choice or not (cf. step 6), they consider the degree to which the NNS did not make a NNS choice. While the latter is arguably more precise, they do not show that the former also yields useful results. Thus, in this chapter, we will test whether their MuPDAR is also useful if one only explores NNS choices in a binary fashion, i.e., whether they correspond to the predicted NS choices.

Second, in their proof-of-concept chapter, they do not utilize the fullest potential of statistical analysis for their data. Specifically, they analyze the choice of *may* vs. *can* with a binary logistic regression model but, while the results are admittedly very promising, their case might have been stronger if they had analyzed the data with a generalized linear mixed-effects model (GLMEM). These models have become increasingly popular in linguistics over the last few years (cf. Baayen 2008; Jaeger 2008) given their ability to

– handle unbalanced designs, i.e., the type of unequal-cell-frequency problems that are emblematic of corpus-linguistic research;
– handle the fact that the data points entered into a corpus-linguistic analysis are often not independent of each other, since one speaker/writer may contribute multiple data points.

Thus, in this chapter, we will test whether the initial success of their MuPDAR approach can be replicated once more advanced GLMEMs are used. Incidentally, this will also be methodologically interesting on its own because of how GLMEMs work. In order to address the relatedness of data points, GLMEMs can provide (speaker-specific) adjustments to the overall intercept of the regression model, the contrasts between levels of categorical predictors, the slopes of numeric predictors, and interactions of predictors. However, the MuPDAR approach involves applying a model that was fit on data from one set of speakers – the native speakers – to a different set of speakers – the non-native speakers – so our analysis will have to take special steps to take this into consideration.

Finally, while Gries & Deshors studied a lexical choice (*may* vs. *can*), we will explore a pragmatic/grammatical choice – subject realization in conversational Japanese.

Section 2 will discuss our corpus data, their annotation, and their statistical analysis using the extension of the MuPDAR approach with GLMEMs. Section 3 will then turn to the results of the analyses. Specifically, Sect. 3.1 provides the results of the first regression model R_1 on the basis of the NS data; Sect. 3.2 briefly discusses the results of applying R_1 to the NNS data, and Sect. 3.3 is then concerned with the second regression model R2, which explores the non-nativeness of non-native speaker choices. Section 4 concludes.

2 Data and Methods

To further explore the MuPDAR approach, we decided to explore the phenomenon of subject realization in Japanese. Subject arguments are not expressed in all Japanese clauses; in fact they are quite often left unrealized, in what has been discussed as "pro-dropping," "ellipsis," or "zero anaphora" (e.g., Clancy 1980; Hinds 1982); cf. (1) for examples of one clause with a realized subject ((1)a) and one without ((1)b).

(1)	a.	*uchi-no*	*ryoushin-wa,*	
		1SG-GEN	parents-TOP	
		Shizuoka-ni		*sunde-i-te,*
		Shizuoka-LOC		live-PROG-CONJ
		'my parents live in Shizuoka, and ...'		
	b.	*muzukashi-i*	*to*	*omo-u.*
		difficult-NPST	QUOT	think-NPST
		'(I) think (it)'s difficult.'		

Shibatani (1985: 839) describes "PRO-dropping" as a process in Japanese – and Romance languages – in which "pronouns are omitted [...] because of their recoverability from the context." Ono and Thompson (1997: 484) have proposed that predicates should not be seen as having "obligatory" arguments or "slots" calling for either a mentioned referent or a "zero" (although the intended referents may be easily inferred from pragmatic context). Subsequent studies, claiming that unexpressed referents can usually be inferred from context, have therefore argued for the importance of examining this phenomenon only in the discourse contexts of interactional or conversational environments (Takagi 2002); in Sect. 2.1.1 we will discuss the corpus data that we will analyze in the present chapter.

Native speakers' realization of subjects in Japanese is based on many nuanced discourse-pragmatic factors which are likely to be difficult for NNS, particularly those with less experience speaking conversational Japanese. Given how speakers have to navigate information-structural demands and the recoverability and/or inferrability of referents in conversational real time, we assume that NS' patterns of subject realization are influenced by discourse-pragmatic factors such as givenness and contrast; accordingly, in this chapter we will explore if and how choices of subject realization or non-realization differ between NS and NNS speakers of Japanese and how these are affected by, or at least correlated with these two factors; in Sect. 2.1.2 we will therefore discuss our annotation of the corpus data.

2.1 Data

2.1.1 The Corpus Data

Data for this corpus of Japanese NS and NNS conversations was collected in various cities across Japan in the fall of 2011. The corpus consists of four hours of conversational data, comprising twelve 20-min conversations, each between one NS and one NNS of Japanese. The 12 conversations were carried out by 24 unique subjects, who volunteered to participate in pairs of two; in all cases these pairs were self-described "friends" (eight pairs), "close friends" (three pairs), or spouses (one pair).

In Japanese, speakers' relationships and social status are relevant to the style or register of spoken language used; by selecting only volunteer pairs of friends or spouses, we could ensure the near-consistent use of casual-register Japanese, rather than the distinct polite-register Japanese, throughout the corpus. While many Japanese language textbooks or L2-learning approaches focus primarily on formal or polite registers of the language (typically used among people who have only recently met), communication that takes place in such social settings likely constitutes only a small fraction of the total amount of linguistic interaction in which Japanese native speakers – and many non-native speakers – participate. Previous Japanese L2 speaker corpora have consisted of formal Japanese in artificial interview settings (Hypermedia Corpus of Spoken Japanese; cf. http://www.env.kitakyu-u.ac.jp/corpus/docs/index.html), as well as written Japanese compositions (Learner's Language Corpus of Japanese; cf. http://cblle.tufs.ac.jp/llc/ja/), but no corpus to our knowledge has attempted to capture casual everyday conversation among NNS and NS speakers who are already well-acquainted with each other (e.g., close friends or spouses), in more natural settings.

The 12 NNS participants had as their L1s either English (8), Korean (2), or Chinese (2); the native English speakers were from the U.S., the UK, Canada, and Australia. Most had taken some coursework in Japanese, while some had learned the language primarily through self-study with textbooks and conversations with Japanese friends. All but one of the NNS participants had been studying Japanese for at least 4 years (self-reported study times ranged from 2 to 41 years). All but two of the NNS participants had been living in Japan for at least 3.5 years (self-reported time spent living in Japan ranged from 1 month to 26 years).

Eight of the twelve NS participants came from the central Chuubu and Kansai regions of Japan; two others were from Okayama prefecture, and two were from Tokyo. Most described themselves as speaking regional dialects, with the two from Tokyo reporting that they spoke *hyoujungo*, or standard Japanese based on the Tokyo dialect.

The recordings in the corpus were transcribed in Romanized Japanese in a slightly-adapted version of DT2 (cf. Du Bois 2006) by the second author; each transcription was thoroughly double-checked by a native Japanese speaker.[3] The corpus contains a total of 13,555 intonation units, and a total of 6,873 clauses (55 % verbal predicates; 24 % nominal predicates; 21 % adjectival predicates).

[3] We thank Nobutaka Takara and Mikuni Okamoto for their help in transcribing the corpus data.

2.1.2 The Annotation

We included most of the clauses from the corpus in our sample, excluding those for which a particular subject referent could not be identified, as explained further below. This resulted in a sample of 5,952 sentences. These were then annotated with regard to the following set of variables. First, every clause was coded for the variable Speaker, i.e., a variable indicating whether the speaker of the clause is a native Japanese speaker (*N(J)S*) or a non-native speaker (*NNS*).

Second, every clause was coded for the variable Givenness, i.e., an interval-scaled variable reflecting the givenness of the subject referent on a scale from 0 to 10. High values (*10*, *9*, *8*, etc.) indicate that the referent is highly given (e.g., the referent has been mentioned directly or indirectly (mentioned overtly or referred to implicitly) in the previous clause (*10*), one clause back (*9*), two clauses back (*8*), etc.), while lower values reflect a greater distance to the last mention (e.g., the referent has been mentioned 9 clauses back (*2*), the referent has been mentioned 10 clauses back (*1*), and a value of *0* indicates that the referent has not been mentioned at all in the 10 preceding clauses). However, given the nature of the data – conversations of two speakers – the referents of first- and second-person expressions were always coded with a *10*.

Third, every example was annotated for Contrast, i.e., a variable representing whether the subject is contrastive (*yes*) or not (*no*). The annotation of Contrast required a detailed inspection of the clauses' contexts. For example, whereas some *wa*-marked NPs act as topics, two *wa*-marked NPs in two clauses in a row leads each of those clauses to have a contrastive structure (Iwasaki 2002: 244), as exemplified in (2), where both clauses are coded as having contrastive subjects.

(2)	*de*	*hitori-wa*	*tabete-i-mashi-ta.*
	and	one.person-TOP	eat-PROG-POL-PST

'so one person person was eating.'

hitori-wa	*matte-i-mashi-ta.*
one.person-TOP	wait-PROG-POL-PST

'another(/one) person was waiting.'

We did not code for "propositional contrast," meaning instances where the entire clause is contrasted with another proposition, rather than one particular element in the clause being marked as contrastive (Kuno 1973: 46–47). We coded only for contrastive subject/topic arguments (rather than contrastive object arguments or propositional contrast), i.e., only for when two or more subjects/topics were being contrasted with each other, usually with respect to the same predicate. For example, both of the following clauses were coded as contrastive because of the affirmative/negative polarity contrast of only one particular element in each clause against the other (this is not an example of propositional contrast because both clauses have the same predicate).

(3)	*nanka*	*kekkou*	*shaber-u*	*ko*	*mo*	*i-tari,*
	DM	quite.a.bit	speak-NPST	kid	too	exist-REP

‘like there are students who speak quite a bit, and,’

shaber-e-nai	*ko*	*mo*	*i-tari*	*shite,*
speak-POT-NEG	kid	too	exist-REP	light.verb

‘and there are also students who can’t speak.’

In addition, arguments were marked as contrastive when they involved contrastive topics, when they were subjects of the inherently contrastive construction *(no) hou ga*, or when they involved the *yori* ‘compared to’ construction.

For some predicates whose subjects are not realized, it is impossible to identify a particular referent as the intended subject; this is sometimes – though not always – due to predicates being part of “fixed expressions with different degrees of lexicalization” (Ono and Thompson 1997: 485). For any predicates for which we could not identify a particular referent as the subject, we labeled those clauses as “uncodeable” and did not include them in our sample.

Finally, each clause was coded for the variable SubjReal, to reflect whether a subject was realized (*yes*) or not (*no*).

In addition to these fixed-effects predictors, we also included annotation for a random effect representing the identity of the speakers (SpeakerID) so that the fact that the data points are not independent but may involve speaker-specific effects is taken into consideration. The corpus consists of 12 recordings, each between a native and a non-native speaker; thus there are 24 individual speaker IDs.

2.2 Statistical Evaluation

In this section, we outline the statistical evaluation of the above-mentioned corpus data. We proceed in three steps: the description of the model fitting/selection process of R_1 using the NS data, its application to the NNS, and the model fitting/selection process of R_2.

2.2.1 Regression R_1: Exploring the Choices Made by NS

Our model fitting/selection process follows the logic outlines by Zuur et al. (2009: Ch. 5).[4] That is, we first determine the random-effects structure of the model, then the fixed-effects structure. As for the former, we begin with a maximal model that

[4] By virtue of the complexity of the statistical methods involved, this section can only be rather technical in nature, plus space constraints do not permit exhaustive definitions and discussion of all the statistical technical terms. We therefore refer the reader to Baayen (2008: Ch. 7), Crawley (2013: Ch. 9, 19), Faraway (2006: Ch. 8–10), and Zuur et al. (2009: Ch. 5).

was fit to the 3263 NS data points only (using REML estimates and the function lmer from the R package lme4 (version 0.999999-2); cf. Bates et al. 2013) and included

- SubjReal: *no* vs. *yes* as the dependent variable;
- Givenness: *0*, *1*, ..., *9*, *10* and Contrast *no* vs. *yes* and their interaction as fixed-effects predictors, where, to allow for possible curvature in the effect of Givenness, the maximal model included Givenness as a polynomial to the third degree;
- random intercepts and slopes for all predictors and their interactions as random effects.

Using likelihood-ratio tests, the random-effects structure of this model is reduced to the minimal adequate one, i.e., the one that did not allow further simplification. After that, we proceed with an analogous reduction of the complexity of the fixed-effects structure using likelihood-ratio tests of ML fits to the final minimal adequate model. The quality of this model is then assessed by means of an overall likelihood-ratio chi-squared significance test, the model's classification accuracy, and its *C*-score; the nature of the effects of this final model is interpreted with plots of predicted probabilities of subject realization both separately for each speaker and as an overall trend.[5]

2.2.2 Applying R_1 to the NNS Data

The next step involves applying the regression model R_1 to the NNS data. Crucially, R_1 involves speaker-specific effects, but since the NNS data stem from different speakers, we only use the fixed-effects part of R_1 to answer the following question for every NNS data point: "would a native speaker have realized the subject here, yes or no?" The fit of the NS model to the NNS data is also quantified with a classification accuracy and a *C*-score.

2.2.3 Regression R_2: Exploring the Choices Made by NNS

Given the results from Sect. 2.2.2, we can determine for each of the 2689 NNS data points whether the NNS chose what was predicted as the most likely NS choice. The results of this comparison are represented in a variable called Correct: *no* (the NNS made the predicted NS choice) vs. *yes* (the NNS did not make the predicted NS choice). The variable Correct then is the dependent variable in the second

[5] Strictly speaking, if one does a MuPDAR analysis in which R1 is really only used for prediction, then one does not really have to apply Occam's razor rigorously to eliminate non-significant/collinear predictors that much because, within MuPDAR, the point of R1 is not to actually interpret R1's coefficients.

Table 1 Results of R_1 (predicted level of SubjReal: *yes*)

Fixed effects				
Predictor	Estimate/coefficient	Std. error	z	p_{deletion}
Intercept	1.25353	0.17581	7.130	<<0.0001
Givenness	−0.33058	0.02043	−16.180	<<0.0001
Contrast (*no* → *yes*)	0.07159	0.36953	0.194	0.846
Givenness * Contrast (*no* → *yes*)	0.34734	0.05130	6.771	<<0.0001
Random effects				
Adjustment to overall intercept (Speaker)	$sd = 0.477134$			
Adjustment to slope of Givenness	$sd = 0.054543$			

regression model fitting/selection process R_2, which proceeds as before: we include the same fixed-effects predictors and random effects as for R_1, first determine the minimal adequate random-effects structure (with likelihood-ratio tests of REML fits), and then the minimal adequate fixed-effects structure (with likelihood-ratio tests of ML fits). Finally, we compute the final model's significance test and classification accuracy and visualize its results in terms of the predicted probabilities of the NNS making the choice that the NS would have made.

3 Results

In this section, we present the multitude of results of the statistical analyses; we proceed analogously to Sect. 2.2.

3.1 *The Results of* R$_1$*, the Regression on the NS*

The results of the first regression, R_1, applied to the NS data only, indicate a good fit. The minimal adequate model we arrived at after the model selection process reflects a highly significant correlation between its predictors and the NS choices of subject realizations: likelihood-ratio chi-squared = 192.13, $df = 3$, $p < 0.0001$. Table 1 represents the results for the fixed and random effects in the model.

The results in Table 1 already indicate that

- if Givenness increases, the probability of a subject being realized decreases (note the negative sign of the coefficient of Givenness);
- Contrast on its own has no effect on subject realization;
- the interaction of Givenness and Contrast is highly significant and in fact annuls the effect of Givenness in isolation when Contrast is *yes*

As usual, however, these effects are much easier to comprehend from a visual representation such as Fig. 2. In both panels of Fig. 2, Givenness is represented on

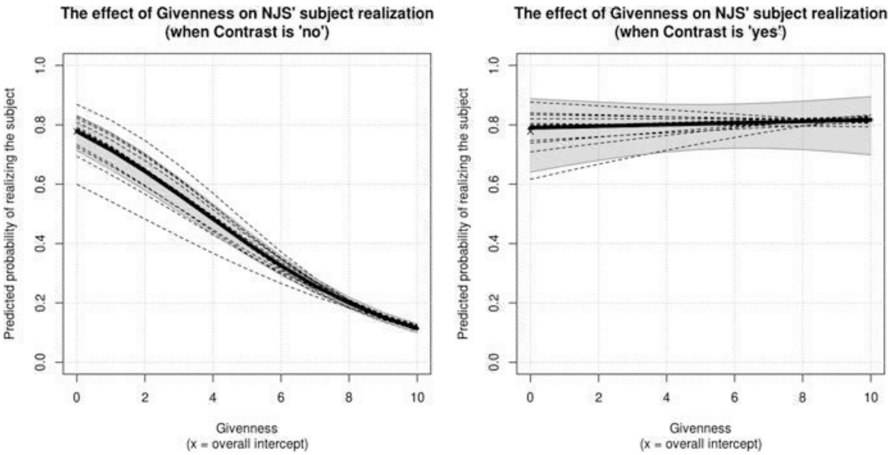

Fig. 2 The effect of the interaction Givenness * Contrast on the predicted probability of NJS' realizing the subject

Table 2 Classification accuracy of R_1 when applied to the NS

	Pred. SubjReal: no	Pred. SubjReal: yes	Totals
Obs. SubjReal: no	2,279	148	2,427
Obs. SubjReal: yes	345	491	836
Totals	2,624	639	3,263

the x-axis while the predicted probability of a subject being realized is represented on the y-axis; to provide a fine-grained resolution of the results, we indicate both the results for every speaker individually (with dashed grey lines) and the results for all speakers (a heavy black line with its grey confidence interval). The left panel shows the effect of Givenness when Contrast is *no*, and there is a strong and clear trend such that, the more given the referent of the subject, the less likely it will be expressed overtly; the speaker-specific results show that this effect holds for all speakers (but of course to varying degrees). The right panel shows the effect of Givenness when Contrast is *yes*; the essentially flat regression line indicates that Givenness has no effect on subject realization when Contrast is *yes* – whatever the value of Givenness, in contrastive settings subjects are very likely to be realized. In this panel, we do find some subject-specific variation: some slopes exhibit an upward trend, some a downward trend, but since a random effect Givenness: Contrast|Speaker did not reach standard levels of significance, the overall conclusion – Givenness has no effect on subject realization when Contrast is *yes* – still stands (Fig. 2).

Even though the final model contains only one significant highest-level predictor, the classification accuracy of the model amounted to 84.9 %, which is highly significantly better than the chance-level baseline of 61.9 % ($p_{\text{binomial test}} < 10^{-100}$); consider Table 2 for the classification matrix resulting from the predictions of R_1. The more precise C-value for this model is 0.82, thus exceeding Harrell's (2001: 248) threshold of 0.8 for good models.

Table 3 Classification accuracy of R_1 when applied to the NNS

	Pred. SubjReal: no	Pred. SubjReal: yes	Totals
Obs. SubjReal: no	1,767	124	1,891
Obs. SubjReal: yes	305	493	798
Totals	2,072	617	2,689

Table 4 Results of R_2 (predicted level of Correct: *yes*)

Fixed effects				
Predictor	Estimate/coefficient	Std. error	z	p_{deletion}
Intercept	1.76	0.1423	12.369	<<0.0001
Givenness	11.4229	2.6543	4.304	<<0.0001
poly(Givenness, 2)	10.8993	2.3960	4.549	<<0.0001
Contrast (*no* → *yes*)	0.4872	0.2552	1.909	0.04573
Random effects				
Adjustment to overall intercept (Speaker)	$sd = 0.44668$			

3.2 *The Results of Applying* R$_1$ *to the NNS Data*

Given the good fit of R_1 to the NS data, we proceeded by generating predictions of subject realizations for the NNS data. Crucially and as mentioned above in Sect. 2.2.2, the predictions for the NNS were based only on the fixed effects listed in Table 1, i.e., the speaker-specific random effects of R_1 were not included given that the NNS are different speakers. Nevertheless, R_1 was able to predict the subject realizations of the NNS nearly exactly as well as those of the NS; consider Table 3 for the classification matrix; the accuracy of the model is 84 %, which is highly significantly better than the chance-level baseline of 58.3 % ($p_{\text{binomial test}} < 10^{-100}$), and $C = 0.8$.

3.3 *The Results of* R$_2$*, the Regression on the NNS*

The results from Table 3 then lead to the final step, the regression R_2 that was fit to predict when the NNS would make a choice differing from that predicted from the NS data; that is, the dependent variable here was Correct: *no* (1,767 + 493-2,260 cases) vs. *yes* (124 + 305 = 429 cases). The minimal adequate model was again highly significant: likelihood-ratio chi-squared = 37.18, $df = 3$, $p < 0.0001$. The results for all fixed and random effects are represented in Table 4; interestingly, the effect of Givenness is not best represented with a straight line but rather with a curved line resulting from a polynomial to the second degree.

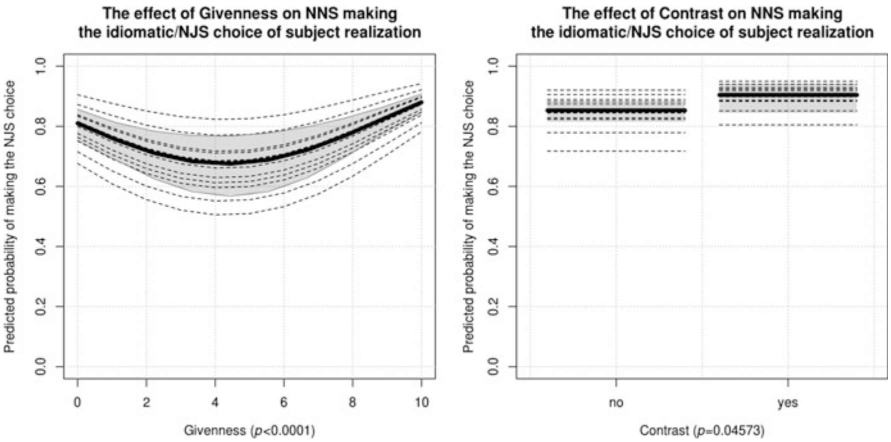

Fig. 3 The significant main effects of Givenness and Contrast on whether NNS make the same subject realization choice a NJS is predicted to have made

Given the curved nature of the effect of Givenness and its being fit with orthogonal polynomials, it is necessary to visualize the results as in Fig. 3. In both panels, the predicted probability of the NNS making the same choice that the NS would have made is represented on the *y*-axis, and again we provide each speaker's prediction with dashed lines and the overall estimate with a heavy line and a grey confidence interval. In the left panel, the predictor Givenness is on the *x*-axis and the result shows that the NNS are most likely to make the NS choice with extreme values of Givenness: i.e., when the referent of the subject is completely new or completely given (in the sense of 'having been mentioned just before'). On the other hand, when the referent of the subject is intermediately given, then the NNS are more likely to not make the subject realization choices a NS would have made. In a nutshell, the NNS can handle the extreme cases, but not (yet) the middle ground.

As for the effect of Contrast, it is relatively weak and only just about significant, but again its results make sense: In the more marked communicative situation with a contrastive subject referent, which can be considered 'more extreme' than the unmarked case, the NNS make choices that are more in line with what NS would have done. On the other hand, when the referent of the subject is not marked (in the sense of 'not being contrastive'), the NNS struggle more with making NS choices.

4 Discussion and Concluding Remarks

4.1 Interim Summary and Implications of the Analysis

The results of the present analysis are strong evidence for the feasibility of the LCR method proposed here, the MuPDAR approach. All regression results are at least significant and come with high degrees of predictive power/classification accuracy. R_1 shows that NS choices to realize the subject are strongly influenced by Givenness and its interaction with Contrast in ways that are compatible with previous findings regarding discourse givenness or inferrability in a wide variety of languages – given or highly inferrable referents are often not lexically realized – and with what can reasonably be expected for Contrast – referents that are to be highlighted contrastively are realized no matter their givenness. At the same time, R_2 shows that the learners in this study have been able to extrapolate these NS preferences, but not fully yet. Extreme values of givenness/inferrability pose few problems to the learners, as do the distinctions of Contrast: the NNS speakers know what to do with subjects when their referents are completely given, completely new, and contrastive – they still struggle with intermediate degrees of givenness/ inferrability, which not only makes sense since this is an 'uncomfortably grey middle area' on the givenness continuum but also because this kind of scenario happens least often. For discourse cohesion reasons, referents are usually introduced but then also used immediately afterwards, which would result in high values of givenness. But when that does not happen and a referent has been introduced but then left in limbo for 4–6 clauses, then the NNS have problems. The MuPDAR approach has revealed this quite clearly and we submit it is hard to imagine how traditional LCR would have found this (so clearly; cf. below). Follow-up analyses could now also explore the random-effect structure to determine, for example, whether the random intercepts/slopes correlate with relevant characteristics of the speakers, such as their L1s. We did this for the present data but, unlike in Miglio et al. (2013), no correlations between random effects and the speakers were found.

4.2 Where to Go from Here

We hope to have shown that the MuPDAR approach is a powerful and flexible tool for LCR. This second proof-of-concept study shows that (i) MuPDAR cannot only be used with traditional regression modeling but is also naturally extended to advanced mixed-effects modeling and that (ii) MuPDAR yields precise and meaningful results regardless of the resolution of R_2 – categorical deviations of NNS choices from NS choices as in this study or differences in degree as in Gries and Deshors (2014). That being said, there are several obvious next steps. One is that we clearly need more applications of this approach; in an ideal world, this would mean that traditional work in LCR would be re-analyzed to determine its validity.

Second, the method may be further refined. Dan Lassiter (p.c.) suggested considering not only the (categorical or numeric) differences between NNS and NS choices, but also the differences between predicted probabilities of NNS and NS choices, which would make this method relate more seamlessly to variationist sociolinguistic studies. While we have no particular hypotheses about how this perspective would play out, it is certainly worth exploring in future work. In addition, various ways of making the analytical results more robust – crossvalidation with bootstrapping approaches are one possibility – should be explored in due course.

Third, it is also worth pointing out that both existing MuPDAR studies involved a final, minimal adequate regression model R_1 (from which insignificant predictors were trimmed following Occam's razor). The reason for this is that the results of R_1 are then also useful in their own right and can be interpreted linguistically/theoretically. However, if R_1 is really only used for prediction then it would theoretically be possible to not trim the maximal R_1 model and make full use of the fact that its classification accuracy will be slightly higher than that of the minimal adequate model we used here.

At this point, it is instructive to briefly discuss the relation of MuPDAR to error analysis. We believe that the present approach is at least a complement of, if not also a massive improvement over, traditional kinds of error analysis. For instance, some studies – Rogatcheva (2012) is a case in point – explore over- and underuses by having linguists/native speakers perform error-tagging on learner data. This is generally a useful approach given how it allows for, technically speaking, *true positives* (present perfects by NNS where they should be), *false positives* (present perfects by NNS where they should not be), *true negatives* (no present perfects by NNS where they should not be), and *false negatives* (no present perfects by NNS where they should be), and on the basis of such data, one can then compute statistics such as *SOC* (suppliance-in-obligatory-contexts) and *TLU* (target-like-use). On the one hand, this approach is undoubtedly more comprehensive than many previous LCR studies that do not include any context in their counts or that cross-tabulate just a single contextual feature in that the error coders will take more context into consideration in their coding decisions.

On the other hand, the process also suffers from some problems, which have to do with the distinctions that the coders/raters will make. A first problem that may arise is concerned with rater reliability. It has been known for many years now that judgment tasks like these are not only affected by a huge variety of factors (cf. Schütze 1996 for the most authoritative overview showing that) but can also be affected by the stimuli themselves over very short periods of time. For instance, Gries and Wulff (2009) discuss a weak but marginally significant within-subject priming effect that appears to indicate how subjects' preferences for sentence completion change over the course of just a short experiment (even when all other significant predictors are still considered). Similarly, Doğruöz and Gries (2012) find that, over the course of only eight acceptability judgments, subjects became more comfortable with unconventional morphological and lexical patterns. Thus, it is likely that raters' judgments/predictions will be affected as they go over and code many learner choices; at the very least, it is possible that they will and the degree to which they will is unknown. The precision of the MuPDAR approach, by contrast, is not affected by learning, habituation, or fatigue, and given the way that, in this chapter, we used mixed-effects modeling, it even accounts for speaker-specific effects that raters will most likely not be able to attend to.

The above is not to downplay the potential of error analysis, especially not if multiple coders are involved, coding protocols are rigorous, order effects etc. are

Table 5 Observed frequencies of SubjReal ~ Speaker * Contrast

Speaker	Contrast	SubjReal: *no*	SubjReal: *yes*	Totals
NJS	*no*	2,404	741	3,145
	yes	23	95	118
NNS	*no*	1,871	671	2,542
	yes	20	127	147
Totals		4,318	1,634	5,952

controlled, and careful interrater reliability statistics are computed. Nevertheless, even if all of these issues were addressed, MuPDAR still has advantages to offer. For instance, an additional problem of error-coding types of analyses is that most coders will not make as fine-grained distinctions/predictions as the regression because their judgments will at best be binary or categorical predictions about what will or should be used. On the other hand, when R_1 is applied to the NNS data, the MuPDAR approach makes very fine-grained predictions on a continuous probability scale, and when R_2 is computed on the basis of the deviations of NNS choices from the NS predictions as in Gries and Deshors (2014), then this regression, too, operates on a continuous scale. Thus, MuPDAR offers more a precise analysis of the data.

Finally, the error analysis and the resulting identification of, say, false positive and false negatives, in and of itself brings one no closer to an explanation of why the NNS did what they did. In the terminology of the present chapter, what the error-analysis approach does is ‘computing R_1 on the NS data and applying it to the NNS data.’ However, one then still needs to do R_2 to understand what it is that is responsible for the NNS making choices that are slightly or very much less idiomatic than those of the NS and on that topic, for example, Rogatcheva (2012) does very little. There are undoubtedly many different factors that jointly determine whether or not NS use the present perfect, but her chapter, while (laudably) computing *SOC* and *TLU*, does nothing to shed light on how many such factors there are, what they are, how strongly they affect speaker choices, and what their interactions might be.

Applying MuPDAR to native and learner corpus data is undoubtedly a complex and technical process, which may seem insurmountable to some and off-putting to even more. However, LCR scholars on the whole seem to agree that the corpus-based analysis of NNS language is, if anything, *more* complex than the analysis of NS language, which we already know from decades of alternation research to involve highly complex interactions of factors in multifactorial models. It is therefore utterly illogical to assume that the more complex set of questions regarding NNS language can be tackled with simple over-/underuse frequencies and pairwise chi-squared/log-likelihood ratio tests – complex data sets need techniques that can handle complex data, not methods that reduce the complexity to a level that has nothing to do anymore with what is really happening in the corpus. As a thought experiment, consider the fact that the currently most frequent type of over-/underuse kind of analysis of our data – recall the many studies cited as using such an approach in Sect. 1 – would reduce the analysis of everything that we found in our data to Table 5, presumably coupled with two chi-squared tests which, strictly speaking, one is in fact not even allowed to compute given that nearly all learner corpus studies are based on data points that are *not* independent, as the chi-squared test would require (which is why we pursued the GLMEM approach).

Against this background, we think it is high time that researchers in LCR begin to embrace tools that do more justice to the complexity they (correctly) claim their data come with. MuPDAR is but one approach to that end, but we believe we have demonstrated it is a powerful one and we hope that it will stimulate many applications exploring the intricacies of NNS language.

References

Aijmer, K. (2005). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 55–76). Amsterdam/Philadelphia: John Benjamins.

Altenberg, B. (2005). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 37–54). Amsterdam/Philadelphia: John Benjamins.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4. Linear mixed-effects models using S4 classes*. http://lme4.r-forge.r-project.org/

Clancy, P. M. (1980). Referential choice in English and Japanese narrative discourse. In W. Chafe (Ed.), *The pear stories: Cognitive and linguistics aspects of narrative production* (pp. 127–202). Norwood: Ablex.

Collentine, J., & Asención-Delaney, Y. (2010). A corpus-based analysis of the discourse functions of *ser/estar* + adjective in three levels of Spanish as FL learners. *Language Learning, 60*(2), 409–445.

Cosme, C. (2008). Participle clauses in learner English: The role of transfer. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 177–200). Amsterdam/Atlanta: Rodopi.

Crawley, M. J. (2013). *The R book* (2nd ed.). Chichester: Wiley.

Doğruöz, A. S., & Gries, S. T. (2012). Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics, 10*(2), 401–426.

Du Bois, J. W. (2006). *Representing discourse*. Ms., University of California, Santa Barbara.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed-effects and non-parametric regression models*. Boca Raton: Chapman & Hall/CRC.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam/Philadelphia: John Benjamins.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123–145). Amsterdam: Rodopi.

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

Gries, S. T., & Wulff, S. (2009). Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics, 7*, 163–186.

Gries, S. T., & 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.

Harrell, F. E., Jr. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. Berlin/New York: Springer.

Hasselgård, H., & Johansson, S. (2012). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 33–61). Amsterdam/Philadelphia: John Benjamins.

Hinds, J. (1982). *Ellipsis in Japanese*. Edmonton: Linguistic Research, Inc.

Hundt, M., & Vogel, K. (2011). Overuse of the progressive in ESL and learner Englishes – Fact or fiction? In J. Mukherjee & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 145–165). Amsterdam/Philadelphia: John Benjamins.

Hypermedia Corpus of Spoken Japanese. (2010). http://www.env.kitakyu-u.ac.jp/corpus/docs/index.html. Accessed Fall, 2010.

Iwasaki, S. (2002). *Japanese*. Amsterdam/Philadelphia: John Benjamins.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.

Jarvis, S., & Crossley, S. A. (Eds.). (2012). *Approaching language transfer through text classification explorations in the detection-based approach*. Bristol: Multilingual Matters.

Krzeskowski, T. (1990). *Contrasting languages: the scope of contrastive linguistics*. Berlin & New York: Mouton de Gruyter.

Kuno, S. (1973). *The structure of the Japanese language*. Cambridge, MA: MIT Press.

Learner's Language Corpus of Japanese. (2013). http://cblle.tufs.ac.jp/llc/ja/. Accessed Spring, 2013.

Miglio, V. G., Gries, S. T., Harris, M. J., Wheeler, E. M., & Santana-Paixão, R. (2013). *Spanish lo(s)-le(s) clitic alternations in psych verbs: A multifactorial corpus-based analysis*. Somerville: Cascadilla Press.

Neff van Aertselaer, J. A., & Bunce, C. (2012). The use of small corpora for tracing the development of academic literacies. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honour of Sylviane Granger* (pp. 63–83). Amsterdam/Philadelphia: John Benjamins.

Ono, T., & Thompson, S. A. (1997). Deconstructing 'Zero Anaphora' in Japanese. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society, 23*, 481–491.

Pery-Woodley, M.-P. (1990). Contrasting discourses: Contrastive analysis and a discourse approach to writing. *Language Teaching, 23*(3), 143–151.

Rogatcheva, S. (2012). Perfect problems: A corpus-based comparison of the perfect in Bulgarian and German EFL writing. In S. Hoffmann, P. Rayson, & G. Leech (Eds.), *English corpus linguistics: Looking back, moving forward* (pp. 149–163). Amsterdam: Rodopi.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: The University of Chicago Press.

Shibatani, M. (1985). Passives and related constructions: A prototype analysis. *Language, 61*(4), 821–848.

Takagi, T. (2002). Contextual resources for interferring unexpressed referents in Japanese conversations. *Pragmatics, 12*(2), 153–182.

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. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 45–66). Amsterdam/Philadelphia: John Benjamins.

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