Using regressions to explore deviations between corpus data and a standard/target: two suggestions

Stefan Th. Gries[1] and Sandra C. Deshors[2]

Abstract

The main goal of this study is to develop more appropriate ways to study variation between corpus data that instantiate a linguistic standard or target on the one hand, and corpus data that are compared to that standard, or that represent speakers that may aspire to approximate the target (such as second- or foreign-language learners). Using the example of SLA/FLA research, we first, briefly, discuss a highly influential model, Granger's (1996) Contrastive Interlanguage Analysis (CIA), and the extent to which much current research fails to exploit this model to its full potential. Then, we outline a few methodological suggestions that, if followed, can elevate corpus-based analysis in SLA/FLA to a new level of precision and predictive accuracy. Specifically, we propose that, and exemplify how, the inclusion of statistical interactions in regressions on corpus data can highlight important differences between native speakers (NS) and learners/non-native speakers (NNS) with different native linguistic (L1) backgrounds. Secondly, we develop a two-step regression procedure that answers one of the most important questions in SLA/FLA research – 'What would a native speaker do?' – and, thus, allows us to study systematic deviations between NS and NNS at an unprecedented degree of granularity. Both methods are explained and exemplified in detail on the basis of over 5,000 uses of *may* and *can* produced by NSs of English and French and Chinese learners of English.

Keywords: English, Chinese, corpus-based multifactorial methods, French, *may* and *can*, interlanguage, IL, regression modeling

[1] Department of Linguistics, University of California, Santa Barbara, Santa Barbara, CA 93106-3100, USA.
[2] Department of Languages & Linguistics, New Mexico State University, MSC 3L, Las Cruces, NM 88003, USA.
Correspondence to: Stefan Gries, *e-mail*: stgries@linguistics.ucsb.edu

Corpora 2014 Vol. 9 (1): 109–136
DOI: 10.3366/cor.2014.0053

www.euppublishing.com/cor

1. Introduction

1.1 Corpus studies in SLA/FLA: the state of the art and central questions/desiderata

The study of second/foreign language acquisition/learning is currently one of the most lively areas of research in corpus linguistics. This is in large part due to (*i*) the many pertinent corpus resources that have become available over the past fifteen or so years, and (*ii*) the on-going development of frameworks that shape and focus our analytical views of, and strategies for, handling corpus data. Both of these reasons have been particularly influenced by the work of the research group run by Granger at the Catholic University of Louvain: the corpus resources they have created or whose creation they have overseen and the theoretical proposals they have put forward. One particularly influential framework is Granger's (1996) Contrastive Interlanguage Analysis (CIA). According to Granger (1996: 44), CIA can be summarised as per Figure 1, where CIA involves both the comparison of native language (NL) to interlanguage (the left side of Figure 1) and the comparisons of different ILs of the same language (the right side of Figure 1). Some of the central goals are to 'uncover factors of "foreign-soundingness" ' (Granger, 1996: 43), which, of course, requires that the corpora involved in the relevant comparisons are comparable. The analysis of foreign-soundingness involves 'bring[ing] out the words, phrases, grammatical items or syntactic structures that are either over- or underused by the learner' (Granger, 2002: 132).

In other words, learners' linguistic choices are not completely in line with NS choices, as Krzeskowski (1990: 206, cited by Granger, 1996: 45) points out:

> 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.

Granger herself adopts a more nuanced picture, clarifying – correctly, we think – that over-/underuses make up, or contribute to, the 'foreign-soundingness *even in the absence of downright errors*' (Granger, 2004: 132, emphasis added). This approach has been very influential and has led to a wealth of results. A case in point is 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.

In spite of the successful application of this perspective/method to (native and learner) corpus data, we believe that, methodologically speaking,

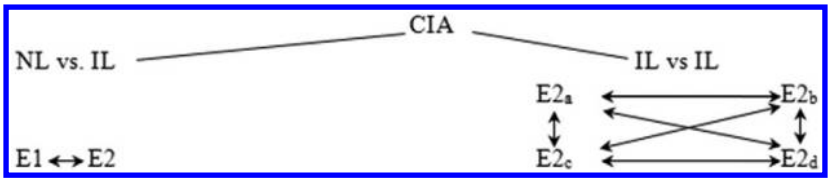

Figure 1: CIA as per Granger (1996: 44). NL = native language, IL = interlanguage, E1 = English as an L1, E2 = English as a foreign language (FL), $E2_l$ = English as a foreign language by speakers with *l* as L1

CIA and related approaches under-utilise corpus data to a considerable degree. While Granger (1996: 45) herself recognises that '[t]he contrastive investigation of raw frequencies [is . . .] undoubtedly the least sophisticated type of quantitative comparison', most studies have done just that: they have compared frequencies of use of *x* in NL to frequencies of *x* in IL(s). Such studies can certainly be revealing, but they often run the risks of relying on rather decontextualised frequencies of use and on rather coarse-grained analyses of what exactly makes learners choose a particular form of expression. Thus, such studies also run the risks of making hasty claims regarding how learners' L1s or non-linguistic determinants (such as teaching (styles)) affect foreign-soundingness (see Aijmer, 2002, for an example, and Gilquin and Paquot, 2007, for a discussion).

By way of an interim summary, it seems that, while CIA does not in principle exclude the use of much more sophisticated analytical approaches, the way researchers have worked within CIA is in need of fresh thinking. Based on work in other linguistic sub-disciplines and some other ideas/developments, we will outline some suggestions as to how the corpus-based analysis of IL can be brought to new levels of both comprehensiveness and detail. These suggestions go beyond much previous work in that they help to address the following five questions, which we believe are central to SLA/FLA research and to whose pursuit this paper contributes:

(*i*) What are the factors that determine a particular linguistic choice (by both NSs and NNSs)?
(*ii*) What are the differences between the NL data and the IL data in how the above factors affect the linguistic choice?
(*iii*) What are the differences between the different IL varieties in how the above factors affect the linguistic choices?
(*iv*) How much does the behaviour of the NNS resemble that of the NS's target language/variety (considering a whole range of factors simultaneously)? And,

(*v*) What are the areas where the behaviour of the NNS exhibits the largest difference from that of the NS's target language (again, considering a large number of factors at the same time)?[3]

Let us begin with the first three questions, which bring us back to the 'massive statistical research' called for by Krzeszowski (1990: 212; plus recall the above quote). Question (*i*) points to the first major way in which corpus-based SLA/FLA research should evolve: it must become 'multifactorial'. Much linguistic study has begun to explore linguistic choices on the basis of large data sets that are (*a*) annotated for many different linguistic and contextual features and (*b*) analysed with correspondingly comprehensive multifactorial statistical procedures that have, since Gries (2000, 2003a, 2003b), become particularly commonplace in the study of alternation phenomena, (see Grondelaers *et al*., 2002; and Szmrecsanyi, 2006; for additional examples). More boldly, an analysis of phenomenon *x* in NL and IL that is only based on mere over-/underuse counts rather than the many factors known to influence *x* in NL can impossibly be really as revealing as a study that takes factors from many different levels of linguistic analysis into consideration. Unfortunately, such multifactorial statistical analyses of corpus data are yet to be widely adopted in SLA research[4] – rather, there is still a preponderance of studies that under-utilise all the contextual information provided by learner corpora, despite the probably uncontroversial fact that learners' lexical and constructional choices are determined by multiple factors (for recent exceptions, see Deshors, 2010; Gries and Wulff, 2013a; or Deshors and Gries, forthcoming).

Once multifactorial studies are performed, they need to be of a kind that addresses question (*ii*), which means they must involve systematic and rigorous tests of how the fifteen factors known to affect *x* play out in NL data and how, if at all, the fifteen factors play out differently in learner data. This immediately leads us to question (*iii*) and, thus, to something central to CIA: not only must we compare systematically and rigorously how the

[3] We intentionally phrase these questions with regard to speakers' overt behavior – their objectively observable linguistic choices – in order to avoid the potential misunderstanding that we are jumping from corpus data to online psychological processing.

[4] We are disregarding here the large body of multifactorial work done by Crossley, Jarvis, and collaborators (see, in particular, Jarvis and Crossley, 2012). While this work involves sophisticated statistical analyses, its focus is not so much on understanding any one particular lexical or grammatical choice in detail; rather, the approach proceeds from an often large number of automatically generated statistics describing texts to the detection of L1s or the perceived quality of non-native writings. Put differently, and as will become clearer below, in many of these studies, the native language is the dependent variable to be predicted: in our work, it is a predictor. By the same token, we are disregarding work such as Szmrecsanyi and Kortmann (2011) on the question of whether a given learner variety's typological profile can be predicted based on the profile of the learner's L1. Similarly to the work done by Jarvis and collaborators, Szmrecsanyi and Kortmann's focus is not one particular linguistic choice but the analysis of part-of-speech frequencies and Greenberg-inspired index values in order to gauge typological profiles of learner varieties.

fifteen factors known to affect *x* affect *x* differently in NL as opposed to IL, but we must also make analogous comparisons between different ILs. These considerations lead to the second major way in which corpus-based SLA/FLA research should evolve: it must involve the study of 'relevant interactions' – a notion that will be discussed in detail below in Section 3.

Let us now turn to the last two questions, which, in some sense, should perhaps be the most important ones for the contrastive (IL) type of analyses. As mentioned above, much of the traditional CIA work involves the (reasonable) assumption that the native and learner corpora have to be somewhat comparable and then base their analysis of differences between NL and IL on the different frequencies with which the phenomenon in question occurs in the NL and IL corpora.

While still the dominant approach, this approach is in fact severely lacking. The most extreme objection to this strategy would be to argue that over-/underuse counts *per se* do not even speak to the issue because any over-/underuse by a learner may be due to the learner being more/less often in linguistic/contextual situations requiring the supposedly over-/underused choice. What one really needs to know was already mentioned more than two decades ago: one needs to be 'comparing/contrasting what non-native and native speakers of a language do *in a comparable situation*' (Péry-Woodley, 1990: 143, cited by Granger, 1996: 43, emphasis added). Thus, most previous research has so far adopted a very lax interpretation of 'in a comparable situation' – namely, the interpretation that the corpora are comparable because the NSs and the NNSs were in a similar language-production setting. It is easy to see that this seems quite unrealistic: for example, the choice of the modal verbs *can* versus *may* is determined by fifteen or so different factors, $F_{1\text{-}15}$, including the syntactic characteristics of the clause and various morphological and semantic features of the subject (see Table 2), and perhaps 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 $F_{1\text{-}15}$ and only share the single factor that they were produced in an essay-writing situation in school. This practice does not, of course, even come close to doing justice to the complexity of the many factors that determine *any* linguistic choice and is certainly not what Péry-Woodley (1990) must have had in mind.

Given all this, our proposal as to how questions (*iv*) and (*v*) should be studied involves a much more fine-grained understanding of 'comparable situation': situations that the NSs and the NNSs are in are comparable when they are similar/identical with regard to the features $F_{1\text{-}n}$ that govern a particular phenomenon. That is to say, using the above example of *can* versus *may*, we should look at NSs' choices of *can* versus *may* when the subject is animate, singular, when the clause is interrogative, … and then compare this to NNSs' choices of *can* versus *may* when the subject is animate, singular, when the clause is interrogative. In this view, 'comparable situation' is now defined much more comprehensively in terms of linguistic/contextual

features, and the traditional decontextualised over-/underuse counts give way to what we think should be one of the fundamental questions of SLA/FLA research: 'in a situation, *S*, characterised by features $F_{1\text{-}n}$ that the learner is now in, what would a native speaker do (and is that what the learner did do)?'

From this perspective, it is obvious that mere over-/underuse counts are lacking: if a learner used *may* 10 percent less often in a corpus file than a native speaker did, that discrepancy may be entirely 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 percent less often than the native speaker says nothing about proficiency regarding *can*/*may* or over-/underuse as it is traditionally regarded: that 10 percent difference is entirely due to the learners' use of negations and, crucially, had the native speaker chosen negations as well, they would have exhibited the same perceived dispreference for *may*. In Section 4, we will outline and exemplify a two-step regression procedure that will allow corpus linguists to address questions (*iv*) and (*v*) by asking 'What would a native speaker do in the exact situation the learner is in?' and 'How do native speaker choices differ from what the learners did?'

1.2 Our example: the use of *may* and *can* in native and learner English

As an example to discuss our two corpus-based statistical strategies, we will use the example of modal verb choice which we mentioned above. This is a fitting example because modality is a semantic field that different languages carve up in different ways and thus gives rise to many potentially conflicting form–function mappings, thereby posing particular challenges to English as a second language (ESL) / English as a foreign language (EFL) learners. The case in point is the contrast between the modal verbs *may* and *can*, which can express different but overlapping meanings of possibility, permission and ability (see, for example, Coates, 1983; and Leech, 1969), but we are really only using it here as a platform for the methodological discussion; Deshors (2010) and Deshors and Gries (forthcoming) discuss the linguistic aspect in much more detail.

Corpus-based studies on modal verbs in native English, such as Leech (2004), showed that the uses of modal forms can be distinguished based on their linguistic contexts since characteristics of their contexts correlate with speakers' choices of one modal form over another. More specifically, and in a way that strikingly aligns with our plea for multifactorial analyses, Klinge and Müller (2005: 1) argued that, in order to capture modal meaning, analyses must 'cut across the boundaries of morphology,

syntax, semantics and pragmatics and all dimensions from cognition to communication are involved'. This argument has also been made in the context of learner language. For example, Guo (2005) noted that 'modal verbs do not occur randomly but with a strong tendency to co-occur with other lexical or grammatical words to form a systematic relationship within a wider environment', but also demonstrates that these co-occurrence patterns present an acquisitional challenge for English learners: '[non-native] writers are not found to be able to understand the colligational requirement and the semantic prosody of a particular pattern even though the forms are correctly produced' (Guo, 2005: 23).

Fittingly, for our present purposes, the study of modality in general, or the modal verbs *may* and *can* in particular, requires an approach that integrates a variety of linguistic factors from different levels of linguistic analysis. It is, therefore, a prime example of the kind of study that stands to benefit from the multifactorial approach advocated above. However, most existing corpus-based work on modals in IL does not pursue such a strategy. For example, Aijmer (2002) analysed advanced learners' use of key modal words based on a corpus of Swedish English writers. She adopted Granger's ICM framework and compared the frequencies of some key modal words. With regard to the modal auxiliaries specifically, she conducted two comparisons that involve NNSs with different linguistic backgrounds. In the first, she compared the frequencies of occurrence of one group of modals in native English and advanced Swedish–English IL; in the second, she compared the frequencies of occurrence of a second group of modals in Swedish, French and German learner English and NS English. Overall, Aijmer's (2002) study revealed that advanced learner writing yields 'a generalized overuse of all the formal categories of modality' (Aijmer, 2001: 72), that only German learners significantly overuse *can* and *could*, that only French learners overuse *may*, and that Swedish NSs have an extremely high use of epistemic *may*.

While we do not debate Aijmer's findings, her methodology did not equip her for the necessary exploration of the linguistic mechanisms involved in learners' choices. Put differently, her design cannot really address any of the five questions we listed under Section 1.1. She does not take all factors that may contribute to the learners' uses into consideration, which means it is even possible that the learners in her data used the modals exactly as NSs would have in exactly the same context and that any of the frequency differences are exclusively due to different context frequencies. While this may not be the only reason for the distributional differences she obtained, Aijmer's design does not – in fact, cannot – tease apart the effects of potentially conflicting forces; our proposal in Section 4, however, will help explore just that possibility.

Neff *et al*.'s (2003) study went beyond Aijmer's in that it investigated the potential (pragmatic) meaning in the target language (L2) of the association of a subject pronoun (e.g., *we*), a modal verb (e.g., *can*, *will*) and a lexical verb. Like Aijmer, Neff *et al*. (2003) used a contrastive

methodological framework to investigate the uses of modals verbs (*can*, *could*, *may*, *might* and *could*) by writers from several L1 backgrounds. Also, like Aijmer, Neff *et al*. (2003) used data extracted from the International Corpus of Learner English (ICLE) but they base their analysis on a wider selection of learner subcorpora including Dutch, French, German, Italian and Spanish learner data. Neff *et al*. (2003) used the American subsection of the Louvain Corpus of Native English Essays (LOCNESS) corpus as the control native corpus. Generally, Neff *et al*. (2003: 215) identified the case of *can* as potentially interesting 'since it is overused by all non-native writers'. They also reported that the frequency of *may* by French learners stands out in comparison to the frequencies by all other NNSs in the study. However, since their study only compared raw frequencies of occurrence with little regard to contextual features, they, like Aijmer or Collins (2009), can not, ultimately, address questions (*i*) to (*v*). Thus, such studies remain at a purely descriptive level; but even at that level, they are extremely coarse-grained.

The remainder of this paper is structured as follows: in Section 2, we will discuss, briefly, the corpus data that we analyse in our case study and their annotation. Section 3 is then concerned with exemplifying the first methodological suggestion, the notion to combine multifactorial analysis and the study of relevant interactions, and how it addresses questions (*i*), (*ii*) and (*iii*). Section 4 then develops and exemplifies a new procedure that addresses questions (*iv*) and (*v*). We conclude our paper in Section 5.

2. Data: corpora and annotation

2.1 Corpus data

In this paper, we exemplify our methodological suggestions on the basis of data on the uses of *may* and *can* in two language varieties (i.e., native and learner English) and across two learner English varieties (i.e., French- and Chinese–English IL). In order to identify, first, characteristic patterns within the two learner varieties and then contrast those patterns across learner varieties, *may* and *can* were analysed on the basis of a total of 81,408 data points, following from the annotation of 5,088 occurrences of *may* and *can* in their respective contexts in corpora of written native, French- and Chinese–English IL according to the features described below.

The data consist of instances of *may* and *can* produced by native English speakers, non-native English speakers whose first language is French and non-native English speakers whose first language is Chinese. The occurrences of *may* and *can* were extracted from three untagged corpora: the LOCNESS corpus, and the French and Chinese subsections of the ICLE. The three corpora included in the study are comparable in that they each consist of essays of approximately 500 words, all dealing with similar topics

may/can	Modal form	Native English (LOCNESS)	French–English IL (ICLE-FR)	Chinese–English IL(ICLE-CH)	Total
may	*may*	410	343	333	1,086
	may not	56	23	21	100
	Total	466	366	354	1,186
can	*can*	1,072	983	1,139	3,194
	cannot	157	212	102	471
	can't	58	50	19	127
	can not	35	45	30	110
	Total	1,322	1,290	1,290	3,902

Table 1: Summary of the occurrences of *may* and *can* in our corpus data

such as crime, education, Europe and university degrees, among others. The non-native data consists of essays written by advanced English learners in their third and fourth year at university as students of English. Table 1 lists the number of occurrences of the *may* and *can* throughout the entire data set, both in their affirmative and negated forms.

2.2 Data extraction and annotation

The data were extracted using the software R (see R Development Core Team, 2012). An R script was written to retrieve all occurrences of *may* and *can* from the data, and to import the data into a spreadsheet to allow for the annotation process. The annotation process involved coding each occurrence of *may* and *can* according to a total of sixteen co-occurring semantic and morpho-syntactic features and operationalised as variables. Each grammatical feature or variable was annotated for a range of levels. Table 2 presents an overview of the range of variables included in the study and their respective levels.

To ensure a thorough treatment of the data, each variable was encoded according to an encoding taxonomy established to allow for its measurement and its consistent treatment across the three sub-corpora. Throughout the annotation process, the assignment of semantic features to each occurrence of the two modals represented a crucial methodological step, particularly in relation to the variables VERBTYPE, VERBSEMANTICS and ANIMTYPE. For example, the variable VERBTYPE marks the types of lexical verbs used alongside *may* and *can* following Vendler's (1957: 143)

Type	Variable	Levels
Data	CORPUS	native, Chinese, French
	GRAMACC (acceptability)	yes, no
Syntactic	NEG (negation)	affirmative, negated
	SENTENCETYPE	declarative, interrogative
	CLTYPE (clause type)	main, coordinate, subordinate
Morphological	FORM	*can*, *may*
	SUBJMORPH (subject morphology)	common noun, proper noun, demonstrative/relative pronoun, other
	SUBJECTPERSON	1, 2, 3
	SUBJECTNUMBER	singular, plural
	VOICE	active, passive
	ASPECT	neutral, perfect/progressive
	SUBJREFNUMBER (subject referent number)	singular, plural
Semantic	SENSES	dynamic, other
	VERBSEMANTICS	abstract, general action, action incurring transformation, action incurring movement, perception, *etc*.
	SUBJECTANIMACY	animate, inanimate
	ANIMTYPE (type of subject animacy)	human/social role, abstract/place/time, man-made object, effected state, mental state/emotion, linguistic expression, *etc*.
	VERBTYPE (type of modalised lexical verb)	achievement/accomplishment, process, state

Table 2: Overview of the variables used in the annotation of the native, French- and Chinese–English data and their respective levels

point that the notion of time is, crucially, related to the use of a verb and is 'at least important enough to warrant separate treatment'. VERBSEMANTICS also targets lexical verbs used alongside *may* and *can*, identifying the type of information that they convey in terms of abstraction, action, communication, *etc*.; and ANIMTYPE describes the animacy of the subject.

The annotation of these latter two variables results from a careful bottom-up approach rather than any particular theoretical framework. Examples 1 and 2 illustrate the annotation of the levels 'abstract' and 'mental/cognition' for VERBSEMANTICS:

(*1*) for we may also let our imagination wander, disregarding the external concrete reality that imprisons us (ICLE-FR-UCL-0036.3), abstract
(*2*) her search for the final touch can be seen as a search for harmony (ICLE-FR-UCL-0039.2), mental/cognitive

3. Multifactoriality and interactivity

3.1 Regressions and interactions in SLA/FLA research: a gentle introduction

In this section, we will explore the data on the choice of *can* versus *may* by NSs and NNSs, and we will discuss how to improve corpus-based SLA/FLA research by addressing questions (*i*), (*ii*) and (*iii*) using the statistical technique of multifactorial regressions involving interactions. Let us begin by explaining briefly the relevant statistical terms, 'multifactorial regression' and 'interaction'. A multifactorial regression is a statistical model that tries to predict a dependent variable – an outcome: here, a binary linguistic choice between *can* versus *may* – on the basis of multiple independent variables, or predictors, using a regression equation. This regression equation quantifies each predictor's importance ('Does this predictor help make the prediction more accurate or not?') and direction of effect ('Which of the two possible outcomes does this predictor make more likely?') and is, thus, a mathematical embodiment of a sentence such as 'If predictor *A* is "m" and predictor *B* is "n", then the speaker will probably choose *x*'.

The null hypothesis in such a regression is that the predictors do not interact. This means that, in a regression equation involving two predictors, *A* and *B*, the default assumption is that *A* always has the same effect on the choice of *x* regardless of what *B* is and *vice versa*. Let us look at an example on the basis of the *can* versus *may* data. For the sake of simplicity we only study three predictors: ASPECT (neutral *versus* perfect/progressive), NEGATION (affirmative *versus* negative) and CORPUS (native *versus* French *versus* Chinese). A multifactorial regression model with these predictors, but without interactions, would be written as follows:

(*a*) Form ~ Corpus + Aspect + Negation

For our data, this multifactorial regression model is highly significant (likelihood ratio = 184.88, $df = 4$, $p < 0.0001$), but note what its results mean.

Predictor	Level of predictor	Predicted probability of *may*
ASPECT	Neutral	0.211
	Perfect/progressive	0.808
NEGATION	Affirmative	0.241
	Negative	0.122
CORPUS	Native	0.25
	French	0.204
	Chinese	0.196

Table 3: Predicted probabilities of *may* in a model without interactions

As explained above, the regression results state for each predictor in the model how strongly and in which direction it affects the choice of *can* versus *may*. In a way that is not relevant here (see Gries, 2013), this allows the user to compute predicted probabilities for *may*, which are shown in Table 3.

While this model is multifactorial and, thus, answers question (*i*) and also satisfies the first above desideratum, it is not as revealing as it should be from the perspective of contrastive analysis. Table 3 shows that the choices of *can* versus *may* differ depending on ASPECT (*may* is more likely with perfect/progressive), NEGATION (*may* is more likely in affirmative clauses), and the speakers' L1 (*may* is more likely in NL than in both ILs). What is then the problem of this model? The problem is that it does not reveal whether there are interactions. An interaction is when the assumption that the effect of a predictor on the choice of *can* versus *may* is the same regardless of the other predictors. Maybe *A* has a particular effect on *x* when *B* is "n", but another effect on *x* when *B* is not "n", but neither model (*a*) shown above, nor Table 3, can reveal this.

Two interactions scenarios, which are statistically identical, but conceptually/linguistically very different, may be relevant for such cases: first, the predictors that have the potential to interact may all be linguistic/conceptual descriptors of the situation in which a speaker made a linguistic choice. This is the case here with ASPECT and NEGATION. Thus, if those predictors interact, then that would mean that the probability of *may* in negated sentences is not the same in both aspects (0.122), but different. To determine whether such an interaction exists, one could fit, for instance, model (*b*), which includes the relevant interaction term:

(*b*) Form ~ Corpus + Aspect + Negation + Aspect:Negation

		Predicted probability of *may*	
Predictor	Predictor levels	ASPECT:neutral	ASPECT:perf/progr
NEGATION	Affirmative	0.234	0.819
	Negative	0.117	0.737

Table 4: Predicted probabilities of *may* in a model with the interaction ASPECT:NEGATION

As the result in Table 4 indicates, this makes a huge difference:[5] the predicted probability of *may* with neutral aspect in Table 3 (0.211) is only similar to the one for affirmative clauses (0.234), but is quite dissimilar to what is predicted for negated clauses (0.117). An analysis without interactions reporting the results of Table 3 would fail to note that the relationship between ASPECT and NEGATION is more complex than Table 3 suggests: the effect of ASPECT on *can* versus *may* seems to depend also on NEGATION. This is, thus, a more precise way of answering question (*i*).

The second interaction scenario is at least as important as the one above, so far as contrastive (IL) analysis is concerned. In this scenario, one predictor is of the same conceptual/linguistic nature as above, but the one it may interact with is the L1 of the speaker, here CORPUS. What does it mean if a predictor such as ASPECT interacts with CORPUS? It means that the effect of ASPECT is not the same across all L1s, and if that is not the question that any contrastive (IL) analysis would want to answer, what is? To determine whether this second type of interaction exists, one could fit, for instance, model (*c*), which includes both of the interactions that CORPUS can enter into:

(*c*) Form ~ Corpus + Aspect + Negation + Corpus:Aspect + Corpus:Negation

The results are shown in Table 5 and indicate very clearly the huge loss of information any researcher incurs who does not include this second type of interaction in their analysis.

Now that these interactions have been included, we see that the large probability of *may* in perfective/progressive aspect listed in Table 3 (0.808) is a huge over-generalisation: the interaction CORPUS:ASPECT shows that the strong preference of perfective/progressive aspect for *may* really only holds

[5] We leave aside here the issue of whether the interaction is significant since this beside the point we are trying to make, namely that one needs to check the interaction, not blindly assume its absence.

		Predicted probability of *may*		
Predictor	Level of predictor	Native	French	Chinese
ASPECT	Neutral	0.241	0.193	0.196
	Perfect/progressive	0.958	0.564	0.416
NEGATION	Affirmative	0.268	0.237	0.214
	Negative	0.18	0.066	0.128

Table 5: Predicted probabilities of *may* in a model with interactions with CORPUS

true for the NSs. The NNSs, by contrast – and, in particular, the Chinese learners – behave completely differently. On the other hand, the predicted probability of *may* in affirmative clauses listed in Table 3 (0.241) is not too unproblematic: all three speaker groups have predicted probabilities of *may* that are not too far away from 0.241, 0.268, 0.237 and 0.214. Thus, some results in Table 3 can be taken at face value (the ones for affirmative clauses), some cannot at all (the ones for perfective/progressive aspect) but, needless to say, we hope, this cannot be inferred from Table 3 but only from a regression that is multifactorial and that includes the right kind of interactions, which is, unfortunately, hardly ever undertaken in SLA/FLA research. Only this type of regression can answer questions (*ii*) and (*iii*).

The next logical step would now be to fit model (*d*) to see whether the possible interaction of ASPECT:NEGATION shown in Table 4 is, in fact, the same in all three speaker groups.

(*d*) Form ~ Corpus + Aspect + Negation + Corpus:Aspect + Corpus:Negation + Corpus:Aspect:Negation

However, rather than discussing this model, we will immediately proceed to the much more realistic scenario – namely, an analysis that does not just include two predictors and CORPUS, but all the predictors we annotated the data for (recall Table 2). Some aspects of this analysis will be discussed in the next section.

3.2 Regressions and interactions in SLA/FLA research: a more realistic example

In this section, we will discuss the results of a regression-based approach to *can* versus *may*, but this time based on all of the data discussed above.

Specifically, we set the reference level of every categorical predictor to the most frequent level (to reduce the risk of collinearity, the ordinal predictor SUBJECTPERS was left in its ordinal ordering: 1, 2, 3) and we then used an automatic iterative model-fitting procedure to find the best regression model for the corpus data. This process:

- Was bidirectional in the sense that, at every step of the model selection process:
  - All predictors in the model were tested for whether their deletion would improve the model;
  - All predictors not in the model were tested for whether their inclusion would improve the model; and,
  - Predictors were only deleted if they did not participate in higher-order terms in the model.
- AIC was used as the test criterion to decide how the model should be improved;
- Was constrained such that it would only be applied to cases without any missing data points and would only allow models between:
  - The minimal model of only an overall intercept: FORM $\sim$ 1; and,
  - The maximal model of every predictor and its interaction with CORPUS. (For explanations of modelling, see Crawley, 2007: Chapter 9; or Gries, 2013: Chapter 5.)

This iterative model selection process resulted in an overall significant model: likelihood ratio $= 911.99$, $df = 60$, $p < 10^{-75}$. There was no sign of over-dispersion ($p_{\chi 2} > 0.9$) and all variance inflation factors were < 10. The classificatory power of the model was acceptable: Nagelkerke's $R^2 = 0.26$, $C = 0.78$, and its classification accuracy was 0.804. (This model, and its summary statistics, was nearly identical to a model resulting from a traditional backwards model selection process based on p-values.)

Given the above argumentation, this type of final model is interesting in two different ways. First, and trivially, it provides descriptive information on how well choices of *may* and *can* can be predicted on the basis of morpho-syntactic and semantic characteristics of their contexts in all of the corpora under consideration. This part of the results shows up as a set of significant main effects and interactions between them that do not involve CORPUS and that show how predictors influence the choice of *can* versus *may* in the same direction and strength in both the NL and the IL data. Consider, briefly, two of these main effects illustrated in Figure 2.

The data show that *may* is more likely in declarative clauses (which differ significantly from interrogatives in this regard) and with third-person subjects (which differ significantly from first- and second person, which in turn are quite similar to each other in their dispreference for *may*). Since these results do not involve CORPUS, this effect is true of both the native and the learner data, which is also obvious from the closeness of the predicted

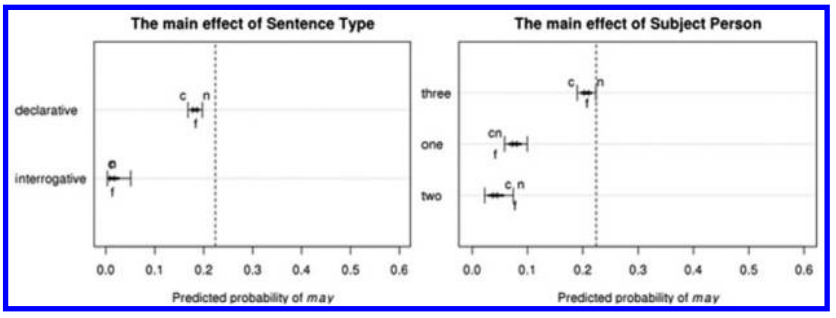

Figure 2: The main effects of SENTENCETYPE (left panel) and SUBJECTPERSON (right panel) on the predicted probability of *may* (versus *can*) with all other predictors in the model: × indicates the predicted probability; the heavy line and the error bars indicate one-standard error and 95 percent confidence intervals. The vertical dashed line marks the overall observed percentage of *may* in the data

probabilities of *may* for each corpus as indicated by the first letters of the corpus ('native', 'french', 'chinese'); thus, again, this part of the results only answers question (*i*).

What about questions (*ii*) and (*iii*), which are more important from a contrastive perspective? What do the data reveal with regard to where and how the choices differ from each other in the three corpora? Results answering these questions show up as significant interactions of a predictor with CORPUS; four such interactions are illustrated in Figure 3.

For instance, the interaction CORPUS:NEGATION reveals that, considering all annotated features at the same time, in affirmative clauses, the three types of NS are rather close together, but the Chinese learners use *may* somewhat less; in negative clauses, instances of *may* become less frequent in each corpus, but particularly for the Chinese and even more for the French. A similar result is obtained for CORPUS:ASPECT: with neutral aspect, the three corpora look very similar, but with perfect/progressive, the French and particularly the Chinese learners exhibit a much greater avoidance of *may* than the native speakers. With CORPUS: SUBJECTANIMACY, the main component of the interaction is that the French learners use *may* with inanimate subjects much more than with animate ones and that that change is greater than for the NSs. Finally, the interaction CORPUS:VERBTYPE shows that different verb types do not seem to result in great differences for the Chinese learners; the French learners have a similar usage profile for accomplishments/achievements but use *may* more with states, but the NSs' preference for *may* with states is much higher.

This level of precision is impossible to come by without the kind of multifactorial regression with interactions advocated here. Thus, it is only this approach that really allows us to speculate on what the motivations for the different findings are. Deshors (2010) argued that many of the

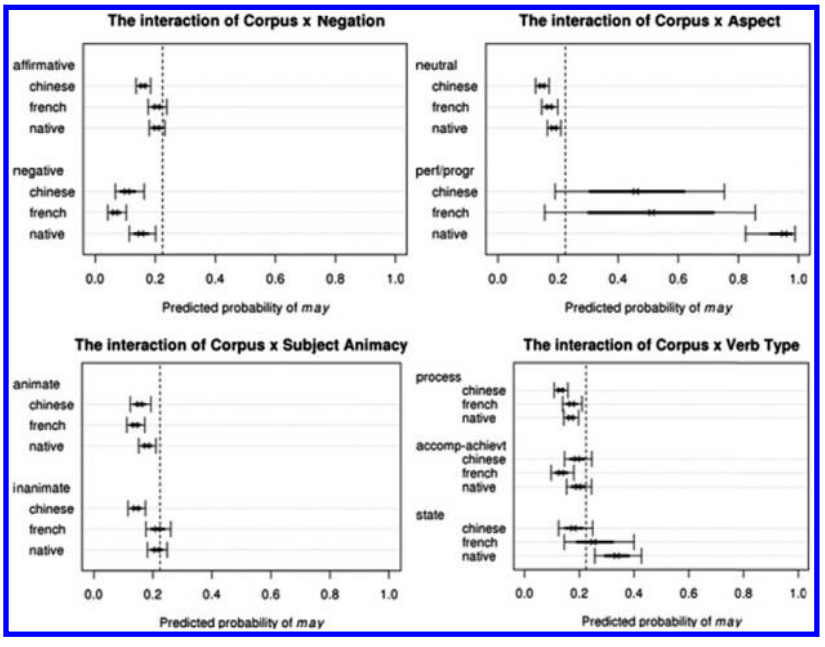

Figure 3: The interactions of CORPUS with NEGATION (top left panel), ASPECT (top right panel), SUBJECTANIMACY (bottom left panel), and VERBTYPE (bottom right panel) on the predicted probability of *may* (vs. *can*) with all other predictors in the model

differences between the NSs and the French learners could be related to Rohdenburg's (1996) complexity principle, which postulates that, in more complex environments, speakers favour more explicit variants over less explicit ones. In her data, and also the data here, we can note the tendency of learners in more complex situations to resort to the overall more default modal *can*. This is, arguably, the default, given that it is:

- Generally more frequent in terms of its token frequency of occurrence in corpora;
- Used in a larger variety of contexts (i.e., exhibits a larger type frequency of uses); and,
- Is acquired earlier.

For instance, negated clauses are arguably more complex than their affirmative counterparts, if only for the additional morpho-syntactic material. A similar argument can be made for CORPUS:ASPECT: once morphologically more complex aspectual structures are used, both learners fall back on *can* more often than the NSs. In the case of Chinese English learners, and in line with Ellis and Sagarra's (2001) finding on L2 learners' attentional biases towards language, differences between the Chinese and

English aspect systems may explain the existence of *may*/*can* co-occurrence patterns characteristic of Chinese–English IL. This view is mainly based on Xiao and McEnery's (2004: 3) statement that '[e]ven though both languages [English and Chinese] mark aspect, the aspect system in these two languages differs significantly.' More specifically,

> While English and Chinese both have a progressive viewpoint, it is used differently in the two languages (. . .). Chinese does not have the perfect, yet English does. Also, the English simple aspect does not correspond to the perfective viewpoint in Chinese.
>
> (Xiao and McEnery, 2002: 3)

With regard to French, progressive aspect is generally marked using the phrase *en train de* (i.e., *in the process of*) instead of a morphological marker, as is the case in English. In common with Chinese learners, the lack of a verbal morphological marker for progressive aspect in French could explain: (*i*) French English learners' tendency to use *can* more frequently than *may* in such contexts; and (*ii*) why French and Chinese English learners behave similarly in relation to aspect.

Thus, the combination of a multifactorial approach with the necessary interactions with CORPUS has a lot to offer and should replace the descriptively much simpler and, thus, lacking mere frequency-based descriptions of over- and underuse. These are still too frequent in the literature – but see Tono (2004) for a rare, exemplary approach that is similar in spirit to ours, but is not equipped to deal with all the complexities and details of learner corpora.

4. The MuPDAR approach towards differences between NL and IL

4.1 The MuPDAR approach: an overview

This section is concerned with our methodological suggestions concerning questions (*iv*) and (*v*) and focusses on what an NS would do in the exact situation the learner is in and how, if at all, NSs' choices differ from the learners' choices. The statistical approach, which we refer to as Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR) is somewhat complex but builds conceptually upon the regression approach discussed in the previous section. Given the complexity of the method, we provide an explanatory flowchart in Figure 4.

The first steps are old hats: we first generated a concordance of *can* and *may* in both NL and NNS/IL data and annotated it for the relevant features that govern the choice of modal verb. Then we fit a first regression R_1 only on the NS data, from which we derive a regression equation that allows us to make predictions of modal verb choices. If that fit is good, R_1 is

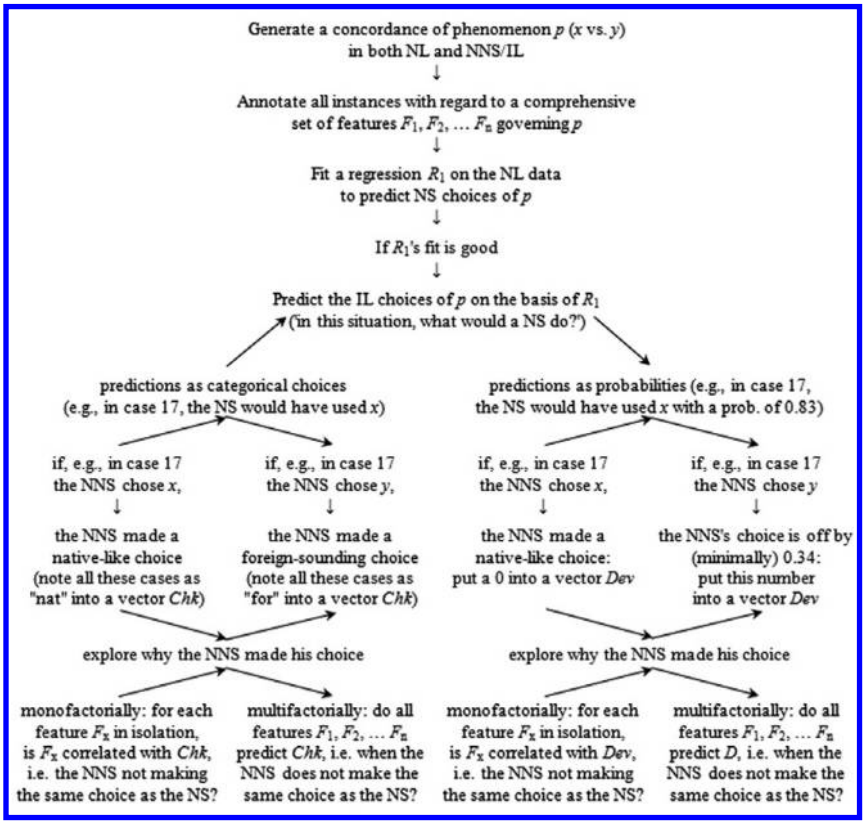

Figure 4: Flowchart of the MuPDAR approach

then applied to the IL data (either one or more IL varieties). This returns for every choice in the IL data a prediction of *can* or *may* and, thus, answers the question, 'Given all the features of the linguistic/contextual situation that the NNS is in right now, what would an NS use, *can* or *may*?'

Now, each prediction can be interpreted on two levels of granularity. The left part of Figure 4 takes the predicted probability of *may* and turns it into a categorical prediction: when the predicted probability of *may* is $\geq$0.5, one says the analysis predicts *may*, otherwise it predicts *can*. Thus, for every case, there are two possibilities: the learner chose what is predicted as an NS choice (which we note in a new vector Chk as 'nat') or not (which we note in Chk as 'for'). Then we have two analytical possibilities. First, (*i*) we can compare Chk either to every annotated feature in isolation – that is to say, cross-tabulate (*i*) ASPECT with Chk, (*ii*) NEGATION with Chk, and even (*iii*) CORPUS with Chk (if we had more than one learner variety). This would show us (*i*) which levels of ASPECT give rise to learners not doing what an NS would do, (*ii*) which levels of NEGATION give rise to learners not doing what an NS would do, and, (*iii*) crucially, which learner variety is

more or less successful at doing what NSs would do. The second analytical possibility would be a logistic regression in which all features are used to predict when NNSs do not choose what NSs would have, which can then be interpreted as reflecting the difficulty of features for learners of particular varieties.

The second level of granularity is the more precise one and is shown in the right part of Figure 4. In the approach above, the predicted probability of *may* was transformed into a categorical choice, which means that a prediction that an NS would use *may* with a probability of 0.51 is treated in the same way as a prediction that an NS would use *may* with a probability of 0.92 (because both values are $\geq$0.5). The second approach here proceeds differently. For every case, we record the probability p with which *may* is predicted. Then, a vector Dev (for deviations) is created, which is set to 0 whenever the NNS did what was predicted for the NS. But whenever the NNS did not do what was predicted as what an NS would do, the value of *Dev* is set to p–0.5.

Let us briefly explain the reason for this step: if the NS model R_1 makes a relatively weak prediction of *may* with a predicted probability of *may* of say 0.6, then, if a NNS chose *can*, (i.e., not what the NS would have chosen), the vector Dev's element for this case is set to 0.6–0.5 = 0.1. If, on the other hand, the NS model R_1 makes a very strong prediction of *can* with a predicted probability of *may* of, say, 0.1, then, if the NNS chose *may*, (i.e., not what the NS would have said), the vector Dev's element for this case is set to 0.1–0.5 = –0.4. Thus, what this step does is to create a vector Dev that quantifies how much the NNS's choice was off from what an NS would have chosen in an identical multifactorial situation: Dev-values of 0 mean the learner got it right; Dev-values other than 0 mean the NNS did not make the predicted NS choice, and the deviation of each Dev-value from 0 indicates how much the NNS was 'off', with the maxima being 0.5 and –0.5. For instance, if the prediction from R_1 is 99.9 percent that the NS would choose *may* but the NNS chose *can*, then the NNS is off by nearly the theoretical maximum: 0.999–0.5 = 0.499.

Once all values of Dev have been defined as above, we fit a linear regression, R_2, that tries to predict Dev – that is to say, where and how much the NNSs deviate in their modal verb uses from the NSs – on the basis of all the annotated linguistic parameters. If the fit of R_2 is good (i.e., if one can say that R_2 captures well where the NNS go wrong and how) then determine which of the annotated features yield the largest and smallest Dev-values because it is those features that reveal where the NNSs still deviate most from the NSs (in the context of all predictors involved) and that can, thus, be seen as being the most difficult for them.

After this complex explanation, we will now discuss the results of this approach when applied to our data. To keep the level of complexity manageable, we will only apply R_1 to the data of the French learners, not also (in contrast) to those of the Chinese learners (but see below), and only pursue the more fine-grained analysis.

	Prediction from NS model: *can*	Prediction from NS model: *may*	Totals
French learners: *can*	**1,159**	39	1,198
French learners: *may*	232	**67**	299
Totals	1,391	106	1,497

Table 6: Cross-classification matrix: NNS choices in the rows, NS predictions in the columns (accuracy: 81.9 percent)

4.2 The MuPDAR approach: some results

We first fit the NS model, R_1, to only the NS data and obtained a highly significant logistic regression model with a good fit: likelihood ratio = 449.34, $df = 32$, $p < 10^{-75}$. There was no sign of over-dispersion ($p_{\chi 2} > 0.9$) and all variance inflation factors were < 10. The classificatory power of the model was good – better, in fact, than that of the model on all three corpora: Nagelkerke's $R^2 = 0.33$, $C = 0.8$; and its classification accuracy was 0.79. Given the good fit, we undertook the next step and applied R_2 to the 1,497 complete cases of the French learners. As a result of this, we obtained the predicted probabilities of *may* and, thus, the choices of *may* and *can* that the NS model says the French learners should make (to sound native-like). On the whole, and this is reassuring, the French learners behaved a lot like the NS data suggest, as is shown in Table 6.

Following the procedure, we then generated a vector Dev which, given how well the French learners did as a whole, had the expected distribution: its median is 0, its mean is very close to 0 (–0.037), and the central 80 percent of the data are 0s. Thus, we then fit a linear regression model (with the same approach as above) to determine how well the learners' deviations from the native-like behaviour can be predicted from the factors that govern the use of *may* and *can*. As it turns out, they can be predicted very well: adj. $R^2 = 0.94$; $F = 359.6$; $df = 64$, 1,432 and $p < 0.001$, which means we could begin to explore which features give rise to the non-nativeness of the French NNSs. The first and most general finding is that the French learners had far fewer difficulties with *can* than with *may*, which is not surprising given *can*'s status as the more general and less restricted form. This also means that the most interesting findings of the above procedure involves interactions of predictors with FORM – in other words, interactions that can reveal which factors were unproblematic with *can* but problematic with *may*, which is why, for the sake of comparability, we discuss four significant interactions that are related to those discussed in Section 3. Consider Figure 5.

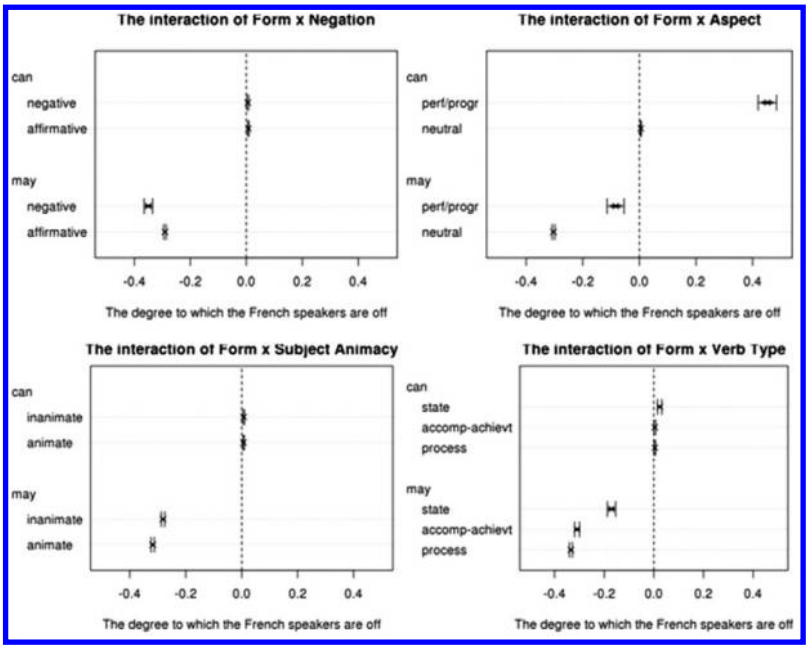

Figure 5: The interactions of FORM with NEGATION (top left panel), ASPECT (top right panel), SUBJECTANIMACY (bottom left panel) and VERBTYPE (bottom right panel) on the deviations of the French learners' uses with all other predictors in the model

The results are quite interesting, both in general and in how they relate to, but also complement, the results for the French learners from Section 3. With regard to FORM:NEGATION, we saw above that French speakers disprefer *may* much more in negative clauses than in affirmative ones, and we could see that that was how they differed from the NS. Now we also see that this is one of the main areas where they go wrong: in this case, both types of results lead to the same conclusion.

The interaction FORM:ASPECT, however, shows that this second approach does not merely replicate the results of the first. We saw above that, with neutral aspect, the French learners' predicted probability of *may* was the same as that of the NSs, but that with perfect/progressive aspect, they dispreferred *may* more than the NSs. The present analysis offers a more precise picture: the top right panel shows that the French learners are really only on target with *can* in neutral aspect, but their choices of *can* with perfect/progressive aspect are often somewhat off target. When the French learners use *may* with perfect/progressive aspect, they are on average a bit off, but their uses of *may* with neutral aspects again differ considerably from what NSs would have chosen in those very same examples.

What about FORM:SUBJECTANIMACY? Again, the uses of *can* are very much on target, but when it comes to *may*, the French learners are still a

long way from the NSs' patterning, especially with animate subjects. Finally, FORM:VERBTYPE: as usual, *may* is the problem for the learners, but not uniformly across verb types. For instance, learners' choices of *may* with state verbs are significantly more on target than those with other verbs. If we compare this result to Figure 3, we also note how the current approach results in a finer resolution. Figure 3 showed that the predicted probability of *may* with process verbs did not differ between NSs and French NNSs. However, R_1 in Section 3 and its results in Figure 3 do not reveal how many of these choices then were wrong and by how much; Figure 5, however, shows that very clearly. While the overall predicted probability of *may* with process verbs is the same for English and French speakers, the latter are much more often wrong in that context.

To summarise, the procedure developed in this section nicely complements the more standard regression approach outlined in Section 3. The former served to highlight how analytical approaches from research on alternations in L1 can be incorporated into SLA research, with a particular emphasis on: (*i*) including many contextual factors at the same time; and (*ii*) the idea that NL and IL varieties must be allowed to interact statistically with all other factors/predictors. This was then shown to reveal the factors that drive speakers' choices in the data as a whole and, more importantly, in contrast between NL and IL varieties. However, what it does not show particularly well is to what degree the different weights of the factors lead the learner to make wrong decisions on a case-by-case basis; this is what the MuPDAR approach in this section targets, and we have demonstrated how well deviations from the native target can be identified and how they can reveal the linguistic characteristics that result in learners' foreign-soundingness. One natural extension of this approach involves not just studying one IL variety but more, and then adding the predictor CORPUS to R_2. Gries and Wulff (2013b) applied our MuPDAR logic to the phenomenon of prenominal order ('big red ball' versus 'red big ball') and showed how interactions of CORPUS and other predictors in R_2 then reveal how learner varieties differ in how they deviate from NS choices.

5. Concluding remarks

Contrastive studies occupy a central place in SLA research and its importance to the analysis of over-/underuses cannot be overestimated. To reiterate: in this paper, we attempted to raise its methodological profile by (*i*) formulating five questions that we believe should govern contemporary corpus-based SLA research, and then (*ii*) proposing three main improvements which help to address these questions at a level of detail that has hitherto been rare:

– **Multifactoriality**: instead of the still common monofactorial analysis of over-/underuses, we propose to use the type of

multifactorial statistical tools that have taken corpus-based research on choices, or alternations, by storm. More specifically, instead of the mere counting of (over-/under-) uses of a particular expression and comparing their frequencies across native and interlanguage varieties (sometimes in connection with other individual linguistic features), we propose to annotate the relevant uses for a multitude of characteristics and subject them to multifactorial methods. This multifactorial perspective is also more appropriate from a cognitive or usage-/exemplar-based or Competition Model perspective in which language learning, representation and processing is characterised as clouds in, or movement through, multi-dimensional exemplar space, or weightings of cues and should, therefore, particularly appeal to scholars who approach SLA from this theoretical perspective.

- **Interactivity**: rather than just exploring one predictor or entering all predictors into a regression without interactions, we propose that it is absolutely essential to include a predictor representing the varieties in question (CORPUS, in the present study) in the regression model, as well as, crucially, its interactions with all other predictors. This is analogous to how TIME or TIMESTAGE should be a predictor in diachronic analyses that is allowed to interact with all other predictors so that changes over time can be tracked reliably (see Jankowski, 2004, for a case where this has not happened and Gries and Hilpert, 2010, for an example of where it has).
- **MuPDAR**: the statistical analysis of how NNSs deviate from choices NSs would probably have made. This method helps to identify – again in a multifactorial way that is compatible with contemporary exemplar-based approaches – the areas or, more technically, combinations of predictors or cues (to take the parlance of the Competition Model) where learners deviate from the desired native-like performance most drastically, which, in turn, informs contrastive analysis.

Note also that the MuPDAR approach is quite versatile and can yield completely new findings in many different research scenarios –whenever one 'thing' can be considered a standard or target against which other 'things' are compared. Apart from SLA/FLA applications, such as the one above, the following applications come to mind:

- How (much) do different, say, English varieties differ from BrE as the mother variety with regard to phenomenon *P* (e.g., verb-construction associations) and how do these findings relate to evolutionary models of New Englishes (see Schneider, 2003);
- How (much) does the Spanish of heritage speakers differ from 'standard' peninsular or, in the US, Mexican Spanish speakers and

learners of Spanish with regard to phenomenon *P* (e.g., the use of reverse constructions with *gustar*) and how does this relate to theories that compare NSs, HSs and NNs of Spanish?
– How (much) does a child's production of phenomenon *P* differ from the patterns in child-directed speech (at different points in time)?

At this point, there is no reason why this approach cannot be extended to matters of pronunciation or lexical choices: the comparison of different groups of data by two subsequent regressions is an extremely powerful tool.

While the main point of this paper is not to downplay previous work, we do believe that corpus-based SLA research needs to follow the most recent developments in quantitative corpus linguistics. If anything, learner corpus research is more complex than native-language corpus research, given how learner patterns are less stable/consistent and less predictable than those of NSs while all are factors that make NS corpus research difficult still apply. In addition, and bearing in mind that beyond learners' L1, extra-linguistic factors such as level of proficiency or time spent abroad also interfere with learners' linguistic choices, the use of statistical methods such as regression (which is compatible with the analysis of such types of factors) can help us to investigate linguistic and non-linguistic aspects of SLA alongside one another in an integrated and unified fashion. Ultimately, it is clear that the use of powerful techniques is increasingly important for the field to evolve and advance – a goal to which we hope this paper has contributed.

Acknowledgments

W e would like to thank Stefanie Wulff for discussion and Gaëtanelle Gilquin for detailed feedback on an earlier draft of this paper. Also, we are grateful to two anonymous reviewers for their very helpful comments. The usual disclaimers apply.

References

Aijmer, K. 2002. 'Modality in advanced Swedish learners' written interlanguage' in S. Granger, J. Hung and S. Petch-Tyson (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 55–76. Amsterdam: John Benjamins.

Coates, J. 1983. The Semantics of the Modal Auxiliaries. London: Croom Helm.

Collins, P. 2009. Modals and Quasi Modals in English. Amsterdam: Rodopi.

Cosme, C. 2008. 'Participle clauses in learner English: the role of transfer' in G. Gilquin, S. Papp and M. Belén Díez-Bedmar (eds) Linking up Contrastive and Learner Corpus Research, pp. 177–200. Amsterdam and Atlanta: Rodopi.

Crawley, M.J. 2007. The R Book. Chichester: John Wiley and Sons.

Deshors, S.C. 2010. A Multifactorial Study of the Uses of 'may' and 'can' in French–English Interlanguage. Unpublished PhD thesis. Sussex: University of Sussex.

Deshors, S.C. 2011. 'Towards a corpus-based identification of linguistic structure in learner language', paper presented at the 'Cognition and Context: Empirical approaches to social cognition and emergent language structure' workshop on the occasion of the forty-fourth Annual Meeting of the Societas Linguistica Europaea, Universidad de la Rioja. 8–11 September 2011. Logrono, Spain.

Deshors, S.C. and St.Th. Gries. Forthcoming. 'A case for the multifactorial assessment of learner language: the uses of *may* and *can* in French–English interlanguage' in D. Glynn and J. Robinson (eds) Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics. Amsterdam and Philadelphia: John Benjamins.

Ellis, N.C. and N. Sagarra. 2011. 'Learned attention in adult language acquistion: a replication and generalization study and meta-analysis', Studies in Second Language Acquisition 33 (4), pp. 589–624.

Gilquin, G. and M. Paquot. 2007. 'Making the most of contrastive interlanguage analysis', paper presented at the ICAME 28 Conference. 23–27 May 2007. Stratford-upon-Avon, UK.

Granger, S. 1996. 'From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora' in K. Aijmer, B. Altenberg and M. Johansson (eds) Languages in Contrast: Text-based Cross-linguistic Studies, pp. 37–51. Lund University Press.

Granger, S. 2002. 'A bird's eye view of learner corpus research' in S. Granger, J. Hung and S. Petch-Tyson (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 3–33. Amsterdam and Philadelphia: John Benjamins.

Granger, S. 2004. 'Computer learner corpus research: current status and future prospects' in U. Connor and T. Upton (eds) Applied Corpus Linguistics: A Multidimensional Perspective, pp. 123–45. Amsterdam: Rodopi.

Gries, St.Th. 2000. Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement. PhD thesis. Hamburg: University of Hamburg.

Gries, St.Th. 2003a. Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement. London and New York: Continuum Press.

Gries, St.Th. 2003b. 'Towards a corpus-based identification of prototypical instances of constructions', Annual Review of Cognitive Linguistics 1, pp. 1–27.

Gries, St.Th. 2013. Statistics for Linguistics with R: A Practical Introduction. (Second edition.) Berlin and New York: De Gruyter Mouton.

Gries, St.Th. and M. Hilpert. 2010. 'Modeling diachronic change in the third person singular: a multi-factorial, verb- and author-specific exploratory approach', English Language and Linguistics 14 (3), pp. 293–320.

Gries, St.Th. and S. Wulff. 2013a. 'The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research', International Journal of Corpus Linquistics 18 (3), pp. 327–56.

Gries, St.Th. and S. Wulff. 2013b. 'Differences in prenominal adjective order between native speakers and learners: a two-step regression-analytic procedure', paper presented at AACL 2013. San Diego. San Diego State University.

Grondelaers, St., D. Speelman and D. Geeraerts. 2002. 'Regressing on *–er*: statistical analysis of texts and language variation' in A. Morin and P. Sébillot (eds) Sixth International Conference on the Statistical Analysis of Textual Data, pp. 335–46. Rennes: INRIA.

Guo, X. 2005. 'Modal auxiliaries in phraseology: a contrastive study of learner English and native speaker English' in P. Danielsson and M. Wagenmakers (eds) proceedings from the Corpus Linguistics Conference Series. University of Birmingham.

Jankowski, B. 2004. 'A transatlantic perspective of variation and change in English deontic modality', Toronto Working Papers in Linguistics 23 (2), pp. 85–113.

Jarvis, S. and S.A. Crossley (eds). 2012. Approaching Language Transfer through Text Classification Explorations in the Detection-based Approach. Bristol: Multilingual Matters.

Klinge, A. and H.H. Müller. 2005. 'Modality: intrigue and inspiration' in A. Klinge and H.H. Müller (eds) Modality Studies in Form and Function, pp. 1–4. London and Oakville, Connecticut: Equinox.

Krzeszowski, T.P. 1990. Contrasting Languages: The Scope of Contrastive Linguistics. Berlin and New York: Mouton de Gruyter.

Leech, G. 1969. Towards a Semantic Description of English. Indiana: Indiana University Press.

Leech, G. 2004. Meaning and the English Verb. (Third edition.) London: Longman.

Neff, J., E. Dafouz, H. Herrera, F. Martínez and J. Pedro Rica. 2003. 'Contrasting the use of learner corpora: the use of modal and

reporting verbs in the expression of writer stance' in S. Granger and S. Petch-Tyson (eds) Extending the Scope of Corpus-based Research: New Applications, New Challenges, pp. 211–30. Amsterdam: Rodopi.

Péry-Woodley, M.P. 1990. 'Contrasting discourses: contrastive analysis and a discourse approach to writing', Language Teaching 24 (3), pp. 205–14.

R Development Core Ream. 2012. R: A Language and Environment for Statistical Computing. Foundation for Statistical Computing. Vienna, Austria. Available online, at: http://www.R-project.org

Rohdenburg, G. 1996. 'Cognitive complexity and increased grammatical explicitness in English', Cognitive Linguistics 7 (2), pp. 149–82.

Schneider, E.W. 2003. 'The dynamics of new Englishes: from identity construction to dialect birth', Language 79 (2), pp. 233–81.

Szmrecsanyi, B. 2006. Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Pyscholinguistics and Discourse Analysis. Berlin and New York: Mouton de Gruyter.

Szmrecsanyi, B. and B. Kortmann. 2011. 'Typological profiling: learner Englishes versus indigenized L2 varieties of English' in J. Mukherjee and M. Hundt (eds) Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap, pp. 167–87. Amsterdam: John Benjamins.

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 and D. Stewart (eds) Corpora and Language Learners, pp. 45–66. Amsterdam and Philadelphia: John Benjamins.

Vendler, Z. 1957. 'Verbs and times', Linguistics in Philosophy 66 (2), pp. 143–60.

Xiao, R. and T. McEnery. 2002. 'A corpus-based approach to tense and aspect in Chinese-English translation' in the first International Symposium on Contrastive and Translation Studies between Chinese and English. 8–11 August 2002. Shanghai, China. Accessed 16 August 2011, at: http://eprints.lancs.ac.uk/68/

Xiao, R. and T. McEnery. 2004. Aspect in Mandarin Chinese: A Corpus-based Study. Amsterdam and Philadelphia: John Benjamins.