A case for the multifactorial assessment of learner language

The uses of *may* and *can* in French-English interlanguage

Sandra C. Deshors and Stefan Th. Gries
New Mexico State University / University of California, Santa Barbara

In this study, we apply Gries and Divjak's Behavioral Profile approach to compare native English *can* and *may*, learner English *can* and *may*, and French *pouvoir*. We annotated over 3,700 examples across three corpora according to more than 20 morphosyntactic and semantic features and we analysed the features' distribution with a hierarchical cluster analysis and a logistic regression. The cluster analysis shows that French English learners build up fairly coherent categories that group the English modals together followed by *pouvoir*, but that they also consider *pouvoir* to be semantically more similar to *can* than to *may*. The regression strongly supports learners' coherent categories; however, a variety of interactions shows where learners' modal use still deviates from that of native speakers.

Keywords: Behavioral Profiles, hierarchical cluster analysis, logistic regression, modal verbs

1. Introduction and overview

Acquiring a foreign language is one of the most cognitively challenging tasks, given how languages differ in every level of linguistic analysis. From a cognitively and psycholinguistically-oriented perspective, learning a language requires identifying a very large amount of co-occurrence data – tense *t* and number *n* require subject-verb agreement with morpheme *m*, idiom *i* consists of word *w* and word *x*, communicative function *f* is communicated with intonation curve *c*, etc. – as well as storing and retrieving them. Crucially, these types of co-occurrences are typically probabilistic only rather than absolute/deterministic and, thus, hard to discern and learn: usually, learners need to cope with many-to-many mappings between forms and functions,

and often it is only the confluence of differently predictive information on several levels of linguistic analysis that narrows down the search for a particular meaning (in comprehension) or a particular form (in production). In the Competition Model by Bates and MacWhinney (1982, 1989), for example, this situation is modeled on the assumption that forms and functions are cues to functions and forms, respectively, and many different cues of different strengths, validities, and reliabilities must be integrated to, say in production, arrive at natural-sounding choices.

Semantics is a particularly tricky linguistic domain in this regard, in native language, but even much more so in foreign language learning. Not only do languages often carve up semantic space very differently (so that the categories of the language acquired first will influence category formation in the following), but semantic differences are also often much less explicitly noticeable (than, say, the presence or absence of a plural morpheme), which makes the identification of probabilistic co-occurrence patterns all the more difficult. In order to allow for a precise description of semantic, or more generally functional, characteristics of synonyms, antonyms, and senses of polysemous words, Gries and Divjak developed the so-called Behavioral Profile (BP) approach (cf. Gries and Divjak 2009). This approach, to be discussed in more detail below, is highly compatible with a psycholinguistic perspective of the type outlined above and involves a very fine-grained annotation of corpus data as well as their statistical analysis.

The method of behavioral profiles has been successfully employed in a variety of contexts – synonyms, antonyms, and word senses of polysemous words have been studied both within one L1 or across two different L1s – as well as having received first experimental support, but so far there have been no studies that test the BP approach's applicability to L1 and L2 data, which is what we will undertake here. The semantic domain we will explore is one that has proven particularly elusive, namely, modality. While many semantic phenomena can be clearly delineated and, to some degree, explained by the linguistic analyst, modality has been much more problematic; in fact, even the scope of the notion of modality has not really been agreed upon yet. In this chapter, we specifically focus on the semantic domain of POSSIBILITY as reflected in:

- the choices of *can* vs. *may* in essays written by native speakers of English;
- the choices of *can* vs. *may* in essays written by French learners of English;[1]
- the use of *pouvoir* in essays written by native speakers of French.

In Section 2, we discuss in what sense these modals pose a particular challenge to the analyst as well as present previous corpus-based work on *can* and *may* and highlight

1. Following Bartning (2009), the term "advanced learner" is henceforth assumed to refer to "a person whose second language is close to that of a native speaker, but whose non-native usage is perceivable in normal oral or written interaction" (Hyltenstam *et al.* 2005: 7, cited in Bartning 2009: 12).

some of the shortcomings of such work. In Section 3, we discuss the BP approach in general as well as our own data and methods in particular. Section 4 presents the results of our exploration, and Section 5 concludes the chapter.

2. Setting the stage

2.1 What is problematic about the modals?

As near synonyms in the domain of modality, *may* and *can* have fueled much theoretical debate with regard to their semantic relations. As a pair, both forms have overlapping semantics which cover simultaneously the meanings of possibility, permission and ability (cf. Collins 2009). This means that both forms can be used to express epistemic, deontic and dynamic types of possibility. It follows that the semantic investigation of *may* and *can* triggers two problematic questions: first, to what extent the various senses of each form can be distinguished, and second, to what degree both forms are semantically equivalent?

With regard to the first question, studies such as Leech (1969) and Coates (1983) have illustrated the difficulty in distinguishing between the senses of *may* and *can*. Leech (1969: 76), for instance, notes that "[t]he permission and possibility meanings of *may* are close enough for the distinction to be blurred in some cases". Similarly, Coates (1983: 14) identifies a "continuum of meaning" – i.e. gradience – in which possible modal uses shade into each other. In the case of the meanings of *can*, for instance, Coates notes that while permission and ability correspond to the core of two largely intersecting fuzzy semantic sets, possibility, on the other hand, is found "in the overlapping peripheral area" (p. 86).

With regard to the issue of the semantic equivalence of *may* and *can*, the literature reveals similarly debated standpoints. While some studies recognize the similarities of the two forms, others do not. In the former case, for instance, Collins (2009: 91) states that "[t]he two modals of possibility *may* and *can*, share a high level of semantic overlap" (despite their differing frequency of occurrence and different degrees of formality), and Leech (1969: 75) notes that "[i]n asking and giving permission, *can* and *may* are almost interchangeable". Conversely, studies such as Coates (1983) have clearly distinguished the two forms. For instance, while Coates (1983) does recognize that the English modals share certain meanings and can be organized into semantic clusters, she generally denies the synonymy of *may* and *can* by classifying the two forms into two distinct semantic groups. Although she accepts that the two forms may have overlapping meanings in some cases, she claims that even then, the two forms do not occur in free variation.

The occurrence of one form over the other has been shown to be influenced, to some extent, by its linguistic context. It has indeed been illustrated that particular

co-occurring grammatical categories interfere with the interpretation of the modals. Leech (2004: 77), for instance, notes that certain uses of *may* are only to be found in particular grammatical contexts: "only the permission sense, for instance, is found in questions (…) and the negation of the possibility sense is different in kind from the negation of the permission sense". Generally, several grammatical categories have been recognized as interacting with the uses of *may* and *can*. While negation is one category that has commonly been identified (cf. Hermerén 1978; Palmer 1979; Coates 1980, 1983; De Haan 1997; Huddleston 2002; Radden 2007; Byloo 2009), voice and sentence types have also been shown to have similar influences on the forms.

Overall, the above-mentioned studies all provide clear illustrations of the complexity of the semantic relations between *may* and *can* on the basis of empirically gathered evidence. However, they all tend to be based on generalized observations of idiosyncratic behavioral tendencies. In that respect, they all raise the issue of how to provide a more systematic account of the modals' semantic characteristics and how to integrate qualitative findings into a quantitative and empirically-grounded approach.

2.2 Previous corpus-based work on the modals

2.2.1 *Native English*

As already mentioned above, Hermerén (1978) has shown that the semantics of the modals in native English are morphosyntactically motivated to a considerable degree such that linguistic categories such as voice, grammatical person, type of main verb (action, state, etc.), aspect and sentence type influence the interpretation of the modals: "if these categories can be shown to modify the meaning of the modal […] it is important that this should be accounted for in the description of the semantics of the modals" (p. 74). While this claim calls for empirical validation, one implication of Hermerén's (1978) argument is that the quantitative study of modal forms will require a powerful and versatile methodological approach. In a very similar fashion, Klinge and Müller (2005: 1) argue that, to capture the essence of modal meaning, "it seems necessary to cut across the boundaries of morphology, syntax, semantics and pragmatics and all dimensions from cognition to communication are involved".

A second corpus-based study of the modals in native English is Gabrielatos and Sarmento (2006). This study illustrates an attempt to account for syntactic contextual information while using a quantitative corpus-based approach to investigate core English modals (i.e. *can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*). Although their study does not involve the comparison of English varieties, it presents, however, a comparative analysis of the frequencies of uses of the modals in an aviation corpus and a representative corpus of American English. Generally, it raises the following questions:

– To what degree do syntactic structures and modal forms interact contextually?
– To what degree does such interaction affect investigated modal forms semantically?
– How can such interaction be quantitatively investigated in a corpus including cross-linguistic and interlanguage data?

The authors acknowledge that the modals' distribution varies as a function of their syntactic contexts and they show that frequencies of occurrence of core English modals reflect the type of syntactic environment in which they feature: "there is a great deal of variation in the use of modal verbs and the structures they occur in, depending on the context of use" (p. 234). However, their lack of a suitable cognitively-motivated theoretical framework prevents them from providing a meaningful interpretation of the data and to further explore their findings.

To this date, Collins (2009: 1) presents:

> the largest and most comprehensive [study] yet attempted in this area [modality] based on an analysis of every token of the modals and quasi-modals (a total of 46,121) across the spoken and written data.

Collins (2009) investigates the meanings of the modals in three parallel corpora of contemporary British English, American English and Australian English. Despite the author's recognition that a corpus quantitative approach "typically combined with a commitment to the notion of 'total accountability' may influence hypotheses applied to the data, or formulated on the basis of it" (p. 5) and despite the large size of his data set, his analysis is of limited informative value due to:

– a theoretical framework that does not allow for the full exploitation of the linguistic context of the modals, and;
– a statistical approach that inhibits rather than unveils linguistic patterns at play in the data.

With regard to the first point, Collins (2009) restricts his approach to the identification of the forms' lexical meanings. His theoretical framework consists of a traditional tripartite taxonomy including epistemic, deontic and dynamic senses. Regrettably, while he recognizes that some uses of the modals can yield preferences for particular syntactic environments, his analysis does not address that fact in a systematic quantitative fashion. As for the second point, while, statistically, Collins (2009) limits his investigation to providing frequency tables of modal forms, his overall approach is problematic because it is based on the erroneous assumption that the frequent occurrence of a modal form warrants its linguistic relevance. In the case of *may* and *can*, for instance, Collins uses raw frequencies to show that deontic *may* is the "least common" sense of the three as it is chosen 7% of the time over epistemic *may* (79%) and dynamic *may* (8.1%). However, he does not show whether the (low) frequency of deontic *may* is significantly different from the also low frequency of dynamic *may*, and our

analysis of his data shows that, excluding the indeterminate cases, the distribution of *may*'s senses across the American, Australian, and British data is highly significant ($\chi^2 = 42.68$; $df = 4$; $p < 0.001$). This, in turn, raises the questions of:

- To what extent are Collins' (2009) frequencies of the occurrences of modal forms in each corpus comparable?
- Since the observed frequency discrepancies are not a matter of chance, then what motivates, linguistically, the different uses of each form in each independent corpus?

So in sum, while studies such as Gabrielatos and Sarmento (2006) and Collins (2009) provide many descriptive results, they are often merely or largely form-based alone and are lacking in terms of determining which of the many frequencies are statistically and/or linguistically relevant. As a result, such studies do not come close to allow us to develop a characterization of modals that essentially allows us to classify/predict modal use.

2.2.2 *Learner English and contrastive approaches*

From a cross-linguistic and an interlanguage perspective, investigating the modals raises two related issues, namely (i) the possibility of a lack of (direct) semantic equivalence between the modal forms in the learner's native language (L1) and his/her target language (L2), and (ii), the fact that such cross-linguistic semantic dissimilarity will affect the uses of the forms in L2. The modals *may* and *can* and native French *pouvoir* illustrate the case in point. Despite the fact that all three forms contribute to the expression of the semantic notion of POSSIBILITY, *pouvoir* synchronically covers the whole range of the modal uses of *may* and *can*.

One corpus-based study of learners' use of modals is Aijmer (2002), which is based on a corpus of Swedish L2 English writers. She compares (i) the frequencies of key modal words in native English and advanced Swedish-English interlanguage, as well as (ii) frequencies encountered in Swedish learner English with those from comparable French and German L2 English. Aijmer's study indicates "a generalized overuse of all the formal categories of modality" and she further points out that "it is only at a functional level that any underuse was detected, with the learner writers failing to use *may* at all in its root meaning" (p. 72).

Similarly, Neff *et al.* (2003) investigate the uses of modal verbs (*can*, *could*, *may, might* and *could*) by writers from several L1 backgrounds. Neff *et al.* (2003) use a learner corpus including Dutch-, French-, German-, Italian-, and Spanish-English interlanguage, which they contrast with a reference corpus of American university English. Neff *et al.* (2003: 215) identify the case of *can* as potentially interesting "since it is overused by all non-native writers". They further report that the frequency of *may* by French native speakers stands out in comparison to the frequencies by all other non-native speakers included in the study, but since their study does basically nothing

but compare raw frequencies of occurrence regardless of any contextual features , it is not particularly illuminating.

Generally, and similar to Gabrielatos and Sarmento (2006) and Collins (2009), both Aijmer (2002) and Neff *et al.* (2003) made the disadvantageous methodological decision to conveniently, but ultimately problematically, rely on information that is retrievable without human effort. In addition, even the studies that address learner use do not relate their findings to the wider context of (second) language acquisition.

In a corpus-based contrastive study, Salkie (2004) investigates the nature of the semantic relations between the three forms in native English and native French. He uses a subpart of the parallel corpus INTERSECT (cf. Salkie 2000), and focuses on three working hypotheses, namely that:

- "*pouvoir* corresponds more closely to one of the English modals rather than the other" (p. 169);
- "*pouvoir* is less specific than the English modals" (p. 170);
- "*pouvoir* has a sense which is different from both the English modals but is not just a general sense of possibility" (p. 170).

While Salkie (2004) concludes in favour of the third hypothesis, it is worth pointing out, however, that his results were based on only 100 randomly extracted occurrences of each English modal form (i.e. *may* and *can*) and their respective French translations.

By way of a more general summary, it is probably fair to say that corpus-based approaches to modality in L1 and L2s leave things to be desired. Some studies point to the immense complexity of the subject but do not choose multifactorial or multivariate methods that are capable of addressing this degree of complexity. In addition, some studies are based on large numbers of modals but, frankly, do not do very much with the vast amount of data other than present arrays of statistically under-analyzed frequency tables. On the other hand, the analytically much more interesting studies of the kind of Salkie (2004) are based on very small samples. Finally, many studies are largely if not exclusively form-based and focus only on learners' over-/underuse of modals in particular examples or kinds of contexts.

2.3 Characteristics of the present study

2.3.1 *Methodological considerations*

The above discussion fairly clearly indicates what kinds of steps would be desirable, an approach that:

- can integrate linguistic information and patterning from many different levels of linguistic analysis in a way alluded to by Hermerén (1978), as well as Klinge and Müller (2005);

- involves not only a sample that is studied with regard to more linguistic parameters, but at the same time also larger than the previous studies that aimed at more than description;
- explores similarities and differences of L1 uses of *can* and *may*, but also explores the way these English modals are used in L2 language (here from French learners) as well as how the same concept is used by the learners in their L1 (here *pouvoir*).

Given these demands, we decided to use the so-called Behavioral Profile approach, which fits the above wish list very well. It combines the statistical methods of contemporary quantitative corpus linguistics with a cognitive-linguistic and psycholinguistic perspective or orientation (cf. Divjak and Gries 2006, 2008, 2009; Gries 2006, 2010b; Gries and Divjak 2009, 2010; and others). As such, it diverges radically from the above-mentioned more traditional corpus-based approaches to modality in both L1 and L2. Methodologically, it involves four steps:

- the retrieval of all instances of a word's lemma from a corpus in their context;
- a manual annotation of a number of features characteristic of the use of the word forms in the data; these features are referred to as ID tags and typically involve morphosyntactic and semantic features in particular. Each ID tag contributes to the profiling of the investigated lexical item(s);
- the generation of a table of co-occurrence percentages, which specify, for example, which words (from a set of near-synonymous words) or senses (of a polysemous word) co-occur with which morphosyntactic and/or semantic ID tags; it is these vectors of percentages that are called *profiles*;
- the evaluation of that table by means of statistical techniques.

Given how this approach is completely based on various kinds of co-occurrence information, it comes as no surprise that, just like much other work in corpus linguistics, the BP approach assumes that "the distributional characteristics of the use of an item reveals many of its semantic and functional properties and purposes" (Gries and Otani 2010: 3). While these previous studies have investigated a variety of different lexical relations (near synonymy, polysemy, antonymy) both within languages (English, Finnish, Russian) and across languages (English and Russian), the present study will add to the domains in which Behavioral Profiles have been used in two ways: (i) so far, no non-native language data have been studied, and (ii) we will add French to the list of languages studied.

As the first BP study focusing on learner data, and only the second BP study that compares data from different languages, this paper is still largely exploratory. We will mainly be concerned with the following two issues:

- To what degree can the Behavioral Profiling handle the kind of learner data that are inherently more messy and volatile than native data and provide a quantitatively adequate and fine-grained characterization of the use of *can* and *may* by

native speakers and learners, and how does that use compare to the use of French speakers' use of *pouvoir*?
– As a follow-up, and if meaningful groups of uses emerge, to what degree do the distributional characteristics that BP studies typically include allow us to predict native speakers' and learners' choices of modal verbs, and how do these speaker groups differ?

The former question will be explored with the kind of cluster-analytic approach usually employed in BP studies; for the latter question, we will turn to a logistic regression (cf. Arppe 2008 for another BP approach using (multinomial) regression).

2.3.2 *Theoretical orientation*

In previous studies, the BP approach was used for more than just the quantitative description of the data. Rather, it is firmly grounded in, and attempts to relate the results of the statistical exploration of the data to usage-based/exemplar-based approaches within Cognitive Linguistics and psycholinguistics. While this orientation is also compatible with our current goals, there is one particular earlier model in L2/FLA research that is especially well-suited to, or compatible with, our current objectives, namely the Competition Model (CM) by Bates and MacWhinney (cf. Bates and MacWhinney 1982, 1989). This model is "a probabilistic theory of grammatical processing which developed out of a large body of crosslinguistic work in adult and child language, as well as in aphasia" (Kilborn and Ito 1989: 261). MacWhinney (2004: 3) himself characterized it as a "unified model [of language acquisition] in which the mechanisms of L1 learning are seen as a subset of the mechanisms of L2 learning".

The CM is characterized by the two following assumptions:

– Linguistic signs map forms and functions onto each other (probabilistically) such that forms and functions are cues to functions and forms respectively.
– In language production, forms compete to express underlying intentions or functions, and in language comprehension, the input contains many different cues of different strengths, validities, and reliabilities, which must be integrated: native speakers "depend on a particular set of probabilistic cues to assign formal surface devices in their language to a specific set of underlying functions" (Bates and MacWhinney 1989: 257).

As a usage-based and probabilistic model, the CM assumes that both frequency and function determine the choice of grammatical forms in language production; as with most usage-based and/or corpus-linguistic approaches, we too consider frequency in a corpus as a proxy for frequency of exposure (in both comprehension and production). Cross-linguistically, this is an important assumption because across languages cues are instantiated in different ways and speakers assign them varying degrees of strength. It is therefore important to describe and explain L1 statistical regularities as

"[t]hey are part of the native speaker's knowledge of his/her language, and they are an important source of information for the language learner" (Bates and MacWhinney 1989: 15).

Overall, Kilborn and Ito (1989: 289) conclude that existing psycholinguistic studies have successfully demonstrated that the CM is appropriate for the characterization of learner language through cue distributions and they report "extensive evidence for the invasion of L1 strategies into L2 processing". In addition, it is also obvious how much the CM is compatible with a BP approach. The main notions that drive the Competition Model are cue strengths, validities, and reliabilities, and all of these are essentially conditional probabilities, i.e. percentages. While the BP approach as such does not cover the full complexity of how conditional cue strengths, validities, and reliabilities can interact, it is a useful and experimentally validated (cf. Divjak and Gries 2008) approach employing a similar logic.

> A theory of language transfer requires that we have some ability to predict where the phenomena in question will and will not occur. In this regard contrastive analysis alone falls short; it is simply not predictive. (Gass 1996: 324)

3. Data and methods

3.1 Retrieval and annotation

The data are from three untagged corpora: the French subsection of the *International Corpus of Learner English* (henceforth ICLE-FR), the *Louvain Corpus of Native English Essays* (LOCNESS), and the *Corpus de Dissertations Françaises* (CODIF). All corpora included in the present work were collected by the Centre for English Corpus Linguistics (CECL) at the Université Catholique de Louvain (UCL) and made available to us by the Director of the Centre, Professor Sylviane Granger. ICLE-FR has a total of 228,081 words, including 177,963 words of argumentative texts and 50,118 words of literary texts. LOCNESS is a 324,304-word corpus that includes three sub-data sets: a 60,209-word-sub-corpus of British A-Level essays, a 95,695-word sub-corpus of British university essays and a sub-corpus of American university essays that has 168,400 words. The CODIF is a corpus of essays written by French-speaking undergraduate students in Romance languages at the Université Catholique de Louvain (UCL). CODIF also includes argumentative and literary texts and has a total of 100,000 words.[2]

2. Information on the total number of words featuring in each individual text type (i.e. argumentative, literary) is not available.

Table 1. Excerpt of an annotation table including selected variables

Case	Match	Corpus	ClType	Use	VerbSemantics	Neg	RefAnim
5	*may*	native	coordinate	process	ment/cog/emotional	affirmative	animate
133	*may*	native	main	state	copula	affirmative	inanimate
1760	*may*	native	main	process	ment/cog/emotional	negative	animate
1886	*can*	il	coordinate	process	ment/cog/emotional	affirmative	animate
2876	*cannot*	il	subordinate	state	abstract	negative	inanimate
3540	*peut*	fr	main	process	ment/cog/emotional	negative	animate
3645	*peuvent*	fr	subordinate	process	abstract	negative	inanimate

Given the corpora's compositions, the three corpora included in our study are highly comparable. They all consist of written data produced by university students (ICLE, CODIF, the LOCNESS British and American university sections) or by students approaching university entrance (i.e. the LOCNESS British A-Level section).[3] All participants' contributions are in the form of an essay of approximately 500 words long. In terms of content, all essays deal with similar topics such as: crime, education, the Gulf War, Europe, or university degrees.

The data we subjected to the BP approach consist of instances of *may* and *can* in native English and French-English interlanguage as well as *pouvoir* in native French from the above corpora. Using scripts written in R (cf. R Development Core Team 2010), we retrieved 3,710 occurrences of the investigated modal forms from all sub-corpora, which were imported into a spreadsheet software and annotated for 22 morphosyntactic and semantic variables.[4] Table 1 exemplifies this database with a very small excerpt of these data, and Table 2 presents the total range of variables included in the study and their respective levels.

For each variable, an encoding taxonomy was designed prior to annotation. Due to the large number of variables included in this study and the absence of a number of them from previous studies on the English modals, not all encoding taxonomies were theoretically motivated. In cases where the annotation is not based on accounts from the existing literature, a bottom-up approach was adopted for the identification of recurrent features in the data. This procedure, for instance, was carried out in the case of the variable VerbSemantics where, prior to annotation, recurrent semantic features were identified as characteristic of the lexical verbs used alongside the modals.

3. The inclusion of the LOCNESS British A-Level section alongside sub-corpora solely including university participants is not judged problematic as LOCNESS only involves English native speakers whose level of English is not expected to develop any further.

4. Although the annotation process included a variable encoding the semantic role of the subject referent of the modals, this study does not account for that variable due to its high correlation with VOICE.

Table 2. Overview of the variables used in the study and their respective levels

Type	Variable	Levels
data	Corpus	native, interlanguage, French
	GramAcc (acceptability)	yes, no
syntactic	Neg (negation)	affirmative, negated
	SentType (sentence type)	declarative, interrogative
	ClType (clause type)	main, coordinate, subordinate
morphological	Form	*can*, *may*, *pouvoir* (and negated forms)
	SubjMorph: subject morphology	adj., adv., common noun, proper noun, relative pronoun, date, noun phrase, etc.
	SubjPerson: subject person	1, 2, 3
	SubjNumber: subject number	singular, plural
	Voice	active, passive
	Aspect	perfect, perfective, progressive
	Mood	indicative, subjunctive
	SubjRefNumber: subject referent number	singular, plural
semantic	Senses	epistemic, deontic, dynamic
	SpeakPresence	weak, medium, strong
	Use	accomplishment, achievement, process, state
	VerbSemantics	abstract, general action, action incurring transformation, action incurring movement, perception, etc.
	RefAnim: subject referent animacy	animate, inanimate
	AnimType: subject referent animacy type	animate, floral, object, place/time, mental/emotional, etc.

Because of space restrictions, we are not able to provide a more comprehensive account of the annotation process (but cf. Deshors 2010 for details). However, three variables – Senses, VerbType, and VerbSemantics – require some brief explanatory comments.

3.1.1 *The variable Senses*

As for Senses, the semantic category of modality includes a wide range of heterogeneous meanings that many scholars have attempted to unite under a variety of categorization systems (cf. Palmer 1979; Coates 1983; Bybee and Fleischman 1995; Huddleston 2002; Nuyts 2006; Byloo 2009). While Depraetere and Reed (2006: 277) note that "in classifying modal meanings, it is possible to use various parameters as criterial to their classification", this study assumes a coding taxonomy based on a traditional tripartite distinction between epistemic, deontic and dynamic meanings.

Following Nuyts (2006: 6), epistemic senses concern "an indication of the epistemic estimation, typically, but not necessarily, by the speaker, of the chances that the state of affairs expressed in the clause applies in the world". Consider (1) as an illustration of epistemic *may*:

(1) indeed, Europe 92 *may* lead to the disappearance of cultural differences

Following Palmer (1979: 58), deontic modality refers to cases where "[b]y uttering a modal, a speaker may actually give permission (*may*, *can*)". (2) illustrates deontic *can*:

(2) if all public schools started to say you *can* only come here if you are Hispanic or if you are Polish, our schooling system would be in great chaos

Finally, dynamic meanings denote "an ascription of a capacity to the subject-participant of the clause (the subject is able to perform the action expressed by the main verb in the clause)" (Nuyts 2006: 3). Generally, dynamic modality expresses the potentiality of an event occurring. Nuyt's type of dynamic modality includes *ability/capability* cases where the possibility of event occurrence stems from the ability of the (grammatical) subject to carry out the event. In that regard, the term *ability* is not restricted to a 'physical' interpretation and equally applies to mental and technical types of ability. Example (3) illustrates dynamic *can*:

(3) Mrs Ramsay is the central character because she *can* see the whole personality of the other ones

Generally, our frequencies of use of *may* and *can* in their different senses match those previously encountered in existing studies solely concerned with the native use of the modals, such as Coates (1980) and Collins (2009). While Coates (1980: 218), for instance, reports that "by far the most common usage of *may* is to express epistemic possibility", she stresses the distinctive nature of the uses of *may* and *can*:

> The patterns resulting from my analysis of the data (…) leads me to conclude that in normal everyday usage *may* and *can* express distinct meanings: *may* is primarily used to express epistemic possibility, while *can* primarily expresses root possibility.[5]

3.1.2 *The variable VerbType*

The variable VerbType targets the lexical verbs with which the forms are used and characterizes their telicity. Conceptually, the variable VerbType follows Vendler (1967) in its recognition that the notion of time is crucially related to the use of a

5. Coates (1980, 1983) categorizes modal meaning according to a two-way distinction that includes epistemic and non-epistemic modality. She refers to the latter type as "root" modality.

verb and is "at least important enough to warrant separate treatment" (p. 143). This variable assesses:

– whether *may* and *can* have preferences for lexical verbs denoting a *state*, a *process*, an *accomplishment* or an *achievement*,[6] and if so,
– it identifies in which type of corpus preferential patterns occur.

3.1.3 *The variable* VerbSemantics

Similarly to the variable VerbType, VerbSemantics identifies the type of semantic information conveyed by the lexical verbs used with the modals. The internal organization of this variable results from a bottom-up approach and does not follow any particular theoretical framework. This variable consists of the levels denoting abstract process, physical actions, actions incurring movement, actions incurring some physical transformation, communicative processes, mental/cognitive/emotional processes, perception processes and verbal statement involving a copula verb. Example (4) illustrates a case where the lexical verb expresses a mental/cognitive/emotional process:

(4) Her search for the final touch can be *seen* as a search for harmony

Once all matches were annotated, the resulting data table was evaluated statistically.

3.2 The BP approach in this study: Statistical analysis

As mentioned above, the data were evaluated in two different ways.[7] The first of these involved the type of cluster analysis that is characteristic of much work using the BP methodology. In this first part, we used Gries 's (2010a) R script Behavioral Profiles 1.01 and computed five behavioral profiles, one for each modal form as occurring in each language variety, i.e. native *can*, native *may*, interlanguage (IL) *can*, IL *may*, and native *pouvoir* (FR). Such profiles consist of vectors of co-occurrence percentages of a single modal form with each level of all independent variables and provide form-specific summaries of their semantic and morphosyntactic behavior in each sub-corpus. In a second step, the profiles were assessed statistically with a hierarchical cluster analysis to explore the similarity and differences between the modal forms, and in keeping with previous studies (cf. Divjak and Gries 2006), we chose the Canberra metric as a measure of (dis)similarity and Ward's rule as an amalgamation strategy.

---

6. Accomplishment verbs encode verbal statements that imply a unique and definite time period; achievement verbs encode verbal statements that imply a unique and definite time instant; process verbs identify statements that reflect non-unique and indefinite time periods; state verbs identify statements that reflect non-unique and indefinite time instants.

7. All statistical computations and plots were performed with R (for Linux), version 2.11.0 (see R Development Core Team 2010).

Following Gries and Otani (2010), we computed different cluster analyses, one involving all variables that the uses of the modals were annotated for, one for only the syntactic variables, and one for only the semantic variables.

The second analytical step involved a binary logistic regression including the following variables and predictors:

- Form as the dependent variable with only two levels here: *can* vs. *may*;
- GramAcc, Neg, SentType, ClType, SubjMorph, SubjPerson, SubjNumber, Voice, Aspect, Mood, SubjRefNumber, Senses, SpeakPresence, Use, VerbSemantics, RefAnim, AnimType as independent variables in the form of main effects;
- all these variables' interactions with Corpus as additional predictors (to see which variables' influence on modal use differs the most between L1 English and L2 English).

The logistic regression was then performed with the model selection process during which insignificant predictors were discarded from the model: first insignificant interactions, then individual variables that were not significant and did not participate in a significant interaction.

4. Results and discussion

4.1 Cluster analysis

Our first cluster analysis yielded the results shown in Figure 1. The left plot is a dendrogram of the five modal forms that were clustered; the right plot represents average silhouette widths for assuming two, three, and four clusters. The average silhouette widths point to a two-cluster solution, maybe a three-cluster solution, but the difference is minor since the former would result in a French-vs.-English clustering, and the latter in a French-vs.-*can*-vs.-*may* clustering. This is compatible with Salkie's analysis, who argued that *pouvoir* is very different from both *can* and *may*, and intuitively, both these solutions "make sense", which provides first evidence in favor of the approach. To anticipate the potential objection that this may seem trivial, let us mention that it is in fact not. The data in Figure 1 show that the BP vectors are good and robust descriptors of how the modals behave because many other theoretically possible cluster solutions, such as the ones listed in (5), would not have made linguistic sense at all.

(5) a. {{{can_{il} may_{native} *pouvoir*} can_{native}} may_{il}}
b. {{{can_{native} may_{il} *pouvoir*} can_{il}} may_{native}}
c. {{can_{il} may_{native}} {*pouvoir* may_{il}} can_{native}}

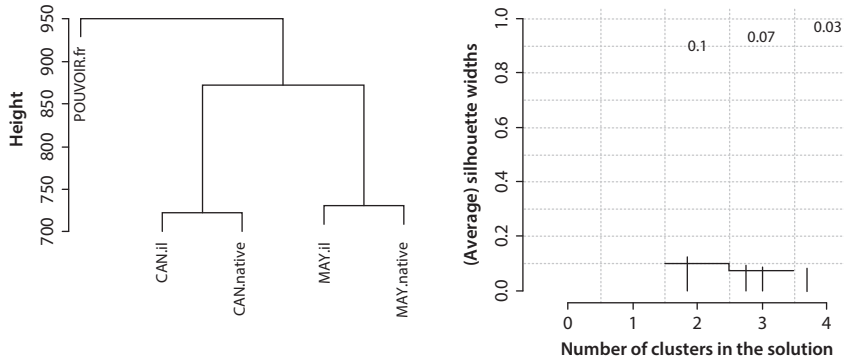

Figure 1. Dendrogram for all independent variables (il = interlanguage)

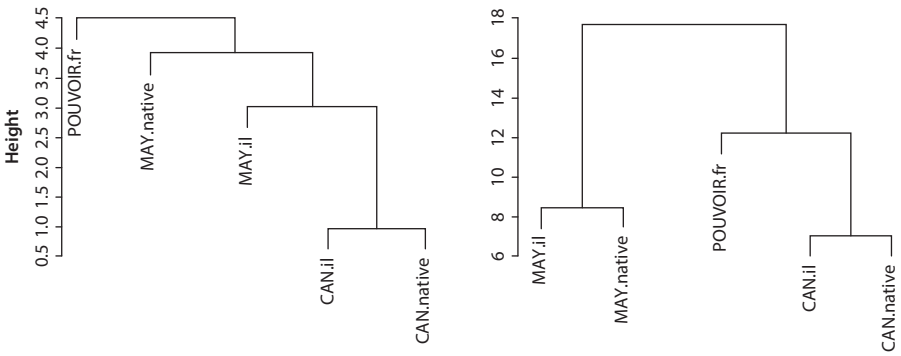

Figure 2. Dendrograms for all morphosyntactic variables (left panel) and all semantic variables (right)

However, in what follows we show that a fine-grained comparative description of cross-linguistic language varieties can be obtained by focusing on differences between the independent variables used for clustering. Consider Figure 2, which shows the dendrograms for all morphosyntactic variables and all the semantic variables in the left and right panel, respectively.

Interestingly, the results show that the intuitively very reasonable dendrogram in Figure 1 is not replicated by looking at morphosyntax or semantics alone, which to some extent at least contrasts with Gries and Otani's results, where the results did not differ very much between the three clusterings. The reasonable similarities of Figure 1 emerge only when all variables are combined. In particular, in both panels of Figure 2 *can*$_{\text{il}}$ and *can*$_{\text{native}}$ are grouped together, but then the remaining forms are grouped differently. In the morphosyntactic dendrogram, the two kinds of *may* are successively amalgamated and the French *pouvoir* is only added after all English forms have been

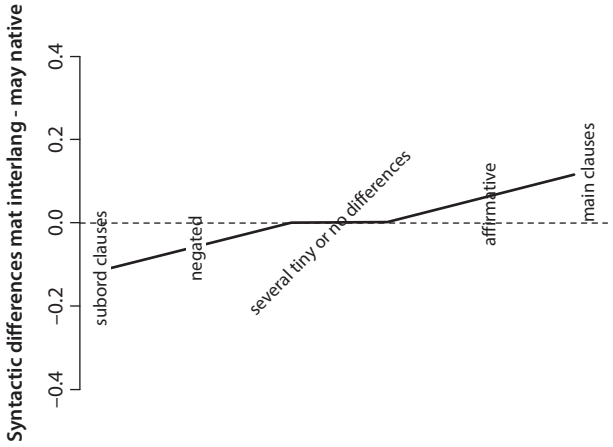

Figure 3. Snakeplot for most extreme differences between syntactic ID tags of *may*

clustered. In other words, morphosyntactically, we find a clear English-French divide, but interlanguage *may* is too different from native *may* to be grouped together. To identify the source of this difference, we used what in BP approaches has been called a snakeplot, namely a plot of the pairwise differences between the percentages for, in this case, may_{il} and may_{native} (cf. Divjak and Gries 2009 or Gries and Otani 2010 for more examples).

As indicated in Figure 3, the main morphosyntactic ways in which learners deviate from native speakers are that learners underuse *may* in subordinate clauses and in negated clauses. This is in fact an interesting finding because it means that learners disprefer the rarer of the two modals – *may* – in those contexts which are already morphosyntactically more challenging, as if using *can* is the default they resort to when they are already under a higher processing load (cf. the so-called complexity principle).

In the semantic dendrogram, by contrast, we find a different patterning. Semantically, can_{il} and can_{native} are again very similar and grouped together early, but then the next clustering step groups the two forms of *may* together. However, interestingly, it is not the English forms that are then all grouped together – rather, contrary to Salkie's earlier analysis, *pouvoir* is semantically more similar to *can* than *may* is.

4.2 Logistic regression

The model selection process involved thirteen steps during which insignificant predictors were discarded. The final and minimally adequate model includes 16 significant variables and 6 significant interactions and returned a highly significant correlation: loglikelihood chi-square = 3296.47; $df = 60$; $p < 0.001$; the correlation between the

Table 3. Overview of the results of the final GLM model

Predictor	Chi-square (*df*)	Predictor	Chi-square (*df*)
Corpus	24.9 (1) ***	AnimType	98.2 (11) ***
GramAcc	13.8 (1) ***	Voice	55.0 (1) ***
Use	67.9 (1) ***	SentType	47.2 (1) ***
Elliptic	100.0 (2) ***	Negation	87.2 (1) ***
ClType	10.9 (1) ***	SpeakPresence	29905.9 (2) ***
VerbType	97.4 (2) ***	Corpus:ClType	60.0 (2) ***
VerbSemantics	384.9 (6) ***	Corpus:VerbSemantics	32.2 (6) ***
SubjPerson	26.6 (2) ***	Corpus:SubjNumber	37.4 (1) ***
SubjNumber	1.3 (1) ns	Corpus:RefAnim	122.2 (1) ***
SubjMorph	49.1 (4) ***	Corpus:AnimType	118.2 (11) ***
RefAnim	59.2 (1) ***	Corpus:Negation	12.0 (1) ***

observed forms – *may* vs. *can* – and predicted probabilities is very high: $R^2 = 0.955$. Correspondingly, the model's classificatory power was found to be very powerful with a classification accuracy of 99%. Table 3 summarizes all the significant variables and interactions yielded in the final model.

Overall, the final model includes one significant interaction involving a morphological variable (out of seven morphological variables), two significant interactions involving syntactic variables (out of three syntactic variables) and three significant interactions involving semantic variables (out of eight semantic variables). But what do the interactions reflect? Let us begin with Corpus:ClType, as represented in Figure 4.

The frequencies of *may* and *can* differ with regard to the type of clauses in which they occur in native and learner English. The (weak!) effect is that, in interlanguage

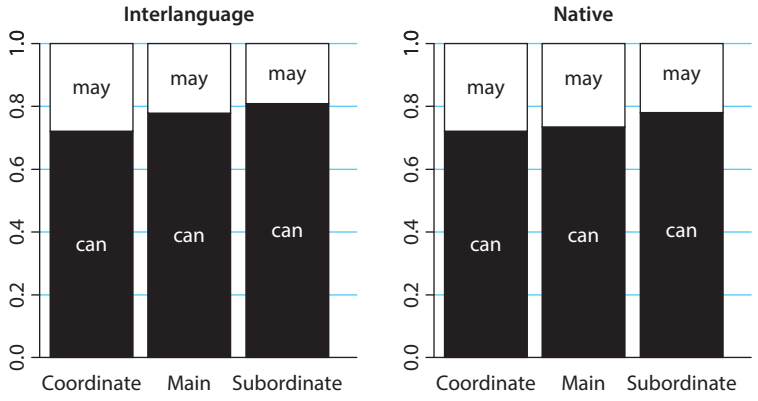

Figure 4. Bar plots of relative frequencies of Corpus:ClType

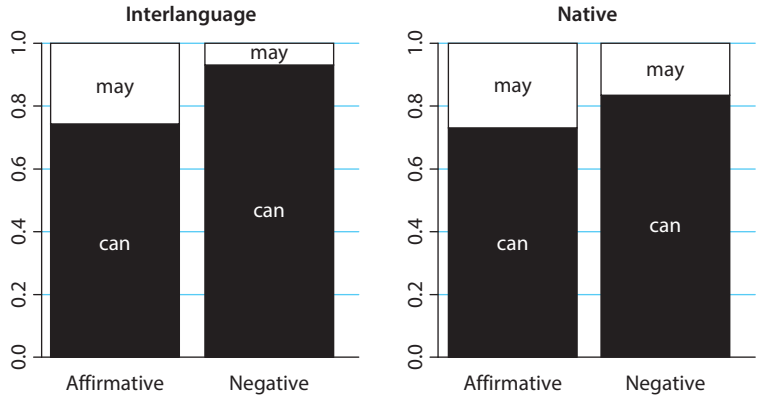

Figure 5. Bar plots of relative frequencies of Corpus:Neg

English, *can* is more strongly preferred over *may* in main clauses than it is in native English.

While, as previously noted, existing literature concerned with the native use of the modals commonly recognizes negation as "an important aspect of modal meaning" (Hermerén 1978), our study not only confirms the need to include negation in an investigation of the uses of the modals but further recognizes its significance as a morphological criteria to assess interlanguage (dis)similarity. Consider Figure 5 for the interaction Corpus:Neg.

Figure 5 shows that, while all speakers prefer to use *can* in negated clauses, the interlanguage speakers do so more strongly. This result does not come as a surprise: On the one hand, this is also compatible with the complexity principle – negated clauses are more complex and preferred with the more frequent modal. On the other hand, where epistemic *may not* would be used in English, French speakers would tend to use a lexical verb along with the adverb *peut-être* to indicate the speaker's uncertainty, as illustrated in (6):

(6) a. This may not be the case
b. Ce n'est peut-être pas le cas

Consider Figure 6 for the interaction Corpus:SubjNumber.

While native speakers use *can* more often with singular subjects than with plural subjects, it is the other way round with the learners, again a result compatible with the complexity principle.

While the native speakers' choices of *may* and *can* do not vary much between animate and inanimate subjects, the learners' choices do: with animate subjects, they prefer *can* much more strongly. Figure 7 represents the interaction Corpus:RefAnim.

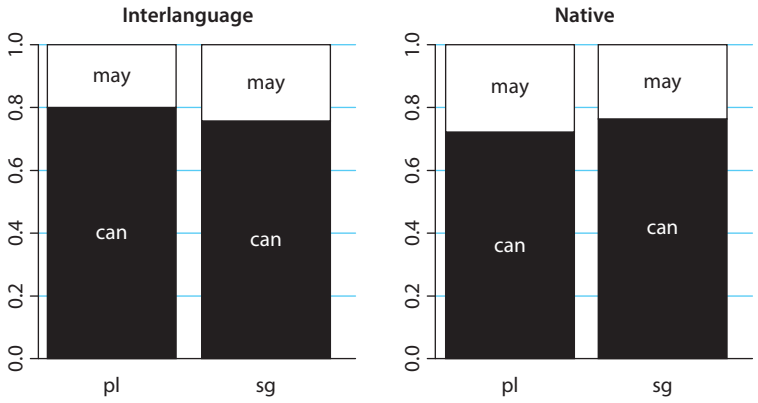

Figure 6. Bar plots of relative frequencies of Corpus:SubjNumber

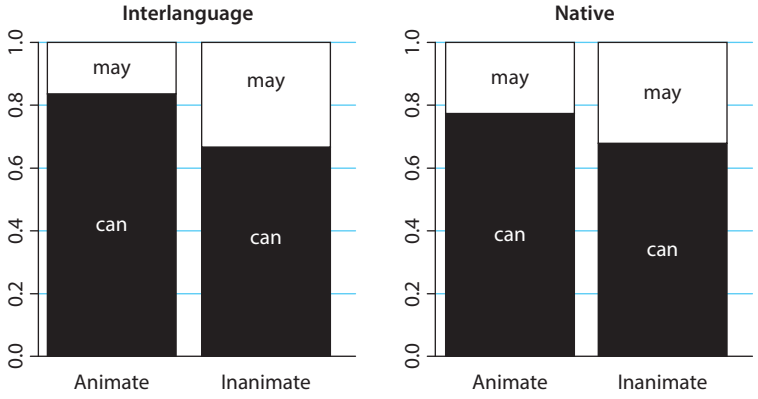

Figure 7. Bar plots of relative frequencies of Corpus:RefAnim

Consider Figure 8 for the interaction Corpus:VerbSemantics; the upper panel represents the interlanguage data, the lower panel represents the native speaker data, and the bars are sorted from large absolute pairwise differences (left) to small absolute pairwise differences (right).

The learners and the native speakers differ most strongly with semantically more abstract verbs and time/place verbs, as in *He thinks that if he can achieve one impossible act, then this will change everything.*

The learners prefer *can* with abstract verbs more strongly than the native speakers, but they prefer *may* more strongly with time/place verbs. However, there are also (less pronounced) differences for verbs that would typically have a human agent.

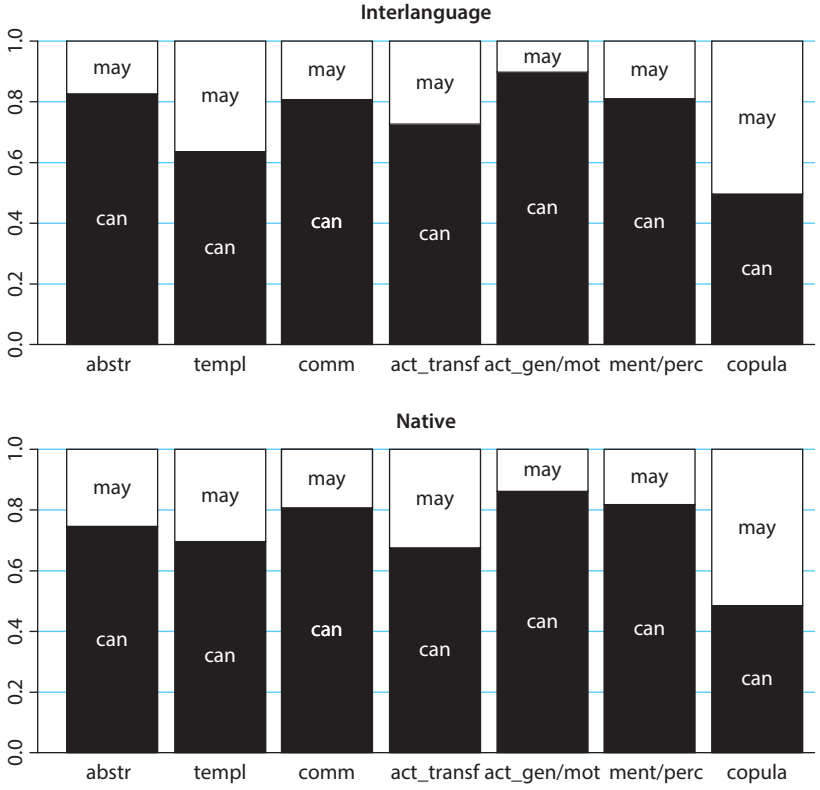

Figure 8. Bar plots of relative frequencies of Corpus:VerbSemantics

For instance, the learners prefer *may* with communication verbs and *can* with action-transformation verbs. Virtually no difference at all is found with copulas.

As for the final interaction, Corpus:AnimType, we do not represent it here graphically. While it is significant, the large number of categories plus the fact that the most pronounced differences occur with a small number of very infrequent categories does not yield much in terms of interesting findings.

As for the main effects, we will not discuss them here in detail. This is because these main effects by definition do not tell us anything about the *can* and *may* variables across languages (since these variables do not interact with Corpus). However, since they *do* tell us something about which modal verb is preferred by both native speakers and learners, we summarize them here visually in Figure 9. The *x*-axis lists the main effects, on the *y*-axis we show the percentage of *can* obtained for levels of these main effects, and then the levels are plotted at their observed percentage of *can*; the dashed line represents the overall percentage of *can* in the data.

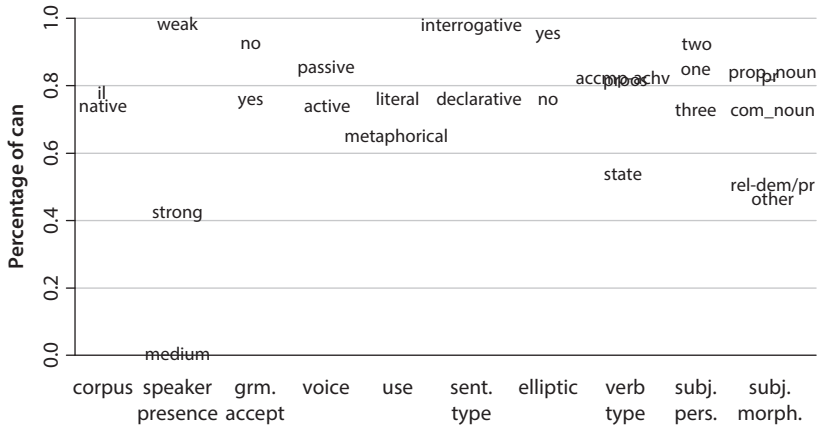

Figure 9. Main effects of the logistic regression

Finally, a brief look at the regression's misclassifications seems to indicate that they did not occur randomly. While all 34 misclassifications occurred in the interlanguage data, 29 of them occurred with *may* in a form characteristic only of the French-English learner language. In the large majority of those misclassifications, *may* is found to express a possibility that results from some sort of theoretical demonstration. Consider the examples in (7) and (8). While the ones in (7) illustrate our current point, (8) provides an additional example of an atypical occurrence of learner *may*, which clearly denotes a strong sense of possibility and whose interpretation is heavily reminiscent of that of *can*.

(7) a. **So** we *may* say that …
b. **To conclude**, we *may* say that …
c. **As a conclusion**, we may say that …
d. **This is why** we may now speak of the stupefying effect
e. **This is the reason why** we may say that …

(8) "Dresden is an old town", we *may* read of its history

5. Concluding remarks

By way of a summary, the BP approach and the subsequent logistic regression allows us to recognize how *can* and *may* (in native and learner English), as well as *pouvoir*, relate to each other as well as what helps determine native speakers' and learners' choices. On the whole, distributionally we do find the expected groupings: the *can*s, then the *may*s, and only then *pouvoir*. However, it is interesting that, semantically,

English *can* is more similar to French *pouvoir* than to English *may*, and the subsequent regression results provided some initial information on why that is so. More specifically, the way learners choose one of the two verbs is often compatible with a processing-based account in terms of the complexity principle – they choose the more basic and frequent *can* over *may* when the environment is complex – but is also strongly influenced by the animacy of the subject and the semantics of the verb: *can* is overpreferred by learners with animate subjects and with abstract verbs, and underpreferred with time/place verb semantics.

With regard to the modals *per se*, our results confirm previous studies' recognition of the influential role of the linguistic context in the uses of *may* and *can*. Indeed, while the main effects included in our final logistic regression model support studies that have identified morphosyntactic components such as Voice and SentType as particularly influential categories (Leech 1969, 2004; Huddleston 2002; Collins 2009), our results reveal the necessity to also take the semantic context of modals more seriously, as reflected by the strong effects of VerbType and VerbSemantics.

More generally speaking and in the parlance of the Competition Model, the cluster analysis and the high classification accuracy of the regression suggest that, on the whole, the learners have built up mental categories for *can* and *may* that are internally rather coherent. However, the interactions in the regression show that these cues are weighted incorrectly and sometimes trigger a verb choice that is not in line with native speaker choices, but that even this kind of incorrect choice is largely predictable (because the regression can still make the correct classifications (cf. Deshors 2010 for more detailed discussion as well as a distinctive collexeme analysis revealing additional verb-specific preferences). In other words, even though this is the first study involving learner data (and only the second involving different languages), the BP approach and especially the follow-up in terms of the logistic regression are therefore an interesting diagnostic: (i) the overall results can testify to the strength of the categories that are being studied, and (ii) the regression with its inclusion of the interactions of all variables with "native speaker vs. learner" exactly pinpoints where interactions become significant, i.e. where the categories of the learner are still substantially different from the native speaker. For further applications and extensions, see Gries and Wulff (2013) for a similar application to the choice of (*of*- and *s*-) genitives by native speakers and learners, and Gries and Deshors (to appear) for an even more advanced approach to precisely pinpoint where non-native speakers' choices deviate from those of native speakers and how much so. Needless to say, more and more rigorous testing is necessary, but to our knowledge this is the first study proposing this kind of approach more generally and the use of a regression with a native-learner variable as a measure of L2 "proficiency"; the results illustrate that learners' "non-nativeness" manifests itself at all linguistic levels simultaneously.

References

Aijmer, K. (2002). 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: John Benjamins.

Arppe, A. (2008). Univariate, bivariate and multivariate methods in corpus-based lexicography: A study of synonymy. Unpublished PhD dissertation, University of Helsinki. Available at: <http://urn.fi/URN:ISBN:978-952-10-5175-3>.

Bartning, I. (2009). The advanced learner variety: 10 years later. In E. Labeau, & F. Myles (Eds.), *The advanced learner variety: The case of French* (pp. 11–40). Frankfurt/Main: Peter Lang.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner, & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 173–218). Cambridge: Cambridge University Press.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney, & E. Bates (Eds.), *The cross-linguistic study of sentence processing* (pp. 3–73). Cambridge: Cambridge University Press.

Bybee, J., & Fleischman, S. (1995). *Modality in language and discourse*. Amsterdam: John Benjamins. DOI: 10.1075/tsl.32

Byloo, P. (2009). Modality and negation: A corpus-based study. Unpublished PhD dissertation, University of Antwerp.

Coates, J. (1980). On the non-equivalence of *may* and *can*. *Lingua*, 50(3), 209–220. DOI: 10.1016/0024-3841(80)90026-1

Coates, J. (1983). *The semantics of the modal auxiliaries*. London: Croom Helm.

Collins, P. (2009). *Modals and quasi modals in English*. Amsterdam: Rodopi.

De Haan, F. (1997). *The interaction of modality and negation: A typological study*. New York: Garland.

Depraetere, I., & Reed, S. (2006). Mood and modality in English. In B. Aarts, & A. MacMahon (Eds.), *The handbook of English linguistics* (pp. 268–287). London: Blackwell.

Deshors, S. C. (2010). A multifactorial study of the uses of *may* and *can* in French-English interlanguage. Unpublished PhD dissertation, University of Sussex.

Divjak, D. S., & Gries, St. Th. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2(1), 23–60. DOI: 10.1515/CLLT.2006.002

Divjak, D. S., & Gries, St. Th. (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, 3(2), 188–213. DOI: 10.1075/ml.3.2.03div

Divjak, D. S., & Gries, St. Th. (2009). Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In K. Dziwirek, & B. Lewandowska-Tomaszczyk (Eds.), *Studies in cognitive corpus linguistics* (pp. 273–296). Frankfurt/Main: Peter Lang.

Gabrielatos, C., & Sarmento, S. (2006). Central modals in an aviation corpus: Frequency and distribution. *Letras de Hoje*, 41(2), 215–240.

Gass, S. (1996). Second language acquisition and linguistic theory: The role of language transfer. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 317–340). San Diego: Academic Press.

Gries, St. Th. (2006). Corpus-based methods and cognitive semantics: The many meanings of *to run*. In St. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis* (pp. 57–99). Berlin: Mouton de Gruyter. DOI: 10.1515/9783110197709

Gries, St. Th. (2010a). Behavioural Profiles 1.01: A program for R 2.7.1 and higher.
Gries, St. Th. (2010b). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*, 5(3), 323–346.
Gries, St. Th., & Deshors, S. C. (To appear). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*.
Gries, St. Th., & Divjak, D. S. (2009). Behavioral profiles: A corpus-based approach to cognitive semantic analysis. In V. Evans, & S. Pourcel (Eds.), *New directions in cognitive linguistics* (pp. 57–75). Amsterdam: John Benjamins.
Gries, St. Th., & Divjak, D. S. (2010). Quantitative approaches in usage-based cognitive semantics: Myths, erroneous assumptions, and a proposal. In D. Glynn, & K. Fischer (Eds.), *Quantitative cognitive semantics: Corpus-driven approaches* (pp. 333–354). Berlin: Mouton de Gruyter.
Gries, St. Th., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*, 34, 121–150.
Gries, St. Th., &Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, 18(3), 327–356.
Hermerén, L. (1978). *On Modality in English: A study of the semantics of the modals*. Lund: LiberLäromedel/Gleerups.
Huddleston, R. D. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Hyltenstam, K., Bartning I., & Fant L. (2005). *High Level Proficiency in Second Language Use*. Research program for Riksbanken Jubileumsfond. (Stockholm university) http://www.biling.su.se/~AAA.
Kilborn, K., & Ito, T. (1989). Sentence processing strategies in adult bilinguals. In B. MacWhinney, & E. Bates (Eds.), *The cross-linguistic study of sentence processing* (pp. 257–291). Cambridge: Cambridge University Press.
Klinge, A., & Müller, H. H. (2005). Modality: Intrigue and inspiration. In A. Klinge, & H. H. Müller (Eds.), *Modality studies in form and function* (pp. 1–4). London: Equinox.
Leech, G. (1969). *Towards a semantic description of English*. Bloomington, IN: Indiana University Press.
Leech, G. (2004). *Meaning and the English verb*. London & New York: Longman.
MacWhinney, B. (2004). A unified model of language acquisition. Retrieved from <http://psyling.psy.cmu.edu/papers/CM-general/unified.pdf> [Accessed 18 June 2010].
Neff, J., Dafouz, E., Herrera H., Martínez, F., & Rica, J. P. (2003). Contrasting the use of learner corpora: The use of modal and reporting verbs in the expression of writer stance. In S. Granger, & S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges* (pp. 211–230). Amsterdam: Rodopi.
Nuyts, J. (2006). Modality: Overview and linguistic issues. In W. Frawley (Ed.), *The expression of modality* (pp. 1–26). Berlin: Mouton de Gruyter.
Palmer, F. (1979). *Modality and the English modals*. London & New York: Longman.
Radden, G. (2007). Interaction of modality and negation. In W. Chłopicki, A. Pawelec, & A. Pokojska (Eds.), *Cognition in language: Volume in Honour of Professor Elżbieta Tabakowska* (pp. 224–254). Kraków: Tertium.
R Development Core Team (2010). *R: A language and environment for statistical computing. Foundation for statistical computing*. Vienna, Austria. <http://www. R-project.org>.

Salkie, R. (2000). Corpus linguistics: A brief guide to research in French language and linguistics. *AFLS Cahiers*, 6, 44–52.

Salkie, R. (2004). Towards a non-unitary analysis of modality. In L. Gournay, & J.-M. Merle (Eds.), *Contrastes: mélanges offerts à Jacqueline Guillemin-Flescher* (pp. 169–182). Paris: Ophrys.

Vendler, Z. (1967). Verbs and times. In Z. Vendler (Ed.), *Linguistics in philosophy* (pp. 97–121). New York: Cornell University Press.