A multifactorial study of the uses of *may* and *can* in French-English interlanguage

Sandra C. Deshors

November 2010

Department of Linguistics
University of Sussex
Brighton, BN1 9QJ, UK

A thesis submitted for the degree of Doctor of Philosophy

"One can only truly
wonder about the
limits of the
Number amidst the
power of the Word
and the stretching
of Linguistics from
philological and
humanistic origins
into a quantitative
science"

Steven Clancy

Abstract

This study contributes to the understanding of how language learners make use of a second language, specifically how French English learners make use of the English modals *may* and *can*. The study is based on the assumptions that (i) acquiring a new language is a cognitively demanding task which requires the acquirer to identify large amount of co-occurrence data, that (ii) those data are probabilistic in nature rather than absolute, and (iii) that semantic differences are particularly hard to discern and learn as they are not explicitly noticeable. This study applies Divjak and Gries's (2008) Behavioural Profile approach to semantic analysis to a corpus of native and learner English and native French in order to offer a fine-grained quantitative investigation of the co-occurrence patterns of *may* and *can* in both English varieties. It shows not only that *may* and *can* can be characterised and differentiated on the basis of their co-occurrence patterns, but also that such co-occurrence patterns vary systematically in native English and French-English interlanguage. This finding is supported by monofactorial and multifactorial statistical results indicating that (i) the meanings and the functions of *may* and *can* in both English varieties are correlated with the distributions of formal elements within their contexts of occurrence and (ii) that the uses of *may* and *can* activate different linguistic levels simultaneously. Generally, these results suggest that the grammatical context of the forms' occurrences presents processing constraints that influence and ultimately characterise learners' choices of *may* and *can*. More specifically, the study identifies six grammatical components that systematically trigger use of *may* and *can* in a non-native fashion. Overall, the study shows that (i) it is possible to predict learner language on the basis of corpus-based investigation and (ii) that the use of multifactorial statistical methods facilitates the formulation of corpus-based and psychologically-informed hypotheses on the processing and the acquisition of lexical items by second language learners.

Acknowledgements

I am very grateful to Dr. M. Lynne Murphy, Prof. Stefan Th. Gries and Prof. Raphael Salkie who all, in various degrees of involvement, took a supervisory role in this Ph.D. project. I am particularly grateful to Lynne for granting me the geographical and intellectual freedom to seek out knowledge beyond the University of Sussex and for supporting my numerous endeavours. I am very grateful to Stefan who is not only an outstanding scientist but also an incredible mentor. I wish to thank him for his generous input and for shaping my way of thinking in most exciting ways. I am also grateful to Raf for sharing with me his expertise on the modals.

I wish to thank the Linguistics Department at the University of California at Santa Barbara and particularly the Chair of the department, Prof. Patricia Clancy, for welcoming me as a visiting scholar throughout the 2009-2010 academic year. The completion of this thesis would not have been possible without the specialist training that I was able to receive while at UCSB or my extensive use of the library there. From the University of Sussex, I thank the former School of Humanities for financially supporting me over the past three years. I also thank Mrs Margaret Reynolds for her unparalleled sympathy and emotional support in the first two years of this project.

I also wish to thank Miss Anna Jordanous, fellow graduate student and dear friend, for her never-ending curiosity and interest in my work and I thank my parents for their support and for never doubting the success of this venture.

Finally, a heartfelt 'thank you' to my partner for his remarkable patience and precious support.

Table of Contents

Chapter 6 Results

Chapter 7 General discussion

List of Figures

List of Tables

List of Abbreviations

BP	Behavioural Profile
CA	Contrastive Analysis
CIA	Contrastive Interlanguage Analysis
CODIF	Corpus de Dissertations Françaises
CM	Competition Model
DCA	Distinctive Collexeme Analysis
EFL	English as a Foreign Language
ESD	English as a Second Dialect
ESL	English as a Second Language
GLM	Generalised Linear Model
HAC	Hierarchical Agglomerative Cluster analysis
ICLE	International Corpus of Learner English
ICM	Integrated Contrastive Model
IL	Interlanguage
LDA	Linear Discriminant Analysis
LOCNESS	Louvain Corpus of Native English Essays
L1	First (native) Language
L2	Second Language
NL	Native Language
SLA	Second Language Acquisition
TL	Target Language

Chapter 1 An introduction to the study

1.1 The scope of the study

This study investigates the uses of *may* and *can* by French learners of English. The use of competing linguistic variants such as *may* and *can* in second language (L2) is a cognitively demanding task: from a cognitive-linguistic and exemplar-based perspective, learners need to identify and acquire the forms' co-occurrence patterns as they occur in native English. In order to understand and explain what motivates learners' lexical choices between *may* and *can*, the analysis must reach beyond the linguistic description of the two modals as isolated linguistic forms and focus on how the two forms relate to their linguistic contexts. This study shows that by shifting the analytical scope of the modals from single lexical items to grammatical contexts, one is able to analyse how the uses of *may* and *can* reflect general principles of second/foreign language acquisition and thereby provide a more cognitively grounded explanation of the uses of *may* and *can* by non-native speakers of English.

1.2 Background of the study

At the core of this study is the general methodological issue of how to include *may* and *can*'s grammatical contexts into their quantitative analysis. This issue brings together three questions that have so far remained unanswered in studies on modality and second language acquisition:

- how to empirically investigate modality as a semantic category;
- how to investigate the modals both quantitatively and in a way that maximally accounts for their linguistic contexts;
- how to provide an empirical analysis of the uses of the modals that contributes to L2 acquisition theory.

Due to its notoriously evasive nature, modality is a complex semantic category to investigate. In the words of Nuyts,

> [modality] remains one of the most problematic and controversial notions: there is no consensus on how to define it and to characterise it, **let alone on how to apply definitions in the empirical analysis of data** (Nuyts 2005: 5) [my emphasis]

In sharp contrast with Nuyts 2005, work on English modals by Kennedy (2002) and work on linguistic variation by Reppen, Fitzmaurice and Biber (2002: vii) have shown that empirically-grounded approaches provide "adequate" descriptions of language use. According to Reppen, Fitzmaurice and Biber, for instance, a crucial aspect of such descriptions lies in their integration of linguistic contextual features into the analysis of linguistic items. Recent work by Klinge and Müller (2005) has also valued the role played by linguistic contextual features in conveying modal meaning. In that regard, Klinge and Müller note that to capture the essence of modality, "it seems necessary to cut across the boundaries of morphology, syntax, semantics and pragmatics and all dimensions from cognition to communication are involved" (2005: 1).

Generally, the notion that the modals' co-occurring grammatical features contribute to their semantics has not equally been exploited in research on native English and learner English. In fact, the relevant literature reveals a clear division between corpus-based work on the modals in native English and learner English: while, in native English, for instance, Kennedy (2002) is concerned with predicting variability in the choice of specific modal verbs on the basis of their register and their grammatical contexts of occurrence, interlanguage studies, such as Aijmer (2002) on the uses of the modals in Swedish-English interlanguage, take hardly any account of the influence of linguistic features on the learners' modal choices. It follows that there is currently no homogeneity in the way variability in the uses of the modals is investigated across English varieties. The tradition in corpus-based interlanguage research to approach modal forms as isolated lexical items is to some extent surprising, given Meunier's (1998) call for syntactically-grounded investigations of interlanguage. According to Meunier, such investigations are "certain to reveal many interesting features of non-nativeness which are yet to be discovered" (1998: 36). Emerging from Meunier's view is the question whether, and if so to what extent, syntactic structures in L2 contribute to the non-nativeness of the use of modal forms by English learners. In the same spirit, it remains to be shown whether other linguistic levels, such as, for instance, the semantic or morphological levels, interfere with the non-native use of *may* and *can*.

According to Granger (1998), exploring this question quantitatively is beneficial to second language acquisition (SLA) theory. In fact, Granger claims that

> [b]y offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will […] [c]ontribute to SLA theory by providing answers to some unresolved questions such as the exact role of transfer (Granger 1998: 17)

The research presented here is a step towards realizing Granger's goal. More specifically, the current study recognises the need to establish whether co-occurring grammatical features of *may* and *can* play a part in the L2 acquisition of the forms and one way to investigate this issue is by integrating contextual grammatical features into the quantitative analysis of the modals in L2.

1.3 Methodological and theoretical assumptions

This study is exclusively based on naturally occurring data as well as a set of assumptions that have been held in corpus linguistics for many years but that, as I show below, have not yet been combined in a study of the modals. According to Reppen, Fitzmaurice and Biber (2002: vii), adequate description of linguistic variation and use "must be based on empirical analyses of natural texts" and "should be based on multiple texts collected from many speakers, so that conclusions are not influenced by a few speakers' idiosyncrasies". Following Reppen, Fitzmaurice and Biber, the analysis presented in the current work is based on the fine-grained quantitative analysis of natural learner corpus data collected in natural contexts. The suitability of this type of data for the investigation of variability in the uses of the modals has been demonstrated both in native and learner English studies. In a context of his study on the modals in native English, Kennedy (2002: 73) notes, for instance, that "[m]odern electronic corpora now make it possible to explore the nature and use of linguistic phenomena in a much wider variety of texts". This is indeed demonstrated in Biber (1999), for instance, who reports the distribution of modal verbs across registers (i.e., conversations, fiction, news, academic prose), showing that the distributional differences provide good indicators of the forms' individual characteristics. In the same spirit but with regard to learner language, Granger (2002: 28) writes that "[c]omputer learner corpora are a very rich type of resource which lends itself to a very wide range of analyses", and she agrees with McEnery and Wilson (1996: 12) who claim that such corpora are "the only reliable source for such features as frequency". Similarly, Aijmer (2002: 56) demonstrates that "[m]odality is just one example of an area of learner language which has not previously been well described and where computer learner corpora can make a contribution".

Given the purpose of the present study, it is crucial to note that descriptions based on electronic corpora "go beyond exploring what is grammatically and semantically possible, and add a distributional dimension which characterizes linguistic features in terms of probability of occurrence" (Kennedy 2002: 73). In the case of the English modals, this is a particularly important point as Kennedy's analysis of the modal verbs in a large corpus demonstrates that

> linguistic variation is characteristically a probabilistic phenomenon rather than an absolute one (…) [and] [i]t can be anticipated that sociolinguistic or regional varieties of the language are similarly likely to show not the presence or absence of particular linguistic phenomena, but a tendency for them to be used more or less than in other varieties (Kennedy 2002: 90)

Although in his study Kennedy (2002) takes into consideration the distribution of the modals in nine identified verb phrases, he primarily focuses on their occurrence as lexical items. The current study takes Kennedy's approach one step further by testing the variability of the uses of the modals as constructions (i.e., *may/can* + lexical verb) and, more generally, as part of co-occurrence patterns--that is, the ways *may* and *can* are used alongside the semantic, syntactic and morphological features in their co-text.

For Reppen, Fitzmaurice and Biber (2002: vii), "[c]orpus-based analyses provide a means of handling large amounts of language and keeping track of many contextual factors at the same time" (p. viii). In addition, the authors claim that the empirical investigation of linguistic variability "must simultaneously consider the influence of a range of contextual factors" (p. vii). However, they note, analysing such influence "presents difficult methodological challenges" (vii). In the context of interlanguage research, quantitative studies of the modals have yet to address such methodological difficulties emerging from the combination of a systematic analysis of modals and their co-occurring grammatical features. Once such quantitative studies are undertaken, however, this will open up new possibilities for exploring qualitative factors and patterns. According to Meunier (1998: 36), quantitative measures are recognised as "essential in language analysis" and "pave the way for more qualitative analysis" (p. 19) as"[s]urface differences – or similarities – between aspects of native and non-native language always require further qualitative investigation" (Meunier 1998: 36). The current study follows this spirit and shows that fine-grained description of corpus data can be carried out by integrating highly detailed annotations of corpus data into linguistic analyses. In fact, one can unveil linguistic patterns of use characteristic of L2 *may* and *can* by annotating a large number of semantic, syntactic and morphological features that co-occur with *may* and *can* in native and learner English and with *pouvoir* in native French and by analysing the resulting data points using a variety of sophisticated statistical techniques. In line with Kennedy (2002: 86) who claims that "[t]here is a substantial and multidimensional variation in the use of the modal verbs and the structure they occur in", this study shows that quantitatively assessing the uses of *may* and *can* in relation to their co-text contributes significantly to our understanding of how and why grammatical contexts affect the structure of learner language.

To explore the multidimensional aspect of the uses of *may* and *can* in a quantitative fashion as well as to fulfil its explanatory goal, this study takes a cognitive linguistic perspective on second/foreign language acquisition and language use. This study is based on the assumption that systematic patterns, structures and functions of language reflect aspects of general cognition. More particularly, and in line with usage-based and exemplar-based theoretical approaches, this study assumes that (i) linguistic knowledge and processing is based on how language is used, (ii) that the language system is built bottom-up (from specific instances to generalisations) and is based on recurring patterns, and (iii) that linguistic structure emerges from language use. Finally, in line with Bates and MacWhinney's (1982, 1989) function-oriented Competition Model, I also assume 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 arrive at natural-sounding choices. Assuming the interrelation between the use of linguistic items and the shape of their grammatical environments bears a major methodological implication for the corpus linguist, namely that sophisticated statistical techniques are required to assess simultaneously the behaviour of, in this case, *may* and *can* in different English varieties and on the basis of a large number of variables across different linguistic levels (i.e., semantic, syntactic and morphological).

1.4 Outline of the study

The current work is organised according to the following overall structure: introduction, methods, results and discussion. Chapters 2, 3 and 4 provide the background information from which the study draws its main motivation. Chapter 5 introduces the data and the methods that were employed for their analysis. Chapters 6 and 7 present the results of the study and discuss their implications. More specifically, individual chapters are structured as follows:

Chapter 2 is concerned with *may* and *can* and discusses the problematic aspects of the two forms as a semantic pair. Throughout the discussion, past theoretical approaches to *may* and *can* are considered with an eye to how they contribute to the establishment of an adequate empirical method of investigation. The discussion is structured as follows. Section 2.1 presents an overview of the chapter. Section 2.2 presents some theoretical preliminaries: first *may* and *can* are presented as modal auxiliaries and second, *may* and *can* are discussed in terms of their traditional meanings. Section 2.3 presents two descriptive studies, Hermerén 1978 and Perkins 1983, that investigate the English modals in relation to their grammatical environments. Section 2.4 presents three quantitative studies: Coates 1983 and Gabrielatos and Sarmento 2006, which attempt to account for the modals' grammatical contexts, and Collins 2009 investigates different native English varieties. Section 2.5 summarises chapter 2.

Chapter 3 concerns interlanguage (IL) and Second Language Acquisition (SLA). In this chapter, I define the nature of interlanguage and discuss related implications for corpus-based second-language-acquisition research. Section 3.1 presents a general overview of the chapter. Section 3.2 presents two approaches to interlanguage, namely those of Selinker (1969, 1972) and Adjemian (1976). Section 3.3 considers methodological implications for the investigation of interlanguage. Section 3.4 discusses second language corpus work, with a particular focus on methods and applications to the English modals. Section 3.5 presents Salkie's (2004) contrastive study of *may, can* and native French *pouvoir.* Section 3.6 presents the theoretical motivations for adopting a cognitive, usage-based approach to investigating IL *may* and *can.* Section 3.7 summarises chapter 3.

Chapter 4 presents an overview of relevant previous work in corpus linguistics. Section 4.1 provides a general introduction to the treatment of linguistic similarity and alternation phenomena in corpus linguistics. Section 4.2 presents the Behavioural Profile approach as a way to investigate semantic similarity. Section 4.3 focuses specifically on methodologies dealing with the description and prediction of alternation phenomena, including monofactorial and multifactorial approaches. Section 4.4 summarises chapter 4.

Chapter 5 sets out the methodology of the current study. Section 5.1 briefly introduces the two major characteristics of the analyses presented in the current work: their fine granularity and their quantitative nature. Section 5.2 presents the corpora and the data. Section 5.3 presents the independent variables according to which the uses of *may* and *can* are assessed and explains how the data are annotated. Section 5.4 explains the statistical procedures to which the data were subjected.

Chapter 6 presents and discusses the results of the analyses. Basic monofactorial results are presented first in Section 6.1. The behavioural tendencies of *may* and *can* in relation to the semantic, morphological and syntactic variables are described and compared to find out which variables affect the uses of *may* and *can* and how and to what extent they characterise the uses of the two modals. Section 6.2 covers the multifactorial results: the uses of *may* and *can* by native speakers and learners are distinguished by means of a hierarchical cluster analysis, a logistic regression and a collexeme analysis. Overall, the analysis sheds light on how to predict learners' choices between *may* and *can* in authentic discourse situations.

Chapter 7 discusses the general implications of the study. Section 7.1 introduces the main line of discussion in the chapter. Section 7.2 revisits the studies discussed in Chapter 2. Section 7.4 demonstrates how the study surpasses existing accounts of the two modals. Section 7.5 assesses the extent of the success of the BP method for the investigation of non-native use of *may* and *can*. Section 7.6 illustrates how the BP method provides a reliable way to formulate corpus-based and psychologically-informed hypotheses on the processing and acquisition of lexical items in second language. Section 7.7 presents a number of recommendations for future work and concludes.

Chapter 2 *May* and *can*: some preliminaries

2.1 Introduction

Both *may* and *can*, as expressions of the semantic category of possibility, are widely discussed in descriptive grammars, which provide detailed accounts of the variety of their uses and interpretations. Across these grammars *may* and *can* have been approached from many perspectives; these include semantic-, lexical- and context-based approaches. Palmer (1990), for instance, analyses *may* and *can* around the two semantic categories of epistemic possibility and deontic possibility. Leech (2004), on the other hand, approaches the modals as lexical forms. The variety of treatments of the modals across grammars reflects their semantic complexity (Palmer 1990). A particular aspect of *may* and *can* that has fuelled much theoretical debate is their degree of semantic similarity: for some scholars, the "two modals of possibility, *can* and *may*, share a high level of semantic overlap" (Collins 2009: 91) and for others, "[i]t is safe to say that *may* and *can* do not mean the same, and the difference in meaning, lodged in the semantic spectrum of the words themselves, transcends the particularities of usage" (Bolinger 1989: 2). Beyond *may* and *can* as isolated lexical forms, scholars such as Hermerén (1978) and Perkins (1983) have brought forward the notion that the linguistic context of the two modals contribute to their various interpretations and, in the case of Perkins, is part of the modals' underlying system. Although Hermerén 1978 and Perkins 1983 have provided a broader outlook on the English modals, the theoretical implications of those studies are based on the qualitative assessment of the modals and those implications remain to be supported quantitatively. This is an important point as Perkins recognises the existence of a correlation between the ways of investigating the modals and the forms' theoretical implication:

> [t]he question of where the line is to be drawn between what belongs to the meaning of the modal and what belongs to the context and situation is at least to some extent determined by the context of the investigation and the material. (Hermerén 1978: 71)

To investigate *may* and *can* from a grammatically-grounded perspective raises methodological issues for the corpus linguist: quantitative studies such as Coates (1983) and Gabrielatos and Sarmento (2006) provide an illustration of the challenges that a fine-grained quantitative investigation of the modals incur. In what follows I identify a possible way to maximally and comprehensively investigate the uses of *may* and *can*. My account starts with a definition of the three types of modality assumed in this study. As a second step, I present the formal and the semantic characteristics of *may* and *can* as modal auxiliaries and show that for quantitative purposes, a polysemous theoretical approach to the modals is necessary. Thirdly, while I show how, due to their overlapping semantics, *may* and *can* present a challenge for the corpus linguist, I draw on existing quantitative studies, namely Coates (1983) and Gabrielatos and Sarmento (2006), to show that

(i) the grammatical context of *may* and *can* influences their interpretation, and

(ii) modal meanings can be systematically assessed through the quantitative treatment of such grammatical contexts.

(iii) Once these facts are established, the uses of *may* and *can* can be investigated with the degree of granularity necessary for an empirically-based theoretical account of both forms.

Finally, I illustrate with Collins 2009 the limitations of excluding linguistic contexts from the quantitative analysis of modal verbs.

2.2 *Theoretical preliminaries*

2.2.1 *May* and *can* as modal auxiliaries

As modal auxiliaries, *may* and *can* display a number of formal characteristics which, as listed in Coates (1983: 4), include the following:

(a) direct negation (*can't*, *mustn't*)

(b) inversion without DO (*can I?*, *must I?*)

(c) 'code' (*John can swim and so can Bill*)

(d) emphasis (*Ann COULD solve the problem*)

(e) no *-s* form for the third person singular (**cans*, **musts*)

(f) no non-finite forms (**to can*, **musting*)

(g) no co-occurrence (**may will*)" [in standard varieties of English]

As members of the grammatical class of modal auxiliaries, *may* and *can* also share a number of semantic characteristics. In that regard, Palmer (1979: 18) notes that while "[t]here is no doubt about the central position of *may, can* and *must*", he adds that "[t]hey are both formally modals and clear exponents of **possibility** and **necessity**" [my emphasis]. With the purpose of defining the core of the category of modality using a prototype approach, Salkie (2009) provides a list of semantic and pragmatic criteria for assessing potential modals' degree of membership to that category. I adopt Salkie's criteria as a starting point, providing evidence of the semantic characteristics shared by *may* and *can*. Salkie approaches modality from a topological perspective and he points out the semantic distinctions between modal forms in their various contexts of occurrences by showing that (i) the modal aspect of the forms is gradable and (ii) the modals' degrees of modality are usage-dependent rather than lexically motivated. In other words, the context of use of a modal contributes to, or is at least correlated with, its varying degrees of modality. In what follows, I present and define Salkie's (2009) criteria for membership to the semantic category of modality.

Salkie (2009) proposes four criteria for the category of modality. Core members of the category typically

(1) express possibility or necessity;

(2) are epistemic or deontic;

(3) are subjective in that they

  (a) denote speaker commitment,

  (b) involve primary pragmatic processes,

  (c) yield a sharp distinction between the modal expression and the propositional content;

(4) are located at one of the extremes of a modal scale.[1]

Salkie (2009) presents the four criteria as tests to be applied on uses of modals in context. According to Salkie, a single modal form can pass or fail different criteria depending on the specifics of its context of utterance and, as a result, that particular form yields degrees of modality ranging from low to high. So ultimately, Salkie's approach provides a way to describe contrasts among

i. occurrence of *may* and their respective degrees of modality
ii. occurrences of *can* and their respective degrees of modality
iii. occurrences of *may* and *can* and their degrees of modality

For instance, with examples (1) and (2) below, Salkie shows how two uses of *may* differ in the number of criteria they each pass.[2]

(1) I don't know for sure but there *may* be milk in the fridge.

(2) Nursing and medical staff *may* also need psychological support to cope with the intense nature of the treatment and the uncertainties in outcome (BNC EE8 800)

While in (1), *may* is reported to match all the criteria, *may* in (2) "does not express speaker uncertainty but generalises over the members of the class of nursing and medical staff" and thereby fails criterion 3a. Salkie also considers the *may* in (2) to fail criterion 3b on the basis that "there is no need to invoke here the knowledge of the speaker in order to interpret *may*" (p. 10). While the form passes all other criteria, Salkie concludes that the two instances of *may* yield different degrees of modality, with (1)

[1] Salkie (2009) assumes that "modal expressions can often be located on a scale" (Salkie 2009:8). Generally, studies concerned with the role of the speaker in the construction of modal meaning tend to 'measure' and assess the degree of presence of the speaker in a scalar fashion. Typical scales applied to the English modals include a ranking of the modal forms according to their increasing (or decreasing) identified degrees of 'strength'. Van der Auwera (1996:185) notes, for instance, that "[t]o call something a 'scale' rather than a 'diagram', the element making up the scale must not merely be ordered, there must also be a dimension along which the elements have increasing values".

[2] Salkie (2009) uses his own examples as well as examples extracted from the British National Corpus, labelled *BNC*. I report BNC corpus sources as they appear in Salkie 2009.

demonstrating a higher degree of modality and (2) demonstrating a lower degree of modality.

Salkie's modal semantic criteria can be applied to *can*, as shown with examples (3) and (4) below:

(3) I *can* track him, and he won't know (BNC cm4 2289)

(4) It *can* be cold in Stockholm.

According to Salkie, *can* in (3) fails criteria 1, 2 and 3 and thereby yields a low degree of modality. Similarly, *can* in (4) fails criteria 1, 2 and 3b, however it passes (3)a, as the occurrence denotes a degree of speaker commitment. Overall, Salkie concludes that *can* in (4) indicates a higher degree of modality than *can* in (3), although both (3) and (4) show a lower degree of modality than *may* in (1) and (2). *Can* is thus considered to have uses that are more peripheral to the modal category than those of *may*, as uses of it "usually fails one or more of the criteria" (2009: 20).

Beyond setting out semantic criteria, Salkie (2009) suggests that a polysemantic theoretical approach to *may* and *can* allows for fine-grained semantic comparisons. Such an approach assumes that (i) individual forms are polysemous in nature, and (ii) that their semantic interpretations are influenced by their utterance context. Given the high degree of semantic similarity of *may* and *can*, the polysemantic approach allows us to (i) contrastively assess the degrees of modality expressed by both modal forms while denoting equivalent senses and (ii) to provide descriptive contrastive accounts of the uses of both forms on the basis of their grammatical context of occurrence.

After a brief overview of the three types of modality relevant to this study, the next section provides an account of the senses of *may* and *can*, followed by a description of the grammatical contexts in which those senses are commonly found.

2.2.2 *May* and *can* and their traditional meanings

The semantic notion of 'modality' includes a wide range of heterogeneous concepts that many scholars have attempted to unite under a variety of categorisation systems (cf. Palmer 1979, Coates 1983, Bybee and Fleischman 1995, Huddleston and Pullum 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", dynamic, deontic and epistemic modality are traditionally recognised as referring to basic semantic modal dimensions (Nuyts 2006). The current study assumes that those three semantic dimensions identify distinctive senses of *may* and *can*. In what follows, I define each of the three modal categories as they are used in this study: epistemic, deontic and dynamic modality.

This study assumes Palmer's (1990) definitions of different kinds of modality. According to Palmer (1990: 6), epistemic modality is essentially concerned with “making a judgement about the truth of a proposition. More specifically, Palmer writes:

> [t]he function of epistemic modals is to make judgements about the possibility, etc. that something is or is not the case. Epistemic modality is, that is to say, the modality of propositions, in the strict sense of the term, rather than of actions, states, events, etc. (Palmer 1990: 50)

Epistemic possibility is thus concerned with "the speaker's assumptions or assessment of possibilities and, in most cases, it indicates the speaker's confidence (or lack of confidence) in the truth of the proposition expressed" (Coates 1983: 18). In the occurrences below, taken from the data of the present study, the speaker has mentally assessed the chances of an event occurring. Consider (5) and (6):

(5) for example, everyone *may* suddenly become vegetarian (ICLE-ALEV-0004.9)

(6) indeed, Europe 92 *may* lead to the disappearance of cultural differences (ICLE-FR-UCL-0079.1)

In (5) and (6), although the speaker acknowledges the possibility of an event occurring (i.e. everyone becoming a vegetarian, Europe 92 leading to the disappearance of cultural differences), the actualisation of the event cannot be asserted due to limited speaker knowledge.

In this work, deontic modality is understood as follows:

> the kind of modality that we call deontic is basically performative. By uttering a modal, a speaker may actually give permission (*may, can*), and make a promise or threat (*shall*) or lay on obligation (*must*) (Palmer 1979: 58)

Examples (7) and (8) illustrate deontic modality through the uses of *may* and *can*:

(7) people in all parts of the European continent speak of 1992 as a "magic" date, if I *may* say so (ICLE-FR-UCL-0093.1)

(8) 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 (ICLE-US-SCU-0013.4)

Finally, dynamic modality is defined as follows:

> dynamic modality is subject-oriented in the sense that it is concerned with the ability or volition of the subject of the sentence, rather than the opinions (epistemic) or attitudes (deontic) of the speaker (and addressee) (Palmer 1990: 36)

Generally, dynamic modality expresses the potentiality of an event occurring and the term 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. In (9) below, the modal refers to the physical ability of an animate grammatical subject: the seeing event is possible because the grammatical subject has the ability to see. Similarly in (10), the grammatical subject is lacking an ability, namely that of acceptance which prevents him to find happiness. Although (9) refers to a physical ability and(10) refers to a mental ability, both cases however, deal with grammatical subject-inherent features that enable the event to be carried out.

(9) those places where they *can* see, feel and even smell beautiful pictures (ICLE-FR-UCL-0010.3)

(10) Caligula is unable to accept his situation and therefore *can* find no happiness in it (ICLE-BR-SUR-0005.1)

Ability types of dynamic modality include occurrences where the grammatical subject is inanimate but refers to inherent properties of the subject. In (11) below, for instance, microwaves are mentioned in general terms and they are recognised as potentially efficient for heating up leftovers. They have, by their very design, a defining feature that enables them to efficiently heat up leftovers.

(11) Microwaves *can* be extremely efficient for heating up left overs (ICLE-US-MICH-0043.1)

The use of the modal auxiliaries is one possible way to express modality in English and as a semantic pair, *may* and *can* cover deontic, epistemic and dynamic meanings. More concretely, both forms can be used to express two types of possibility, namely epistemic possibility and dynamic possibility as well as permission and the possibility distinction applies to epistemic, deontic and dynamic senses. In what follows, I first discuss *may* and *can*'s possibility uses followed by their permission uses.

With regard to epistemic possibility, *may* is commonly preferred over *can*. Consider the use of epistemic *may* in (12) below (the example below is borrowed from Leech 2004):

(12) Careful, that gun *may* be loaded

As Leech (2004) points out, the epistemic possibility sense of *may* in this example can be captured in the paraphrase 'it is possible that it is loaded'. With regard to *can*, it is generally recognised that "*can* may serve as a marker of epistemic possibility, albeit restricted to non-affirmative contexts" (Collins 2009: 98). Leech (2004: 74), for instance, notes that "[the epistemic] sense of *can* is often found in the negative with *cannot* or *can't*. (…) sometimes *can* (= 'possibility') has a habitual meaning, which can be paraphrased by the use of the adverb *sometimes*". Consider Collins' (2009) examples below as an illustration of epistemic possibility *can.*

(13) No, it *can't* be hundred percent wrong 'cause the program um if you don't have your exact time of birth you set it to oblique orbit of zero for the time and PM for the hour (ICE-AUS S1A-096 95)

With regard to dynamic possibility, Leech (1969: 220) points out that "[g]rammarians have sometimes noted that *may* and *can* are not exactly interchangeable in the sense of possibility". For Leech (1964, 2004), this use illustrates the lack of semantic equivalence between *may* and *can*. Further, their semantic difference can be captured through the two grammatical constructions *it is possible that* and *it is possible for*: while *can* is synonymous with the structure *it is possible for*, *may*, on the other hand, is synonymous with the structure *it is possible that*. This semantic difference is illustrated below in (14) and (15):

(14) The pound *can* be devalued [=it is possible for the pound to be devalued]

(15) The pound *may* be devalued [=it is possible for the pound to be devalued]

In (14), "the notion of possibility is general and theoretical" (Leech 1969: 221) in the sense that it is common knowledge that 'technically' the pound can be devalued. In (15), however, the devaluation of the pound is not seen as a possible event but rather as a practical course of action under consideration. In that respect, *can* in (15) illustrates a case of "theoretical possibility" and *may* in (15) illustrates a case of "factual possibility".

Collins (2009) reports a notable difference between the frequencies of occurrence of dynamic *may* and dynamic *can* and he recognises that although dynamic possibility is a minor meaning for *may*, it is a major one for *can* (p. 96). Collins (2009) also reports three types of dynamic possibility, namely theoretical possibility, dynamic implication and ability, and he further shows that while *may* and *can* share the uses of dynamic possibility and dynamic implication, only *can* is used in the ability sense.

The dynamic implication refers to cases of implied directive speech acts: "[t]his category differs from theoretical possibility to the extent that it requires an expansion of the semantic framework into pragmatics" (Collins 2009: 96). In (16) and (17), dynamic implication is illustrated first with *may* and then with *can*. In (16), *may* denotes a suggestion and in (17) *can* denotes an instruction. Both examples are from Collins 2009.[3]

(16) And you *may* remember that the organisations the republics that were in the Soviet Union competed in the recent Winter Olympics under the title Commonwealth of Independent States (ICE-AUS W2D-001 67)

(17) Well, I'll pour the ladies', you *can* pour the men's (ICE-AUS S1A-004)

In its ability sense, *can* is semantically equivalent to *be capable of* or *know how to* (Leech 2004). The distinction between the senses of capability and possibility "tends to become blurred in actual usage" (Leech 1969: 222). In addition, Collins (2009: 103) notes that "the ability meaning may be close to actualization, though it may not be realised".

The second meaning shared by *may* and *can* is that of permission. This meaning can be expressed either with deontic *may* or deontic *can*, as shown below in (18) and (19):

(18) you *may* use my desk
Well, wait a minute, it's royal mess, isn't it (C-US SBC-019 138-140)

[3] Collins (2009) uses examples extracted from the British component of the *International Corpus of English* ('ICE-GB'), the Australian component of the *International Corpus of English* ('ICE-AUS'), and a specially assembled corpus of American English ('C-US'). More details about Collins' (2009) American sub-corpus are provided further below in Section 2.4.3

(19) You *can* come back for a second helping (ICE-AUS S1A-004 257)

The semantic equivalence of *may* and *can* in the permission sense is a matter of controversy: some scholars promote their semantic equivalence (despite their differing frequency of occurrence and different degrees of formality; Leech 2004, Collins 2009), others have distinguished the two forms. On the similarity of the two forms, Leech notes that "[t]he permission and possibility meanings of *may* are close enough for the distinction to be blurred in some cases" (p. 76). However, while "[i]n asking and giving permission, *can* and *may* are almost interchangeable" (p. 75), "[*m*]*ay* characteristically signal[s] permission given by the speaker of writer, or (in questions) by the hearer" (p. 76). On the forms' semantic dissimilarity, Vanparys (1987) argues that *can* generally makes a statement of permission. *May*, on the other hand, grants or requests a permission. Bolinger (1989) uses the notion of *inclusion* to illustrate the distinction between the two forms and claims that *may* "refers to the external, to what transcends the entity or situation" (p. 7) and *can* "refers to what a person, or situation is endowed with, whether naturally in place (physical, mental) or implanted (authorized, permitted)" (p. 7).

While the above has summarily discussed different senses of the modals, the situation is more complex than is obvious from that discussion. This is because the senses of the modals are correlated with their uses in grammatical contexts, as for example, Leech (1969, 2004), Huddleston and Pullum (2002) and Collins (2009) show. Leech, for instance, notes certain uses of *may* are to be found in particular grammatical contexts:

> There are important grammatical differences between the senses of *may.* 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. (Leech 2004: 77)

Voice has a similar influence on the uses of the forms. In the case of *can*, for instance, "[i]f we understand an active sentence in the sense of ability, the corresponding passive sentence has to be understood in the possibility sense" (Leech 1969: 223), that is its dynamic sense. Examples (20) and (21) from Leech (2004: 92) illustrate this point.

(20) he *can* (= is able to) beat the world champion

(21) the world champion *can* be beaten by him

In (20), *can* refers to the subject's intrinsic physical ability to beat the world champion, whereas in (21), the modal refers to the general possibility to beat the world champion, regardless of anyone's ability. *May* also takes on different interpretations according to whether it is used in questions or statements. In questions, the modal indicates a request for permission whereas in statements, it communicates a granting of permission (Leech 2004: 90). Further, in its epistemic possibility sense, *may* does not occur in questions "where its function is usurped by *can* or *could*" (Leech 2004: 92). Leech's (2004: 92) examples in (22) and (23) illustrate that point

(22) **May* they be asleep?

(23) *Can/could* they be asleep?

Aspect is another grammatical category that interferes the uses of the modals. Huddleston and Pullum (2002) point out that English modals that precede the perfect marker have are sometimes within the scope of the perfect aspect, and sometimes not. In the case of epistemic modal senses, Leech (2004) observes that the modal is never within the scope of perfect or progressive aspect.

Negation is the grammatical process that is most commonly recognised as interacting with the modals (cf. Hermerén 1978; Palmer 1979; Coates 1980, 1983; Palmer 1995; De Haan 1997; Huddleston and Pullum 2002; Radden 2007; Byloo 2009). Generally, the relevant literature presents negation as a twofold phenomenon (Halliday 1970, Hermerén 1978, Palmer 1979, Huddleston and Pullum 2002). Because of the invariant location of the negative *not* in the English verb string, negated modal sentences are often ambiguous as to whether the modality or the event is being negated. Consider (24) and (25) below:

(24) they believe that they *cannot* change their fate (ICLE-BR-SUR-0006.1)

(25) the literary man *may not* do without the progress and research of the scientist (ICLE-FR-UCL-0031.3)

In (24), the modal form refers to the possibility of changing one's fate and the proposition refer to the process of changing one's fate. In (25), the modal form refers to a degree of probability of one's capacity to do with or without the progress of the scientist and the proposition refers to the process of doing with or without the progress of the scientist. So in modalised utterances negation can potentially be applied to either the modal form or the proposition. However, the deceivingly straightforward dichotomy between (negated) modal forms and (negated) propositions hides semantic implications:

> [s]ince I take modality and negation to be two categories that change the basic meaning of the sentence, it stands to reason that they also affect a change of meaning on themselves (De Haan 1997: 11-12)

Coates (1980: 213) highlights a semantic contrast between the negated forms of *may* and *can.* With *may*, the negation "affects the proposition and not the modality. (...) That is, while *can* + *not* = 'it is not possible for x', *may* + *not* = 'it is possible that not x'". In that regard, Palmer (1979) remarks that the location of the negation depends on the type of modality expressed (i.e. epistemic, deontic or dynamic). For instance, in the case of epistemic possibility, Palmer observes that

> the proposition is negated by *may not* ('It is possible that … not...'), while the modality is negated by *can't* ('It is not possible that … ') (Palmer 1979: 53)

In other words, epistemic *may not* negates the modality and equates to a 'possible-not' type of negation, and epistemic *can't* negates the proposition/event and equates to a 'not-possible' type of negation. However, in deontic cases, Palmer (1979) reports that both *may not* and *cannot (can't)* negate the modality by refusing permission.

Comparatively to *can*, *may* is deficient in terms of the number of negated form that it has to offer. *Can* offers three (i.e. *cannot*, *can not* and *can't*), *may* only one (i.e. *may not*). While Palmer (1995: 468) states that "in English, it is to be assumed that the negative is formally associated with the modal since it is generally cliticised", however,

> there is, perhaps, a problem with English epistemic *may not*: this is not normally cliticised, so that the negative is not so clearly associated formally with the modal (Palmer 1995: 469)

In this section, we have seen how intertwined the meanings of *may* and *can* are and how descriptive grammars recognise degrees of interaction between the interpretations of the modals and their sentential grammatical components. In the next section, this notion of interaction is taken one step further with Hermerén (1978) and Perkins (1983) whose qualitative studies point towards a more radical investigation of modal verbs and their grammatical contexts.

2.3 Qualitative studies

2.3.1 Hermerén (1978)

Hermerén's (1978) overall approach to the semantics of the modals attempts to strike a balance between assuming "the autonomy of the individual linguistic item or the over-all importance of the context and the situation" (p. 71). Generally, for Hermerén (1978), including the linguistic context in which modal forms are used is an indispensable part of their semantic investigation. Although Hermerén recognises the usefulness of Palmer's (1974) approach which uses formal criteria such as tense, voice and negation to support semantic distinctions between the modals, he nevertheless points out Palmer's failure to relate the syntax and the semantics of the modals due to many exceptions that have to be made. For Hermerén, relating semantic forms to their syntactic possibilities is an approach that must be carried out in a systematic fashion. One benefit of such an approach is that subsequent accounts can be used to determine whether the frequency of occurrence of the modal forms is context-dependent or not. It follows that such an approach implies to take into consideration not only the total number of the various investigated modals but also the difference in content and style in their texts of occurrence. According to Hermerén, this allows us “to see if there [is] any relation between the number and the meanings of the modals and the content and style of the texts in which they occur” (p. 23). A second advantage of considering the linguistic context of utterance of the modals is that while it allows the researcher to investigate all the meanings of the forms, a context-based approach entails investigating the forms from the perspective of their uses. This is an important point because such a theoretical approach ultimately allows the researcher to include, at an initial stage of analysis, non-semantic linguistic levels such as syntactic or morphological levels.

In contrast with the majority of the studies of his time, Hermerén (1978) investigates modal forms from a grammatical-contextual perspective. Hermerén proposes a semantic classification of the modals by means of a paraphrasing technique. His method involves the identification of various types of grammatical contexts and the assessment of their relevance with regard to the meanings of the modals. Hermerén's account is of substantial interest for the current study because it suggests a possible methodology to (i) address the semantic complexities involved in an investigation of *may* and *can*, and (ii) to investigate modal forms in a more data-driven way than is usually adopted in studies on modality. Broadly, Hermerén's approach to the interpretation of modal forms relies on the hypothesis that the semantic variations of the forms are grammatically, rather than lexically, motivated. This standpoint stems from the observation that when used in different sentence types such as questions and statements or passive and active clauses, a modal does not always preserve its meaning (p. 71). Hermerén's innovative contribution lies in the claim that the modal forms are semantically sensitive to their co-occurring "basic sentence units", i.e. what in corpus linguistics are collocations and colligations/collostructions.[4] In other words, the semantic import of a given modal emerges from both the lexical item itself and the grammatical context in which it occurs. Although Hermerén explicitly lists the linguistic components voice, grammatical person, type of main verb (action, state, etc.), aspect and sentence type as playing an influencing role in the interpretation of the modals, the list is not exhaustive as it includes "any general linguistic category (whether semantic or syntactic)" that is compatible with the

[4] With the term "basic sentence unit" Hermerén (1978) refers to the linguistic elements that constitute the co-text of the modal, namely verbs, subjects, etc. Henceforth I refer to "basic sentence units" as *linguistic components*.

modal and that "can be shown to modify the meaning of the modal" (p. 74). For instance, Hermerén's quote below illustrates how the linguistic category voice affects the meaning of *can*:

> When a sentence with a modal is changed from the active into the passive (or vice versa), the change of sentence type often affects the meaning of the modal. Thus *can* has the Possibility sense in 'A dog can be chased by young children' which is to be interpreted as 'It is possible for a dog …', whereas *can* in the corresponding active sentence 'Young children can chase a dog' could also be interpreted in the Ability sense. (Hermerén 1978: 72)

Hermerén further demonstrates how changing the grammatical person of the surface subject influences the meaning of *may*:

> [t]wo different interpretations of *may* are seen in the following two sentences: 'You may do it' and 'Sonia may do it'. In the first sentence, with a second-person subject, *may* normally expresses Permission, while in the second sentence, with a third-person subject, *may* can also have the Possibility sense. (Hermerén 1978: 73)

In addition, Hermerén claims that the semantics of a main verb modified by a modal may influence the interpretation of that modal:

> *may* in 'You may do it' expresses Permission. However, if the activity verb do is replaced by a verb expressing a state, as in 'You may be late', may expresses Possibility rather than Permission. Similarly, Fillmore (1969: 113) mentions that *may* is usable in a construction with an achievement verb (i.e. a verb that has a built-in reference to the outcome of an activity) only in the Possibility sense as in 'He may find the eggs'. This is, as Fillmore points out, because the 'by-chance' relationship between the activity and the outcome. *May* in a sentence like 'he may look for the eggs' can, however, be interpreted either in the Possibility or the Permission sense (Hermerén 1978: 73)[5]

[5] Hermerén's argumentation raises the issue of identifying appropriate types of lexical verbs. As a starting point, Hermerén's argumentation is useful in the sense that it allows one to recognise the potential explanatory power of main lexical verbs with regard to the meanings of the modifying modals. Unfortunately, however, it does not reflect any attempt to operationalise them as much as possible. Lexical verbs represent an important grammatical component in the modals' linguistic contexts because they can provide useful information with regard to (i) the possible types of verbal events preferred by each modal form (e.g. states, achievements, processes, etc.) and (ii) whether or not the forms yield preferences for particular verb semantics (e.g. action, communication, cognition, etc.) and, if so, which. While Hermerén takes (i) into consideration, he fails to recognise the potential usefulness of (ii). Given Hermerén's (1978) examples, it can be inferred that both possibility and permission senses are compatible with *process*-type verbs such as *look for* or *do*. However, lexical

Hermerén's argumentation raises the question of whether *can* is similarly sensitive to the semantics of the lexical verbs it modifies. An interesting question would be whether corpus data would yield contrasting preferential co-occurrence patterns for *may* and *can* with regard to the types of verbs they respectively modify. Methodologically, the systematic investigation of a large enough data set would establish the existence (or non-existence) of a semantic correlation between a modal form and its modified lexical verb. Such an investigation could identify the types of lexical verbs that affect *may* and *can* individually; I return to this issue further below in this section. One benefit of the inclusion of main lexical verbs in a study on modals is that it allows one to check for the conceptual groundedness of observed (dis)preferences of *may* and *can* for particular types of lexical verbs.

The outcome of Hermerén's semantic investigation consists of a three-scale system of modal meaning, each modal scale referring to a number of semantically related modalities. He identifies three scales, or types of modalities, namely *internal modalities*, *neutral modalities* and *external modalities*. The distinction between these types of modality "is based on other sentence elements than the modal. Most often, these elements are to be found in the purely lexical context of the modal" (p. 97). In what follows, I first provide a brief overview of each modal scale, then discuss the place of *may* and *can* within Hermerén's three-scale system. Finally, I highlight the limitations of Hermerén's framework for the purpose of a systematic study of *may* and *can*.

The *scale of internal modalities* includes four modalities: DETERMINATION, INTENTION, WILLINGNESS and ABILITY. Hermerén characterises *internal modalities* as being "inherent in their surface subjects" (p. 95). This means that

> the action, quality or state expressed by the main verb of the modal is inherent in the surface subject of the modal, in such a way that in an active sentence "the subject is actor with respect to the modality

verbs also refer to processes that differ in nature and that can thereby denote, for instance, cognitive processes (e.g. *think*, *remember*) or physical processes such as actions (e.g. *walk*, *drink*). In turn, physical processes referring to actions may be sub-categorised according to the type of action that they denote for instance actions including geographical movement (e.g. *drive away*) or physical transformation (e.g. *break*). Hermerén's claim that the types of lexical verbs that are modified by *may* and *can* are useful to investigate calls for further and more in-depth investigation.

> [expressed by the modal]" (Halliday 1970: 339, quoted in Hermerén 1978: 99)

The *scale of neutral modalities* also includes four modalities: CERTAINTY, PREDICTION (including CUSTOM and PRESUMPTION), PROBABILITY, POSSIBILITY. Because Hermerén (1978: 103) recognises that the epistemic senses of the modals "differ from both the *internal* and *external* modalities", he categorises all epistemic senses as modalities of the neutral modal scale. The *scale of external modalities* includes six modalities: NECESSITY, SUGGESTION, APPROPRIATENESS, WANT, HOPE, and PERMISSION. External modalities refer to cases where "someone or something outside the subject decides what the subject is obliged or permitted to do or be" (p. 96).

Although *can* is present in all three scales of Hermerén's modal system, *may*, on the other hand, does not feature in the scale of internal modalities. Table 1 below illustrates the distribution of *may* and *can* across Hermerén's three modal scales. The form included in parentheses is mentioned by Hermerén as a possible alternative, but is not illustrated as such.

Table 1 Distribution of *may* and *can* across Hermerén's (1978) modal scales

modalities	internal modality	neutral modality	external modality
ability	*can*		
possibility		*may / can*	
suggestion			*may (can)*
permission			*may / can*

Table 1 shows that *may* and *can* are the only exponents of the POSSIBILITY modality and as such, they represent the neutral scale of modalities. For Hermerén, "each of these two modals can express two different kinds of possibility according to context" (p. 110) and he accordingly identifies POSS(1) and POSS(2): POSS(1) "indicates the speaker's view of the likelihood of an event occurring or having occurred (a state existing or having existed) (p.110) and POSS(2) "indicates that there is an (ungraded) possibility of the occurrence of an event or the existence of a state"/"expresses the situation that the speaker considers possible" (p. 110).[6] While both *can* and *may* are exponents of POSS(1) and POSS(2), grammatical features such as negation and tense can both distinguish between the two types of POSSIBILITY and characterise the use of *may* or *can* within each type of POSSIBILITY.[7]

The external scale of modalities includes the modalities permission and suggestion and *may* and *can* are the sole exponents of the PERMISSION modality. With regard to PERMISSION, Hermerén is concerned with which features qualify for inclusion of the modals in that modality and he reports that both *may* and *can* yield a "surface subject (…) [that] is represented as being affected by the permission" (p. 128).[8] With regard to SUGGESTION, the modality includes both polite requests and peremptory demands. While *may* and *can* are both recognised as exponents of that category, Hermerén's account of the forms is of limited reliability since, as the author points out, the only two provided examples for *may* and *can* "do not occur in the material investigated" (p. 121). However, Hermerén's observes that when denoting SUGGESTION (i) both *may* and *can* occur in statements or in interrogative statements and (ii) *may* (without *can*) occurs in interrogative statements.[9]

[6] A relation of inclusion exists between between POSS(1) and POSS(2) in the sense that POSS(1) entails POSS(2). In other words, for an event to be possible to happen, it has to be feasible in the first place.

[7] Hermerén notes that "when *can* expresses POSS(1), it seems to occur almost exclusively in negative contexts" (1978: 111). *May*, on the other hand, in the same sense of POSS(1), "is not restricted to negative contexts" (p. 111). With regard to POSS(2), while *can* is reported to occur in active clauses with both animate and inanimate subject referents, *cannot*, on the other hand, tends to occur in passive clauses and only with animate deep subjects. Still in the sense of POSS(2), *may* is reported to have a similar grammatical context to that of *can* since it "occurs with both animate and inanimate surface subjects, in both active and passive clauses" (p. 114).

[8] Although Hermerén acknowledges previous observations made by grammarians that "[i]n those cases where both *can* and *may* are expression of the modality PERM [PERMISSION], there seems to be a difference in their use in negative and interrogative clauses" (p. 126), he does not provide any descriptive account of the syntactic behaviour of the modals of the PERMISSION modality. In fact, his only contribution on the matter consists of the general observation that "even with examples taken from a corpus, it seems rarely to be the case that *can* and *may* can be rendered only with a paraphrase containing PERMISSION to the exclusion of the other modalities" (p. 128).

[9] Although Hermerén provides no syntactically-motivated account of the behaviour of the forms, it would have been of interest to contrast (i) the behavioural patterns of *may* and *can* when used in statement structures only, and (ii) the behavioural patterns of *may* only when used in statement structures and interrogative structures, respectively.

Within the internal scale of modalities, only ABILITY is of relevance to us.[10] Two exponents represent ABILITY, namely *can* and *could*. According to Hermerén, this modality refers to subjects' physical and mental ability and can be found in positive and negated forms of *can*.

Despite the fact that Hermerén convincingly illustrates the need to integrate contextual linguistic features into the investigation of the meanings of the modals, his study does not provide solid empirically-grounded evidence in support of this claim. In fact, Hermerén's study (1978) is not set up for the contrastive investigation of modal pairs. This is reflected in the fact that (i) *may* and *can* are not contrasted in a systematic fashion and (ii) when approached as a semantic pair, the grammatical criteria against which they are compared vary according to the semantic modality that they express. Overall, Hermerén's account of the meanings of the modals consists of a description of the semantic modalities included in each of the three modal scales and only modal forms sharing a semantic modality are contrasted with one another. This means that in the case of *may* and *can*, the two forms cannot be contrasted against each other in all their senses since they do not have all their modalities in common. For instance, as seen in Table 1, (i) *may* does not feature in the internal scale of modalities and (ii) Hermerén only briefly mentions their semantic equivalence on the external scale and does not contrast the two forms' grammatical behaviour. Furthermore, in cases where *may* and *can* do share semantic modality (i.e. POSSIBILITY and PERMISSION, SUGGESTION) and can therefore be contrasted, it is not necessarily done. In other words, Hermerén does not perform consistent comparisons of the two forms even in cases that would allow for it. For instance, in the case of the POSSIBILITY modality, although the grammatical behaviours of *may* and *can* are contrasted according to voice and subject animacy, when contrasted with regard to the PERMISSION and the SUGGESTION modalities, no syntactic description is provided and it is the modalities that are contrasted, not the syntactic behaviour of the modals.

[10] Hermerén's (1978) internal scale of modalities consists of four modalities: DETERMINATION, INTENTION, WILLINGNESS and ABILITY. There is no relation of implication between the first three modalities and ABILITY.

An additional limitation of Hermerén's study lies in its failure to consider the possibility that modal meanings may be affected by not only single grammatical components, but also the mutual interaction of several of those components. Underestimating this possibility has limiting implications for a fine-grained investigation of the modals.

Despite its above-mentioned shortcomings, by proposing a theoretical approach based on the investigation of the modals from the perspective of their syntactic behaviour, Hermerén allows us to consider a new methodological way for the quantitative treatment of the semantics of the English modal auxiliaries. On the basis of Hermerén's approach, it is indeed conceivable to extract each occurrence of a modal from a corpus and annotate it not only according to its contextual meaning but also according to contextual linguistic features, namely the syntactic units included in a modal's grammatical context. Such a methodological approach would allow for the systematic account of the frequency of occurrence of both *may* and *can* with each linguistic component included in the investigation, thus making the contrastive comparison of the forms methodologically reliable. Hermerén's study concludes that if modals are indeed compatible with general linguistic categories and "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". This suggests that a grammatically-grounded approach could have theoretical implications for an overall approach to the semantics of the English modal auxiliaries. Such a strong claim calls for empirical validation.

The notion that modal meaning is construed with reference to other sentential components is not limited to Hermerén's study. In the following section, I introduce a more recent study by Perkins (1983) that takes that notion further by claiming that the way individual modal verbs behave within their linguistic contexts reflects the modal system.

2.3.2 Perkins (1983)

In his study of the English modal expressions, Perkins (1983: 161) writes that "[a]lthough form is an obvious and necessary focus for linguistics, it can never provide anything like a complete picture of language, and in fact it may not even be the best starting point". While I am sympathetic to this line of approach, it raises the following questions: in relation to what other components should the modals be investigated so that a complete picture can be achieved, and what motivates the existence of such relations? With regard to both issues, Perkins 1983 provides a useful platform for discussion as the study illustrates (i) the extent to which the importance of grammar in relation to the uses of the modals has so far been underestimated and (ii) the lack of consistency with which has been treated. Generally, Perkins' study is set up to capture the core meanings of a number of English modal expressions by (i) isolating "a single core meaning for each of the English modals which is independent of its context of use" (p. 26) and (ii) by "identify[ing] and account[ing] for the properties that different uses of modal expressions obviously share" (p. 28).[11] Thus Perkins' method consists of extracting commonalities between all the uses of a single modal in order to abstract a more general meaning, comprehensive of all contextualised instances of that modal.

Although Perkins' study focuses on the semantics of modal forms rather their uses (i.e., the way they are assembled within the structure of their clauses), he nonetheless generally recognises the semantic motivation of syntax and acknowledges that "syntax is as it is by virtue of the semantic and pragmatic constraints upon individual expressions" (p. 160). Despite this recognition, Perkins admittedly excludes from his analysis the systematic treatment of syntax and only considers cases where there is an "obvious" (p. 26) semantic motivation. I question the methodological approach implied by the term *obvious*. First, although Perkins is not explicit as to what counts as "obvious", one can imagine that the term refers to cases such as the use of permission *can* in interrogative sentences. In light of the current study, an approach that relies on the analyst's subjective judgement to identify meaningful relations between modal forms and syntactic components is unreliable.

With specific regard to *may* and *can*, an interesting aspect of Perkins' study concerns the two related notions (according to Perkins) that (i) both forms share the same core meaning, and (ii) that *may* and *can*, as modals, "most certainly" (p. 104) combine with

[11] With the term *context*, Perkins refers to both the linguistic environment of an expression as well as its non-linguistic environment, that is, its pragmatic environment.

their clausal syntax. Generally, according to Perkins, both *may* and *can* are "contextually determined formal variants which realise the same core meaning" (p. 41). In the case of *can*, for instance, Perkins notes that "postulating an invariant core meaning that can interact with one or more of the three different systems of laws according to its context of use" (p. 35) allows the analyst to show that "many of the problems connected with giving an adequate semantic definition of *can* (such as its alleged polysemy and semantic indeterminacy) may be plausibly regarded as contextual" (p. 35). At a more general level, Perkins finds that formally, the modals verbs are the least explicit of all modal expressions in that

> they convey no more information than there exists a certain relationship between the truth of some proposition *p* or the occurrence of some event *e* and some circumstance *c* relative to some set of principles *k* (…). All they specify is the nature of the relationship between *c* and *x* (where *x* represents either the truth of *p* or the occurrence of *e*) without including any direct information about the actual identity of *c* or *x* (Perkins 1983: 104)

In a final point, Perkins adds that "their [English modals'] lack of markedness is most certainly bound up with the fact that of all modal expressions they are the most fully integrated within the structure of the clause" (p. 104, my emphasis). Perkins does not dwell on the notion of integration, nor does he make explicitly clear what this notion actually involves or what specific implications it could have for the modals. One can infer, however, that he means that the grammatical structure of the linguistic context of the modals plays a role in the underlying system of the forms. While, given the nature and the scope of his study, Perkins is not in a position to assess this level of integration or support his statement with empirical data, his study calls for a quantitative investigation into whether modal verbs can clearly be distinguished and subsequently individually characterised on the basis of their grammatical behaviour.

In summary, this section has shown, from a theoretical perspective, the potentially crucial role of the linguistic context in the uses of the English modal verbs. Supporting this view, Herméren 1978 and Perkins 1983 complement each other and in the case of Herméren's study, there is an emphasis on the wide range of grammatical features interfering with the use of modal verbs across grammatical levels. In the case of Perkins' study, the author takes a slightly broader outlook by saying that clausal structures play a part in the semantics of the modals. In what follows, I draw from three quantitative studies of the English modals, Coates 1983, Gabrielatos and Sarmento 2006 and Collins 2009, and I identify, for the purpose of the present study, ways to integrate a fine-grained approach characteristic of the above-mentioned descriptive studies into a large scale corpus-based investigation of *may* and *can*.

2.4 Quantitative studies

2.4.1 Coates (1983)

In the words of Collins (2009: 6), "Coates (1983) remains the most detailed and widely referred to corpus-based study of the English modals". Generally, Coates investigates the semantics of the modals and she particularly focuses on indeterminacy as a characteristic semantic feature of the modals.[12] Coates recognises that a semantic investigation of the modals requires to establish connections between semantic categories (e.g. epistemic uses) and syntactic categories (e.g. negation, passive, grammatical subject person) and, crucially, her analysis takes such associations into account in an unprecedented quantitative fashion. She notes, for instance, that "[t]hese associations can be quantified by using the computer to scan the coding of each example (…), to check the co-occurrence of any given syntactic feature" (p. 37). So overall, Coates provides quantitative evidence that the modals' syntactic environments provide useful information for their semantic investigation. Methodologically, Coates' study proceeds in two steps: first, she identifies a network of meanings and forms, and second, she organises the modal forms into semantic clusters on the basis of corpus data.[13] [14] More specifically, Coates' cluster approach to the semantics of the modals, consists of an

[12] Coates (1983) identifies three types of indeterminacy: *gradience*, *ambiguity* and *merger. Gradience* refers to "a continuum of meaning" in which possible uses of a modal form shade into each other. *Ambiguity* is concerned with cases where "it is not possible to decide which of two meanings is intended" (p. 15) and *merger* refers to cases where although two uses are different, they are not mutually exclusive.

[13] Coates' (1983) clusters were experimentally validated by Coates herself. She ran two informant tests using a card sorting method.

[14] Coates (1983) uses a 545,000-word corpus that includes both written and spoken material. For the written data, she uses the Lancaster corpus which includes a variety of genres. The written data also include unprinted material such as private letters and diaries which are extracted from the corpus of the Survey of English usage. The spoken data, was also extracted from the corpus of the Survey of English usage and includes both private and public discourse such as private conversations between friends and radio discussion programmes and sport commentaries.

investigation of the interactions of the modals' meanings based on the identification of semantic groups. In other words, Coates (1983) recognises that the English modals share a number of senses amongst themselves and, while investigating cases of indeterminacy, she organises the modals into clusters reflecting their underlying semantic structures. Generally, the clusters are indicative of the uses that are shared by different modals and the inclusion of individual modals as members of particular clusters of uses is based on their frequency of use. Interestingly, unlike existing studies contemporary to her own (see Coates 1980 for details of the studies), Coates makes a clear distinction between the uses of *may* and *can* and she segregates the two modals by placing their respective semantic concepts into two different sets, as can be seen in Table 2 below which summarises Coates' clusters.

Table 2 Modal clusters according to Coates (1983)

Clusters	Semantic concepts	Modals
1	obligation, necessity	*must*, *need*, *should, ought*
2	intention, prediction, futurity	*will, shall*
3	possibility, ability, permission	*can, could*
4	epistemic possibility	*may, might*

This clustering phenomenon is not surprising however, since:

i. the most frequent use of *can* is that of possibility (Leech 2004);
ii. its least frequent use is that of epistemic possibility (Ehrman 1966); and
iii. *May* behaves in the reverse way.

With regard to *may* and *can*, Coates' (1983) cluster results confirm those of an earlier study, Coates (1980), on the non-equivalence of *may* and *can* and where she concludes 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, and cannot be used to express epistemic possibility".[15] She further adds that

[15] For Coates (1980, 1983), *root possibility* refers to non-epistemic possibility and includes both deontic and dynamic uses.

"[w]hile it is true, as linguistic theoreticians have observed that *may* and *can* can both express root possibility and permission, (…) they are not in free variation" (p. 219).

According to Coates, the semantic distinctions between the uses of *may* and *can* can be seen through their respective syntactic co-occurrence patterns. Her quantitative analysis of *may* and *can* in relation to their syntactic environments shows that in the case of possibility/ability/permission *can*, for instance, stative verbs, passive voice and inanimate subjects are characteristic of Root *can*. In the case of epistemic *may*, her data yield strong associations between syntactic and semantic features. Table 3 below summarises and illustrates Coates' modal-syntactic context associations.

Table 3 Summary of the syntactic co-occurrence patterns of epistemic *may*

Syntactic pattern	Example
Perfective aspect	*I may have put it there out of the way*
Progressive aspect	*They may be reading something by Shakespeare*
Existential subject	*January I suppose there may be an interview*
Quasi modal	*I may be able to leave here and still owe them my notice*
Stative verb	*I think he may be a very violent man*
negation	*They say he may never work again because he's got a schizophrenia*

Coates' study represents a step forward in the semantic investigation of the English modals as it successfully combines, in a systematic fashion, the semantic and the syntactic linguistic levels of analysis and thereby provides empirical support towards the view that the semantics of the modals reaches beyond modal forms *per se*. Two shortcomings, however, emerge from Coates' study: first, that in order to fully validate Hermerén's claim, her study would have to include additional linguistic levels, namely morphology and semantics, in order to check whether within their clustered groups, the modals can be characterised by particular morphological or semantic environments. The second shortcoming in Coates' study is methodological in nature. It is concerned with her decision to dissociate the initial semantic clustering from the subsequent form-syntax analysis. One may question, at this point that should the syntactic co-occurring patterns of the modals be influential in the uses of the modal forms, then it would be necessary to include the syntactic components at the initial stage of the analysis rather than integrate them after completion of the clustering process. This is an important point because although Coates' approach provides a more reliable descriptive account of the syntactic environment of the modals than previous descriptive studies, crucially, her semantic analysis is initially biased as it does not account for the modals' potential sensitivity to other linguistic levels.

2.4.2 Gabrielatos and Sarmento (2006)

Gabrielatos and Sarmento (2006) is a corpus-based study that attempts to account for the syntactic contextual information using a quantitative approach to investigate core modals in native English (i.e., *can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will* and *would*). It involves a comparative analysis of the frequencies of uses of the modals in an aviation corpus and a representative corpus of American English and, generally, it raises the following questions:[16]

- 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?

In line with Hermerén's and Coates' accounts, the authors acknowledge that the modals' distribution varies according to their syntactic contexts.[17] Voice, in that regard, is reported to potentially correlate with the uses of the modals. One particular grammatical category identified as potentially causing such variation is voice. In that regard, the authors stress the need to engage in more detailed investigation of the uses of the modals with passive and active voice infinitives. Beyond the study's syntactic considerations,

[16] Gabrielatos and Sarmento's 2006 study investigates The Aviation Corpus (AC) which includes three Boeing 737 manuals: a Maintenance Manual, a Quick Reference Handbook and an Operations Manual.

[17] The selected syntactic contexts for Gabrielatos and Sarmento's 2006 study include: 1) modal alone, 2) modal +infinitive (present infinitive, active voice), 3) modal+*be*+present participle (present infinitive, passive voice), 4) modal + *be*+present participle, 5) modal+*have*+past participle, 6) modal+*be*+*being*+past participle, 7) modal+*have*+*been*+past participle, 8) modal+*have*+*been*+present participle, 9) modal+*have*+*been*+*being*+past participle (or adjective).

the authors suggest the need to extend the scope of similar studies to the semantic domain: "[f]urther studies should also focus (…) on the collocation patterns of modal expressions in the corpus, such as verb collocations of central modals" (p. 236-237). Despite these insights, Gabrielatos and Sarmento's study has two shortcomings. The first is the small size of the investigated data set. While, to agree with the authors, their study constitutes "a starting point" for the quantitative investigation of the modals' distributional variations in aviation corpora, the reliability of their results needs to be validated through the analysis of more data. Secondly, despite their interesting descriptive account, the authors' largely form-based monofactorial approach prevents them from characterising the uses of the modals in a way that would allow for classification and prediction. One possible way to consider further studies of the type of Gabrielatos and Sarmento (2006) is to not only investigate, independently of each other, the syntactic and the semantic factors potentially causing the modals to behave differently, but also to investigate whether (and if so, to what degree) the interaction of the two affects the uses of the forms in any way. To return to the interaction of modals with voice and lexical verbs, it is conceivable that certain modal forms prefer lexical verbs denoting, say an abstract process, but only do so in cases where a passive voice is used. Ultimately, this kind of multifactorial approach is needed in order to move away from the traditional descriptive outlook and equip the analyst for finer-grained analyses.

2.4.3 Collins (2009)

To date, Collins (2009) presents the most comprehensive quantitative study of the modals (and quasi-modals) and it investigates the modals *can*, *could*, *may, might*, *must*, *need*, *ought to*, *shall*, *should*, *will* and *would*. Collins' study is based on the analysis of a 1,2 million-word corpus of spoken and written English and investigates the meanings of the modals in three parallel corpora of contemporary English, namely the British and the Australian subsections of the *International Corpus of English* and an especially assembled corpus of American English. Spoken data consist of monologues and dialogues. Written data include non-printed material such as student writing and letters and printed material falling under the following categories: academic, popular, reportage, instructional, persuasive and creative. The American sub-corpus used in Collins (2009) consists of the spoken *Santa Barbara Corpus* and a selection of texts extracted from the *Freiburg-Brown Corpus of Written American English*.

With regard to *may* and *can* specifically, Collins' study is based on 9,924 occurrences of the two forms, that is 7,663 occurrences of *can* and 2,261 occurrences of *may*.

Collins' overall methodological approach to modals is form-based, and his study is limited to the distributional differences of the forms and their senses across the three sub-corpora. So unlike Gabrielatos and Sarmento (2006), Collins is not concerned with the possible correlations between the senses of modal verbs and syntactic variables such as voice. In fact, Collins' quantitative account excludes the notion that syntactic components interfere with the senses of the modals. For coding purposes, Collins adopts a traditional tripartite semantic taxonomy including epistemic, deontic and dynamic senses (as defined in Section 2.2.2). In addition, he uses an 'indeterminate' tag for cases where *may* and *can* overlap semantically. To enrich his quantitative analysis, Collins presents qualitative insights on the uses of the modals in relation to the concepts of modal strength, modal degrees and subjectivity/objectivity. For instance, in the case of epistemic *may*, Collins (2009: 93) refers to Verstraete (2001) who "claims that epistemic modality cannot be objective, invoking as an argument the resistance of epistemic modals in interrogatives". Collins shows that "instances of objective *may* do occur, where the estimation is one that is entertained more generally" (p. 93). For Collins, objective modality with *may* can be seen, for example, in cases such as (26) below where "the impersonal extraposition with *it's thought* as matrix clause (…) indicate[s] that the judgment is not limited to the speaker but rather on public record, as it were" (p. 93).

(26) It's thought that the man *may* have committed suicide

Interestingly, (26) shows that while Collins excludes the grammatical contexts of the modals from his quantitative analysis, he includes them in his qualitative input. Given the informative nature of his qualitative insights throughout his analysis, one may ask why contextual features such as voice or clause type were not initially integrated into the main quantitative analysis. This question is reinforced especially by the fact that previous studies such as Hermerén (1978), Coates (1983) and Gabrielatos and Sarmento (2006) have not only demonstrated the relevance of this line of research for the modals but, in the case of Coates and Gabrielatos and Sarmento, they have also provided empirical support towards the need to carry out grammatically-grounded quantitative analyses.

Collins' overall statistical approach to his corpus data adds to his inability to combine a quantitative analysis with one that is grammatically-grounded. A major limitation in Collins' statistical approach is his overall assumption that a high frequency necessarily yields a linguistically interesting phenomenon. Collins' statistical account solely relies on comparing raw as well as normalised frequencies of the modals' occurrences. With regard to *may* and *can*, for instance, Collins compares the occurrences of both forms in the three sub-corpora and in different genres (i.e., in dialogues and monologues for the spoken data and in printed and non-printed texts for the written data). A third type of comparison involves the senses of *may* and *can*: Collins compares the frequencies of *may* in each of its senses of occurrence (i.e., epistemic, deontic and dynamic) across all three varieties of native English and the same for *can*. Despite an extensive catalogue of frequencies, Collins' results remain hard to interpret and the exact implications of his findings are often unclear, which is in no small part due to the fact that his study does not provide any statistical analysis of the frequencies it presents. For instance, the study does not specify whether the observed distributional differences between the senses of *may* and those of *can* across the three sub-corpora are statistically significant, and if so, to what degree they are significant. This is an important shortcoming because ultimately, it leads one to question the (degree of) comparability of Collins' frequencies throughout the corpus. I illustrate this point below on the basis of Collins' frequency table of the meanings of *may* across the three sub-corpora. In Table 4 raw frequencies are bracketed and tokens per one million words are unbracketed.

Table 4 Collins' (2009) frequency table of the meanings of *may*

	ICE-AUS	ICE-GB	C-US	Total	%
Epistemic	651	1023	636 (125)	2310 (1799)	79
Deontic	78	70	56 (11)	204 (159)	7
Dynamic	101	60	76 (15)	237 (176)	8,1
Indeterminate	51	65	56 (11)	172 (127)	5,9
Total	881	1218	825 (162)	2924 (2261)	100

Although Collins uses the above frequencies to show that deontic *may* is the least "common" sense of the three as it is chosen 7% of the time as compared to epistemic *may* (79%) and dynamic *may* (8%), he does not show that the low frequency of deontic *may* is significantly different from the low frequency of dynamic *may*. In fact, the number of indeterminate cases is high enough to allow for the possibility that deontic senses may end up outnumbering dynamic ones. My 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 (χ^2=42.68; df=4; p<0.001). One question emerging from that analysis is that since the observed frequency of the senses of *may* are not randomly distributed, then what motivates the different uses of each form in each independent corpus?

It is becoming clear that Collins' overall statistical approach prevents him from identifying areas of research on the modals that would ultimately offer new perspectives for the study of their semantics. His statistical approach therefore falls short of even matching previous work by Hermerén and Coates: a large sample size alone is no guarantee for new(er) insightful results, especially since investigating quantitatively the linguistic factors that motivate the uses of different modals involves adopting multifactorial statistical methods, which Collins does not do. Indeed, to consider the linguistic context of a lexical item involves first identifying grammatical components that are likely to interact with the senses of the modals and then quantifying the degree of interaction of those components with the forms and in relation to their senses. One advantage of this type of methodological approach is that it would allow the analyst to further the work of Gabrielatos and Sarmento (2006) in two ways. First, it would determine whether or not the distribution of modals' senses vary as a function of their syntactic environment on the basis of (i) a significantly larger data set and (ii) a wider variety of syntactic environments that would include sentence type, negation or/and clause type. Second, a multifactorial statistical approach would also allow the analyst to investigate whether the distribution of the senses of the modals vary as a function of their semantic and morphological environments. Finally, a multifactorial treatment of the modals would provide an unprecedented opportunity to empirically validate Hermerén's claim that grammatical categories play an influential role in the semantics of the modals.

In sum, despite of the large amount of frequency provided by Collins (2009), he has not advanced corpus-based and quantitative research on the English modals. Although studies such as Gabrielatos and Sarmento (2006) are limited with regard to the size of their corpus and the narrow range of English varieties that they investigate, they nevertheless tested Herméren's grammatical approach to the semantics of modal verbs. In contrast, while Collins's analysis is based on the largest data set ever used to investigate modals, due to his form-based quantitative approach of the modals, Collins' data remain unexplored.

2.5 Concluding remarks

In this chapter, I have shown with Hermerén's study that modal verbs should be investigated in a fine-grained fashion and it is crucial for the analysis of modal verbs to account for grammatical components at the semantic, syntactic and morphological linguistic levels. A small number of quantitative linguists, such as Coates and Gabrielatos and Sarmento have tried to include linguistic context in their analyses, and have thereby provided empirical evidence supporting the usefulness of a systematic investigation of the modals' syntactic contexts. However, to date, other additional linguistic levels (e.g., semantics, morphology) have not been integrated simultaneously in a single study. Furthermore, the total number of investigated features remains small.

In addition to discussing the appropriate degree of granularity for investigating the modals, I have also highlighted in this chapter the importance of adopting statistical methods that are powerful enough to cope with the degree of complexity that a multifactorial analysis of the modals requires. In that regard, the discussion of Collins 2009 has shown that in order to be meaningful, a quantitative analysis of the modals needs to provide more than frequencies of occurrence. In fact, this type of investigation requires researchers to select statistical tools that allow them to explore in depth their corpus data and provide meaningful interpretations of their findings. In Table 5 I summarise the findings on the modals of the main studies and their desiderata that have influenced the current work.

Table 5 Overview of the main studies, findings and desiderata guiding the current work in relation to the modals

Studies on the modals in native English	Main findings and desiderata
Qualitative studies	
Hermerén (1978)	**Finding**: the semantic import of a modal emerges from both the lexical item itself **and** the grammatical context in which it occurs. **Desideratum**: although Hermerén's study is based on corpus examples, it is not quantitative.
Perkins (1983)	**Finding**: "of all modal expressions […] [the modal verbs] are the most fully integrated within the structure of the clause" (p. 104). **Desideratum**: Perkins' study is neither empirically- nor grammatically-grounded.
Quantitative studies	
Coates (1983)	**Findings**: - Coates brings quantitative evidence that the modals' syntactic environments provide useful information for their semantic investigation; - the distinction between *may* and *can* can be seen in their respective syntactic-co-occurrence patterns. **Desideratum**: Coates' quantitative assessment of the influence of syntactic contexts on modal verbs is not included at the initial stage of the modal forms' semantic analysis.
Gabrielatos and Sarmento (2006)	**Finding**: the modals' distribution varies according to their syntactic contexts. **Desideratum**: Gabrielatos and Sarmento's study does not include semantic nor morphological contexts.
Collins (2009)	**Finding**: despite a large data set, the interpretation and the implications of Collins' results are hard to draw due to his limited statistical approach. **Desideratum**: Collins' study does not include multifactorial analyses and is not grounded in any theoretical framework.

Chapter 3 Interlanguage and second language acquisition

3.1 Introduction

Learner language has captured the interest of many scholars and pedagogues since approximately the 1940s (cf. Fries 1945, Lado 1957), particularly cross-linguistic influences between speakers' native language and their second/foreign language. However, it is not until the late 1960s that learner language is recognised and investigated as a linguistic system in its own right:

> An *interlanguage* may be linguistically described using as data the observable output resulting from a speaker's attempt to produce a foreign norm, i.e., both his errors and nonerrors. **It is assumed that such behavior is highly structured**. In comprehensive language transfer work, it seems to me that recognition of the existence of an interlanguage cannot be avoided and that it must be dealt with as a system, not as an isolated collection of errors (Selinker 1969: 71) [my emphasis]

In other words, Selinker (1969), assumes that variability between native language varieties and their matching interlanguage systems is not accidental but is in fact principled in nature. Since the recognition of interlanguage varieties (e.g. French-English IL, Spanish-French IL) as systems in their own right, many studies have attempted to collect linguistic evidence in support of the systematic nature of interlanguage varieties (cf. Bartning 2009 for reviews of recent studies in the domains of syntax and morphology). Overall, interlanguage varieties have been investigated according to the different stages of the second language acquisition process (i.e. beginner, intermediate, advanced) and generally, the body of literature concerned with the acquisitional process of interlanguage varieties suggests that individual developmental stages highlight different facets of interlanguage systems. Observed variability between the language produced by adult second language learners and that of native speakers has led to large collections of studies within both the linguistic and the psycholinguistics traditions and this has led analysts to approach learner language from two perspectives. In linguistically-oriented studies, interlanguage varieties have been investigated as linguistic products which provide evidence of the interaction between two or more linguistic systems (Bialystock and Sharwood Smith 1985) (see. Bialystock and Sharwood Smith 1985).[18] Psycholinguistically-oriented studies, on the other hand, have been concerned with the acquisition of interlanguage systems and particularly the psychological processes that underlie organisational principles of language (Gass 1996: 321). In that regard, Jordens (1997: 291) notes that "[w]ithin this tradition of second language research, it is assumed that L2 learners process L2 data on the basis of language learning mechanisms which are part of the human language faculty". Generally, from a psycholinguistic perspective, it is more as a *system* than as a *product* that interlanguage has triggered most interest (cf. Bialystock and Sharwood Smith 1985). While it is now currently accepted that cross-linguistic influences can be traced both at linguistic and cognitive levels in the sense that "L1 [native language] influences occur not only as *direct* linguistic reflexes, but they also indirectly reflect underlying organisational principles of language" (Gass 1996: 321), the intriguing aspect of interlanguage varieties is how the linguistic and the cognitive are combined during the process of second language production. It is reasonable to believe that interlanguage varieties can be characterised by particular combinations of both linguistic patterns and cognitive behavioural tendencies. In that respect, and to follow Ellis, interlanguage varieties are dynamic systems whose variability is inevitably on-going due to the developmental nature of the second language acquisition process:

> [i]nterlanguage constitutes an unstable system and as such is permeable to invasion by new linguistic forms. Its dynamic quality is reflected in tremendous **variability** in language-learner language and also in overlapping stages of development as one set of variable rules is revised in favour of another. (Ellis 1985: 118)

In what follows, I define the nature of interlanguage varieties and discuss problematic issues related to interlanguage and its corpus-based investigation. I proceed in four steps. First, I use two studies, namely Selinker (1972) and Adjemian (1976), to illustrate the psycholinguistic and the linguistic facets of interlanguage systems. Second, I discuss existing methodological issues for the investigation of learner language. In that regard, I

[18] Bialystok and Sharwood Smith (1985:101) write: "[i]n one sense, like the word 'language' itself, IL [interlanguage] denotes a product: it is the outcome of language use".

particularly focus on Jarvis (2000) to show the usefulness of a statistically-grounded approach for a corpus-based investigation of learner language. Third, I turn to existing second language corpus work and present how the English modals have so far been investigated on the basis of advanced learner language corpus data. I then show how such studies can be improved by adopting a cognitive usage-based theoretical framework. The chapter concludes with a summary section.

3.2 On the nature of interlanguage

Selinker (1972: 214) recognises that the sentences produced by *most* learners of a second language are "not identical to the hypothesized corresponding set of utterances which would have been produced by an active speaker of the TL [target language] had he attempted to express the same meaning as the learner".[19] Furthermore, he claims that this lack of correspondence between native and non-native sentence patterns is psychologically motivated. Selinker's general approach to interlanguage systems is based on the core assumption that adult second-language learners are biologically endowed with a *latent psychological structure* and that they activate that structure "whenever they attempt to express meanings, which they may already have, in a language which they are in a process of learning" (p. 212).[20] Within this theoretical approach, Selinker investigates second language learning by identifying and isolating psychologically relevant second-language data (p. 211). Generally, Selinker (1972: 213) makes two assumptions, namely that (i) while producing second-language sentences, the learner focuses his/her attention "upon one norm of the language whose sentences he[/she] is attempting to produce" and (ii) the TL that the learner is attempting to learn refers to "one norm of one dialect within the interlanguage focus of attention of the learner". For Selinker (1972), although learners activate their latent psychological structure when attempting to produce sentences in the TL, the sentence patterns that they ultimately produce reveal their failure to achieve native-speaker competence.[21] On

[19] Selinker (1972) is only concerned with “adult” learners. The term *adult* refers to learners who are over the age of 12. Selinker's understanding of the term is based on Lenneberg's (1967) notion that a "critical" period of brain maturation is passed after the onset of puberty and, beyond that point, language development is suspended.

[20] To refer to such situations, Selinker uses the term *meaningful performance situations*.

[21] Selinker (1972) identifies two types of learners, namely the “successful” learner who achieves native-speaker competence and the “unsuccessful” learner or the "ideal second language learner" (p. 213) who "fail[s] to achieve native-speaker competence and who is "representative of the vast majority of

that basis the TL sentences produced by the learners provide observable data from which theoretical predictions of behavioural events can be made. In that regard, Selinker writes:

> predictions of behavioral events in a theory of second language learning should be primarily concerned with the linguistic shapes of the utterances produced in ILs. Successful predictions of such behavioral events in meaningful performance situations will add credence to the theoretical constructs related to the latent psychological structure. (Selinker 1972: 214)

Predicting behavioural events involves a process of *interlingual identification*, which consists of two main steps: first, the gathering of observable data from meaningful performance situations and, second, the analysis of "the psycholinguistic processes which establish the knowledge which underlies IL behavior" (p. 214). For the data gathering process, Selinker stresses the need to gather three sets of utterances within one theoretical framework, namely a) "utterances in the learner's native language (NL) produced by the learner", b) "IL utterances produced by the learner", and c) "TL utterances produced by native speakers of that TL" (p. 214). Selinker identifies five psychological processes that are central to the production of second-language sentences and which together affect the shape of IL utterances. Among them is the process of *language transfer* (Selinker 1969) which refers to "a process occurring from the native to the foreign language" (Selinker 1972: 90).[22] For Selinker (1972), one can investigate psycholinguistic processes of language transfer on the basis of learners' linguistic choices in situations where two alternative choices exist. For Selinker, cases of linguistic transfer in relation to binary linguistic choices are identified "if frequency analysis shows that a statistically significant trend in the speaker's native language appears towards one of those two alternatives, which is then paralleled by a significant trend toward the same alternative in the speaker's interlanguage behavior" (p. 90). On the basis of previous syntactic studies on word order in Hebrew-English interlanguage, Selinker (1969) starts from the working assumption that "transfer of the structural

second-language learners" (p. 213). For the purpose of his study, Selinker focuses on the latter type of learners at a point of "attempted learning", that is regardless of their ultimate success (or lack of it) in achieving native speaker competence.

[22] In addition to *language transfer*, Selinker (1972) identifies the following cognitive processes as core processes in the process of second language production: *transfer-of-training*, *strategies of second-language learning*, *strategies of second-language communication* and *over-generalisation*.

patterns of Hebrew into English occurs at all levels of the hierarchy that linguists might isolate" (p. 70) and he investigates syntactic patterns of L1 Hebrew into L2 English.[23] Selinker's study (1969) involves eliciting specific types of sentence from 132 Israeli and 31 American schoolchildren in seven experiments. The material for the experiments consists of an interview containing approximately 50 questions each in Hebrew and English. All interview questions are organised around semantic areas matching the participants' experience and interest. The purpose of the interview is to elicit sentences which, in turn, are used to compile questionnaires for the experimental part of the study. Overall, on the basis of the interview, Selinker identifies four types of syntactic strings that follow the particular pattern where one verb is followed by one these combinations below (I illustrate each string using Selinker's examples):[24]

- Object string (Ob)[25] + Time string (Ti) → I see *him* (Ob) / *a year ago* (Ti)
- Object string (Ob) + Place string (Pl) → I saw *him* (Ob) / *in his apartment*(Pl)
- Object string (Ob) + Adverb string (Ad) → I like *English and geography* (Ob)/ *best* (Ad)
- Pl + Ti → I live in *Forest Park Apartments* (Pl) / *now* (Ti)[26]

On the basis of the data gathered in his experiments, Selinker identifies "a [statistically] significant Hebrew norm of syntactic string behavior" (p. 86) for Ob-Ti, Ob_s-Pl, Ad-Ob and Ob_s-X. He also found statistically significant parallel interlanguage trends in each of these four cases. These findings lead Selinker to conclude that cases of transfer can be identified "when parallel nonchance arrangements (…) result from the statistical operations performed" (p. 86). Generally, Selinker's approach to assessing cross-

[23] Selinker (1969) notes that native speakers of Hebrew commonly make syntactic mistakes such as *I like very much cats* and that such mistakes are influenced by the Hebrew pattern of the type *ani ohev meod xatulim*.

[24] Selinker (1969) stresses that the four syntactic strings are *combinations* in that their members can occur in either arrangement or order. This means that, for instance, the string Object string (Ob) + Time string (Ti)' may equally occur under this form: Ob-Ti or under this form Ti-Ob. In Selinker's parlance, Ob-Ti and Ti-Ob are *arrangement* and the hyphen indicates the fixed order of occurrence of the syntactic members.

[25] To avoid cases of anticipated parallels between the two strings Ob-Ti and Ob-Pl, two different types of Ob were distinguished later on in the analysis: Ob_n cases were identified where the object is a noun and Ob_s in cases of substitute objects.

[26] Selinker (1969: 84) notes that "[s]plitting the data another way produced a fourth absolute and a sixth English norm, Ob_s-X, with 76 occurrences and no counterexamples".

linguistic transfer phenomena is attractive in that it is comprehensive and provides "an operational definition of language transfer in terms of any native and foreign language situation, no matter what linguistic level is identified and isolated" (Selinker 1969: 89). Furthermore, Selinker's statistically-based approach allows one to evaluate the data in a way that is systematic and objective and that ultimately allows the researcher to explain and predict learners' binary linguistic choices.

However, despite these attractive characteristics, Selinker's approach has two major limitations. First, it relies on the existence of direct translational correspondences of linguistic items in NL and TL. As Selinker himself puts it: "[a] preliminary step […] is for the descriptive analyst to judge that he[/she] is facing a situation in which two alternate choices exist for the speaker **in each of the two languages**" (Selinker 1969: 89, my emphasis). From this perspective, the applicability of Selinker's approach is limited as the chances of finding cross-linguistic binary sets of forms/structures to express similar meanings are not high. It is easy to imagine, for instance, that language *X* would have a single form to express meanings *a* and *b* and language *Y* would have two distinct forms to express the meanings *a* and *b*. In such cases, while parallel comparisons of frequencies in IL and TL are still possible, it is difficult to envisage how transfer from the NL could be assessed. A second major shortcoming lies in the fact that while Selinker (indirectly) recognises that language transfer phenomena are likely to occur at all linguistic levels, he ignores the fact that during the language production process speakers deal with all linguistic levels simultaneously. This is important because while Selinker offers one way to assess transfer at one specific linguistic level, he does not show how to do that in a way that is psychologically realistic.

A second study that approaches interlanguage varieties as systems in their own right is Adjemian 1976. For Adjemian, "the differences between learner speech forms and the corresponding acceptable TL forms cannot be always explained by transfer" (p. 297). On the basis of existing studies, Adjemian argues that a linguistically-based investigation of IL is more appropriate than one that is cognitively-based. While he recognises that learning strategies and linguistic rules both contribute to the characterisation of IL, he stresses that, contrary to the direct influence of linguistic rules, learning strategies are only indirectly involved:

> [o]ne of the consequences of the position supported here is that the researcher must gather data that will yield a well-supported linguistic analysis before tackling the problems of studying learning strategies. Within the other framework both goals are pursued simultaneously, which can lead to generalizations that may not be totally supported by a broader analysis (Adjemian 1976: 304).

For Adjemian, the crucial difference between native and non-native language varieties is the *permeability* of interlanguage grammars. According to Adjemian, interlanguage grammars (unlike native grammars) are interim grammars that, by their very nature, are not fixed. Instead, they develop and change. This adaptable structure of IL systems allows cross-linguistic transfer to occur. In other words, the linguistic structure of IL is such that learners can apply specific linguistic rules in the L2 in linguistic contexts where such rules would not normally be applied. Adjemian further claims that the permeability of IL plays a major role during the communication process because at that stage the learner encounters problematic structures that he/she tries to overcome by means of simplification and streamlining strategies. In Adjemian's words:

> [i]t is at this point that the permeability of the IL will permit violation of its internal systematicity by the use of production, communicative, or other strategies to "improperly" generalize, simplify or otherwise modify a linguistic function of the IL (Adjemian 1976: 309)

Thus variability between IL and TL can be most clearly observed in communicative situations. In these situations, learners produce forms and structures that are most typical of IL. While such structures are produced spontaneously, their structure may differ from the forms in both the learner's NL and the TL (p. 299). This suggests that grammatical structures in interlanguage varieties are functionally motivated. In other words, it is through their use of the TL language and their on-line negotiation of form-function mappings that learners carve the backbone of interlanguage grammars. While more recent studies such as Perdue 2000 and Bartning 2009 have recognised the difficulty with which learners discover form/function correspondences in the TL, it is conceivable that the "dynamic quality" (p. 118) of interlanguage varieties (to use Ellis' 1985 term) lies in the on-line 'creation' of innovative form-function mappings.[27]

While the implications of Adjemian's study are potentially enlightening, the author's methodological approach remains conservative by prescribing the use of both error analysis and contrastive analysis. Generally, Adjemian neither supports the notion that NL grammars are fixed nor does he show how, methodologically, the permeability of IL can be best investigated and demonstrated: "[t]he data base used for this linguistic analysis must therefore be broad and yet specifically oriented: geared toward one linguistic function or set of related functions, but gathered from differing contexts and by a variety of methods" (p. 300). Regrettably, Adjemian does not make explicit (i) whether variability between IL and TL is to be assessed qualitatively or quantitatively, (ii) whether (and if so, how) to integrate the different grammatical levels into a single analysis, and (iii) whether (and if so, how) to account for the interaction between the different grammatical levels.

As an interim summary, based on this analysis of Selinker's (1969, 1972) and Adjemian's (1976) studies, it is clear that an effective investigation of learner language should combine

i. a rigorous quantitative approach that explains and predicts learners' lexical and/or grammatical choices during the (second) language production process,

ii. large-scale linguistic data evidence to identify linguistic patterns characteristic of learners' IL,

[27] Perdue writes that:

> The communication limitations of the variety, in a way, push the learner to further acquisition. The process is constrained, however, by the present functioning of the variety. In particular, new and targetlike forms that a learner acquires are not necessarily used in targetlike ways (with targetlike functions); mastering the form-function correspondences of the target represents a specific and complex learning task (Perdue 2000: 301)

Bartning writes that:

> [T]he advanced learners (…) have difficulty in systematising functional morphology. It is evident that these difficulties do not only concern the automatisation of forms but also the discovery of correspondences between complex notions and their respective forms (Bartning 2009: 19)

iii. linguistic accounts that reach beyond the description of learner language by investigating the factors potentially interfering with learners' lexical choices.

While such a tripartite approach raises a number of complex methodological issues that few studies have yet addressed, the current study takes first steps to implement such an approach. I present and discuss some of the complexities of this enterprise.

3.3 Methodological considerations

It is not uncontroversial to rely solely on quantitative methods to investigate variability in synchronic learner language data. Studies such as Bialystock and Sharwood Smith (1985: 110), for instance, have claimed that synchronic variability in learner language requires qualitative treatment.[28] Jarvis (2000), on the other hand, has convincingly illustrated how learner language studies can benefit from rigorous statistical approaches. In that regard, Odlin notes that the methods exemplified by Selinker and Jarvis, amongst others, have shown that

> it is possible to subject claims about cross-linguistic influence to rigorous tests [and that] such testing has often indicated language transfer to be at work, and the reality of the phenomenon is undeniable even though much remains to be understood (Odlin 2005: 448)

Generally, such studies tend to be experimental in nature, and they usually do not address both issues of grammatical variation and of form/function mappings in learner language. As he explores the effect of learners' native language in their lexical choices, Jarvis (2000) argues for the adoption of a unified and rigorous methodological framework for the study of L1 influence on L2. According to Jarvis, establishing such a framework provides a way to address "the existing confusion over L1 influence" (p. 248) while overcoming "inconsistencies or incompatibilities between the empirical methodologies of different transfer studies" (p. 249). Similarly to Selinker (1972), Jarvis' study is concerned with the influence of learners' native language in their lexical

[28] Bialystock and Sharwood Smith (1985: 110) write: "A second kind of variability is that a learner's speech shows variability *synchronically*, i.e., at a particular point in time. Certain linguistic forms, for example, may be used one way in some situations and another way in others. The explanation for this type of variability requires qualitative models which reflect not the amount of knowledge that the learner has, but the psycholinguistic conditions under which the knowledge may be demonstrated".

choice in IL. However, it is not centred on binary lexical choices (i.e. learners' preference of form X over from Y). Rather it is focused on learners' immediate lexical choices in response to a given denotatum (i.e. which is the lexical item that best suits a given denotatum out of all possible choices). Generally, Jarvis' experimental approach allows him to show that "the learners' word choices in English are more similar to their L1 lexical patterns than to the lexical choices of native English speakers" (p. 288).

While Jarvis' approach is based on the investigation of group tendencies, his framework relies on the identification of three potential group-related effects of L1 influence: (i) intra L1-group similarities, (ii) inter-L1-group differences and (iii) L1-IL performance similarities. Jarvis' study involves 537 Finnish students and 98 speakers of American English, both groups ranging from 11 to 16 years of age. Participants performed three tasks: (i) a written narrative task, (ii) a lexical listing task and (iii) a receptive counterpart of the lexical listing task. Jarvis identifies nine non-linguistic variables that should ideally be controlled for a transfer analysis.[29] This is important as Jarvis' results generally "suggest that learners' referential word choices pattern better according to L1 background than according to other variables" (p. 298). Jarvis' study provides two major outcomes that are directly relevant for the current study. First, it demonstrates some degree of empirical rigour "is not only possible but also essential in this area of research" (p. 298). Second, compared to Selinker's approach, Jarvis provides a broader framework that does not restrict the analyst to cases of lexically and/or grammatically "parallel arrangements":

> L1 influence refers to **any** instance of learner data where a statistically significant correlation (or probability-based relation) is shown to exist between some feature of learners' IL performance and their L1 background (Jarvis 2000: 252) [my emphasis]

Despite its solid methodological approach, Jarvis' study presents one major shortcoming, namely that it is not set up to investigate cross-linguistic transfer at the level of form-function mappings. This is because Jarvis treats participants' lexical

[29] Jarvis (2000: 260-261) identifies the nine following non-linguistic variables: 1) age; 2) personality, motivation and language aptitude; 3) social, educational and cultural background; 4) language background (all previous L1s and L2s); 5) type and amount of target language exposure; 6) target language proficiency; 7) language distance between L1 and TL; 8) task type and area language use; 9) prototypicality and markedness of the linguistic feature.

choices in isolation from their linguistic contexts of occurrence. While a context-based approach such as the *co-occurrence* approach in corpus linguistics, assumes that "the distributional characteristics of the use of an item reveal many of its semantic and functional properties and purposes" (Gries and Otani 2010: 2), this kind of context-based approach would be better-suited to pinpoint traces of "non-nativeness" at the lexical level. This is because the approach would allow the researcher to explain to what degree his/her identified cases of transfer are semantically or functionally relevant. Furthermore, this type of approach reaches further than just the quantitative description of the data, as it ultimately requires the analyst to relate the results of his/her statistical exploration of the data to psycholinguistic theory or usage-based and exemplar-based approaches within Cognitive Linguistics. A clear advantage of this type of approach is that it helps the analyst gain insight into how learners' use of the TL shapes the structure of their interlanguage varieties.

As an interim summary, this section has so far shown that SLA research has mainly used experimental and introspective methods of data investigation (Granger 2000: 5). While this is explained by the "difficulty of controlling the variables that affect learner output in a non-experimental setting", reliance on such methodologies means that conclusions are made "on a relatively narrow empirical base, focusing on the language of a very limited number of subjects, which consequently raises questions about the generalizability of the results" (Granger 2002: 6). Learner corpus research provides a way to combine non-experimental and quantitative approaches to learner language. What is more, a corpus-based approach to learner language allows the researcher to identify the characteristics of particular interlanguage varieties (i.e., the interactions of particular L1s and L2s). As Hanks (2000: 211) notes, "[w]hat a corpus gives us is the opportunity to study traces and patterns of linguistic behaviour". In the next section, I focus on past corpus-based studies of English modals by second and foreign language users from the perspective of Granger's (1996) model for the investigation of learner language. In turn, this allows me assess to what extent such studies have been successful in characterising IL varieties and, crucially, to what extent they have contributed to the wider issue of understanding how learners shape IL grammars.

3.4 Second language corpus work and the English modals

3.4.1 CA and CIA approaches to learner language

Over the past fifteen years, Granger's (1996) Integrated Contrastive Model (ICM) approach has been successfully adopted in learner language studies and, overall, its contribution to learner language research has led to innovative studies which have allowed analysts to characterise individual IL varieties in terms of their general lexical or morphosyntactic behavioural tendencies. More particularly, such studies have allowed analysts to (i) draw general profiles of individual IL varieties, (ii) identify types of IL on the basis of their learner population, and (iii) carry out contrastive studies of various IL varieties.

The purpose of this section is twofold. First, it presents and illustrates Granger's (1996) influential ICM approach and discusses recent EFL/ESL studies that have applied it to the English modals. Second, it shows that such studies still call for methodological improvement although they tend to involve frequency counts of machine-retrievable linguistic items and provide descriptive accounts of the distributional differences of single linguistic items in native and non-native language. However, they do not tend to allow analysts to take an analytical and explanatory outlook towards phenomena of investigation of a semantic or a more conceptual nature such as, as Salkie (2004) suggests, the potential conceptual influence of *pouvoir* in the use of *may* and *can* in French-English interlanguage (I provide the details of Salkie's study further below in Section 3.5).

Granger's (1996) ICM approach aims to explore the phenomenon of cross-linguistic transfer and combines the Contrastive Analysis (CA) approach, which involves the comparison of original data in different native languages and the Contrastive Interlanguage Analysis approach (CIA), which involves the comparison of a native language with a non-native variety of that language. For the purpose of this study, I assume Gilquin's (2008: 5) definition of the term *transfer*, namely "the influence, within an individual's linguistic system, of one or more languages over another". Recent studies, including Gilquin (2008), have recognised the need to combine both the CA and

CIA approaches "for a sound and systematic assessment of the role of transfer in second language acquisition" (p. 3). Generally, the ICM framework assumes that the notion of transfer bridges the two approaches and that a constant "to-ing and fro-ing between CA and CIA data (…) helps analysts to formulate predictions about interlanguage which can be checked against CIA data (Granger 1996: 46).[30] It follows, as Gilquin (2008: 13) notes, that cases of transfer can only be identified as transfer if similarities between the learner's behaviour in interlanguage and in his/her native language can be established.

As expected of a framework specifically designed for the purpose of quantitative corpus-based analysis, the ICM framework assumes that the distribution of formal elements in learner language allows researchers to describe and in turn characterise individual types of interlanguage. From the learning perspective, the ICM framework assumes that the second-language learning process is probabilistic in nature, in that

> [f]requency is an aspect of language of which we have very little intuitive awareness but one that plays a major part in many linguistic applications which require a knowledge of not only what is possible in language but what is likely to occur (Granger 2002: 4)

So while Granger assumes that acquiring a second/foreign language involves gaining probabilistic knowledge of the TL, her approach can also be seen to share Jarvis' (2000) view that distributional differences of formal elements in native and learner language allow the analyst to capture traces of non-nativeness. According to Granger's approach, in order to profile interlanguage varieties, the researcher needs to retrieve linguistic items from the learner corpus data and assess whether those items are over- or under-used in comparison to how they are used by native speakers. For Granger (2002: 132), who defines the process of "bring[ing] out the words, phrases, grammatical items or syntactic structures that are either over- or under-used by learners" as the aim of computer learner research, such frequency assessment serves "many purposes, both theoretical and applied" (Granger 2004: 127). In fact, for Granger, assessing the over- or under-use of formal elements in interlanguage is essential as distributional differences contribute to the "foreign-soundingness [of advanced interlanguage varieties] (…) even

[30] Granger stresses that the term 'prediction' refers to "mere hypotheses which can be confirmed or refuted by corpus investigation" (1996: 46).

in the absence of downright errors" (Granger 2004: 132), which she notes is particularly typical of advanced learner varieties.

Granger and Rayson (1998: 119) present a methodological approach to interlanguage that is applicable to any learner variety and which "demonstrates a potential of automatic profiling for revealing stylistic characteristic of EFL [English as a foreign language] texts". Overall, the approach illustrates how frequency counts contribute to the characterisation of learner language. The authors introduce the *automated profiling technique* which allows to identify salient lexical behaviour through frequency counts of given lexical items. The technique involves (i) selecting a number of word categories (e.g., nouns, adjectives, prepositions, conjunctions), and (ii) computing frequency counts within each word category.[31] [32] Ultimately, the combination of all frequency counts is expected to yield a word category profile of the particular type of interlanguage under investigation. This means that for each investigated word category, the analyst is able to see which members of that category learners under- or over-use. For instance, in investigating articles, Granger and Rayson (1998) find that French learners over-use the English indefinite article *a* but under-use the definite article *the*. According to Granger and Rayson, the benefits of the automated profiling technique are twofold. First, given the previously mentioned permeable quality of interlanguage grammars, the technique helps in achieving a better understanding of learner grammar and lexis, particularly when applied to a wide range of learner corpora. Second, the method "help[s] researchers form a quick picture of the interlanguage of a given learner population" (p. 131). While Granger and Rayson's profiling technique is a useful exploratory method, it is nonetheless (i) a largely descriptive method, since it is based in the investigation of forms, and (ii) a method that does not straightforwardly allow the analyst to investigate IL variability in terms of potential correlations between learners' lexical choices in IL and their processing of the TL grammar. Finally, Granger and Rayson's technique does not allow for the assessment of potential interactions between lexical forms and their co-occurring grammatical features.

[31] The word category list in Granger and Rayson (1998) includes nine categories: nouns, adjectives, prepositions, articles, determiners, conjunctions, pronouns, adverbs and verbs.

[32] The computation process consists of submitting corpora of native and non-native writing to a lexical frequency software program which uses the chi-square test to discover the lexical items whose frequency distribution across the subcorpora is statistically significant.

With regard to (i), this method inevitably limits the scope of interlanguage research by restricting it to investigations with a lexical, morphological or syntactic focus. By their nature, semantically-driven investigations are not exclusively form-based and require a substantial amount of interpretive effort:

> if the research focus is on a semantic category such as agency or causality, or a syntactic structure, such as complex noun phrases, automatic retrieval becomes more difficult, if not simply impossible. (Granger 2003a: 23)

So while Granger's ICM approach has proved successful with regard to particular form-based linguistic domains, it has become closely associated with the use of methodological tools such as *WordSmith Tool*s. The methodological advantage of those tools, according to Granger (2002: 15) is that "[t]ext retrieval software such as *WordSmith Tools* can count not only words but also word partials and sequences of words which it can sort into alphabetical and frequency order". In the case of Granger and Rayson (1998), the authors use a lexical frequency software developed at the University of Lancaster (see Rayson and Wilson 1996 for details). The association between learner corpus research and text retrieval software has so far limited the application of the ICM approach to studies whose topic of investigation can be dealt with automatically. This methodological shortcoming triggers an important concern, namely that in the case of the modal verbs, an in-depth context-based investigation, as recommended in Hermerén 1978, is difficult to envisage. This is an important point as Odlin (1989: 84) recognises the usefulness of grammatically-grounded approaches for the purpose of semantic investigations in learner language and he notes that "progress with regard to the issue of meaning in second language research will (…) be achieved through a clearer understanding of the interactions between syntax and other subsystems". The recognition that grammatical contexts of linguistic items need to be included in IL semantic investigations emerges from work by Perdue (2000: 300), who emphasises that the structure of interlanguage varieties is determined by a limited set of organizing principles of different levels – syntactic, semantic and pragmatic". He further adds that "[t]hese principles interact and the specific interaction determines the organisation of a given learner variety at a given point of time" (p. 300). It is clear that the methodological requirements for an investigation of these grammatical interactions in learner language are far greater than what existing ready-made tools can cope with.

So far in this section, I have shown that recent learner corpus research has under-performed mainly due to methodological limitations that prevent researchers from

- investigating evasive semantic phenomena;
- integrating contextual feature of a variety of linguistic levels into their analyses;
- providing adequate explanatory and predictive power.

In what follows, I illustrate the above shortcomings by presenting and discussing three corpus-based studies on the modals in EFL/ESL: Aijmer (2002), Neff *et al.* (2003) and Deuber (2010).

3.4.2 EFL/ESL studies

Aijmer (2002) analyses advanced learners' use of key modal words based on a corpus of Swedish L2 English writers. She adopts Granger's ICM framework and her analysis consists of a comparison of the frequencies of some key modal words.[33] With regard to the modal auxiliaries specifically, she conducts two comparisons that involve non-native speakers with different linguistic backgrounds. In the first, she compares the frequencies of occurrence of one group of modals in native English and advanced Swedish-English interlanguage and in the second, she compares the frequencies of occurrence of a second group of modals in Swedish, French and German learner English and native English. Table 6 lists the modals that were included in each comparison.

[33] Aijmer's study is based on data samples of approximately 52,000 words taken from the German, French and Swedish subsections of the International Corpus of Learner English (ICLE) and the Louvain Corpus of Native English Essays (LOCNESS).

Table 6 Modals included in Aijmer's (2002) comparisons of interlanguage varieties

Types of modals	Types of speakers
will ('ll); can, would, could, must, have (got) to, should, may, might, ought to, shall	Swedish native speakers English native speakers
can, could, may, might	French native speakers German native speakers Swedish native speakers English native speakers

To investigate the modals' frequencies of occurrences, Aijmer uses concordance software that uses the chi-square test as a measure of significance in differences. Overall, Aijmer's study indicates that advanced learner writing yields "a generalized overuse of all the formal categories of modality" (p. 72). In the cases of *may* and *can*, she notes that while only German learners significantly overuse *can* and *could*, only French learners overuse *may*. Aijmer also finds that Swedish native speakers have an extremely high use of epistemic *may*. Table 7 presents Aijmer's reported distribution of the root and epistemic *may*.

Table 7 Distribution of root and epistemic *may* in Aijmer's (2002) data

Root *may*		Epistemic *may*	
Swedish native speakers	Engl. native speakers	Swedish native speakers	Engl. native speakers
0	10	46	25

Based on the above distribution of *may* and existing corpus findings from the *Longman Grammar of Written and Spoken English* that report that *could*, *may* and *might* "are used almost exclusively to mark logical possibility" (LGWSE, 6.6.4.1, quoted in Aijmer 2002: 65), Aijmer concludes 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). While her finding invites the analyst to go beyond the simplest of descriptive accounts by at least also investigating the modals' uses at a functional level, her methodology does not equip her for a fine-grained exploration of the linguistic and cognitive/mental mechanisms involved in form/function mapping processes. It is strictly even possible that the learners in her data used the modals exactly as native speakers would have done in the exact same context and that the frequency differences are exclusively due to different context frequencies. While this is unlikely to be the only reason for the distributional differences obtained, Aijmer's study does not test the degree to which this is not the case. Investigating such mappings requires the analyst to include in his/her investigation several linguistic levels (e.g. syntax, lexis) and to statistically analyse them simultaneously. Such a multifactorial approach is compatible with the view that "words and syntactic patterns are represented as qualitatively similar nodes in a network where, in [language] production, lexical and syntactic nodes are activated when they fit the semantic/pragmatic meaning to be communicated" (Gries 2010b: 335). In the case of the modals, this view suggests that the use of modal *X* would trigger the use of particular syntactic structures or the choice of particular lexical items. It is thus essential to retrieve and include as much co-textual information as possible in the analysis of any modal. While this type of multifactorial approach would provide a way to further Aijmer's study, it entails adoption of retrieval and analytical methods that, again, are beyond what most ready-made software currently has to offer.

Although they do not adopt a psychologically informed theoretical framework, Neff *et al.* (2003) explore, to some extent, the potential pragmatic meaning of the modals as part of *we* + *modal* constructions. Their study goes beyond Aijmer's in that it investigates the potential (pragmatic) meaning in 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) use 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) use data extracted from the 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.* use the American subsection of the LOCNESS corpus as the control native corpus. The data was analysed using WordSmith Tools 3.0 and the software's keywords tool is used to perform chi-square tests. Comparisons of statistical significance were then carried out on the basis the matching *p*-values of chi-square scores. 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 in the study. However, since their study only compares raw frequencies of occurrence without regard to contextual features, it is not particularly illuminating. Based on a previous similar study (Neff *et al.* 2000) that identified the *we can* construction as showing the "the highest frequency for two-word clusters within the SUW [Spanish corpus] subcorpus" (p. 221), Neff *et al.* (2003) further investigate the uses of that construction and several other clusters of "*we* + modal verb" (*could*, *might*, *must*). Generally, they find that the construction "*we* + modal verb" is overused by native French speakers, in comparison to American English speakers and they identify *we can say*, *we could say* and *we may say* as representative of that finding. The authors conclude that, rather surprisingly, "these *we* clusters are used by writers with L1 Romance languages to present new topics" and they assign the constructions the pragmatic function of "including the reader in the writer's discourse community and assuming that the information presented is common knowledge" (p. 223). Overall, Neff *et al.*'s exploration of the uses of the modals from a constructional perspective presents a definite step forward from Aijmer's study which approached the forms in total isolation from their co-texts. Still, they limit the interpretation of their data to one context and to a pragmatic orientation, and they investigate the case of one single decontextualised construction (as opposed to investigating the behaviour of the construction within its context of utterance). This means that ultimately, they are not in a position to assess how the pragmatic differences they observed in the uses of the modals manifest themselves grammatically in the wide context of the utterance (e.g. at sentential level). In other words, they are not able to recognise potential behavioural patterns characteristic of the construction when used in particular semantic or morphosyntactic environments. Just as for Aijmer, addressing these shortcomings would require adoption of methodological tools that are able to handle a multifactorial treatment of the corpus data.

A recent quantitative study of the modals in English as a second language (ESL) that takes a more grammatically-grounded approach is Deuber (2010). Drawing on previous observations that Creole strongly influences the uses of the two modal pairs *can/could* and *will/would* in Trinidadian English, the study provides "an expanded comparative analysis of quantitative distributions based on all currently available ICE [International Corpus of English], and, more importantly, [it] presents a detailed analysis of uses and meanings" (p. 113). Of specific interest to Deuber is the investigation of (i) how the coexistence of an ENL (i.e., English as a native language) variety with a different but related system can affect the rather stable English modality system and (ii) how such coexistence is reflected in the choice of one modal form (e.g. *can*) over another (e.g. *could*) by ESL (i.e., English as a second language) speakers. Deuber bases her analysis on a broad data set that allows her to compare a thirty-thousand word corpus of spoken Trinidadian English to three native varieties of English, one variety of ESD (i.e., English as a second dialect) and five ESL varieties.[34] The corpus consists of conversations, class lessons, unscripted speeches and broadcast news.

Like the above-mentioned EFL studies, Deuber (2010: 108) presents an attempt "to contribute to the very limited research on modal verbs in varieties of English outside the inner circle of ENL varieties". However, Deuber's study differs in two major ways. First, as an object of investigation, non-native speakers' lexical choices have so far only rarely been investigated in a quantitative corpus-based fashion. Second, Deuber refers to existing work such as Winford 1980 to recognise that a number of grammatical features in Creole such as zero progressives and *does* as a marker of present habitual influence, to a degree, the behaviour of Trinidadian English. She further stresses the importance of grammatical components in the uses of modal forms in ESL by acknowledging that (i) "perfect forms (of the standard type) [have to] be considered separately from nonperfect forms since they have a different range of meanings and are potentially subject to influence from different forms in the Creole" (p. 115), and (ii), by recognising the possibility that "the negative forms of *can/could* pattern differently from the positive forms" (p. 116).

[34] While all subcorpora included in the study are components of the International Corpus of English (ICE), the ICE-T&T is used to investigate Trinidadian English and, in turn, compared to (i) British English (ICE-GB), Irish English (ICE-Ireland) and New Zealand English (ICE-New Zealand) for comparisons with native English, and with (ii) East African English (ICE-East Africa), Hong Kong English (ICE-Hong Kong), Indian English (ICE-India), Philippine English (ICE-Philippines) and Singapore English (ICE-Singapore).

Deuber's study provides quantitative support for the influence of English-related Creole on Trinidadian English and she illustrates such influence with the use of Creole *could* and *would* "as present tense modals equivalent to standard English *can* and *will*" (p. 134). However, she does not provide an explanation of the influential process. This is mainly due to two reasons, namely (i) an inconsistency in her theoretical and her methodological approaches and (ii) her statistical approach. With regard to the former, while Deuber initially recognises the potential correlation between grammatical features in English-based Creole and their potential impact on Trinidadian English, her study is not set up to assess the extent to which the uses of the modals included in her study reflect such correlation. Deuber's analysis of the uses of the modals is based on a token-based approach which involves extracting and annotating each form for their interpretation in context. Similarly to Collins (2009), the uses and the meanings of the modals in the Trinidadian data are investigated on the basis of both English reference grammars and specialised studies of the modals such as Coates (1983) and Palmer (1990), While, in the case of *can/could*, Deuber identifies nonpast/nonhypothetical uses as one of four text categories, her annotation taxonomy for that particular text category consists of four levels: possibility (dynamic/epistemic), dynamic ability, dynamic perception/cognition verbs and deontic (permission).[35] Like Aijmer (2002) and Neff *et al.* (2003), Deuber 2010 uses statistics based on normalised frequency counts and ratios of one member of a modal pair to the other (i.e., *can* to *could*). Overall, Deuber concludes that although "quantitative distributions of pairs of related modal verbs can give useful indications of the differential patterns of usage in some cases (…) they may also mask considerable differences between varieties" (p. 135). Her methodological approach, however, does not allow her to assess such differences on the basis of the grammatical context of occurrence of the modals (an approach which she herself reports as useful). Ultimately, Deuber is led to conclude that

> [t]he quantitative findings from the ICE corpora analyzed in the present study have indicated that the nonpast modals *will* and *can* are more widely used in ESL varieties. However, closer examination of the use of these modals **in context** is needed to determine **to what extent there may be specific uses of *can* versus *could* and *will* versus *would* in**

[35] The four text categories identified by Deuber (2010) for the uses and meanings of *can*/*could* are: nonpast/nonhypothetical uses, past/hypothetical uses, pragmatically specialised uses and unclear/indeterminate uses (p. 119).

> **different varieties, to what extent choices between members of these pairs and competing forms in the modal system may differ**" (Deuber 2010: 137) [my emphasis]

As an interim summary, it is fair to say that the three studies discussed, among others, have demonstrated the usefulness of the Granger's ICM approach for the investigation of learner language and they have provided quantitative evidence that English learners use of the modals differs from that of native speakers. In addition, differences in modal uses between speakers of different linguistic backgrounds (i.e., French-, Spanish-, and German-English learners) have led to the hypothesis that, in the case of overuse of a particular modal form, cross-linguistic transfer may emerge from the influence of learners' L1. More generally, all three studies present converging evidence that indicates the need to approach the uses of the modals in L2 (i) from a functional perspective, (ii) in a more grammatically-grounded approach, and (iii) as part of constructions rather than as isolated lexical items.

Despite such benefits, form-based approaches such as those discussed above have remained mainly descriptive and hardly any attempts have been made to make sense of descriptive results at a more abstract level by relating them to, say, second language processing strategies. This is an important point because ultimately such an analytical focus is necessary to further our understanding of how non-native speakers shape IL grammars. In addition, it is crucial that corpus-based IL studies develop more appropriate (i.e., not just form-based) corpus-linguistic research methods that allow the analyst to investigate his/her data beyond what is machine-retrievable. In that respect, and while form-based studies do raise the issue of the practicality of investigating, for instance, potential cases of cross-linguistic transfer at conceptual level (i.e. when cross-linguistic translational equivalents trigger different semantic concepts in different language and a learner activates his/her native language's concept for the purpose of using it in L2), cross-linguistic contrastive studies such as Salkie 2004 suggest the existence of conceptual transfers and thereby highlight the urgency to apply more sophisticated corpus methods in IL studies. In the following section, I present and discuss Salkie 2004.

3.5 A contrastive cross-linguistic study: Salkie (2004)

From a cross-linguistic and an interlanguage perspective, investigating *may*, *can* and *pouvoir* raises two related issues: (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 this. 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*.

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 subset of the parallel corpus INTERSECT (Salkie 2000) and analyses one hundred randomly extracted occurrences of *may*, *can* and their corresponding sentences in French.[36] Overall, Salkie's study is motivated by the notion that “any patterns that emerge from looking at large numbers of translations must surely have some kind of semantic basis” (2004: 172). For the purpose of the analysis, an equal number of fiction and non-fiction texts were included in the study.[37] In the study, the direction of the translation (i.e., whether *may* and *can* are translated from *pouvoir* or the other way round) was not taken into consideration. In that regard, Salkie (2004: 172) notes that "for contrastive linguistics it is correspondences between texts in two languages which is (*sic*) crucial, rather than the direction of the translation" (Salkie 2004: 172).

Broadly, Salkie's analysis focuses on the senses of *may* and *can* and whether, for each of their occurrences, *pouvoir* is found as their translational equivalent. Although Salkie incorporates the grammatical contexts in which the modal forms occur into his analysis, only animate subjects and passive voice are accounted for both *can* and *pouvoir* and reflexive and impersonal subjects (i.e., *on*) are noted in the case of *pouvoir* only.

[36] Analysed occurrences of *may*, *can* and their French translational equivalent were extracted using the ParaConc (Barlow 1995) concordancer.

[37] In the case of non-fiction texts, the data consist of newspapers, government documents, instruction for software and domestic appliances, reports from the United Nations and scientific and technical documents.

Occurrences of the modal form *may* are not analysed in relation to any other grammatical components. The results of Salkie's analysis are reported below in Tables 8 and 9 for *can* and *may*, respectively.[38]

Table 8 Summary of Salkie's (2004) results for *can*

Type of sense and equivalent in French	Examples %
Ability sense with animate subject	76
of which = present tense of *pouvoir*	40
of which = future or conditional of *pouvoir*	3
of which not translated by *pouvoir*	33
Passive in English = passive in French	9
Passive in English, reflexive / *on* in French	7
Others	8
Total	100

Table 9 Summary of Salkie's (2004) results for *may*

Type of sense and equivalent in French	Examples %
Epistemic sense with speaker uncertainty	6
Epistemic sense without speaker uncertainty	62
of which = present tense of *pouvoir*	29
of which = other form of *pouvoir*	7
of which no modal expression in French	8
of which other expression in French	18
Permission	10
Concession	8
Ability	3
Others	11
Total	100

[38] I have preserved the order of the results as it features in Salkie's study.

Salkie's analysis is centred on three working hypotheses:

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

At the core of Salkie's study is the assumption that modal sentences yield different degrees of modality, and for the purpose of his analysis, he assumes a scale of degree of modality for the possibility modals. On that scale, the ability sense of *can* belongs at the low end, while the epistemic sense of *may* is placed at the high end. Salkie's analysis consists of assessing and comparing where on the scale the different occurrences of *may, can* and *pouvoir* can be placed. As Table 8 shows, Salkie finds that a total of 76 *can*s occur in their ability senses and that in 40 of those occurrences, *pouvoir* is the most direct equivalent. With regard to *may*, Salkie finds "few clear-cut examples which would be placed at the top of the scale of degree of modality, with an epistemic sense expressing uncertainty on the part of the speaker" (p. 6). Furthermore, while *may* is compatible with lower and higher degrees of modality, *pouvoir* "is seldom used as an equivalent for rare cases of high degree of modality" and is used "in half the cases where epistemic *may* yields a lesser degree of modality" (p. 7).

While Salkie (2004) concludes in favour of his third hypothesis, it is important to reiterate that his results were based on only one hundred occurrences of each English modal form (i.e. *may* and *can*) and their respective French translations and a (much) larger corpus would have increased the reliability of the results. However, in the context of the current study, Salkie's results suggest a potentially fruitful direction to investigate L2 uses of *may* and *can*: that is, their possible interference with *pouvoir* during L2 production. In order to further Salkie's study and investigate whether such interference exists, it is reasonable to envisage studying the three modal forms on the basis of two additional parameters, namely (i) an increased number of grammatical components so as to provide a more detailed account of the forms' contexts of use and (ii) a psychologically-informed theoretical framework. The latter allows one to account for the notions (a) that (second) language production results from speakers' processing and storage of information and (b) that L1 information may be integrated (and thereby reflected) in L2 production. In what follows, I show how adopting a cognitive linguistic perspective to studying learner corpus data provides the necessary theoretical tool to explain potential semantic interferences between the uses of *pouvoir* and the uses of IL *may* and *can*. In Table 10 I summarise the findings and desiderata of the main studies on the nature of interlanguage and its methodological implications for investigating L2 uses of English modals that have guided the current work.

Table 10 Overview of the main studies, findings and desiderata guiding the current work in relation to interlanguage and its methodological implications for investigating L2 uses of English modals

Selected publications	Main findings and desiderata
The nature of IL	
Selinker (1969, 1972)	**Findings**: - the behaviour of IL is highly structured; - the lack of correspondence between native and non-native sentence patterns is psychologically motivated; - the sentence patterns that learners ultimately produce reveal their failure to achieve native-speaker competence. **Desideratum**: Selinker's approach is centred around the existence of cross-linguistic binary sets of forms/structures to express similar meanings.
Adjemians (1976)	**Findings**: - a linguistically-based investigation of IL is more appropriate than one that is cognitively-based; - IL grammars are interim grammars that are, by their nature, not fixed. **Desideratum**: Adjemians' study lacks a methodological framework to assess the permeability of IL.
Corpus work in IL	
Granger (2002, 2003a, 2004)	**Finding**: the distribution of formal elements in learner language allows for the description and characterisation of individual IL varieties. **Desideratum**: "if the research focus is on a semantic category (…) automatic retrieval becomes more difficult, if not simply impossible" (2003a: 23).
Jarvis (2000)	**Finding**: there is a need to adopt a rigorous methodological framework for the study of L1 influence over L2. **Desideratum**: such rigorous methodological framework need be applicable in corpus-based studies.
L2 corpus work on the modals	
Aijmer (2002); Neff *et. al.* (2003); Deuber (2010)	**Finding**: there is a need to approach the modals in L2 from a functional perspective, in a more grammatically-grounded way and as part of constructions rather than as isolated lexical items. **Desideratum**: there is a need to move away from form-based corpus methods to investigate L2 uses of the English modals and to adopt multifactorial methods that allow the analyst to include into his/her analysis the modal forms' co-occurring grammatical features.
Cross-linguistic corpus work on the modals	
Salkie (2004)	**Finding**: semantically, there is a clear distinction between *pouvoir* and *may* and *can*. **Desideratum**: small data set.

3.6 *Theoretical motivations for taking a cognitive usage-based approach to IL* may *and* can

So far, I have highlighted the benefits of adopting empirical and probabilistic methods of investigation to study learner language and I have shown how corpus-based IL studies of learner language contribute to the characterisation of IL varieties. I have also highlighted and summarised the methodological limitations of L2 corpus-based studies in that they tend to adopt form-based approaches which limit their scope of investigation in terms of the amount of linguistic context that they include. Such limitations are not only important in the sense of how they affect the scope of the studies, the range of methods employed and the results these methods can yield, they are maybe even more important in how such traditional form-based corpus methods also limit analysts with regard to the range of theoretical frameworks that they are able to apply to their studies: form-based methods, for instance, prevent analysts to adopt cognitive linguistic approaches such as Bates and MacWhinney's (1982, 1989) Competition Model (CM) of language use and acquisition, a grammatically-grounded model which provides a useful functional framework for investigating cross-linguistic transfer in a statistically-grounded approach. As a result, to date, there is no (at least to my knowledge) statistically-grounded corpus study of the English modals in learner language that adopts an empirically-grounded (i.e. usage-based) cognitive-linguistic orientation. This is a regrettable situation as Divjak (2010: 5) notes that "[c]ognitive linguistics is more suited than other frameworks to deal with a phenomenon characterised by high similarity and low contrastivity in meaning" (see also Gries 2010b for discussion). In what follows, I briefly present some basic cognitive linguistic assumptions and I demonstrate the compatibility between the cognitive-linguistic usage-based approach and the BP corpus method applied in the current work, generally following Gries and Otani (2010) and Gries (2010c). In a second step, I show the relevance of adopting a cognitive linguistic perspective to investigate L2 uses of *may* and *can*.

3.6.1 Cognitive approaches to language and grammar: some basic assumptions

To interpret corpus data from a cognitive linguistic standpoint entails (i) examining the data from the perspective of the speaker (Langacker 1999), (ii) accounting for a variety of psychological processes involved in language production (Geeraerts 2006) and (iii) seeking to understand the mechanisms involved in speakers' uses of linguistic forms as well as unveiling speakers' knowledge about those forms (Geeraerts 2006). Cognitive linguistic theoretical approaches are based on the Cognitive Commitment (Lakoff 1990), which posits that linguistic structures and principles are related to cognitive structures and principles. In other words, cognitive approaches assume that language reflects the structure, the organisation and the principles of human cognition (see Lakoff 1990). Generally, understanding how meaning is construed is central to cognitive linguistic frameworks. In that respect, cognitive approaches recognise the importance of studying speakers' grammatical and lexical choices, in that they provide accounts of meaning construal. An important trend in cognitive linguistics is concerned with usage-based approaches to grammar.

In a way that is compatible with the Cognitive Commitment, usage-based approaches to grammar aim to account for aspects of grammatical structure by relating them to general cognition and assume that every encounter with a linguistic pattern in a usage event involves a cognitive event that results from, and feeds into, the linguistic system (Kemmer and Barlow 1999). That is, cognitive approaches to grammar assume a crucial correlation between speakers' knowledge of lexical items and their uses in grammatical contexts. Also, grammar is considered to be meaningful in its own right and to express more schematic meaning than lexical items. Generally, cognitive approaches to grammar seek to uncover the patterns of construals that are realised by the grammatical structure of language. That is, such approaches hold that lexicon and grammar lie on a continuum. To quote the probably most influential early work, “both lexical and syntactic constructions are essentially the same type of declaratively represented data structure: both pair form with meaning” (Goldberg 1995: 7).

3.6.2 Previous cognitive approaches to the English modals

Although the English modals have raised a high degree of interest within cognitive linguistics, both from diachronic and synchronic perspectives, theoretical studies on the modals have so far not facilitated empirical and quantitative applications and for that reason such studies have been excluded from the current work as they would not have been beneficial to its analysis. Generally, existing cognitive linguistics studies on the modals tend to be concerned with investigating modal forms in relation to polysemy/monosemy (e.g., Langacker 1990, Sweetser 1990, Goossens 1992). Goossens (1992), for instance, presents a semantic analysis of the senses of *can* on the basis of a radial category including prototypical cores. According to Goossens, such prototypical cores are determined on the basis of the modals' frequency of use, their centrality in the network of uses and onomasiological contrasts with other modals. For Sweetser, the modals' polysemy is restricted to three senses derived from three domains of human experience: the sociophysical world, the world of reasoning and the conversational world. Another area of research that has raised interest amongst cognitive linguists is how speakers conceptualise modality. Work by Talmy (e.g., Talmy 2000) characterises that line of research. Broadly, Talmy (2000) is concerned with the development of a modal schematic conceptual background based on the notion of force-dynamics. Talmy argues that, as grammatical words, the modals encode the ways entities interact with regard to force and barriers. Although the current work recognises the importance of the above-mentioned work on the modals, it nonetheless favours a perspective on the modals that is not constrained by an exclusively cognitive semantic perspective. This is motivated by the desire to maintain a strong empirical focus throughout the current study and to combine, in direct ways, a cognitive theoretical approach with an empirical analysis.

In contrast with the above-mentioned studies, the current work investigates the modals from the perspective of a usage-based model of language. Broadly, this means that no distinction between linguistic knowledge and the use of language is assumed. In other words, the current study holds the view that knowing a language is knowing how to use that language. An important consequence of adopting such a theoretical standpoint is that it implies a central focus on linguistic 'usage events', that is “instances of a speaker producing and understanding language” (Kemmer and Barlow 1999: viii). In what follows, I elaborate on and define usage-based approaches to language. I proceed in two steps: first, I define in more detail usage-based assumptions and I explain their implications with regard to (second) language acquisition and use. For that purpose I present Bates and MacWhinney's (1982, 1989) Competition Model as an illustration of a probabilistic usage-based approach (compatible with the objective of the current study). In a second step, I use Rohdenburg's (1996) study to show how linguistic patterns are cognitively motivated.

3.6.3 The usage-based approach and Bates and MacWhinney's (1982, 1989) Competition Model

Usage-based approaches are based on the cognitive linguistic assumption that knowledge of language is experientially based in actual speech which means that meaning and abstract linguistic patterns emerge from speakers' experience of actual speech events. More specifically, the usage-based model assumes that speakers acquire and represent linguistic knowledge on the basis of perceiving and storing very richly 'annotated' (in the sense of 'interpreted') linguistic experiences into a multidimensional knowledge space. In other words, speakers make a linguistic experience such as, hearing a sound or a word and they store this experience along with a substantial amount of pragmatic and contextual information within that multidimensional knowledge space. As a result, the sequential and co-occurrence information that a speaker perceives in an event are stored as points in multidimensional space at coordinates describing that event. According to the usage-based model, speakers process stored linguistic information in such ways that allow them to identify abstract linguistic patterns from specific utterances on specific occasions of use (Tomasello 2000). This stresses the central aspect of syntax-lexicon relations in usage-based approaches and the fact that the ways the syntax and the lexicon relate with each other is in itself meaningful.

One particular usage-based model that combines a processing-based approach to (second) language acquisition and empirical methodological approach is Bates and MacWhinney's (1982, 1989) Competition Model (CM). CM provides a theoretical framework that models the way linguistic items are processed on the basis of their distributional properties.

Broadly, CM is "a probabilistic theory of grammatical processing which developed out of a large body of cross-linguistic work in adult and child language, as well as in aphasia" (Kilborn and Ito 1989: 261). Very much in line with current work in cognitive linguistic approaches, the CM assumes that linguistic signs represent mappings between forms and functions. More specifically,

- linguistic signs map forms and functions onto each other such that forms and functions are cues to functions and forms respectively, and
- 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 probabilistic model, the CM assumes that frequency information as well as function determine the choice of grammatical forms in language production. Cross-linguistically, cues are instantiated in different ways across languages and speakers assign them varying degrees of strength. In the context of investigating interlanguage varieties, Bates and MacWhinney (1989:15) stress that it is 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".

Generally, MacWhinney (2004: 3) characterises the CM 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". According to Kilborn and Ito (1989), the CM provides an adequate theoretical model to investigate learner language. They report on sentence processing experiments involving adults who speak two or more languages. They show that existing psycholinguistic studies have successfully demonstrated that the CM is appropriate for the characterisation of learner language through cue distributions and they report "extensive evidence for the invasion of L1 strategies into L2 processing" (p. 289). Although Kilborn and Ito show that the application of the CM for the investigation of learner language is supported by psychological evidence, from a linguistic corpus-based perspective, it remains to be empirically supported. One way to assess its validity linguistically is to complement the experimental results presented in Kilborn and Ito with corpus data, and which is what this study is set up to do. Inevitably, this task implies the adoption of a methodology that follows the same logic as the CM and is compatible with the notions of cue strengths, validities, and reliabilities, and which are essentially expressed as conditional probabilities.

3.6.4 Rohdenburg's (1996) complexity principle

Usage-based approaches by definition involve linguistic processing: linguistic patterns are cognitively motivated and need to be processed both by speakers and by hearers. That also means that contextual grammatical features present processing constraints that influence speakers' choices, as shown, for instance, by Rohdenburg (1996) or Hawkins (2004). Rohdenburg's study is concerned with the distribution of competing syntactic constructions and how it is influenced by (i) the different degrees of explicitness of those constructions and (ii), the degree of complexity of the grammatical environment in which they occur. According to Rohdenburg (1996), the notion of form explicitness has two implications: first, the existence of mutual contrasts between linguistic variants and, second, that more explicit options can be clearly distinguished from less explicit ones. For Rohdenburg, "[d]ifferences in grammatical explicitness may be expressed in many (…) ways" (p. 151). For the purpose of his study, he focuses on "formal contrasts involving the deletion (or addition) and the substitution of grammatical (…) elements" (p. 151). Example (27) below illustrates the notion of explicitness (the example is extracted from Rohdenburg's study, p. 151).

(27) a. I helped him **to write** the paper
b. I helped him **write** the paper

In (27), the syntactic phenomenon of interest is the use of the infinitive marker *to* as an optional grammatical signal. According to Rohdenburg, the presence of *to* in (27)a makes that variant more explicit than it counterpart in (27)b. This is because "the more explicit variant is represented by the bulkier element or construction" (p. 152). On the basis of his complexity principle, Rohdenburg argues that speakers tend to prefer "more explicit grammatical alternatives (…) in cognitively more complex environments" (p.

149). To return to the above example, while both variants are sensitive to the complexity principle, Rohdenburg proposes that speakers are more likely to use (27)a in more challenging syntactic environments. In his study, Rohdenburg identifies five specific syntactic environments that influence speakers to choose more explicit options such as (27)a. Those environments are:

1. discontinuous constructions of various kinds
2. (the presence of) more or less complex surface objects preceding finite and non-finite clauses
3. heavy subject expressions (including subordinate clauses)
4. complex subordinate clauses
5. passive constructions

From the perspective of corpus-based IL research, the notion that linguistic contexts constrain speakers' linguistic choices remain to be investigated and validated. In the next section, I show that beyond Rohdenburg's complexity principle, adopting a usage-based perspective to investigate L2 *may*/*can* provides a way to explore the two modals in terms of prototype formation as well as their L2 acquisition as constructions (i.e., *may*/*can* + lexical verb).

3.6.5 Prototype formation and construction acquisition in L2

So far in this section, I have provided a brief description of cognitive approaches to language and grammar and I have shown with the CM how usage-based approaches in particular provide a way to observe and explain the cognitive mechanisms involved in the use of language, on the basis of the frequency of occurrence of given linguistic items. Cognitive Linguistics holds that frequency of occurrence influences the shape of grammar during language acquisition and use through two psychological processes *schematisation* and *entrenchment*. Broadly, schematisation is a cognitive process during which speakers extract linguistic patterns from the language and generalise commonalities among the uses of given linguistic items. As a result, speakers develop schemas that are abstract representations of more fully specified and contextualised instances. It follows that schemas and instances differ in terms of the fineness of detail in which they are characterised (Langacker 1987). Entrenchment, on the other hand, refers to the strengthening of the mental representation of a given linguistic unit. Frequency of occurrence plays a major part in entrenchment processes as higher frequencies correlate with higher degrees of entrenchment. Ultimately, the deeper entrenched a linguistic unit is, the 'better' acquired it is.

Usage-based, and particularly exemplar-based, approaches recognise the existence of a probabilistic correlation between the frequency of occurrence of a lexical item, its entrenchment and its degree of prototypicality within a given category.[39] More precisely, "[t]he greater the token frequency of an exemplar, the more it contributes to defining the category, and the greater the likelihood that it will be considered the prototype" (Ellis and Ferreira-Junior 2009). In the context of the current study, it is important to stress the relevance of the process of prototype formation as it raises three questions, namely (i) with regard to modal verbs, do learners consider *may* or *can* as a more prototypical modal verb than the other and if so, which one is the more prototypical, (ii) which uses of *may* or *can* are the more prototypical uses from the learners' perspective, and (iii) does (and if so to what extent) prototypicality influence the order of acquisition of *may* and *can* in L2?

The relevance of prototypicality in relation to L2 *may* and *can* can also be seen at the level of constructions with *may*/*can* + lexical verbs cases: according to recent work by Ellis and Collins (2009) and Ellis and Ferreira-Junior (2009), “L2 learning is driven by the frequency and frequency distribution of exemplars within constructions” (Ellis and Collins 2009: abstract). Furthermore, Ellis and Ferreira-Junior's (2009) study suggests that learners may differ from native speakers with regard to their perception of which construction is the most prototypical (for example, which lexical verb goes with which modal) and which lexical verbs are most prototypically associated with *may* and *can* within each construction. Concretely, this line of approach suggests that:

– L2 *may* and *can* may have lexically-specific preferences based on their co-occurrence patterns, and that

[39] Within the exemplar-based model of knowledge representation, *exemplars* refer to individual members of a particular category.

- those preferences may be identifiable when compared to those of *may* and *can* in L1.

Ellis and Collins identify four areas in the acquisition process of L2 constructions that contribute to the construction-learning process.[40] For the purpose of the current work, I exclusively concentrate on three of those areas which I define in turn below: *input frequency* (and particularly *type frequency* and *Zipfian distribution*), *form* (which includes s*alience/perception*) and finally *function*, represented by *prototypicality of meaning*.[41] By *type frequency*, Ellis and Collins (2009: 330) refer to "the number of distinct lexical items that can be substituted in a given slot in a construction". So in the case of *may* and *can*, this determinant is concerned with the numbers of lexical verbs associated with the modal forms. In category learning, Zipfian distribution is relevant in the sense that "acquisition is optimized by the introduction of an initial, low-variance sample centred on prototypical exemplars" and "[t]his low-variance sample allows learners to get a fix on what will account for most of the category members" (Ellis and Collins 2009: 330). The role of salience and perception in learning constructions is based on the notion that "selective attention, salience, expectation, and surprise are key elements in the analysis of all learning, animal and human alike" (Ellis and Collins 2009: 331). As a determinant, salience refers to "the amount of learning induced from an experience of a cue-outcome association" (p. 331). The authors further note that "[m]any grammatical meaning-form relationships, particularly those that are notoriously difficult for L2 learners, like grammatical particles and inflections in many languages, are of low salience in the language stream" (p. 331). Finally, as previously stated, and in addition to input frequency and salience and perception, prototypicality plays a major part in the acquisition of constructions in L2. Central to this approach is that, in the context of construction learning, all determinants are interrelated; according to Ellis and Ferreira-Junior (2009: 382), for instance, "it is the conspiracy of these several different factors working together that drives acquisition of linguistic constructions".

[40] The four areas identified by Ellis and Collins (2009) as playing a part in construction learning are: *input frequency*, *form*, *function* and *interactions between contingency of form-function mapping*.

[41] Other determinants of construction learning identified by Ellis and Collins are *construction frequency*, *frequency*, *recency* and *redundancy*.

3.7 Concluding remarks

So far we have seen that the shape of learner language is by no means accidental and it is in fact, according to Selinker, highly structured. I have shown that learner language has so far been explored from psychological and linguistic perspectives, and both lines of research have helped to further our understanding of the mechanisms at work in producing non-native language. Corpus-based studies with a cognitive-linguistic orientation, such as Rohdenburg (1996), suggest the possibility of reconciling psychological and linguistic approaches in IL research through a usage-based model of language acquisition and use. While this line of inquiry has already proved fruitful in the field of second language acquisition, with the work of Ellis and Ferreira-Junior (2009) and Ellis and Collins (2009) for instance, it remains to be applied in the field of learner corpus research.

To date, corpus-based approaches to modality in L2 have given some indication of how learners use non-native modals, but they leave much to be desired. While some studies have pointed to the immense complexity of the subject, they have not employed multifactorial or multivariate methods that are capable of addressing this degree of complexity. Some studies that have been based on large numbers of modals have not done much with the vast amount of data other than to present arrays of frequency tables, which are under-analysed in the sense of, for example, little has been done to put the observed results to the more rigorous test of prediction. Meanwhile, cross-linguistically, the analytically more interesting studies, such as Salkie (2004), suggest the existence of semantic interference between native and interlanguage varieties which, in turn, call the analyst to adopt analytic framework that allows study of how learners construe meaning. Because it is probabilistic and cognitively-inspired, the usage-based model of language emerges as providing the necessary theoretical tools to conduct a grammatically-grounded investigation of L2 uses of *may* and *can* that reaches beyond existing form-based studies. Table 11 summarises the main studies and findings that have guided the current work with regard to (second) language acquisition and use.

Table 11 Overview of the main studies, their findings and desiderata guiding the current work in relation to usage-based approaches to (L2) language acquisition and use

Usage-based approach to (L2) language acquisition/use	
Selected publications	Main findings and desiderata
Rohdenburg (1996)	**Findings**: grammatical contexts constrain native speakers' linguistic choices. **Desideratum**: the question whether grammatical contexts also constrain English learners' linguistic choices is yet to be investigated.
Bates and MacWhinney (1982, 1989), Kilborn and Ito (1989)	**Findings**: - the Competition Model (CM) is a probabilistic theory of grammatical processing on the basis of which frequency information and function determine speakers' choices of grammatical forms; - CM is appropriate for the characterisation of learner language. **Desideratum**: the CM is yet to be applied to investigate learner corpus data.
Ellis and Ferreira-Junior (2009), Ellis and Collins (2009)	**Findings**: - "L2 learning is driven by the frequency and frequency distribution of exemplar within constructions" (Ellis and Collins 2009: abstract); - L2 construction-learning is influenced by type frequency, Zipfian frequency, salience and prototypicality of meaning. **Desideratum**: how can the influence of type frequency on L2 construction-learning be measured?

Chapter 4 Previous work in quantitative corpus linguistics

4.1 Introduction

In this chapter, I discuss a variety of approaches and methods that have motivated the present study. However, given the large amount and diversity of previous work, I first summarise the main desiderata that emerged from the analysis of previous work, first theoretically and then methodologically.

With regard to theoretical, or conceptual, desiderata, a study of English *may* and *can* in L1 and L2 should:

- go beyond cases where only a binary choice is available (cf. Selinker 1969);
- include all levels of linguistic analysis – morphology, syntax, semantics – simultaneously (cf. Selinker 1969 and following Klinge and Müller 2005);
- reject a fixed native-speaker grammar and assume a permeable learner grammar (following Adjemian 1976); similarly, reject a strict separation of L1 and L2 mechanisms and assume a probabilistic and processing-based model (as in Bates and MacWhinney's (1982, 1989) Competition Model and as compatible with Kilborn and Ito's (1989) work);
- take into consideration effects from L1 and TL on IL (following Jarvis 2000).

With regard to methodological desiderata, a study of English *may* and *can* in L1 and L2 should

- go beyond the descriptive analysis of modal verbs by providing statistical analyses that allow the researcher to capture meaningful co-occurrence patterns from the data (cf. Collins 2009);
- investigate to what extent the distributional differences between *may* and *can* differ in semantic and morphological linguistic environments and in comparison

with syntactic environments (following Gabrielatos and Sarmento 2006);

- seek to explain the influence of grammatical features from one English variety over another by using statistical tools that are powerful enough to handle more than one feature at a time (following Deuber 2010);
- aim for an analytical corpus-based account of how the same concept is used by learners in L2 and in their L1 but on the basis of a large corpus data set (following Salkie 2004).

In recent years, quantitative corpus linguistics has undergone major developments which are compatible with achieving the above desiderata. These developments have been motivated by two factors. The first is the general recognition that "linguistic data are more probabilistic than has been widely recognized in theoretical linguistics" (Bresnan *et al.* 2007: 91) and the second is the growing recognition that corpus-based analyses should assume and be embedded in psycholinguistically informed and (cognitively-inspired) usage-based theoretical frameworks (Gries 2010b). One such framework is the usage/exemplar-based model discussed in the previous chapter. According to Gries (2010b: 337), exemplar-based models are "compatible with our knowledge that speakers/listeners store immense amounts of probabilistic information". Methodologically, such cognitively-inspired approaches involve the use of statistical methods that can model language in use:

> on a methodological level, this kind of model [exemplar-based approach] forces us to turn more towards multifactorial approaches in hypothesis-testing where model selection processes are used to, in the parlance of an exemplar model approach, determine which dimensions for which data are available should be retained (i.e., for which dimensions we need to rotate our multidimensional space to see another important difference) (Gries 2010, p.c.)

Gries (2010b) shows how corpus linguistics, which has to date mainly been considered as a descriptive discipline, can make use of sophisticated statistical tools and thereby provide full-fledged scientific analyses that go beyond description, to explain and predict linguistic phenomena. Such an approach involves the operationalisation of the linguistic levels believed to contribute to the on-line processing (i.e., semantics, syntax, morphology). By integrating those levels into a multifactorial data investigation process, the corpus linguist is then able to model language in use and thereby provide a more realistic analysis.

This kind of approach has been particularly useful for the investigation of semantic similarity as well as lexical and syntactic alternations. In the former case, Divjak and Gries (2008) Behavioural Profile approach provides a way to focus on lexical semantic similarity by summarising and comparing the semantic and morphosyntactic behaviour of given lexical items. In the case of alternation phenomena, existing studies have used different types of statistical approaches. First, through the use of multifactorial techniques such as *Linear Discriminant Analysis* (LDA) in Gries (2003a) or binary/multiple regression models. Broadly, the use of these methods allows the analyst to quantify the extent to which individual independent variables contribute to the choice of one particular member of an alternating pair. Gries and Stefanowitsch's (2004) method of distinctive collexeme analysis, while not multifactorial in nature, also serves as a way to investigate alternation phenomena from a corpus-based, constructional perspective. This method is contrastive in nature and "identifies lexemes that exhibit a strong preference for one member of the pair as opposed to the other and thus makes it possible to identify certain distributional differences between the members of the pair" (Gries and Stefanowitsch 2004: 97). For the purpose of the current study, the distinctive collexeme analysis provides a way to investigate interlanguage *may* and *can* from the perspective of their verbal complementation. In what follows, I survey in turn behavioural profiles, a selection of multifactorial statistical methods and Gries and Stefanowitsch's (2004) distinctive collexeme analysis method.

4.2 An approach to semantic similarity: behavioural profiles (Gries and Divjak 2009)

A corpus-based approach that meets many of the above desiderata is the recently developed Behavioural Profile (BP) approach (Gries and Divjak 2009). It combines the statistical methods of contemporary quantitative corpus linguistics with a cognitive-linguistic and psycholinguistic orientation (cf. Divjak and Gries 2006, 2008, 2009; Gries 2006; Gries and Divjak 2009, 2010c; and others). As such, it diverges radically from the more traditional corpus-based approaches. Theoretically, the method follows the exemplar-based approach, defined in Section 3.1. Given that the exemplar-based approach is completely based on various kinds of co-occurrence information, it comes as no surprise that, like much other work in corpus linguistics, the BP approach assumes that "distributional similarity reflects, or is indicative of, functional similarity" (Gries and Divjak 2009: 59). While previous BP studies have investigated lexical relations (near-synonymy, polysemy, antonymy) both within languages (English, Finnish, Russian) and across languages (English and Russian), the present study adds to the domains in which Behavioural Profiles have been used in two ways: (i) by studying non-native language data, and (ii) by adding 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, it is mainly concerned with the following issues:

- to what degree can behavioural profiling handle learner data, which are inherently more messy and volatile than native data?
- to what degree can behavioural profiling provide a quantitatively adequate and fine-grained characterisation of the use of *can* and *may* by native speakers and learners?
- to what degree do French speakers' use of *pouvoir* compare to the use of *can* and *may* by native speakers and learners?
- 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 notion of *profile* in *Behavioural Profile* is based on the idea that linguistic items can be characterised on the basis of their co-occurrence with other linguistic components such as, for instance, negation or referent animacy. In other words, the profiling process involved in the method requires the analyst to take into account a wide variety of semantic and morphosyntactic features that all together contribute to the profile of a given lexical item. Thus a form's profile refers to

> [a] comprehensive inventory of elements co-occurring with a word within the confines of a simple clause or sentence in actual speech or writing (Gries and Divjak 2009: 61)

According to Hanks (2000), interactions between grammatically co-occurring elements trigger particular senses of a lexical item in use according to conceptual semantic networks. In Hanks' words,

> the meaning potential of each word is made up of a number of components, which may be activated cognitively by other words in the context in which it is used. These cognitive components are linked in a network which provides the whole semantic base of the language, with enormous dynamic potential for saying new things and relating the unknown to the known (Hanks 2000: 214)

This means that the interpretation of a particular lexical item is the result of the interaction of that particular item with other linguistic components in the sentence. To illustrate the point that meanings do not arise from individual words but rather from combinations, Hanks (2000: 211) provides the following examples where the verb *climb* is interpreted differently in each of its occurrences:

(28) the two men who first climbed Mt Everest
(29) he climbed a sycamore tree to get a better view
(30) he climbed a gate into a field

According to Hanks, (28), (29) and (30) incur three different implications. In the case of (28), the verb climb implies that the two men reached the top of Mount Everest. In the case of (29), however, the proposition does not imply that the climber reached the top of the tree. Rather, as Hanks writes, the proposition implies that “the climber stopped part-way up the sycamore tree”. Finally, in the case of (30), the proposition not only implies that the climber reached the top of the gate, but that he also climbed down on the other side. According to Hanks, different implications exist because “[p]ropositions, not words, have entailments. But words can be used as convenient storage locations for conventional phraseology and for the entailments or implications that are associated with those bits of phraseology" (p. 211).

Methodologically, the profiling process involves a rigorous four-step procedure from data retrieval to data analysis (Divjak and Gries 2009; Gries and Divjak 2009, 2010; Gries and Otani 2010). I describe below each step in chronological order:

1. retrieve all instances of a word's lemma from a corpus in their context;
2. annotate manually a number of features characteristic of the use of the word forms in the data; these features are semantic and morphosyntactic in nature and include a number of characteristics referred to in the BP literature as ID tags.[42] Each ID tag contributes to the profiling of the investigated lexical item(s);
3. generate a co-occurrence table; and
4. evaluate the table by means of statistical techniques.

While, to agree with Granger (2003a: 23), quantitative fine-grained semantic analyses in learner language are impossible to carry out on the basis of a form-based methodological approach, the BP method offers a way to combine a high level of objectivity with an unprecedented degree of granularity in learner language analyses. With regard to granularity, Gries notes, for instance, that

> [i]f one really wishes to exploit all the information available in concordances of say, a particular verb, then there are many characteristics that computer scripts usually fail to retrieve with a high degree of precision: example include animacy of participants, clause types, transitivity of the verb, properties of the process denoted by the verb, the metaphoricity of the use of the verb, etc. The number of properties that need to be retrieved manually or at least semi-manually can easily reach 100,000 (Gries 2008: 422)

With regard to objectivity, the degree of the researcher's subjective analytical input is reduced to a minimum through (i) the use of full concordances, (ii) data annotation that

[42] Henceforth, individual semantic and morphosyntactic features are referred to as *variables*. Their respective levels are referred to as *ID tags*.

can be made explicit through coding instructions and/or criteria and tested for consistency and (iii), the use of statistical techniques for the analysis of the data.

With the ultimate objective of modelling language in use, the statistical assessment of the data provides a way to measure the degree of similarity between given lexical forms. Such measurement is based on the forms' combinations of semantic and morphosyntactic features and their varying degrees of strength. The information related to the forms' combinations is accessed through a statistical modelling process which determines which dimensions for which data are available and should be retained.

Finally, given the slippery nature of semantic analyses in native language in general, but even more so in non-native language, the BP approach equips the analyst with a methodology to investigate semantic aspects of learner language with an unprecedented degree of precision.[43]

4.3 Approaches to alternation phenomena: description and prediction

Corpus linguists have investigated linguistic alternation phenomena from the perspectives of description and prediction using both monofactorial and multifactorial statistical approaches. Both types of approaches have triggered the use of different statistical tools. In the remainder of this section, I provide an overview of the statistical methods that have so far been associated with alternation phenomena.

4.3.1 Frequency tables and chi-square tests

A basic way to investigate alternation phenomena is on the basis of *r*x2 frequency tables and their analyses using, for example, chi-square tests which determine whether observed distributional differences between two linguistic alternating variants are caused by chance. By definition, this type of approach is monofactorial as it involves the cross-tabulation of a categorical variable with another categorical variable.[44]

[43] See Janda's (2009) discussion on how languages often carve up semantic space very differently.

[44] Categorical variables refer to classes of entities such as ANIMACY in semantics or CONSTRUCTION in syntax.

4.3.2 Means/medians and *t/U*-tests

Beyond chi-square tests, another (monofactorial) way to investigate alternation phenomena consists of computing and comparing the means of the number of occurrences of alternating variants throughout a given data set. To assess the degree of statistical significance between the means or medians of two independent samples, two tests can be applied: the *t* test and the *U* test. A crucial difference between the two tests is that the data sample in the *t* test is assumed to be normally distributed whereas the *U* test makes no such distributional assumption.

4.3.3 Multifactorial approaches: discriminant analysis and regression models

In this section, I present and discuss briefly two multifactorial statistical approaches: *discriminant analysis* and *binary regressions models*. Discriminating between, say, two syntactic structures involves recognising which linguistic attributes support the choice of one construction over the other. To facilitate such recognition, *linear discriminant analysis* (LDA) is based on (i) the identification of weights that quantify each variable's importance for the best possible discrimination between the two constructions, (ii) the computation of discriminant scores for each instance of the construction in the analysed corpus data and, optionally, (iii) a subsequent sorting process of all the sentences in order of their discriminant scores. The result is a continuum where

> [t]he sentences represented by the two most extreme points are the ideals of the two categories made up by the sets of positive values and the set of negative values and, thus, represent the prototypes of the two constructions: they exhibit exactly those characteristics that have a high cue validity for the construction they instantiate (Gries 2003b: 10)

Overall, the LDA indicates the degree to which a particular linguistic attribute contributes towards the choice of one construction over another. Further, conceptually at least, models underlying LDAs can be considered an appropriate tool in the sense that predictions of speakers' choices are calculated using all independent variables simultaneously, as is the case in the speaker's on-line production process: "human native speakers subconsciously somehow manage to keep track of all the variables in real time" (Gries 2003a: 6). Downsides of the LDA approach, however, are, strictly speaking, that they require multivariate normality and do not address cases well where “pervasive correlations” (p. 76) are at play in the data (Bresnan *et al*. 2007). By contrast, the technique of *binary logistic regression* also allows the researcher to study which particular combinations of variables contribute to the choice of one construction over another and to what degree they do so, but it does not require multivariate normality. The usefulness of logistic regression models for the multifactorial assessment of corpus data is clearly demonstrated by Hoffmann's (2006) modelling of preposition placement in English relative clauses. However, in order to compute his regression model, Hoffmann uses the statistical research tool *Goldvarb* which limits the degree of precision of his results, as *Goldvarb* is not suited for the computation of categorical effects:[45]

> [t]okens exhibiting such factors either had to be eliminated from the data, or grouped together ("recoded") with other non-categorical factors from the same factor group, provided there were sufficient linguistic reasons supporting such a re-grouping (Hoffmann 2006: 170)

The analytical needs for versatile tools suggest the need to resort to powerful tools that are flexible enough to handle the data regardless of the complexity they exhibit.

4.3.4 Distinctive collexeme analysis

Distinctive collexeme analysis (DCA) provides an additional way to investigate alternating linguistic pairs (e.g., active vs. passive voice, ditransitives vs. prepositional datives, etc.). DCA serves as an extension of collostructional analysis [CA] (Stefanowitsch and Gries 2003), which recognises possible associations between words and constructions. Generally, proponents of DCA assume the psychological reality of those associations. In the words of Stefanowitsch,

> [p]ut simply, it [CA] assumes that speakers subconsciously perform a statistical analysis of the input and that the statistical associations found in the data are reflected in psychological associations in the mind if the language user (Stefanowitsch 2006: 258)

[45] In Hoffmann's parlance, the term *categorical* is to be understood in the sense of 'exceptionless'.

The method of distinctive collexeme analysis involves two assumptions, namely that investigated constructions are recognised by the speaker as initially comparable, and the investigated constructions belong to the same linguistic system.[46]

So while conceptually the DCA method is based on the idea that slots in functionally similar syntactic patterns attract and repel particular words, practically, the DCA provides a way to quantify the degree of attraction or repulsion of lexical items in the investigated syntactic slots. The DCA is generally accepted as a reliable and a flexible method to study alternation phenomena. This is based on a number of studies that have investigated a variety of alternation cases such as dative alternation, active vs. passive voice and verb-particle constructions, and also cases that have received less attention where the alternating forms express more or less the same meaning, such as English *-s* genitives versus *of* constructions and the *will* future versus the *be going to* future (cf. Gries and Stefanowitsch (2004) for several case studies).

4.4 Concluding remarks

Overall, this chapter has focused on linguistic alternation phenomena and how corpus linguistics can be used to account for such phenomena. Statistically, alternation phenomena have been approached both from a monofactorial perspective, through the use of, say, frequency tables/chi-square tests, and from a multifactorial perspective, through the application of LDAs or binary logistic regression techniques. Although the benefits of such statistical methods as well as the usefulness of statistical software such as Goldvarb are noted, it emerges that a more appropriate methodological approach for the study of L2 uses of *may* and *can* is provided by logistic regression models that can handle categorical as well as all other types of independent variables.

[46] The term *linguistic system* is to be interpreted "both in an internal psycholinguistic and an external systemic sense" (Stefanowitsch 2006: 258)

Chapter 5 The present study: the corpus data

5.1 Introduction

In this work, I compare three varieties of language: French as a native language, English as a native language and French-English interlanguage (IL). I specifically focus on how the uses of *may* and *can* by native French learners differ from those of native speakers. For that purpose, I aim to pinpoint:

- which contextual components (i.e., morphosyntactic structures, semantic components) contribute to the semantic characterisation of *may* and *can* in French-English interlanguage; and
- to what degree those components interact in context so as to make this use of *may* and *can* characteristic of the French-English type of interlanguage.

While this work requires thorough investigation of the contexts of occurrences of the investigated modal forms, it raises a methodological challenge that has not yet been addressed within the field of interlanguage studies: how best to exploit corpus-based quantitative methods in order to carry out semantic investigations. In the remainder of this section I explain how the corpus data were operationalised. *Operationalisation* refers to the process of deciding which variables are included in the analysis and how they are investigated.

5.2 Corpora and retrieval

The data are derived from three untagged corpora: the International Corpus of Learner English (ICLE), the Louvain Corpus of Native English Essays (LOCNESS), and the Corpus de Dissertations Françaises (CODIF). As the current work is essentially concerned with French-English interlanguage, only the French subsection of ICLE (henceforth ICLE-FR) was used. ICLE-FR, which consists of advanced-level writing, 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 data subsets: 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). The corpus was collected by the Centre for English Corpus Linguistics (UCL) and was made available to me for the purpose of this doctoral project by the Director of the Centre, Professor Sylviane Granger. CODIF includes argumentative and literary texts and has a total of 100,000 words.[47]

The three corpora included in the study are generally very comparable. They all present 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). All participants contributed an essay of approximately 500 words. All the essays have similar topics such as: crime, education, the Gulf War, Europe, university degrees. Following Jarvis (2000) and Granger (2003b), this study recognises the existence of "outside variables" (Jarvis 2000) or "learner variables" (Granger 2003b) that are extra-linguistic in nature and that can potentially affect the degree of influence of L1 in IL. According to Jarvis (2000: 260), the following variables "should ideally be controlled" in view of optimal traceability of potential L1 influence: 1) age, 2) personality, motivation, and language aptitude, 3) social, educational, and cultural background, 4) language background, 5) type and amount of target language exposure, 6) target language proficiency, 7) language distance between the L1 and the target language, 8) task type and area of language use, and 9) prototypicality and markedness of the linguistic feature. The three corpora are comparable in terms of age, proficiency levels, mother-tongue background (in the case of the IL data) and learners' geographical provenance (see Granger 2003b).

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

The data consist of instances of *may* and *can* in native English and French-English interlanguage as well as *pouvoir* in native French, as extracted from ICLE-FR, LOCNESS and CODIF, respectively.

Table 12 below summarises the number of occurrences of *may* and *can* throughout the entire dataset, as featuring in ICLE-FR and LOCNESS, both in their affirmative and negated forms. Table 13 summarises the number of occurrences of *pouvoir* as featuring in CODIF, both in its affirmative and negated forms.

Table 12 Summary of the occurrences of *may* and *can* in ICLE-FR and LOCNESS

may/can	Modal form	Native English and French-English IL(ICLE-FR & LOCNESS)	Native English (LOCNESS)	French-English IL(ICLE-FR)
may	*may*	753	410	343
	may not	79	56	23
	Total	832	466	366
can	*can*	2055	1072	983
	cannot	369	157	212
	can't	108	58	50
	can not	80	35	45
	Total	2612	1322	1290

Table 13 Summary of the occurrences of *pouvoir* in CODIF

pouvoir	Negated *pouvoir*	Total
200	64	264

In order to investigate the data, I used the software R (cf. R Development Core Team 2010) and I wrote R scripts that allowed me to:

- retrieve all occurrences of the investigated modal forms from all sub-corpora; and
- to import the data into a spreadsheet software to allow for the annotation process and the generation of the co-occurrence table.[48]

The resulting spreadsheet consists of 3710 rows of occurrences and 24 columns of annotated variables including case number and preceding and subsequent contexts (of 150 words each). That is, cells within the table contain ID tag levels that describe the annotated match. Overall, the data were manually annotated for 22 semantic and morphosyntactic variables, for a total number of 98 ID tag levels.[49] The resulting data table was evaluated statistically using the interactive R script BP 1.01 (Gries 2010a). The statistical evaluation process is presented below in Section 5.4. Table 14 presents the total range of variables included in the study. Description and discussions of the operationalisation of each variable are provided in Section 5.3. As Table 14 shows, in the case of semantic variables, different sentential parts are targeted: modal forms (MF), modal form subsequent verbs (SV), subject referents (RFT) and grammatical subjects (SBJ). Not all variables in the table are paraphrased.

[48] The software R was used at all stages of the data investigation process to (i) extract all occurrences of *may, can* and *pouvoir*, (ii) compile both an annotation and a co-occurrence table and run all statistical analyses.

[49] The 24 columns of annotations include the encoding of the lemma of each lexical verb occurring with the modal forms and the encoding of the modal forms themselves. Those two components are not included in the total count of manually annotated semantic and morphosyntactic variables.

Table 14 Overview of the variables used in the study

Types of variables	Variables
semantic	Senses (MF)
	SpeakPresence (MF)
	Use (SV)
	VerbType (SV)
	VerbSemantics (SV)
	RefAnim: subject referent animacy (RFT)
	AnimType: subject referent animacy type (RFT)
	SubjRefNumber: subject referent number
syntactic	Neg: negation
	SentType: sentence type
	ClType: clause type
morphological	Form
	SubjMorph: subject morphology
	SubjPerson: subject person
	SubjNumber: subject number
	Elliptic
	Voice
	Aspect
	Mood
data	Corpus
	GramAcc: acceptability

For the purpose of this study, it was judged that in order to successfully identify the nature of the various co-texts from which *may, can* and *pouvoir* derive their specific meanings, as many variables as possible should be employed. This approach allows for close investigation of the possible interactions between all variables. As Table 14 indicates, the current study includes three syntactic variables, eight morphological variables, eight semantic variables and two other variables.

To ensure a thorough treatment of the data, each variable was annotated according to an encoding taxonomy established to allow for its measurement and its consistent treatment across the three corpora. This operationalisation stage involved a decision-making process to identify

- the object of measurement of each variable; and
- how that measurement is to be carried out consistently throughout the data.

Below I provide a detailed account of the operationalisation process for each variable included in the study. I am first concerned with the operationalisation of the semantic variables (i.e. SENSES, SPEAKPRESENCE, USE, VERBTYPE, VERBSEMANTICS, REFANIM, ANIMTYPE,and SUBJREFNUMBER), then the operationalisation of the syntactic variables (i.e. NEG, SENTTYPE, CLTYPE) and the operationalisation of the morphological variables (i.e. SUBJMORPH, SUBJPERSON and SUBJNUMBER, ELLIPTIC, VOICE, ASPECT, MOOD). Lastly, I describe the data-related variables (i.e. CORPUS, GRAMACC).

5.3 Annotation

5.3.1 Semantic variables

The SENSES variable

The variable SENSES encodes the semantic interpretations of *may*, *can* and *pouvoir* as a cross-linguistic group of modal forms that belong to the semantic domain of POSSIBILITY. Each ID tag level included in the SENSES variable reflects a particular facet of that semantic domain. SENSES thereby identifies different types of possibility. Additionally, SENSES accounts for negation as a semantic notion. As discussed in Section 2.2.2, existing literature on modality both indicates that the form-meaning relationship between *may/can* and negation is not straightforward (Palmer 1995) and suggests possible semantic interactions between the modal forms and negation. Generally, SENSES includes an encoding system that allows for:

- the identification of types of possibility as expressed contextually by the modal forms; and

– the identification of cases where negation exclusively affects the interpretation of the modal form.[50]

Table 15 provides an overview of the eight ID tag levels included in the SENSES taxonomy.

Table 15 The SENSES variable and its ID tag levels

Type of variable	Variable	ID Tag levels	Type of negation
semantic	SENSES	epistemic modality	-
		negated epistemic modality	Ext.N
		epistemic possibility + negated proposition	Int.N
		dynamic possibility	-
		negated dynamic possibility	Ext.N
		deontic possibility	-
		negated deontic possibility	Ext.N
		deontic possibility + negated proposition	Int.N

As shown in Table 15, the variable SENSES encodes the three types of possibility: epistemic, deontic and dynamic possibility and the two types of negation. The first type is concerned with cases of external negation (marked as Ext.N in Table 15), where the negation directly affects the modal form and the second type is concerned with internal negation (marked as Int.N in Table 15) cases where the negation is applied to a subsequent proposition and does not semantically interact with the modal form. Further below, I illustrate the two types of negation with examples from the corpora.

With regard to negation, the literature concerned with modality recognises negation as a twofold phenomenon (Halliday 1970, Hermerén 1978, Palmer 1979, Huddleston and Pullum 2002). In Palmer's (1979: 26) words, "we can distinguish between the negation

[50] Negated forms included in the study are *may not*, *cannot* and *can not*.

of the modality and the negation of the event". In the same vein, Huddleston and Pullum identify two types of negation and refer to each type in terms of internal and external negation. In internal cases, "the negation applies semantically to the complement" of the modal (e.g. *he may not have read it:* 'it is possible that he did not read it'). In such cases, the negation is internal to the scope of the modal. In external cases, "the negative applies to the modal itself" (e.g. *he can't have read it:* 'it is not possible that he read it'). Here, the modal does not have scope over the negation. The negation is external to the scope of the modal.

Both types of negation, internal and external, feature in the corpus data. Examples (31) and (32) illustrate cases of external negation. Each example allows for a paraphrase of the type 'it is not possible that X', thus indicating that the modal form is affected by the negation:

(31) These two notions *cannot* be disassociated (ICLE-FR-ULB-0015.1)
'it is not possible that the two notions are dissociated'

(32) However, we *may* not forget that if it occurs it will require some concessions (ICLE-FR-UCL-0008.1)
'it is not possible that we forget that if it occurs it will require some concessions'

The data yield two cases of ambiguity with regard to negation. First, occurrences of *may not* where the distinction between internal and external negation is not clear-cut, and second, occurrences of *can not* where meaning and form are at odds. Example (33) below illustrates the case of *may not* that could ultimately lead to different interpretations of *may.*

(33) you *may* also not forget that at the time the United States were losing a lot of their power (ICLE-FR-UCL-0072.3)

For (33), the two following paraphrases are conceivable: in the case of external negation: 'it is possible that you don't forget' and in the case of internal negation: 'it is possible that you forget'. Throughout the annotation process, the interpretation of *may not* relied heavily on contextual information. Similarly, occurrences of *can not* primarily relied on contextual information for their allocation of SENSES ID tag levels.

With regard to modality and as shown in Table 15, for the purpose of the current work, the taxonomy for the variable SENSES is based on the traditional tripartite model for the categorisation of modality including epistemic, deontic and dynamic possibility. This variable corresponds to Gabrielatos and Sarmento's (2006: 236) call for "a more detailed examination of the distribution of modality types".

The data include examples of dynamic possibility involving perception verbs and verbs of cognition (e.g. *remember*). As discussed in Coates (1983), both types of occurrences represent subcategories of the ability sense. In this study, those occurrences are coded as dynamic:

(34) if one possesses that power of transcending time, one *can* remember the happier moments that life brought (ICLE-FR-UCL-0055.2)

(35) the teaching staff at university has in fact many things in common with what you *can* hear on the place of work when working during the summer (ICLE-FR-ULG-0016.1)

Finally, the dynamic possibility category includes cases of neutral possibility where neither permission nor ability is expressed.

(36) one *can* also identify with these characters bonding with one another in order to gain acceptance that they don't usually and honestly get from society (ICLE-US-PRB-0006.1)

The SPEAKPRESENCE variable

The SPEAKPRESENCE variable encodes the degree of presence of the speaker as reflected by the modal form. As Table 16 shows, the SpeakPresence variable includes three ID tag levels: *weak*, *medium* and *strong*, which refer to a particular level of strength of the degree of speaker presence expressed by the modal form: the *weak* ID tag level indicates a low level of speaker presence, *medium* indicates an intermediate level of speaker presence and *strong* indicates a high level of speaker presence.

Table 16 The SpeakPresence variable and its ID tags levels

Type of variable	Variable	ID Tag levels
semantic	SpeakPresence	weak
		medium
		strong

The role of the speaker in the expression of modality is widely discussed throughout the literature. Some argue that, fundamentally, modality allows the speaker to express various degrees of possibility and necessity (Palmer 1990) and that assessments of the epistemically possible/necessary and the deontically possible/necessary can be undertaken (Lyons 1977). According to Palmer (1986), modality reflects a process of grammaticalisation of speakers' (subjective) attitudes and opinions. A more semantic perspective defines modality as conveying the speaker's comment on the content of a proposition (Altman 1984) and as qualifying states of affairs (Nuyts 2001). Despite the general recognition that the speaker has a role to play in the expression of modal meaning, the nature and the purpose of that role both have never been clearly defined in a unitary fashion. Indeed, Heltoft recognises speaker-relatedness as a controversial aspect of modality:

> The main problem, it would seem, is the relation between modality in the narrow sense: necessity and possibility – as found typically in Germanic modal verb systems – and the area of subjectivity or speaker-relatedness. (Heltoft 2005: 81)

Coding information related to the speaker based on the uses of *may, can* and *pouvoir* incurs a number of both theoretical and empirical hurdles. The first hurdle is defining the notion of subjectivity. Following Herslund's logic, “[i]f modality is the manifestation of the speaker's attitude towards the propositional content of his utterance, it follows that a good deal of the field of modality can be subsumed under the label of 'subjectivity” (p. 39). However, the supposedly defining term *subjectivity* is used across the literature with a variety of definitions and for different purposes. For some, the term is exclusively applied to epistemic modality where the speaker mentally assesses a degree of possibility/necessity (Lyons 1977, Nuyts 2001). The term *subjectivity* is then to be understood relative to the term *objectivity* and partakes in "a distinction between formally reliable evidence and more intuitive guessing" (Herslund 2005: 39). For scholars such as Le Querler (1996), the term *subjective* refers to one of three types of modality, along with *intersubjective* and *objective* modality, and is generally similar to epistemic interpretations. Finally, for Verstraete, the term *subjective* conveys information about the speaker and specifically refers to a degree of explicitness of the presence of the speaker at the core of his definition:

> This different use of subjectivity and objectivity does not refer to the question of whether a linguistic element is related to the speaker or not, but to the question of **how explicitly the speaker is present in an utterance** (Verstraete 2001: 1512) [my emphasis]

The second hurdle is the methodology involved in the measurement of speaker-related information. Indeed, including the variable SpeakPresence in the study raises the methodological issue of operationalising a variable that is subjective in nature. In other words, the degree of presence of the speaker is not a semantic feature that can be quantified straightforwardly. The treatment of SpeakPresence involves an assessment that may indeed vary from one subjective interpretation to another. Furthermore, the treatment of SpeakPresence also involves assessing degrees of modal strength across types of possibilities (i.e. epistemic, dynamic, and deontic possibilities and their negated counterparts). Corpus-based studies with a special interest in the speaker's role in the construction of modal meaning tend to measure and assess degrees of speaker presence in a scalar fashion. Typical scales applied to the English modals include, for instance, a ranking of the modal forms according to their increasing (or decreasing) identified degrees of 'strength'. As noted by van der Auwera (1996: 185), "to call something a 'scale' rather than a 'diagram', the elements making up the scale must not merely be ordered, there must also be a dimension along which the elements have increasing values". In that regard, Salkie (1996: 382), for instance, recognises that on a scale of epistemic modality *may* is 'weaker' than *must*. Such a scalar approach implies the notion of relativity in the sense that each modal form and its identified degree of speaker presence is to be understood relatively to other modals. Verstraete, however, warns that

> a basic scalar organisation that obeys all the traditional criteria of scalarity can still be disrupted by a different type of organisation on another semantic dimension that is associated with the same set of expressions. (Verstraete 2005: 4102)

Verstraete further shows that, at least in the domain of pragmatics,

> deontic modal expressions are not subject to scalar quantity implicatures the way epistemic modal expressions are, in spite of the widespread assumption in the literature that the two types of modality are the same in this respect. (Verstraete 2005: 1416)

Next, I show how, for the purpose of this study, the gap between epistemic and deontic implicature mechanisms is bridged so that a unified encoding system for the variable SpeakPresence can equally be applied to all uses of the modal forms.

Conceptually, the encoding of the SpeakPresence variable is based on the principle that the distinction between types of possibility yields different speaker roles. Following Palmer (1979) and Coates (1983), epistemic uses of modal forms place the speaker in the role of an assessor, as he/she evaluates the degree of possibility of an event occurring. Deontic uses, on the other hand, are basically performative (Palmer, 1979: 58) and for Larreya and Rivière (1999), neutral uses of *can* place the speaker in the position of an observer:

> There is, in the modal system, a key opposition "neutral" / "subjective". Modality is always the expression of a judgement (an opinion, a feeling, etc.) about an event. This judgement, however, can cover two different aspects. It can be (or be presented as) neutral, that is as the possible judgement of any observer: that is the case in *Mary can swim*, where the

> modal judgement expressed by *can* is normally the result of a simple observation of the world. Or, in contrast, the modal judgement can be presented as the expression of the PERSONAL OPINION of the speaker (or his/her WILL, or his/her WISH), and as a result may have what we call a subjective character: thus *you may smoke* expresses the will of the speaker.' (Larreya and Rivière 1999, quoted in Salkie 2001: 3)

The treatment of the SpeakPresence variable assumes Verstraete's (2001) definition of the term *subjectivity* as defined above. Verstraete's (2001) understanding of *subjectivity* allows for a common treatment of the different types of possibility as identified in the corpus data. Indeed, Salkie notes:

> Verstraete argues that English epistemic modals are always subjective, dynamic modals are never subjective, while deontic modals sometimes are and sometimes are not (2001: 1525). If we take subjectivity in his sense as our third criterion for modality, then epistemic modals in general will have a higher degree of modality than dynamic ones, with deontic modals in between. (Salkie 2009: 4)

The three-level taxonomy for the treatment of SpeakPresence is based on Verstraete's (2005) pragmatic approach to modality and speaker-relatedness. Verstraete's (2005) theoretical approach distinguishes between *modal sources* and *modal agents*. He identifies the *modal source* as

> the person (or other entity) responsible for making the assessment encoded by the modal expression. In epistemic modal expressions, for instance, the modal source is the person who judges the event to be possible, probable or necessary (…) while in deontic modal expressions this is the person who gives the permission or imposes the obligation to carry out the action. (Verstraete 2005: 1409-1410)

The *modal agent*, on the other hand, is identified as "the person who is expected to carry out the action" (p. 1410).

Modal agents are typically relevant to deontic types of modality. They "distinguish deontic modal expressions from their epistemic counterparts [which] do not predicate something of one specific participant in the clause, but rather modify the clause as a whole". Generally, Verstraete (2005) attempts to show that deontic expressions not only express weaker or stronger commitment to desirability on the part of some authority but also 'carry different presuppositions about the willingness of the *modal agent* to carry out the action in question'. Table 17 illustrates the correlation between *modal source* and *modal agent* according to Verstraete (2005). The '+' signs indicate the explicit presence of the *modal source* and/or *modal agent*. The '-' signs indicate the absence of explicit *modal source* and/or *modal agent.*

Table 17 *Modal source* and *modal agent* as distinctive features, as represented in Verstraete (2005)

	epistemic	deontic	dynamic
modal source	+	+	-
modal agent	-	+	+

Table 17 shows that *modal source* is present in epistemic and deontic types of modality and it is missing in dynamic modality. This means that (according to Verstraete 2005) while the speaker is explicitly present in epistemic and deontic cases, he/she is not so in dynamic cases. In other words, semantically, the utterance places no emphasis on the speaker (i.e. neither his/her mental assessment of a possible event occurrence nor his/her illocutionary force are semantically put to the foreground of the utterance). *Modal agents*, on the other hand, are, as expected, negatively marked (in a technical sense) for epistemic uses, indicating that they are not to carry out any particular task imposed upon them. They are positively marked for both deontic uses and dynamic uses. The first case is obvious as, according to Verstraete, deontic uses require a *modal agent*. Although in the second case no pragmatic implication is expressed, the (grammatical) subject is presented as the possible recipient of an event occurrence which then puts him/her/it as the modal focus so to speak and thereby shifts the attention away from the speaker.

Table 18 illustrates the adaptation of Verstraete's (2005) assessment model to the present study by showing how the eight identified senses of the investigated lexical items can be categorised into three levels of degree of speaker presence:[51]

Table 18 Cross-tabulating types of modality and degrees of speaker presence

	modal source	modal agent	Degree of speaker presence
dynamic possibility	-	+	weak
dynamic possibility Ext.N	-	+	weak
epistemic possibility	+	-	medium
epistemic possibility Ext.N	+	-	medium
epistemic possibility Int.N	+	-	medium
deontic possibility	+	+	strong
deontic possibility Ext.N	+	+	strong
deontic possibility Int.N	+	+	strong

As shown in Table 18 , weak scores include dynamic possibility and dynamic possibility Ext.N senses, medium scores include epistemic possibility and both its negation types. Similarly, strong scores include deontic possibility and both its negation types. The distinction between the strong and medium levels is made on the basis that a strong speaker presence presupposes a degree of willingness on the part of the modal agent to carry out a particular action. Verstraete (2005) notes for instance that "[p]ermission presupposes that the agent is actively willing to carry out the action and therefore carries an expectation of actualization" (p. 1408). Epistemic uses (i.e. uses that convey a medium degree of speaker presence) imply no presupposition of agent attitude as they are concerned with the speaker's mental assessment process (see Section 2.2.2). In Verstraete's words:

> [Epistemic and deontic expressions] differ in more respects than just modal strength, and this disrupts the implicature mechanism that works well for the weaker and stronger degrees of epistemic modality. (Verstraete 2005: abst)

[51] The SENSES and the SPEAKPRESENCE variables are expected to be highly correlated.

Examples (37), (38) and (39) below illustrate how the three-level scale is applied to the data.

(37) In addition to the robbery charge you *can* add attempted murder, assault with deadly weapon and maybe a few others (ICLE-US-IND-0014.1) – *strong*

(38) Some people *may* argue that man could not do without thinking (ICLE-FR-UCL-0017.3) – *medium*

(39) In order for this world to be one that our children and grandchildren *can* benefit from, recycling needs to become commonplace uniformly (ICLE-US-SCU-0010.3) – *weak*

Example (37) provides a case of 'extended' use of *can* in its deontic possibility sense (i.e. in this case, the force of the sentence is essentially the same as the command *add attempted murder*). In (37) the speaker actively invites the addressee to add attempted murder to an already existing list of charges. *Can* in (37) presupposes the agent's willingness to carry out the "adding" process and the modal form therefore must be positively marked for both modal source and modal agent. Consequently, (37) illustrates a case of strong explicit speaker presence. Example (38), on the other hand, does not imply any sort of negotiation about any action and generally indicates the speaker's reasoning on the likelihood of an argument about a particular fact. For that reason, (38) illustrates a medium degree of speaker presence with a positively marked (i.e., + in Table 18*)* modal source and a negatively marked (i.e., - in Table 18) modal agent. Finally, (39) presents a case of weak speaker presence where *can* denotes dynamic possibility, which requires both a negative marking for modal source and a positive marking for modal agent.

The last hurdle encountered in operationalising the SpeakPresence variable is the comparability of cross-linguistic forms in terms of lexicalised speaker-related information. Two questions arise here: do the cross-linguistic forms lexicalise speaker-related information in the same ways or not, and does the chosen coding framework maximally account for cross-linguistic disparities? Such questions highlight the limitations of approaching SpeakPresence formally, thus requiring for a more pragmatic approach.

Example (40) illustrates the suitability of the SpeakPresence taxonomy for the annotation of the French data. In (40) *pouvoir* denotes a sense of dynamic possibility.

(40) La découverte de nouveaux horizons *peut* également stimuler l'entreprise éconnomique (L1FUEF02)
'discovering new horizons *can* also stimulate economic enterprise' [my translation]

'Discovering new horizons' in (40) refers to a mental scenario in which semantically the speaker plays a major role. The theoretical nature of event occurrence does not allow for the reading of any pragmatic implication in that specific use of *pouvoir.* As a result, (40) shows a case a *medium* level of explicitness of speaker presence.[52]

The Use variable

The Use variable targets the lexical verbs that follow the modal forms and identifies whether they are used metaphorically or literally, as shown in Table 19.

Table 19 The Use variable and its ID tag levels

Type of variable	Variable	ID Tag levels
semantic	Use	metaphorical
		literal

In metaphorical cases, the lexical verb is used figuratively, and in literal cases they are

[52] My decision to include SpeakPresence was motivated by the literature on modals that recognises the part speakers' subjectivity plays in the expression of modal meaning. Within a quantitative context-based exploration of the uses of the modals, I have included as many variables that might influence the use of modals as possible of those identified in descriptive accounts. SpeakPresence was initially included as a variable because it describes something different from Senses *per se*, but since SpeakPresence maps directly on to sense distinctions, it does not contribute in a different way (from Senses) to the picture that emerges in the results. SpeakPresence is included here as part of the complete record of the annotation process, but was discounted in the logistic regression analysis where its redundancy would have affected the outcome and would have incurred an problematic issue of collinearity.

not. Distinctions between the two types of uses were primarily based on *may* and *can*'s contexts of utterance. However, in ambiguous cases where both types of use could apply to a single lexical verb, a dictionary of native English was used as a reference.[53]

Literal and *metaphorical* ID tag levels are illustrated below in (41), (42), (43) and (44):

(41) Without an awareness of your liberty and freedom to act, you *cannot* acknowledge the absurdity of life (ICLE-BR-SUR-0014.1) (*literal*)

(42) A scientist *may* discover how to make a plant grow (ICLE-ALEV-0027.8) (*literal*)

(43) Winning the Lottery jackpot *can* often run peoples' lives (ICLE-ALEV-009.5) (*metaphorical*)

(44) He *may* have been weeded out during the first season of play (ICLE-US-SCU-0006.1) (*metaphorical*)

The VerbType variable

The VerbType variable marks the types of lexical verbs used alongside modal forms. Consider for instance the following sentence in (45) where the verb *begin* is encoded by the VerbType variable:

(45) people *may* begin to think before acting (ICLE-US-MICH-0033.1)

Conceptually, the VerbType variable follows Vendler (1957) in its recognition that the notion of time is crucially related to the use of a verb and is "at least important enough to warrant separate treatment" (p. 143). Vendler (1957) identifies four types of verbs, which are included here as ID tag levels for the Verb Type variable. Those levels are

[53] Ambiguous cases were dealt with using the Merriam-Webster online dictionary. In cases where both types of use could apply to a single lexical verb, the decision to code that particular verb as *literal* or *metaphorical* was based on the examples provided in the dictionary and their resemblance with the corpus occurrences.

listed in Table 20 below:[54]

Table 20 The VerbType variable and its ID tag levels

Type of variable	Variable	ID Tag levels
semantic	VerbType	state
		accomplishment
		achievement
		process

The VerbType variable comprises four types of verb, namely *state*, *accomplishment*, *achievement* and *process*. This verb classification distinguishes between time periods and time instants on the one hand and uniqueness/definiteness and non-uniqueness/ indefiniteness of those time periods and time instants on the other hand. As Vendler (1957: 146) notes, "some verbs can be predicated for single moments in time, while others can be predicated for shorter or longer periods of time". In that respect, accomplishment verbs encode verbal statements that imply a unique and definite time period and achievement verbs encode verbal statements that imply a unique and definite time instant. Similarly, process verbs identify statements that reflect non-unique and indefinite time periods and state verbs identify statements that reflect non-unique and indefinite time instants. Each verb type is illustrated below in (46), (47), (48) and (49), starting from accomplishment and achievement types of verbs, then followed by the process and state types.

(46) a student *can* graduate from college as a doctor of science, and still not be as important, nor make as much money as a professional football player (ICLE-US-SCU-0008.2) - *accomplishment*

Graduate in (46) denotes an event that has a climax or that involves a stage of completeness. The verb predicates a period of time during which the action develops to ultimately reach a climax. It follows that as an *accomplishment* type of verb, *graduate*

[54] The *process* verb type in Table 20 is what Vendler (1957) refers to as *activity* type-verbs. I deviated from Vendler's terminology to include cases where the existence of an agent is implied (i.e., *activity* cases) as well as cases where the existence of an agent is not necessarily implied (i.e., *process* cases).

does not go on for an indefinite period of time. Comparatively, *die* in (47) does not predicate a definite period of time but rather a definite time instant. Indeed, the verb in (47) refers to a specific moment in time during which death occurs. In that respect, *die* in (47) illustrates a case of *achievement* type of verb.

(47) he wants to be awake because if he sleeps, he *may* <u>die</u> (ICLE-FR-UCL-0049.2) - *achievement*

(48) at first he *cannot* <u>use</u> his body, he behaves like a reptile (ICLE-FR-UCL-0070.2) - *process*

(49) Europe 92 *may* <u>mean</u> the birth of a new economic power, but not of a cultural nation (ICLE-FR-UCL-0104.1) - *state*

As a process verb, *use* in (48) predicates an action that takes place continuously over a period of time but for an indefinite length of time. Similarly, *mean* in (49), predicates a string of unspecified time instants.

The VerbSemantics variable

Like the VerbType variable, the VerbSemantics variable targets lexical verbs used alongside modal forms and identifies the type of information that they convey in terms of abstraction, action, communication, etc. Like much existing work using behavioural profiles, the internal organisation of this variable results from a careful bottom-up approach rather than any particular theoretical framework. Table 21 below lists all identified types of verbal information yielded in the corpus data.

Table 21 The VerbSemantics variable and its ID tag levels

Type of variable	Variable	ID Tag levels
semantic	VerbSemantics	abstract
		action general
		action motion
		action transformation
		communication
		copula
		mental/cognitive/emotional
		perception

Examples (50), (51) and (52) illustrate *abstract, action general* and *mental/ cognitive/emotional* ID tag levels, respectively.

(50) for we *may* also let our imagination wander, disregarding the external concrete reality that imprisons us (ICLE-FR-UCL-0036.3) - *abstract*

(51) a mother who works *can* come home, can whip up something in minutes in the microwave (ICLE-US-MICH-0041.1) - *action general*

(52) Her search for the final touch *can* be seen as a search for harmony (ICLE-FR-UCL-0039.2) - *mental/cognitive/emotional*

(In)animacy-related variables

The study includes two animacy-related variables, namely RefAnim and AnimType. Both variables apply to the referents of the subjects of the investigated modal forms.

The variable RefAnim

RefAnim variable identifies whether the referent of the subject of *may*, *can* or *pouvoir* is *animate* or *inanimate*, as Table 22 shows.

Table 22 The variable RefAnim and its ID tag levels

Type of variable	Variable	ID Tags
semantic	RefAnim: subject referent animacy	animate
		inanimate

Examples (53) and (54) below illustrate cases of *animate* and *inanimate* subject referents, respectively.

(53) if you still hesitate whether even today people *may* dream or even be imaginative or not, you ignore the existence of art (ICLE-FR-ULG-0002.1)

(54) this picture *may* frighten some of us, reassure others or let a great deal indifferent (ICLE-FR-UCL-0057.1)

The variable AnimType

AnimType identifies the types of (in)animacy yielded by the subject referents. The included ID tag levels for this variable are shown in Table 23. Given the large number of ID tag levels, individual examples for each level are not discussed in detail. Instead, Table 23 includes an 'example' column that illustrates each tag with a word or a phrase extracted from the data.

Table 23 The variable AnimType and its ID tag levels

Type of variable	Variable	ID Tag levels	Examples
semantic	AnimType: subject referent animacy type	animal	*birds*
		flora	*plant*
		human	*people, guy*
		imaginary being	*fictional beings, character*
		nationals	*Americans, Europeans*
		social role	*shop owners; scientists*
		(pseudo) cleft structure	*it may be predicted that*
		absence	*nothing*
		abstract	*cultural differences, problems, power*
		action	*reading, prayer*
		dummy '*it*'	*it may not sound very patriotic*
		effect	*consequences, results*
		form/substance	*drugs, radioactive materials*
		group	*Parliament, committees*
		measure	*majority, doses*
		mental/emtional	*consciousness, imagination*
		natural entity	*crops, egg*
		object/artefact	*computers, missiles*
		place/time	*1993, countries*
		process	*changes, progress*
		scholarly work	*essay, chapter*
		social conventions	*constitution, tax rates*
		state	*existence, knowing*

In inanimate cases, *process* and *action* differ in that actions are intentionally motivated and processes are not.

The SubjRefNumber variable

As shown in Table 24, the SubjRefNumber variable encodes the number of the subject referent of the investigated modal forms in terms of *singular* and *plural*. Examples (55) and (56) provide illustration of singular and plural, respectively.

Table 24 The SubjRefNumber variable and its ID tag levels

Type of variable	Variable	ID Tag levels
semantic	SubjRefNumber: subject referent number	singular
		plural

(55) the effect *may* be to stop people eating beef (ICLE-ALEV-0003.9)

(56) even though events *may* shape our destiny, we should never let neither technology nor science hold our future in its power (ICLE-FR-UCL-0052.3)

5.3.2 Syntactic variables

The syntactic variables Neg and SentType encode the data in a straightforward fashion according to the ID tag levels listed in Table 25.

Table 25 Syntactic variables and their ID tag levels

Type of variable	Variable	ID Tag levels
syntactic	Neg (negation)	affirmative
		negation
	SentType (sentence type)	declarative
		interrogative
	ClType (clause type)	main
		subordinate
		coordinate

The variable CLTYPE, which locates the investigated modal form in a *main* clause, *subordinate* clause or *coordinate* clause, presented numerous cases of embedded clause types. Examples (57) and (58) illustrate cases where a coordinate clause is embedded with a subordinate clause.

(57) money is incapable of being evil [since[it is a mere piece of paper or coins] and [*can not* make judgements and decision]] (ICLE-US-IND-0026.1)

(58) His logic is that [he *can* do the impossible], [have the moon in his hands], [he *can* change the order of the cosmos] and people will really be happy and immortal (ICLE-BR-SUR-0007.1)

In cases such as (57) the modal form was coded as *coordinate*. (58) is an ambiguous case where the modal form is either at the start of a new clause or covert co-ordination of the bracketed parts. Cases such as (58) were treated as co-ordination despite the subordinate clause level.

5.3.3 Morphological variables

The variable FORM encodes the investigated modal forms. As Table 26 below shows, the study investigates five modal items across the three sub-corpora: *can* in native and learner English, *may* in native and learner English and *pouvoir* in native French. The annotation of the modal forms includes negated forms, as Table 26 shows.

Table 26 The FORM variable and its ID tag levels

Type of variable	Variable	ID Tag levels
morphological	FORM	*can* interlanguage
		can native
		may interlanguage
		may native
		pouvoir

Subject-related morphological variables

Subject-related morphological variables are SubjMorph, SubjPerson and SubjNumber. These variables encode subjects according to the grammatical category they belong to (e.g. adjective, relative pronoun), number and person. Table 27 shows all individual levels.

Table 27 Subject-related morphological variables and their ID tag levels

Type of variable	Variable	ID Tag levels
morphological	SubjMorph: subject morphology	adjective
		adverb
		common noun
		date
		demonstrative pronoun
		interrogative pronoun
		noun-phrase
		proper noun
		quantifier
		relative pronoun
		subject pronoun
	SubjPerson: subject person	one
		two
		three
	SubjNumber: subject number	singular
		plural

Examples (59) and (60) illustrate the SubjMorph variable. (59) shows a *relative* pronoun ID tag level and (60) shows a *common noun* level.

(59) the parallel that *can* be drawn here is obvious (ICLE-FR-ULG-0022.2)

(60) desire *cannot* be measured on some utilitarian scale of pleasure (ICLE-US-IND-0005.1)

With regard to SubjPerson, the *one* ID tag level refers to a first-person subject pronoun or noun, The *two* ID tag level refers to a second-person subject pronoun or noun and the level *three* refers to a third-person subject pronoun or noun. The SubjNumber variable reflects whether the subject pronoun or noun is singular or plural. Example (61) below illustrate both the *one* and *plural* ID tag levels:

(61) we *can* wonder whether there is still a place for dreaming and imagination (ICLE-FR-ULB-0015.1)

Number is treated both semantically, with the SubjRefNumber variable, and morphologically, via the SubjNumber variable. This distinction allows for the treatment of cases where a singular form and a plural referent both feature in the same occurrence.

Other morphological variables

Other morphological variables are: Elliptic (i.e. when the lexical verb used with the modal form is used anaphorically and therefore is not morphologically present directly after the modal), Voice, Aspect, Mood. Table 28 presents the morphological variables and their respective ID tag levels.

Table 28 Other morphological variables and their ID tag levels

Type of variable	Variables	ID Tag levels
morphological	Elliptic	yes
		no
	Voice	active
		passive
	Aspect	perfect
		perfective
		progressive
	Mood	indicative
		subjunctive

A case is coded 'yes' for the Elliptic variable in cases such as (62) below where the modal form *can* is semantically linked to the lexical verb *do*:

(62) he has <u>done</u> all he *can* to show them the way to freedom (ICLE-BR-SUR-0007.1)

Examples (63) and (64) illustrate the *passive* and *progressive* ID tag levels of the Voice and Aspect variables, respectively.

(63) We have to make a balance between material comfort and inner happiness, which I think *can* only<u> be found</u> in our mind, where everything starts (ICLE-FR-ULG-0010.1)

(64) now that you have read this and *may* <u>be trying</u> to figure me out, I will tell you about myself (ICLE-US-IND-0018.1)

The variable Mood is only applicable to the French data which equally includes occurrences of *pouvoir* in the indicative and subjunctive modes, as illustrated in (65) and (66), respectively.

(65) en produisant deux fois plus vite, on *peut* produire deux fois plus (L1FUEF 15)
'by producing twice as fast, one *can* produce twice as much' [my translation]

(66) je ne pense pas *que nous puissions* envisager le monde d'une telle façon (FUMF 32)
'I don't think the world *can* be envisaged in such a way' [my translation]

5.3.4 Other variables

The final two variables are the Corpus and the GramAcc (grammatical acceptability) variables. Corpus encodes which corpus the occurrence has been retrieved from (i.e., native English, native French or French-English interlanguage). GramAcc codes whether I intuitively judged the sentential context of the occurrences to be grammatical or not. Table 29 and Table 30 present a breakdown of the ID tag levels for CORPUS and GRAMACC.

Table 29 The CORPUS variable and its ID tag levels

Type of variable	Variable	ID Tag levels
data	CORPUS	native
		interlanguage
		French

Table 30 The GRAMACC variable and its ID tag levels

Type of variable	Variable	ID Tag levels
data	GRAMACC (ACCEPTABILITY)	yes
		no

Example (67) illustrates an occurrence of *can* annotated as not grammatical. That coding decision is based on the fact that by using *can not* instead of *cannot* or *can't*, the speaker implies a negation of the proposition *break this program* rather than a negation of the modality expressed by *can*. However, the overall meaning of the sentence suggests the opposite interpretation, that is the negation of the modal.

(67) The back bone to a computer is its program, it *can not* break this program (ICLE-ALEV-0009.6)

5.4 The statistical analysis of the data

Two main statistical approaches were employed for the data analysis: a monofactorial approach that assesses the dependent variable FORM in relation to all individual predictors (henceforth I use the terms *predictor* and *independent variable* as equivalents), and a multifactorial approach that explores the effects of the independent variables, their interactions with Corpus and their effects on Form. A variable *interacts* with Corpus if it yields differing patterns of behaviour for *may* and *can* across the native and the non-native English data.

For both the monofactorial and multifactorial approaches, several statistical tests were employed to assess the behaviour of *may* and *can* as (i) contrasting individual lexical items (across the English data, native and non-native alike), and (ii) as a pair of lexical items whose behavioural patterns in French-English IL contrast with those in native English. In the case of (i), lexical items were compared statistically on the basis of individual behavioural profiles computed for the purpose of the analyses. Broadly, the combination of the BP annotation scheme and statistical techniques enables this single study to (i) provide fine-grained descriptions of the uses of *may* and *can*, (ii) generate hypotheses regarding learners' motivations to use *may* or *can*, and (iii) predict learners' uses of *may* and *can*. As can be inferred from Section 5.3, the total number of data points resulting from the annotation of this data is too large to make sense of with the naked eye. In order to identify co-occurrence patterns in the data it is necessary to make use of sophisticated statistical methods and I exploited the dataset as much as possible by combining a variety of statistical techniques (i.e., exploratory methods such as hierarchical cluster analysis and confirmatory methods such as binary logistic regression) that are not normally applied alongside each other in a single study. As this study shows, a primary benefit of employing multiple statistical techniques is that it allows for different kinds of analysis as well as corroboration across tests. Finally, an additional benefit of this approach is that it provides a way to assess the degree of appropriateness of each statistical technique for the purpose of a grammatically-grounded quantitative analysis of the modals.

In what follows, I first briefly define the nature of Behavioural Profile vectors (Section 5.4.1), then describe the monofactorial and multifactorial statistical tests selected for this study in Sections 5.4.2 and 5.4.3, respectively.

5.4.1 Description: Behavioural Profile vectors

While the BP approach provides a solid annotation scheme for exploratory data investigation, it also provides the necessary underpinning for confirming hypotheses. For example:

> The (…) BP approach yields data on the basis of which reliable predictions can be made: the choice of one near-synonym over another can be predicted significantly better than chance would have it if BPs are used as the basis on which to compute (dis)similarity (Divjak 2010: 193)

In preparation for the multifactorial analyses (i.e., both for the HAC and the computation of the graphic representation of the data), I used Gries' (2010a) R script Behavioral Profiles 1.01 and computed behavioural profiles for each modal form's occurrences in each language variety (i.e., *can*$_{\text{native}}$, *may*$_{\text{native}}$, *can*$_{\text{il}}$, *may*$_{\text{il}}$ and *pouvoir*) and in relation to the identified semantic and morphosyntactic predictors. Conceptually, a behavioural profile refers to a "comprehensive inventory of elements co-occurring with a word within the confines of a single clause or sentence in actual speech or writing" (Divjak and Gries 2009: 277). Statistically, such profiles are vectors of co-occurrence percentages of a single modal form with all individual independent variables' levels. Profiles provide form-specific summaries of their semantic and morphosyntactic behaviour in each sub-corpus.

5.4.2 Monofactorial statistical tests

In order to find the degree of statistical significance for each independent variable in relation to Form and to assess the size of its effect, I used the R `table.plotter` function (Gries 2009), a comprehensive function that computes a number of statistical tests. The selected output values for the purpose of this study are chi-square values and their matching *p*-values, degrees of freedom and Cramer's *V* coefficients. I used chi-square values (and their matching *p*-values) to assess whether distributional differences between Form and the individual predictors were significant, that is not attributable to chance. The chi-square results were interpreted on the basis of matching *p*-values which were used to accept or reject independent variables as statistically significant or not. A *p*-value that is larger than 0.001 is considered *highly significant*. A *p*-value that is $0.001 \leq p < 0.01$ is considered as *very significant*, a *p*-value that is $0.01 \leq p < 0.05$ is considered *significant* and a *p*-value that is $0.05 \leq p < 0.1$ is considered *marginally significant.* Finally, the Cramer's *V* value was used as a correlation coefficient to assess the strength of the association between Form and all individual predictors. In other words, the Cramer's *V* value quantifies the effect size of an observed correlation. Cramer's *V* values range from 0 to 1 where 0 indicates no correlation and 1 indicates a perfect correlation. The larger the effect size the more likely it is that the correlation is linguistically relevant. Importantly, as an effect size, sample size does not affect Cramer's *V* values: observed correlations between values are quantified independently of the size of the corpus (see Gries to appear, for a detailed description on the computation of effect sizes).

5.4.3 Multifactorial statistical tests

The multifactorial analysis consists of two main steps, namely (i) the statistical assessment of the modal forms' individual profiles with a HAC (as is customary in many BP approaches), and (ii) the computation of a binary logistic regression (as is customary in many alternation studies) in order to model the uses of *may* and *can*. Below I describe each method in turn.

Exploratory statistical techniques such as hierarchical cluster analysis (HAC) provide a tool to organise large data sets by finding dissimilarities between investigated elements and by grouping similar elements together. Techniques like HAC are hypothesis-generating in nature and null-hypothesis significance testing may, but also may not, be applied (Divjak 2010).

In terms of data requirements, the HAC technique requires the data to be turned into a co-occurrence table and summarised in the form of behavioural profile vectors. Following Gries and Otani (2010), I computed several cluster analyses: one involving all variables that the use of the modals were annotated for, one for only the syntactic variables, and one for only the semantic variables. In keeping with previous studies (e.g. Divjak and Gries 2006), I chose the Canberra metric as a measure of (dis)similarity and Ward's rule as an amalgamation strategy. Generally, the HAC method assumes that each cluster originally consists of one single element cluster and in turn, all elements are successively merged into larger and larger clusters (see Gries 2009: Section 5.5, for a detailed description of the HAC clustering process). The output of a HAC analysis is a dendrogram featuring clusters that exhibit high intra-cluster similarity and low inter-cluster similarity and which are, ultimately, all part of a single cluster, the original data set. The purpose of computing a HAC in this study was to explore the cross-linguistic similarity and the differences between *may*, *can* and *pouvoir* and to establish degrees of similarities between the three forms on the basis of a large number of contextual clues. To carry out the cluster analyses, I used HCLUST in R 2.10.0. The cluster analyses were later validated on the basis of a resampling scheme carried out with the R function PVCLUST (see Section 6.2.1). Conceptually, resampling consists of sampling repeatedly and randomly, with replacement, from the entire data sample (Crawley 2007: chapter 8, Arppe 2008). As Arppe (2008: 149) notes, "[t]he purpose of the repeated resampling is to ensure that all the data is taken into account both in the training and fitting as well as the testing of the models."

In contrast with exploratory statistical methods, confirmatory methods, such as binary logistic regression, allow the analyst to focus on the dependent variable (i.e. *may* and *can* in the current study) and its relation to individual predictors. More concretely, in the current work, the computation of a binary logistic regression provides a way to establish the existence of possible correlations between the predictors and learners' modal choices. Crucially, the logistic regression allows the researcher to assess the extent of the impact of individual predictors on the dependent variable. It is important to note that, in terms of data format, the logistic regression technique requires a different type of data set from the HAC analysis. In the case of the logistic regression technique, the analysis requires that the data be formatted as a raw annotation data table in which all extractions are individually tagged. Furthermore, in terms of data distribution, the binary logistic regression makes no particular assumption except that the data points are independent of one another. Overall, given the nature of the present study, the logistic regression technique presents the best choice of statistical method for binary dependent variables (Gries 2009) and the suitability of this statistical method for the investigation of modality-related phenomena was recently demonstrated in Divjak 2009, a study on the interaction between aspect and modality in Russian. To carry out my analysis, I applied the logistic regression, using GLM (*Generalised Linear Model*) in R 2.10.0, to the English (i.e. native and non-native) corpus occurrences of *may* and *can* which as a semantic pair allow for a choice between modals.

The initial regression model included the following independent variables:

- FORM as the dependent variable with only two levels: *may* and *can*;
- GRAMACC, NEG, SENTTYPE, CLTYPE, SUBJMORPH, SUBJPERSON, SUBJNUMBER, VOICE, ASPECT, MOOD, SUBJREFNUMBER, SENSES, 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 most between L1 English and L2 English).

The logistic regression was then performed with the model selection process. This procedure discarded first insignificant interactions, then individual variables that were not significant and did not participate in a significant interaction. It is important to note that the variables included in the initial logistic regression model were selected with an eye to avoid collinearity problems which could have affected the outcome of the analysis. Finally, in the interest of validating the results, I applied the *leave-one-out* technique to make sure that my predictions were not obtained only when the training and test datasets were identical. The leave-one-out procedure is a cross-validation method that uses a single observation from the original dataset as the validation data and the rest of the observations as the training data. The procedure involves repeating the validation process in such a way that each observation in the data is used once as the validation data. The results of the logistic regression are discussed in Chapter 6.

5.4.4 Tests for lexical specificity: distinctive collexeme analysis

The distinctive collexeme analysis involves four steps:

1. identifying and recording the frequency of all collexemes in each investigated construction

2. identifying the frequency of each construction
3. submitting those frequencies to a Fisher exact statistical test
4. sorting the collexemes according to their distinctiveness value.

The statistical analysis is based on the frequencies presented in Table 31, which are subjected to Fisher-Yates tests.

Table 31 Table of the necessary frequencies for the computation of collexeme distinctiveness

	Collexeme *C* (lexical verb)	Verbs other than collexeme *C*	Row totals
Construction A (*can*)	a	b	a+ b
Construction B (*may*)	c	d	c + d
Column totals	a + c	b + d	total

The distinctive collexeme analysis (DCA) involves the computation of a frequency table such as Table 31 for each individual collexeme. In turn, the results in this table contribute to the computation of two additional input tables, one for each English variety (i.e., native and French-English interlanguage) and consisting of the raw frequencies of *may* and *can* with their respective lemmas of occurrence. I used R In order to compute the two input tables and then I subjected both input tables to Gries' (2007) Coll. analysis 3.2 program for the computation of the Fisher exact test. I present and discuss the results of the statistical analysis in Chapter 6.

Chapter 6 Results

The present study applies a range of statistical techniques which have proved fruitful in varying degrees. Overall, within the context of the current work, the binary logistic regression technique has emerged as the most powerful method in that its results point towards several fruitful directions in the field of interlanguage. Although the results of all the statistical techniques selected for this study are presented and discussed in this chapter, the regression results predominantly motivate the subsequent discussion of the results in Chapter 7. Throughout the current chapter it is important to keep in mind that the different statistical techniques used in this study utilise different sub-datasets. Table 32 below summarises which sub-datasets were included for each type of statistical analysis.

Table 32 Summary table of the sub-datasets included in each type of statistical analysis

Statistical technique	Sub-datasets included in the statistical analyses		
	Native English: LOCNESS (incl. British and US data)	Learner English: ICLE-FR	Native French: CODIF [55]
Monofactorial analysis with TABLE.PLOT	√	√	-
Multifactorial analysis: HAC	√	√	√
Multifactorial analysis: binary logistic regression	√	√	-
Distinctive collexeme analysis	√	√	-

[55] I address the question of the relevance of including *pouvoir* in the current study and to what extent it plays a part in the analysis and the interpretation of the data in Section 6.2.1 while discussing the results of the cluster analysis.

As Table 32 shows, throughout the current study I treated the British and American English sub-datasets together under the umbrella of native English. The decision to conflate both native English varieties was motivated by the desire to sharpen the contrast between native and learner language. Furthermore, although I recognise the potential research interest in distinguishing British and American English and investigating the potential (dis)similarities between British and American English in relation to learner English, the primary goal of the current work is to assess the relevance and the usefulness of the BP method for the investigation of learner language. In that respect, the inclusion of the additional distinction British vs. American native English would have been of secondary interest.

In terms of structure, this chapter reflects the order in which the statistical tests were carried out. The monofactorial results are presented first, followed by the regression results and finally the collexeme results. To maximise readability throughout this chapter, I have only included a selection of the graphic outputs of the analyses. In the monofactorial section of the results, graphic representation of the data is limited to two semantic variables which underwent a conflation process, SENSES and VERBSEMANTICS. In addition, the monofactorial section also includes summary graphs for all the semantic, morphological and syntactic variables. With regard to the logistic regression section, graphic representations are provided for the variables NEG and VERBSEMANTICS and graphic representations of the results involving all other variables can be found in the appendix.

6.1 Monofactorial exploration

6.1.1 Introduction

This section discusses in detail the results of the monofactorial analysis and shows the extent to which each semantic and morphosyntactic predictor contributes to the uses of *may* and *can*. Since the current study strongly argues that (advanced) interlanguage should be approached from the perspective of the interactions of its linguistic components during language production, a monofactorial analysis may come as a surprise. The motivations behind the inclusion of a monofactorial analysis are as follows. First, I aim to provide a full-fledged empirically-based description of the uses *may* and *can* in a way that has not yet been carried out in previous studies, but that should also allow for comparison with the previous studies all of which were monofactorial in nature. In addition, this is an approach adopted both in the first multifactorial corpus study on alternations (Gries 2003a) and one of the most recent studies (Arppe 2008). Secondly, it follows that with the monofactorial results I aim to identify factors that contribute to the uses of *may* and *can* and which, due to the methodological limitations of previous studies, have not so far been identified. In what follows I present and discuss the monofactorial results of the semantic, morphological and syntactic variables, in that order. While the monofactorial results of all the variables are listed below, I only discuss those of the variables that are statistically significant.

Overall, the monofactorial analysis with `table.plotter` yielded a total of eleven predictors with *highly significant p*-values, one *very significant* predictor, one *significant* predictor, two *marginally significant* predictors and only two non-significant predictors. This demonstrates that the monofactorial results support an empirically-grounded approach to the uses of *may* and *can*. Table 33 summarises the results for all independent variables. The table includes computed Cramer's *V* coefficient as an indication of the effect size of each predictor on the dependent variable FORM.

Table 33 Monofactorial results for _may_ and _can_ in relation to each independent variable

Predictor	Chi-square (_df_): sign.	Cramer's _V_		Predictor	Chi-square (_df_): sign.	Cramer's _V_
Senses	3252.25 _(7)_: ***	0,972		SentType	27.7 _(1)_: ***	0,090
SpeakPresence	3252.5 _(2)_: ***	0,972		GramAcc	21.76 _(1)_: ***	0,079
VerbSemantics	240.66 _(8)_: ***	0,135		Corpus	7.36 _(1)_: **	0,046
VerbType	191.04 _(3)_: ***	0,236		ClType	7.89 _(2)_: *	0,048
Aspect	142.44 _(2)_: ***	0,204		Elliptic	3.72_(1)_: MS	0,033
AnimType	169.49 _(22)_: ***	0,222		Use	0.83 _(1)_: NS	0,016
SubjMorph	103.87 _(10)_: ***	0,174		SubjNumber	0.02 _(1)_:NS	0,001
Neg	59.92 _(1)_: ***	0,132		SubjRefNumber	0 _(1)_: NS	0,001
RefAnim	57.04 _(1)_: ***	0,129				
SubjPerson	57.72 _(2)_: ***	0,135				
Voice	38.18 _(1)_: ***	0,105				

In addition to its statistical computations, `table.plotter` outputs a graphical representation of the overall behaviour of predictors in relation to Form. The visual output of the function includes the following information in the form of what Gries (2009: 177) calls a cross-tabulation plot:

i. the observed frequency of all individual ID tag levels of the independent variable under investigation, in relation to _may_ and _can_ individually;
ii. the preference or dispreference of individual levels for _may_ or _can_ (in blue and red respectively)_;_ and
iii. the relative size of the observed preferences and dispreferences (as represented by the physical size of the numbers, which is a function of the Pearson residuals).

Throughout the monofactorial exploration, and in cases of predictors including two or more levels, I used `table.plotter`'s visual output to check for the potential need to conflate ID tag levels. This was motivated by the aim to maximise the statistical power of individual levels and to facilitate, at a later stage of analysis, the identification of potential sources of statistical significance. Based on cross-tabulation plots, I conflated ID tag levels that shared all three of the following properties:

- similar conceptual information (e.g., in the case of AnimType, the levels *flora* and *animal* have in common that they refer to a *non-human* type of animacy);
- similar directional tendencies (i.e., two or more levels share a (dis)preference for *may* or *can*);
- similar effect sizes (i.e., observed shared (dis)preferences between two or more levels are of similar proportion).

Using this set of rules, five of the predictors were subjected to a conflation process (i.e., Senses, VerbType, VerbSemantics, AnimType, SubjMorph) involving different numbers of levels (between 2 and 4) as well as different numbers of conflation layers. That is, in these cases, two or more levels were conflated (e.g. *flora* and *animal* in AnimType) and then the resulting conflated tag (here *non-human*) was combined with (an)other tag level(s) (here *natural entity*). Below, I provide selected illustrations of the conflation process: however, those were ultimately integrated to the result descriptions of individual variables. I consider all significant predictors individually and discuss their results in detail. Semantic variables are treated first, followed by morphological and syntactic variables.

6.1.2 Semantic variables

Overall, the monofactorial results for the semantic variables clearly indicate that this type of predictor plays a crucial role in the uses of *may* and *can*, as they are observed to influence the behaviour of the two modals in several ways. Table 34 below presents an overview of the behavioural tendencies of *may* and *can* across all the levels of the semantic variables. In Table 34, the semantic variables are organised into three types according to the three sentential parts that they apply to, namely the modal form, the modalised lexical verb or the subject referent. In turn, within each type of variable, ID tag levels are grouped on the basis of their similarities in relation to their (dis)preference for *can* or *may*. For instance, across Senses and SpeakPresence, the two variables that apply to the modal forms *may* and *can*, the ID tag levels *dynamic* and *weak* are similar in that they both yield a dispreference for *may* and a slight preference for *can* (see Figures 1, 2 and 3 for graphic summaries of the behavioural tendencies of *may* and *can* across the levels of the semantic variables).

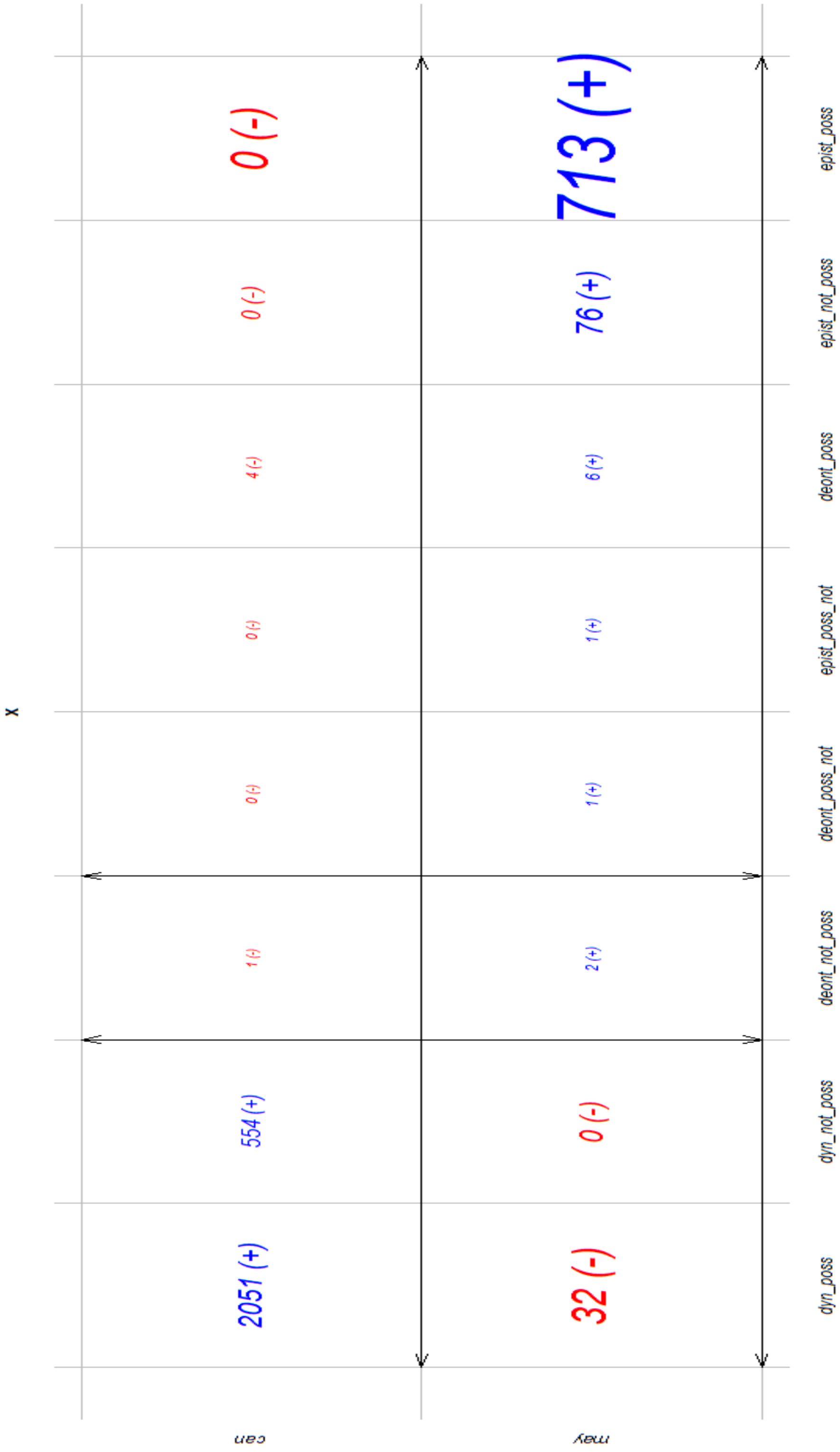
Table 34 Behavioural tendencies of *may*/*can* as a pair and in relation to the semantic variables

Type of semantic variable	ID tag levels	Behavioural description
Semantic variables that apply directly to *may* and *can*: - Senses - SpeakPresence	*dynamic, weak*	*may* and *can* behave identically with both levels, with a dispreference for *may* and a slight preference for *can*
	deontic, strong	*may* and *can* behave identically with both levels, with a dispreference for *can* and a stronger preference for *may*
	epistemic, medium	*may* and *can* behave identically with both levels, with a clear dispreference for *can* and a much stronger preference for *may*
Semantic variables that apply to lexical verbs modalised by *may* and *can*: - VerbType - VerbSemantics	*mental.perception, communication, process, accomplishment-achievement, action-general-motion*	*may* and *can* behave similarly in all those levels with a slight preference for *can* and a stronger dispreference for *may*. In the case of *action-general- motion*, preferential patterns are stronger than with the other levels
	copula, state	*may* and *can* behave relatively similarly with both levels with a clear dispreference or *can* and a very strong preference for *may*
	temporal, action-transformation	Both ID tags yield the same general tendency: a slight dispreference for *can* and a clear preference for *may* (although the preference for *may* in *temporal* is notably stronger than that of *action-transformation*)
	abstract	This ID tag level yields no clear preferential pattern
Semantic variables that apply to *may* and *can*'s subject referent: - RefAnim - AnimType	*animate, human, other*	*may* and *can* behave relatively similarly with all three levels with a slight preference for *can* and a stronger dispreference for *may*. In the case of *other*, the dispreference for *may* is stronger than in the other two levels
	inanimate, place.time, social.convention, natural.nonhum, mentemot, national.group, abstract	Generally, *may* is preferred and *can* is dispreferred. However, this preferential pattern is weak in the cases of *place.time* and *national.group*
	effect.state, linguistic, dynamic	*may* is very strongly preferred and *can* is clearly dispreferred

In what follows, I focus on individual variables and I present and discuss the behaviour of the two modal forms in relation to all significant semantic variables. In turn, I consider the following variables: Senses, VerbSemantics, VerbType, RefAnim, and AnimType.

Senses (along with SpeakPresence described below) yields the highest chi-square value (chi-square = 3252.25, $p < 0.001$) as well as the highest Cramer's *V* value (0.972) by far when all predictors are considered. Senses presents the first illustration of a conflation process. On the basis of the eight original levels (see Section 5.3), a first frequency count was carried out using `table.plotter`, and the cross-tabulation plot is represented below in Table 35.

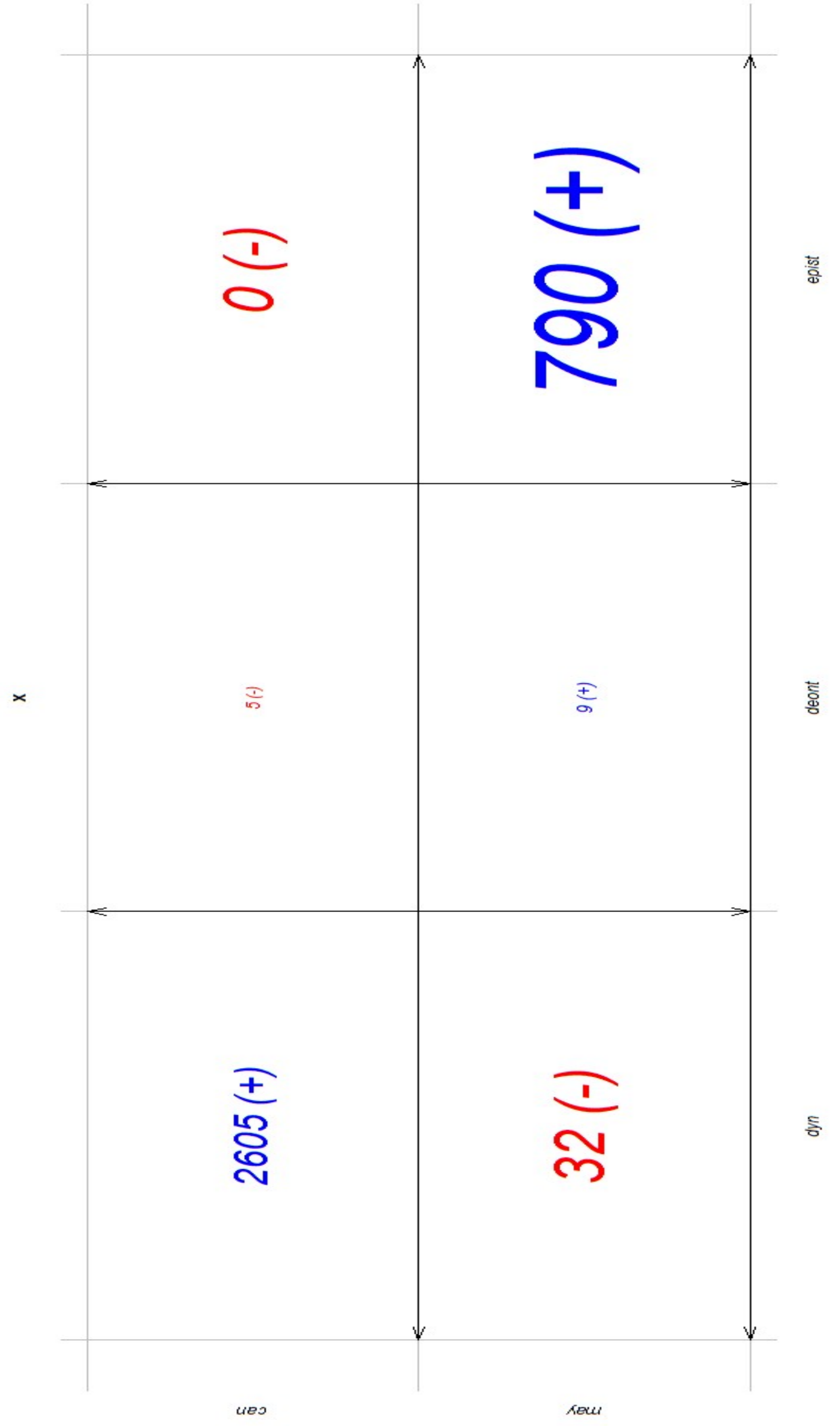
Table 35 Cross-tabulation plot for the semantic variable Senses before conflation



	can	may
epist_poss	0 (-)	713 (+)
epist_not_poss	0 (-)	76 (+)
deont_poss	4 (-)	6 (+)
epist_poss_not	0 (-)	1 (+)
deont_poss_not	0 (-)	1 (+)
deont_not_poss	1 (-)	2 (+)
dyn_not_poss	554 (+)	0 (-)
dyn_poss	2051 (+)	32 (-)

Table 35 highlights the overall dissimilarity between *may* and *can* in relation to the semantic variable SENSES. The general behavioural tendencies of the two modals can first be captured at a glance based on the colour code and the *(+)* and *(-)* signs on the figure: numbers in red followed by *(-)* indicate the number of occurrences of a particular modal (i.e., *may* or *can*) in which the modal form is dispreferred with a given ID tag level. Conversely, numbers in blue followed by a *(+)* sign indicate the number of occurrences of *may* or *can* in cases where the modal form in preferred with a given ID tag level. Table 35 shows that based on effect size (i.e., font size on the figure) *may* is strongly characterised by epistemic uses and *can* by dynamic uses. However, it is interesting to note that neither modal is characterised by its preferred senses in equal proportions. Indeed, Table 35 shows that although *can* and *may* yield a preference for *dyn_poss* and *epist_poss* respectively, effect sizes show that *can* is less characterised by its dynamic uses than *may* is by its epistemic uses. Furthermore, Table 35 indicates that negation contributes to the characterisation of each modal in different measures. This is based on the effect sizes of the levels *dyn_not_poss* and *epist_not_poss* and the fact that the effect size of the former is smaller than the effect size of the latter, thus suggesting that cases of negated epistemic modality contribute more strongly to the characterisation of *may* overall than cases of negated dynamic modality contribute to the characterisation of *can*. In addition, Table 35 shows that all deontic uses (i.e., *deont_not_poss*, *deont_poss_not* and *deont_poss*) behave very similarly, if not identically. Not only do they all indicate the same preference for *may*, they also do so in similar proportion. Further, semantically, all three levels share the notion of permission. This means that, conceptually, it makes sense to conflate them all, as opposed to, say, *deont_poss* and *epist_poss*. Similarly, the two dynamic levels and the three epistemic levels were conflated, respectively. It is important to note at this point that although *epist_pos_not* behaves similarly to the three deontic levels (i.e., *deont_not_poss*, *deont_poss_not* and *deont_poss*), it was not conflated with the deontic uses. This decision is based on the fact that, conceptually, deontic and epistemic uses are not compatible. This means that in the case of this particular conflation, conceptual similarity between conflated levels was judged to be more important than the forms' behaviour in relation to each of those levels. Ultimately, this explains why epistemic and deontic uses may seem to have been conflated differently. As shown further below, the variable VerbSemantics incurred the same type of decision at the conflation stage. Table 36 presents a graphic representation of the behaviour of Senses in relation to *may* and *can*, but this time, after conflation.

Table 36 Cross-tabulation plot for the semantic variable Senses after conflation



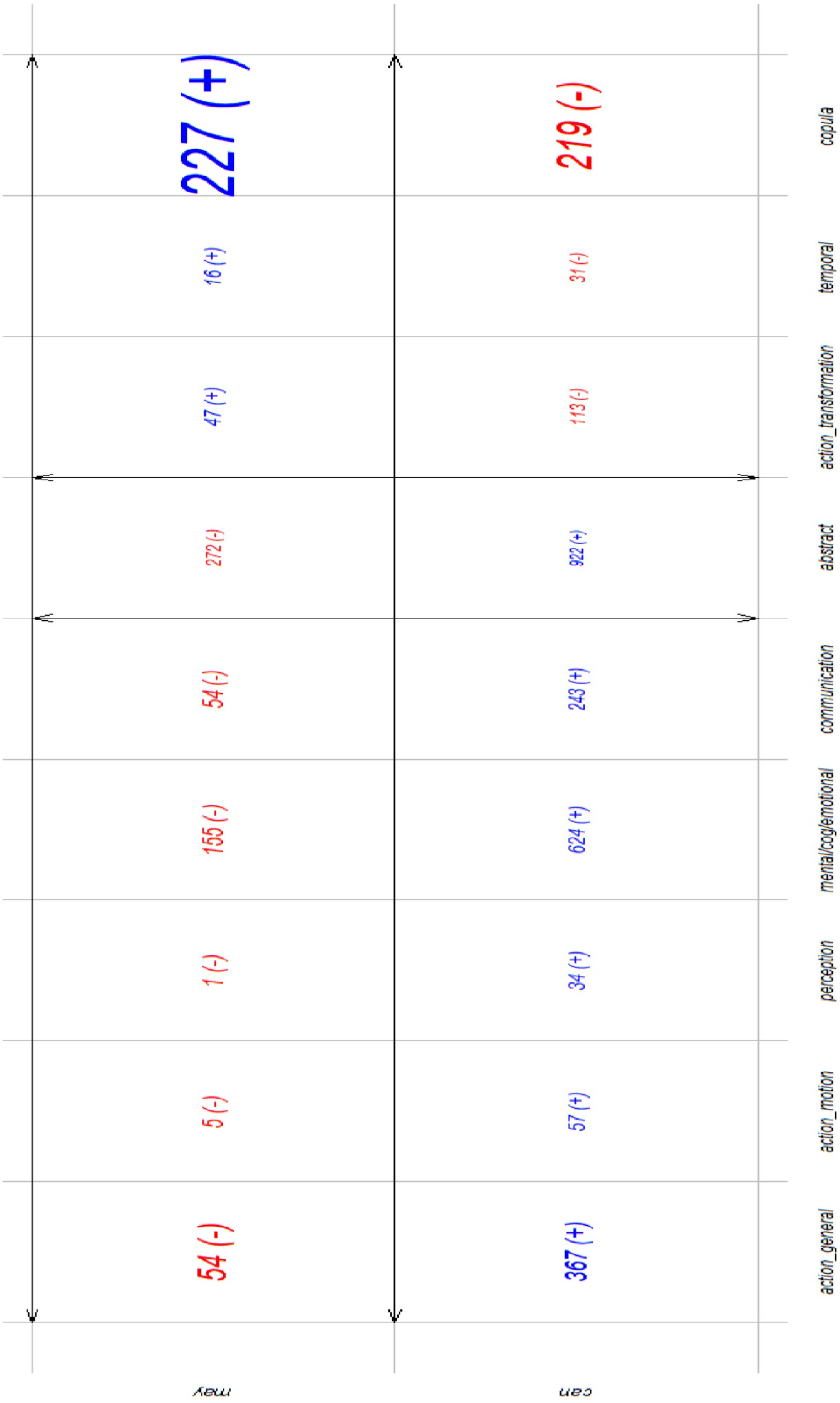
Overall, Table 36 reveals that deontic uses do not play a significant part in my corpus as for both *may* and *can*, deontic uses yield the smallest effect sizes. This result is in sharp contrast with the results for epistemic uses which yield the largest effect sizes in the table and the results for dynamic uses, which also yield relatively large effect sizes overall. As Table 36 confirms, after conflation, *may* and *can* sharply contrast semantically: while *may* is much preferred in its epistemic sense, *can* is clearly preferred in its dynamic sense. This result is very much in line with Coates's (1980, 1983) findings: for Coates (1983: 103), *may* is "most commonly used to express epistemic 'possibility'" and *can* is "most commonly used to express root 'possibility'". Although Coates (1983) recognises that the English modals share certain meanings and can be organised into semantic clusters, she generally rejects the synonymy of *may* and *can* by classifying the two forms into two distinct semantic groups. Despite her acceptance that, in some cases, the two forms may have overlapping meanings, she claims that even then, the two forms do not occur in free variation.

An additional interesting aspect emerging from Table 36 is the notion that the dispreferences of *may* and *can* play a non-negligible part in the characterisation of the two forms and in their distinction. Still on the basis of effect sizes, Table 36 shows, for instance, that *can* is more characterised by its dispreference for epistemic uses than its preference for dynamic uses, as the effect size the form yields in the epistemic is larger than the effect size it yields in the dynamic. Finally, after conflation, it is possible to say that the two modals are not semantically loaded in equal proportions and that although *can* is far more frequent throughout the data, epistemic uses are much more characteristic of *may* than dynamic uses are of *can.*[56]

Let us now consider the results for VerbSemantics, as represented in Table 37 below.

[56] As expected, the monofactorial analysis confirmed the high degree of correlation between the variables Senses and SpeakPresence. Given this correlation, the results of the monofactorial analysis SpeakPresence are not presented nor discussed in this chapter.

Table 37 Cross-tabulation plot for the semantic variable VERBSEMANTICS before conflation

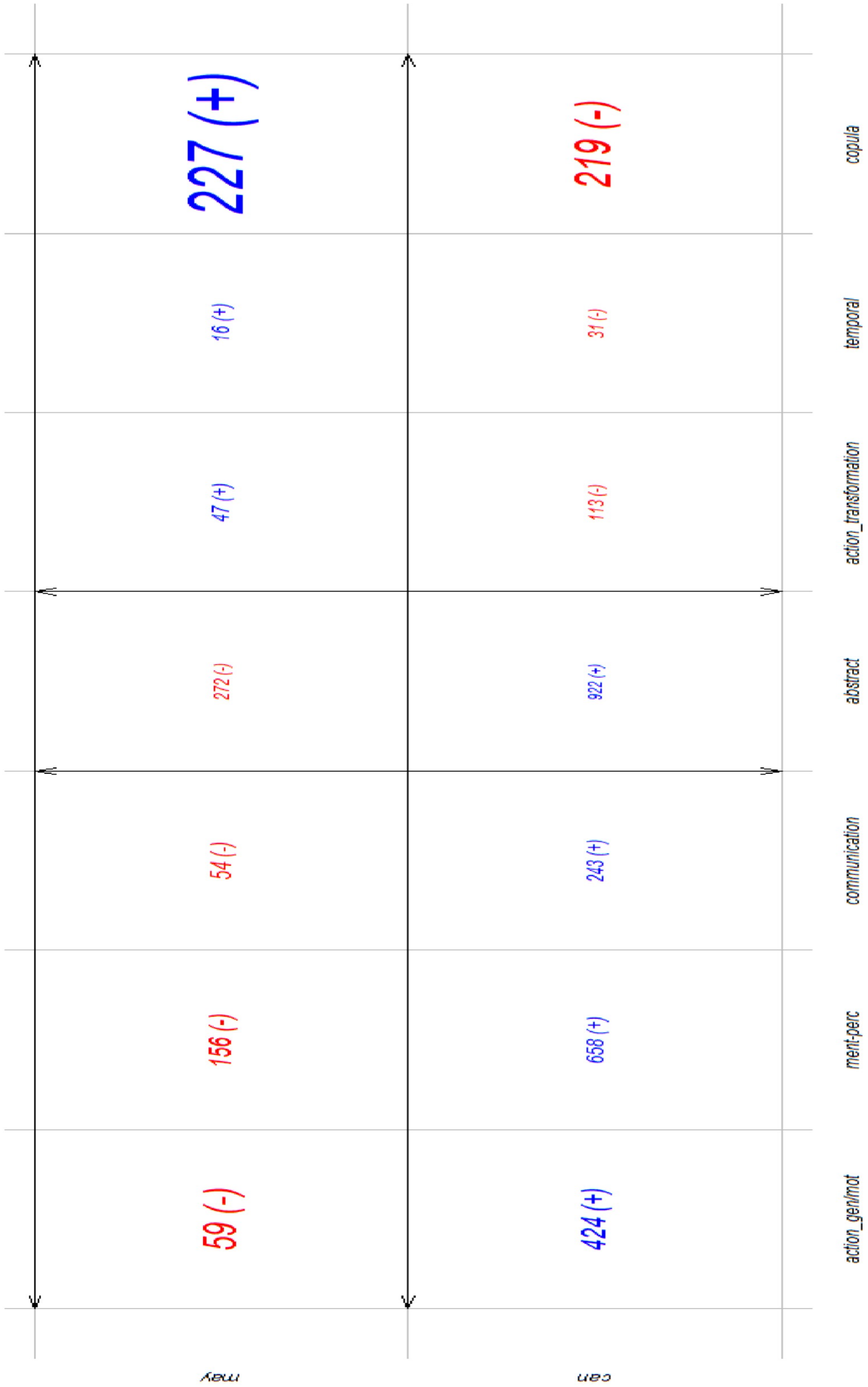


	copula	temporal	action_transformation	abstract	communication	mental/cog/emotional	perception	action_motion	action_general
may	227 (+)	16 (+)	47 (+)	272 (-)	54 (-)	155 (-)	1 (-)	5 (-)	54 (-)
can	219 (-)	31 (-)	113 (-)	922 (+)	243 (+)	624 (+)	34 (+)	57 (+)	367 (+)

Overall, Table 37 shows that from a monofactorial perspective, there is not a great deal of contrast between the uses of *may* and *can* across the variable VerbSemantics. This is based on the notion that despite their individual preferences for particular ID tag levels, in terms of effect sizes, *may* and *can* yield a general tendency towards homogeneous patterns of use. An exception is the case of *may* in relation to the ID tag level *copula*. As Table 37 indicates, *copula* strongly characterises the uses of *may* and stands out in the results table with the largest effect size. In contrast, *can* does not yield any equally proportioned preference for any ID tag level. As an additional point, it is interesting to note that Table 37 also shows that across all ID tag levels, *may* and *can* emerge with different profiles: on the one hand *can* is preferred with the majority of ID tag levels (i.e., six levels out of nine) with effect sizes that are relatively similar across levels and on the other hand, *may* is preferred only with the remaining three levels but, as mentioned above, is characterised more strongly than *can* through the large effect size of the ID tag level *copula*. This result suggests that in relation to VerbSemantics, *can* emerges as a possibly more generic lexical item while *may*, on the other hand, emerges as a more restricted and more specific lexical item.

VerbSemantics is the second variable to have been subjected to a process of ID tag level conflation. Two pairs of conceptually similar levels were identified as behaving similarly, namely *action_general* and *action_motion*, and *mental/cog/emotional* and *perception*. As shown in Table 38, the first pair was conflated into *action_gen/mot*, and the second pair into *ment-perc*.

Table 38 Cross-tabulation plot for the semantic variable VerbSemantics after conflation



	copula	temporal	action_transformation	abstract	communication	ment-perc	action_gen/mot
may	227 (+)	16 (+)	47 (+)	272 (-)	54 (-)	156 (-)	59 (-)
can	219 (-)	31 (-)	113 (-)	922 (+)	243 (+)	658 (+)	424 (+)

In a similar way to the Senses variable, the monofactorial treatment of VerbSemantics required a conflation process which incurred to decide to what extent the levels *action_general*, *action_motion* and *action_transformation* can be grouped together given that (i) conceptually, all three levels share the notion of 'action' and (ii) in terms of behaviour, the levels do not yield identical behavioural patterns. More concretely, although *action-general* and *action-motion* share their preference for *can*, effect sizes are larger for *action-general* and with regard to *action-transformation*. Further, although *action-transformation* yields similar effect sizes to those of *action_motion*, it behaves in the opposite direction with a preference for *may* rather than *can*. On the basis of its contrasting directionality, *action_transformation* was not conflated with the other two motion levels. Furthermore, as an action generally implied a motion more often than not, *action_general* and *action_motion* were treated as conceptually close enough to be treated together.

Overall, Table 38 confirms that after conflation, *may* has a strong preference for copula verbs. While both evidence of frequency and of effect size converge to give this conclusion, in the case of *can*, the cross-tabulation plot shows that lexical verbs denoting general actions and motions are preferred. Interestingly, Table 38 presents *may* as semantically more sensitive to its most characteristic preference (i.e., *copula*) than *can* is with regard to its most characteristic preference, *action_gen/mot*, due to the larger effect size of *copula* on *may* in comparison to the smaller effect size of *action_gen/mot* on *can*. Additionally, Table 38 can be used to illustrate the limitations of studies that depend solely on frequency counts (see Section 3.4.2). While *can* is most frequently used with lexical verbs denoting abstract processes, their linguistic effect on Form is not of great relevance. The most characteristic ID tag level in relation to VerbSemantics and the preferential patterns of *can* is in fact the level with the second lowest frequency count here (*action_gen/mot* with 424 occurrences).

May and *can* relate to the variable VerbType in interesting ways (see Table 45 in the appendix): although *may* has a clear preference for state verbs (e.g. *appear*, *contain*, *exist*, etc.) and *can* has preferences for both process verbs and accomplishment and achievement verbs, *may* is sensitive to the three types of verbs (i.e., *process*, *accomp-*

achievt and *state*) to different, and more obvious, degrees.[57] For instance, in terms of effect size, *may*'s preference for state verbs is much larger than its dispreference for process verbs which, in turn, is larger than *may*'s dispreference for accomplishment and achievement verbs. *Can*, on the other hand, is sensitive to the three levels in similar degrees: the effect sizes of *can*'s dispreference for state verbs and preference for process and accomplishment and achievement verbs are relatively similar. Generally, across VerbType, this result presents *can* with a more homogeneous profile than *may* and suggests that *can* is not as clearly influenced as *may* is by the type of verb that it modalises. More generally, it is interesting to note that overall and throughout the monofactorial analysis, *may* tends to show a higher of variability than *can* which in tends to yield more uniform behavioural patterns.

The RefAnim variable is statistically *highly significant* and thereby contributes to the difference of uses between *may* and *can.* However, due to small effect sizes, it does not yield much that clearly distinguishes the two forms (see Table 46 in the appendix). The biggest effect (although quite small overall) of RefAnim x Form is the dispreference of *may* for animate subjects, as opposed to *can*, which prefers animate subjects. Thus it emerges from the data that a major distinguishing criterion between *may* and *can* is that *may*'s largest effect sizes are regularly those that indicate its dispreferences, while *can*'s largest effect sizes are regularly those that yield its morphological and syntactic preferences. To some degree, this is related to the fact that *may* is less frequent than *can*.

As illustrated in Section 5.3, AnimType includes the largest number of original ID tag levels (i.e., 24). As a result, the variable was subjected to a substantial conflation process, as illustrated in Table 39 below.

[57] VerbType underwent a conflation process which brought together the two levels *accomplishment* and *achievement*.

Table 39 Conflation stages for the semantic variable AnimType

Conflations	ID tag levels to be conflated	ID tag levels conflated into
1	flora, animal	non-human
2	object/artefact, scholarly work	man-made
3	effect, state	eff/state
4	action, process	dynamic
5	absence, measure	quantity
6	dummy '*it*', dummy '*il*', cleft structure	ling
7	imaginary being, quantifier, form/substance, man-made	other
8	nationals, group, social role	natnl/group/socrole

This conflation reveals that AnimType is a crucial semantic feature for distinguishing the uses of *may* and *can*. The most significant effect between the two modals and the type of animacy of their grammatical subjects is *may*'s dispreference for human subjects, which is *can*'s most characteristic subject type, although with a smaller size effect compared with the effect size of *may*'s dispreference (see Table 47 in the appendix). As in the discussion of VerbType, we can see that *may* generally distinguishes itself from *can* by yielding a more varied range of effect sizes. Regardless of the direction of the form's preferences (i.e. whether a form prefers or disprefers a particular type of subject), five groups of effects sizes can be broadly identified. They are, from largest to smallest, one one-member group consisting of the level *hum,* a second group of *dynamic*, *ling* and *eff_state*, a third one-member group consisting of *abstract*, a fourth group of *ment/emot* and *nat/group/socrole*, and a fifth group with *nat-non-human*, *soc_conv* and *pl/time*. Comparatively, *can* yields a higher number of similar and smaller effect sizes. This suggests that, in relation to AnimType and in comparison to *can*, *may* exhibits a slightly more versatile profile. This result is also supported by the lack of a balanced distribution of all the ID tag levels across the two modal forms: overall, while *can* is preferred over *may* for two levels (i.e., *hum* and *other*), *may* is preferred over *can* for the remaining nine levels. This suggests that *may* has greater flexibility in that, unlike *can*, it occurs with a wide variety of referent animacy types.

At the level of ID tag, an obvious distinguishing feature between the two forms confirms the obtained result for RefAnim, namely that while *may* is dispreferred with animate subject referents, *can* is not. Interestingly, in cases where both *may* and *can* have a preference for inanimate subjects, the results indicate the extent to which types of inanimate subjects differ when used with one modal or the other. Overall, *may* is more versatile than *can* in the sense that *may* occurs with inanimate subjects denoting, for instance, a state (i.e., a state of affairs: *eff/state*, a state of mind: *ment/emot*), a process (i.e., *dynamic*) or a semantically empty linguistic item (i.e., *ling*). *Can*, on the other hand, can occur with all those, but is dispreferred with them.

Figure 1, Figure 2 and Figure 3 below summarise *may* and *can*'s (dis)preferences for the statistically significant semantic variables that apply to the modal forms, their modalised lexical verbs and their subject referent, respectively. The graphs quantify the degrees of (dis)preference of *may* and *can* for individual ID tag levels and they result from the computation, for each ID tag level, of the differences between the percentage of observed and expected frequencies of *can* on the one hand and the differences between the percentage of observed and expected frequencies of *may* on the other hand.

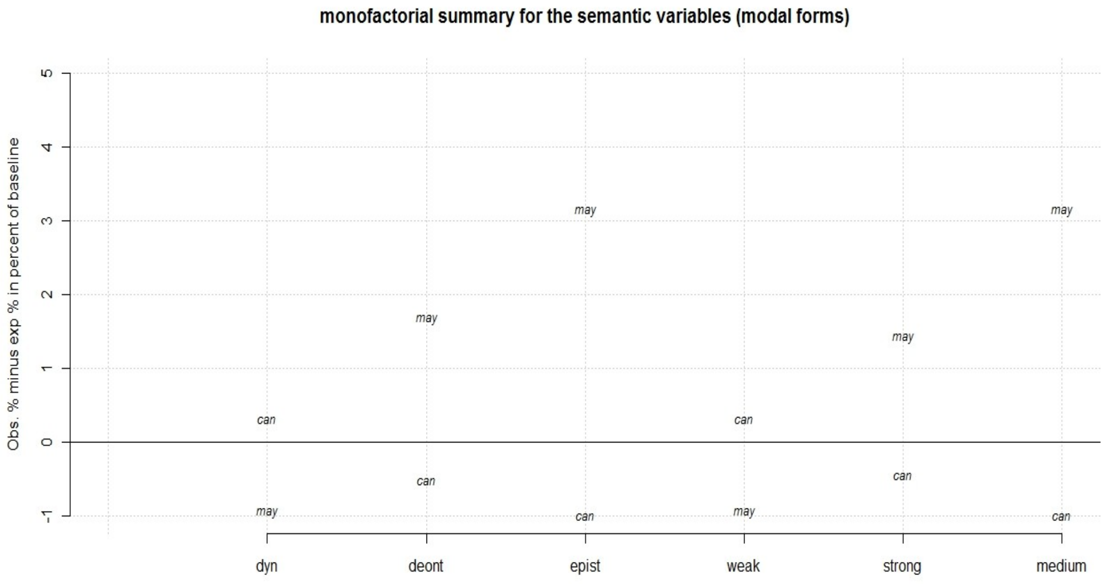


Figure 1 Monofactorial summary graph for the semantic variables that apply directly to *may* and *can*

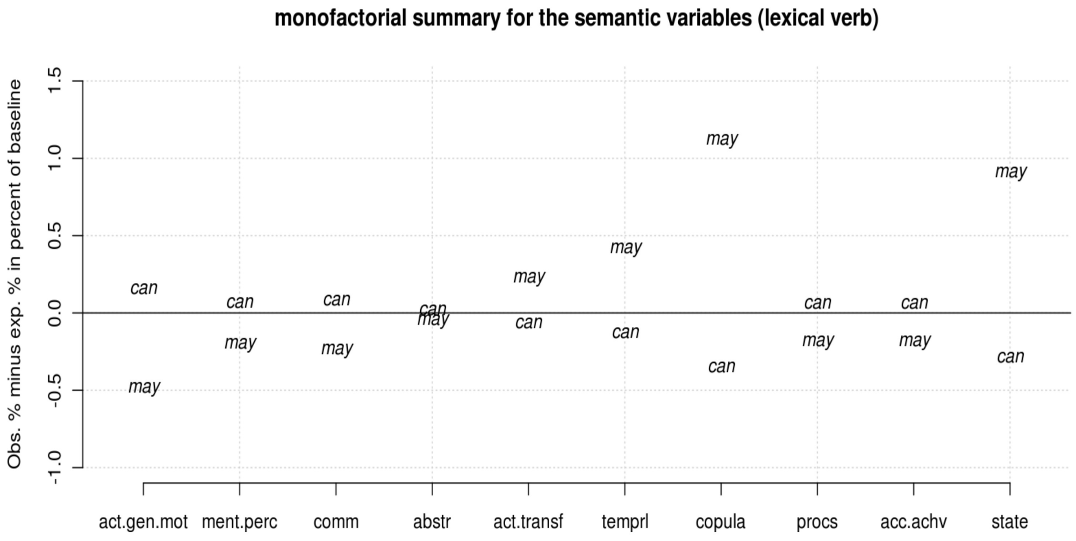


Figure 2 Monofactorial summary graph for the semantic variables that apply to lexical verbs modalised by *may* and *can* (incl. VERBTYPE and VERBSEMANTICS)

Figure 2 summarises the behaviour of *may* and *can* contrastively in relation to all the levels of the predictors VERBTYPE and VERBSEMANTICS. Generally, Figure 2 illustrates the non-equivalence of the two forms for the majority of ID tag levels: only four levels out of 10 yield similar preferential patterns for *may* and *can* (*ment.perc, comm., procs* and *acc.achv*). The levels according to which *may* and *can* contrast most clearly are (in order of the size of the difference, from the largest to the smallest): *copula*, *state*, *temprl, act.gen.mot* and *act.transf*. All those levels except *act.gen.mot* have in common that they feature *may* (over *can*) as their most characteristic form. Furthermore, although both *may* and *can* equally share the same number of levels, *can* tends to be preferred in far less extreme degrees than *may* is. Indeed, Figure 2 shows that in cases where *can* is preferred, it remains very close to the baseline as opposed to *may* which generally reaches much higher levels.

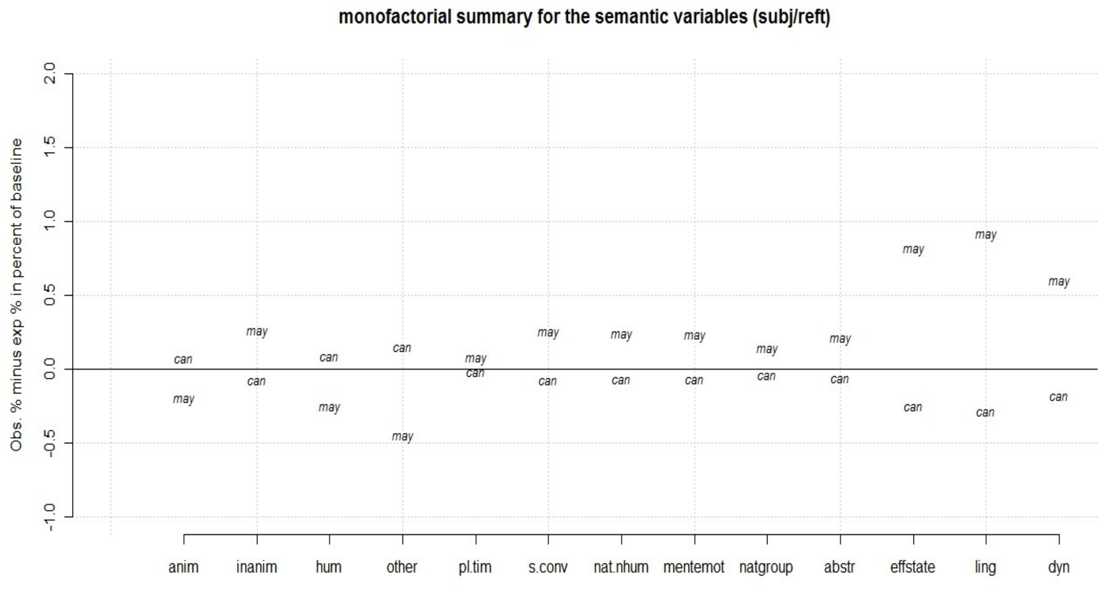


Figure 3 Monofactorial summary graph for the semantic variables that apply to *may* and *can*'s subject referent (incl. REFANIM and ANIMTYPE)

Figure 3 reiterates the lack of equal distribution of the AnimType-related ID tag levels between the two modal forms, as noted above in the discussion on AnimType, and shows that is true across all subject-related levels. Furthermore, for the majority of the levels, contrasts between the two forms are triggered by *may* rather than by *can* which tends to stay closer to the baseline and yields a more homogeneous behaviour.

Overall, the monofactorial results for the semantic variables indicate that *may* yields larger size effects than *can* in both directions (i.e., preferences and dispreferences). While this is correlated with the effect of *can*'s expected frequencies being larger, it is nevertheless suggestive. In addition, on the basis of results of variables including three or more levels and from the perspective of variables that apply to modalised lexical verbs, *can* yields more flexibility in the sense that it is preferred over *may* for a higher number of ID tag levels. In the case of VerbSemantics, for instance, *can* is the preferred choice for four out of the seven (conflated) ID tag levels and *may* is only preferred for the remaining three. VerbType illustrates a similar pattern. In cases where the variable applies to the subject referent of the modalised lexical verb, it is *may* that strikingly shows more flexibility. With AnimType, for example, *may* is preferred for nine ID tag levels, only two characterise *can*. Those behavioural tendencies suggest that semantically, both forms have different 'orientations' so-to-speak.

6.1.3 Morphological variables

In this section, I present the results for the variables: SubjMorph, SubjPerson, Voice and Elliptic, in that order. Table 40 summarises the behavioural tendencies of *may* and *can* across all the ID tag levels of the morphological variables.

Table 40 Behavioural tendencies of *may*/*can* as a pair and in relation to the morphological variables[58]

Type of morphological variable	ID tag levels	Behavioral description
- SubjMorph - SubjPerson - Voice - Elliptic	*perfective; non-elliptic*	The data yield no preferential behavioural patterns for *may* and *can*
	progressive; perfect	*may* and *can* yield their most extreme behavioural tendencies in relation to morphological variables with outstanding preferences for *may* and very strong dispreferences for *can*
	passive; elliptic; one; two; pronoun; proper noun; common noun	*may* and *can* behave relatively similarly in that *can* tends to be slightly preferred and *may* more strongly dispreferred. However, this behavioural pattern is less pronounced in the cases of *pronoun* and *proper-noun* and more pronounced in the case of *elliptic*
	other; relative-demonstrative pronoun	*may* and *can* behave identically with both ID tag levels: *may* is strongly preferred and *can* is slightly dispreferred
	active; three	*may* and *can* behave identically with both ID tag levels: *may* is slightly preferred and *can* is slightly dispreferred

The monofactorial results for SubjMorph illustrate the fine degree of granularity that the BP approach allows (see Table 49 in the appendix). For instance, they identify *may* and *can*'s preferences for specific types of pronouns, namely relative and demonstrative in the case of *may* and subject pronouns in the case of *can*. For the purpose of the current analysis, this distinction is important because *pr* on the one hand and *rel-dem_pr* on the other hand yield *may* and *can*'s largest positive effect sizes, respectively. Overall, and as already observed with regard to semantic variables, *may* yields much greater flexibility than *can* in that it is compatible with a wider variety of subject morphological types.[59]

In relation to SubjPerson, *may* again exhibits larger effects than *can* (see Table 50 in the appendix). While *may* yields a preference for third-person subject pronouns, *can*, on the other hand, is preferred with both first- and second-person pronouns. As is the case for other variables such as VerbType, the variable SubjPerson contributes more sharply to the characterisation of *may* than to that of *can*. That is indicated by larger differences

[58] see Figure 4 for a graphic representation of this table.

[59] After conflation, the level *other* includes: non-noun phrases, quantifiers and dummy 'there'.

between the effect sizes of each individual level, thus indicating which levels contribute the most (negatively or positively) to the characterisation of the modal. In the case of *may*, for instance, one can see that the level *one* gives the largest effect on the form. In the context of SubjPerson, a major characteristic of *may* is its dispreference for first person subjects.

Generally, Voice does not strongly discriminate between *may* and *can* (see Table 51 in the appendix). However, the monofactorial results confirm Gabrielatos and Sarmento's (2006) recommendation to take into account passive and active voice lexical verbs in the investigation of the uses of the modals (see Section 2.4.2 for discussion). With a highly significant *p*-value (*p*-<0.001, *chi-square*=38.18), Voice is a reliable variable to analyse behavioural differences between *may* and *can*. Overall, while *may* is the preferred form in active clauses, *can* tends to be selected in passive clauses.

This behavioural pattern brings up an issue raised in Hawkins (2004) and, more recently, in Hawkins and Buttery (to appear: 13), namely the correlation between frequency and sentential complexity. Hawkins and Buttery write that "[s]tructural complexity and frequency are generally inversely correlated in language use, i.e. the more complex a structure is, the less frequently it is used in general". This view leads us to hypothesise that *can* is selected over *may* as a default form in cases of more cognitively demanding structural contexts. This conclusion is based on the observation that the passive voice is structurally more complex than the active voice, and *may* yields the lowest number of occurrences in the passive. Taking this point one step further, still in the spirit of Hawkins (2004), one may infer from the above results that *may* and *can* differ in terms of the cognitive processing load that they incur and that, generally, *can* is less cognitively demanding than *may*. Hawkins argues that

> [processing] efficiency is increased by selecting and arranging linguistic forms so as to provide the earliest possible access to as much of the ultimate syntactic and semantic representation as possible (Hawkins 2004: 9)

It is thus conceivable that *can* provides a more effective lexical choice than *may* does because it minimises processing effort in the otherwise cognitively demanding passive voice. I return to this issue in more detail in Sections 7.6.1 and 7.6.2 where I discuss how the processing effort incurred by lexical items in L2 can contribute to the emergence of non-native linguistic patterns.

The monofactorial analysis reveals that the perfect aspect influences the behaviour of *may* and *can* more than the perfective and progressive aspects do (see Table 52 in the appendix). Based on effect size, the strongest effect is reflected in *may*'s preference for the perfect aspect. In relation to ASPECT, *may* and *can* are not sensitive to their (most) preferred ID tag levels in equal proportions: *may*'s preference for the perfect aspect is significantly more pronounced than *can*'s preference for the perfective aspect. The significance of this result should be considered with caution as there is a chance that *may*'s clear preference for the perfect may be influenced by its 'probability' sense which is intuitively more compatible with the perfect aspect than the 'possibility' sense of *can*. Indeed, although the sentence *He may have come* is perfectly acceptable, the sentence **He can have come* is not grammatical.[60]

Finally, although ELLIPTIC is statistically only marginally significant, the monofactorial results indicate that *can* is more likely than *may* to code the lexical verb it modalises (see Table 53 in the appendix). This is based on *can*'s (only slightly) larger effect size. *Can*'s (small) preference for occurring with elided verbs suggests that *can* is used alongside grammatical elements that are more difficult to process.

Figure 4 below summarises the behaviour of *may* and *can* in relation to SUBJMORPH, SUBJPERSON, VOICE and ELLIPTIC.

[60] In that respect, the result shown in the cross-tabulation plot for the morphological variable ELLIPTIC may be biased (see Table 53 in the appendix section).

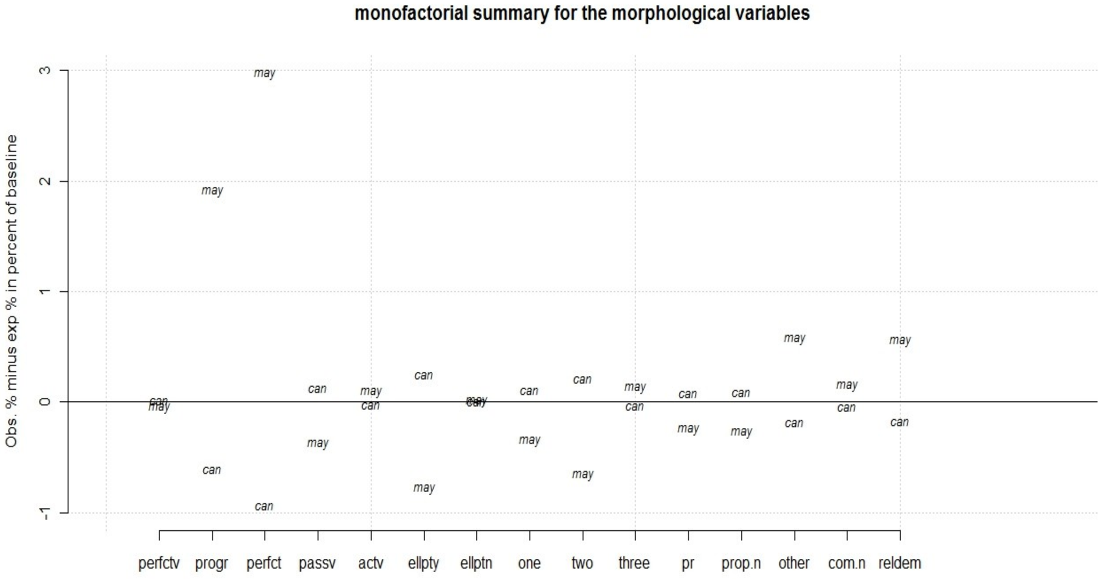


Figure 4 Monofactorial summary graph for the morphological variables (incl. SubjMorph, SubjPerson, Voice and Elliptic)

Aspect, by far, demonstrates the greatest degrees of within-predictor variability between the two forms and SubjPerson and SubjMorph, which both include three or more levels, confirm the need to operationalise predictors maximally in that the great majority of levels reveal a rich account of *may* and *can*'s multifaceted contrasting tendencies. The only exceptions are the two pairs of levels *pr* (i.e., pronoun) and *prop.n* (i.e., proper noun) as well as *comm.n* (i.e., common noun) and *three* within which *can* and *may* behave similarly. Finally, overall, and across morphological predictors, *can* yields, on average a constant behaviour whereas *may* outlines a much more uneven profile.

6.1.4 Syntactic variables

In this section, I present the results of the variables SentType, ClType and Neg, in that order.

Based on effect sizes, interrogative contexts strongly influence the uses of *may* and *can* (see Table 54 in the appendix). However, although such contexts influence *may*'s behaviour to greater extents than they influence *can*'s behaviour, once again unlike *can*, *may* stands out in its dispreference for interrogative contexts. This result should be interpreted carefully as it may be influenced by the very nature of the data. Indeed, while Leech (2004: 77) recognises that certain uses of *may* are only found in particular grammatical contexts and that "only the permission sense, for instance, is found in questions", Leech (1969: 75) notes that "[i]n asking and giving permission, *can* and *may* are almost interchangeable". Coates (1980: 103), however, stresses that "where there is overlap, *may* and *can* are not in free variation but *may* is marked for formality". Both such formality and the fact that, communicatively, it makes very little sense to ask for permission in the course of an essay writing exercise, may have a negative impact on the frequency of use of *may* in relation to SentType.

Compared to all other (statistically) significant variables, ClType yields (i) one of the highest *p*-values ($0.01 \leq p < 0.05$) and one of the lowest Cramer's *V* values (0.048). Generally, *can* is more frequently used than expected in subordinate clauses (see Table 55 in the appendix). *May*, on the other hand, prefers to occur in main and coordinate clauses – that is, in less complex and thus less cognitively demanding syntactic contexts.

To finish with Neg, while the distributional differences of the variable are statistically *highly significant*, results show that both the uses of *may* and *can* are sensitive to negated linguistic contexts (see Table 56 in the appendix): the effect sizes for both forms are larger in negated cases than they are in affirmative cases. However, both forms are sensitive in different directions. While *can* prefers negated contexts, *may* does not. Interestingly, *may*'s dispreference for negated contexts is stronger than *can*'s preference for such contexts. In other words, *may* is marked in the negative but more sharply and *can* is marked positively but with less strength. Finally, as for ClType and SentType, *may* is again preferred in less cognitively demanding syntactic contexts.

Figure 5 below summarises the behaviour of *may* and *can* with all of the above syntactic variables.

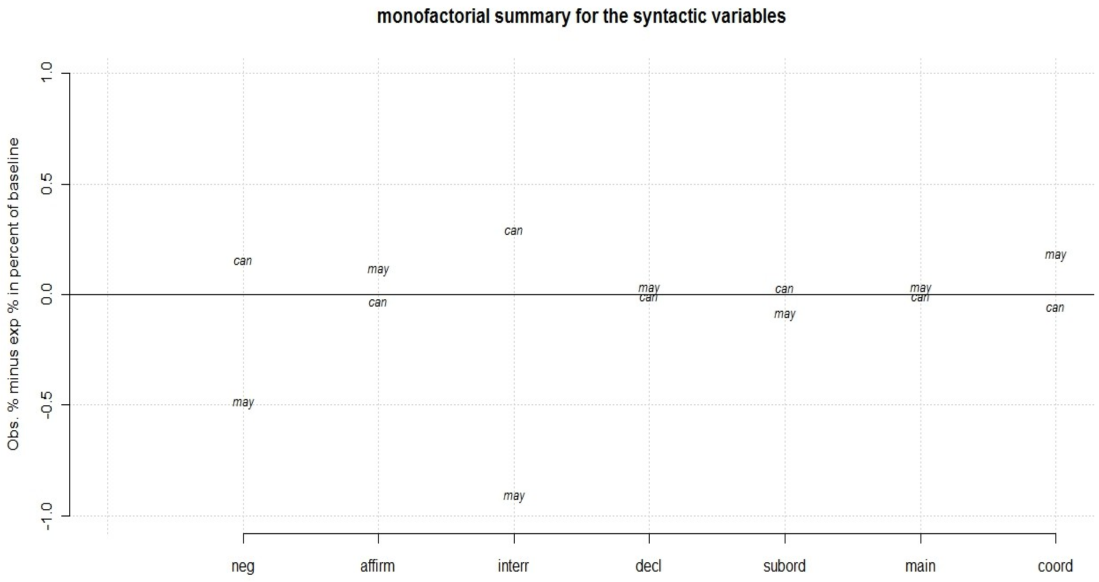


Figure 5 Monofactorial summary graph for the syntactic variables (incl. Neg, SentType and ClType)

Figure 5 suggests that the degree of syntactic complexity in which *may* and *can* occur plays a part in the speaker's choice of one form over the other and that while *can* tends to be selected in more cognitively demanding syntactic contexts, *may* is selected in less challenging syntactic environments. In the present work, syntactic complexity is defined on the basis of the grammar and the structure of the investigated native and learner language varieties and it is assumed, following Hawkins' (1999, to appear), that:

- grammars (implicational universals, hierarchies and distributional preferences) are conventionalizations of the patterns and preferences that one observes in the performance of languages with structural choices (between competing word orders, relative clause structures, morphological alternatives, etc.) (Hawkins to appear: 1),

and that:

- [t]he hierarchies and distributional preferences [of syntactic structures] across languages reflect […] degrees of processing difficulty (Hawkins 1999: 280).

Figure 5 shows that *may* and *can* display four types of behaviour that correspond to four degrees of syntactic complexity. In Table 41, I summarise the four levels from the most simple (i.e. level 1) to the most complex (i.e. level 4).

Table 41 Behavioural tendencies of *may*/*can* as a pair and in relation to sentential syntactic complexity

Complexity level	ID tag levels	Behavioural description
1	*declarative*, *main*	*may* and *can* behave identically
2	*affirmative, coordinate*, *subordinate*	*may* and *can* contrast in similar proportions; while *may* is preferred in *affirmative* and *coordinate* cases, *can* is preferred in more complex *subordinate* cases
3	*negation*	*may* is clearly dispreferred
4	*interrogation*	*may* is very clearly dispreferred and *can* is clearly preferred

As I develop in Section 7.6, the notion that speakers' lexical choices can be explained on the basis of complexity-based hierarchies of syntactic components is in line with Hawkins (2004), who argues that the shape of grammars correlates with language processing demands.

6.1.5 Other variables

The remainder of the section focuses on the variables GramAcc and Corpus, in turn.

Generally, the monofactorial analysis indicates that *can* is more likely to be found in erroneous linguistic contexts than *may* is (see Table 57 in the appendix). Since the unacceptable uses are from the learner part of the corpus data, any finding other than this would probably be surprising. Should *can* be used as a form to fall back on in case the context becomes difficult, then it is only natural that learners may fall back on *can* too often, which corresponds to the above result and also fits in with the data for the final variable. So far, considering individual predictors and their effects on Form has demonstrated the non-equivalence of *may* and *can* with regard to their characteristic affinities for particular contextual linguistic components. However, when all predictors are considered simultaneously to assess the forms' behavioural patterns in native and learner language, an informative result cannot be obtained (see Table 58 in the appendix). Generally, one can say that the two corpora diverge in that learners use more *can* and native speakers more *may.* Thus, from a native speaker perspective, learners overuse *can* to the detriment of *may.* However, the effect sizes are all rather small and therefore this conclusion is not very convincing. While monofactorial results have so far illustrated the non-equivalence of *may* and *can* in terms of their different preferences for particular linguistic contexts, both forms (and *may* particularly) have proved flexible in terms of the kinds of components they can co-occur with. Intuitively, one would expect such flexibility to be reflected through contrasting effect sizes of *may* and *can*. Note that it is exactly this type of approach that has dominated previous approaches to the modals in general and modals in SLA so far.

Let us therefore return to the desiderata listed in Section 4.1 and the discussion in Section 3.6.3 on Bates and MacWhinney's Competition Model with the particular attention to the notion that cues are instantiated in different ways across language varieties and are assigned varying degrees of strength (Gass 1996). On the basis of this perspective, it is reasonable to expect at least some degree of difference between the effect sizes shown in native and learner English. It is necessary to follow up and complement the above type of results by subjecting the data to a multifactorial study, such as a binary logistic regression, which not only computes the individual predictors' main effects on Form but also allows us to determine how these predictors interact with Corpus. These interactions make it possible to see which predictors' effects discriminate between native and learner language. In other words, the benefit of this method is that it pinpoints exactly which predictors cause the two language varieties to differ.

As an interim summary, the monofactorial results indicate that *may* and *can*'s raw frequencies of occurrence with individual ID tag levels are suggestive in terms of how *may* and *can* are used in different contexts in L1 and L2. Nevertheless, the results are also not sufficient to draw an accurate and informative picture of the uses of the two modals: the large majority of the cross-tabulation plots shows that, although *may* is less frequent than *can*, of the two forms it is more often more frequent than expected by chance. Consistently, *may* yields effect sizes that are larger than *can*'s, both in cases where ID tag levels are preferred or dispreferred. In the next section, I present the multifactorial results of the present study. First I present the results of the cluster analysis and then those of the binary logistic regression.

6.2 Multifactorial results

6.2.1 Cluster analysis (HAC)

The first HAC analysis yielded the results shown in Figure 6. The left panel is a dendrogram of the five items that were clustered on the basis of the BP percentages for all five modal forms as co-occurring with all independent variables and across the 3710 sentence data sample. The right panel represents the output of the validation of the cluster analysis with PVCLUST. In what follows, I base my discussions on the validated clusters in the right panels of the dendrogram figures.

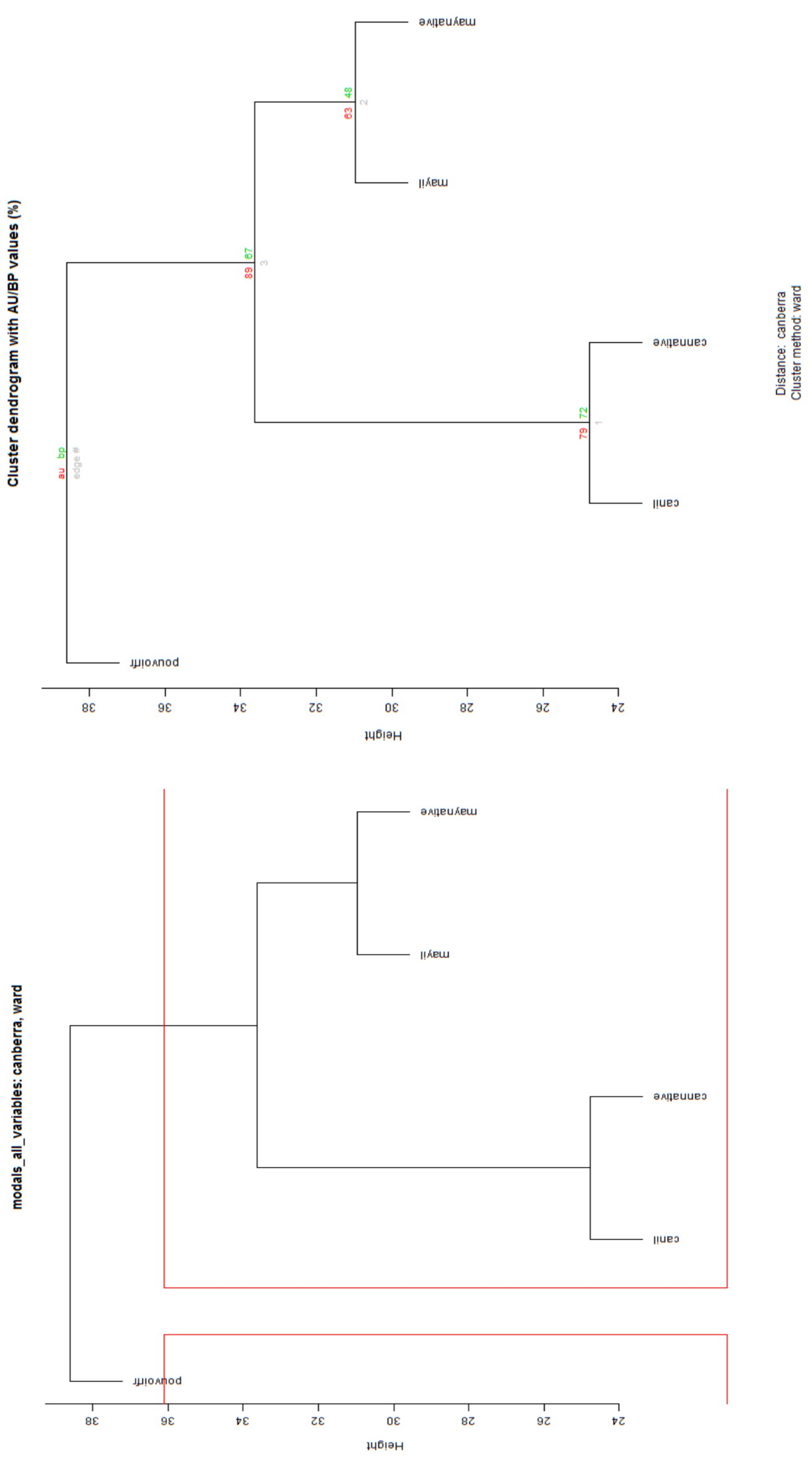


Figure 6 Dendrogram for all independent variables (*can*/*may*$_{\text{il}}$ = interlanguage *can*/*may*)

The two panels of Figure 6 show dendrograms of the five items that were clustered on the basis of the BP percentages for all five modal forms as co-occurring with all independent variables. The dendrogram provides the degree of (dis)similarity between the clustered elements. Reading the tree plot from bottom to top, forms that are clustered early are more similar than the forms clustered late and vertical lines provide an indication of the degree of autonomy of clustered elements. In other words, the longer the vertical line between clusters, the more autonomous the earlier cluster is from the next cluster it is amalgamated with.

Figure 6 shows that all clusters are amalgamated in one overarching cluster at distance 38.5. The main clusters separate at distance 33.5 and the sub-clusters separate at 31 and 25. The package PVCLUST for R allows to assess the degree of uncertainty of those clusters and to establish, on the basis of Approximately Unbiased (AU) *p*-values how strongly the data support the clusters. In general, Figure 6 shows that the {{*can*$_{\text{il}}$ *can*$_{\text{native}}$} *may*$_{\text{il}}$ *may*$_{\text{native}}$}} cluster is the most strongly supported by the data with an AU *p*-value of 100-89=11%. In second place, is the {*can*$_{\text{il}}$ *can*$_{\text{native}}$} sub-cluster (AU *p*-value of 100-79=21%) followed by the {*may*$_{\text{il}}$ *may*$_{\text{native}}$} sub-cluster (AU *p*-value of 100-63=37%). Interestingly, the three-cluster solution (i.e., French *pouvoir*, native and learner *can* and native and learner *may*) is compatible with Salkie's (2004) analysis (see Section 3.5), which argued that *pouvoir* is very different from both *can* and *may*, and intuitively, both these solutions 'make sense'. This provides the first evidence in favour of a multifactorial approach. To anticipate the potential objection that this may seem trivial, it is not. The data in Figure 6 show that the multidimensional BP vectors are good and robust descriptors of how the modals cluster because, unlike my data, many other cluster solutions, such as the ones listed in (68), would not make linguistic sense at all.

(68) a. {{{*can*$_{\text{il}}$ *may*$_{\text{native}}$ *pouvoir*} *can*$_{\text{native}}$} *may*$_{\text{il}}$}

b. {{{*can*$_{\text{native}}$ *may*$_{\text{il}}$ *pouvoir*} *can*$_{\text{il}}$} *may*$_{\text{native}}$}

c. {{*can*$_{\text{il}}$ *may*$_{\text{native}}$} {*pouvoir may*$_{\text{il}}$} *can*$_{\text{native}}$}

However, what follows shows 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. Figure 7 and Figure 8 show the dendrograms for all morphosyntactic and semantic variables.

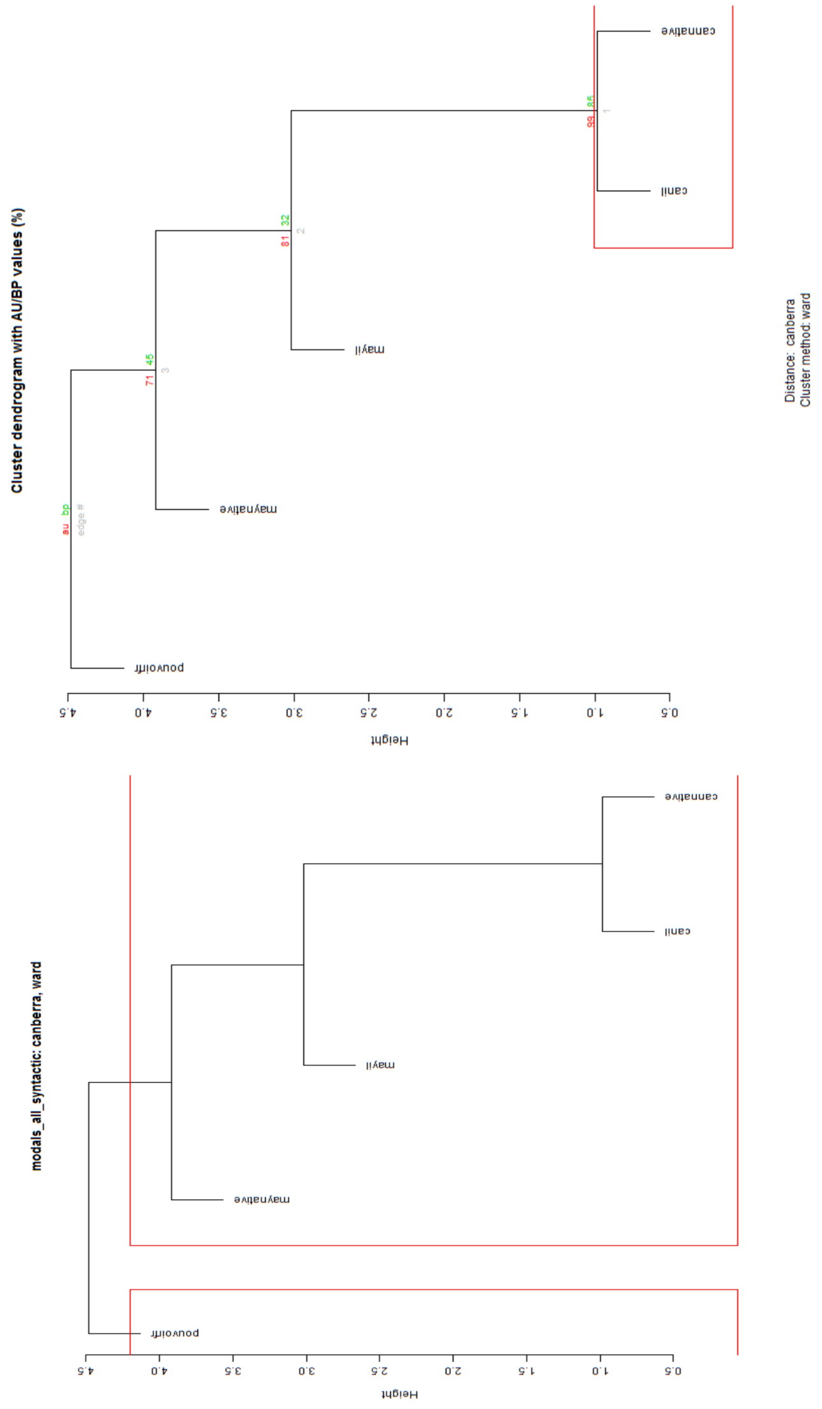


Figure 7 Dendrograms for all morphosyntactic variables

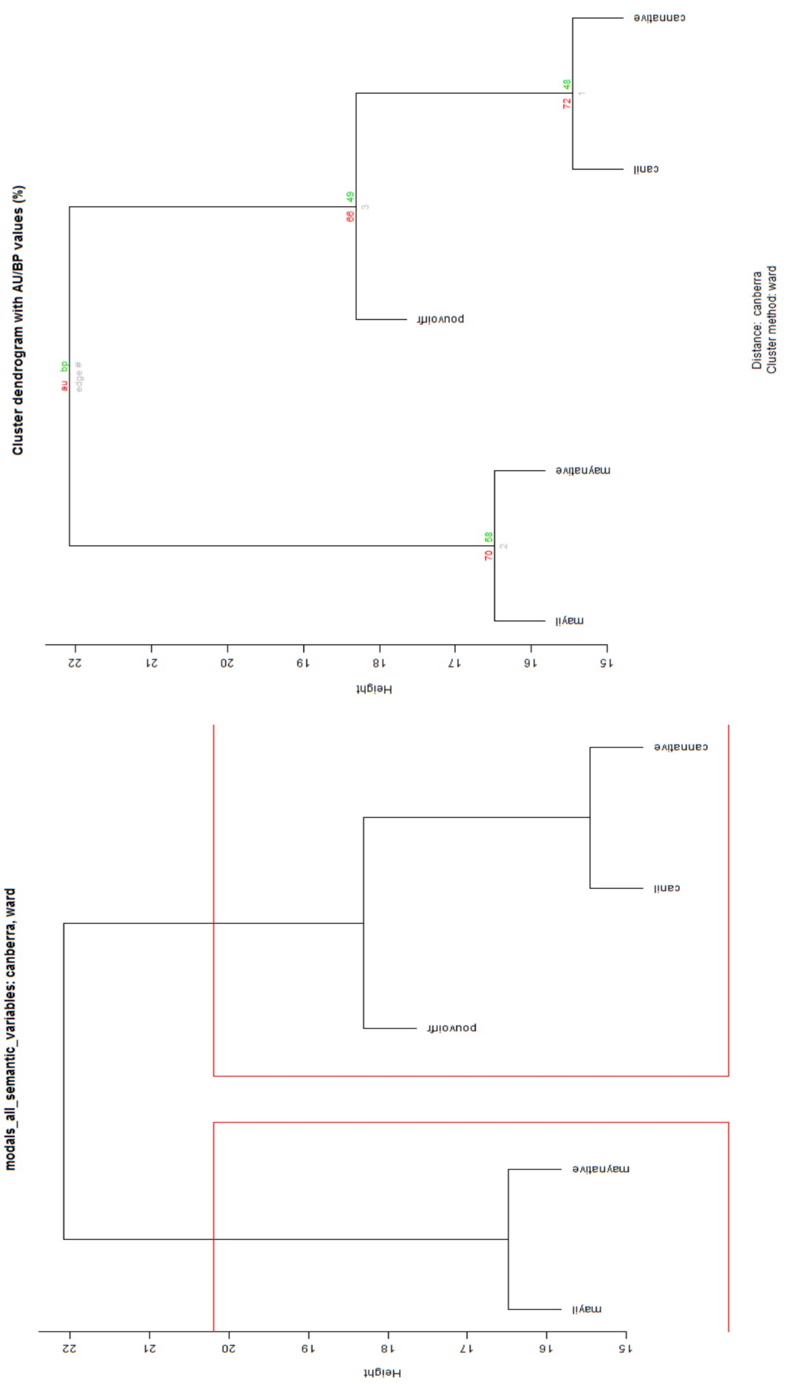


Figure 8 Dendrograms for all semantic variables

In Figure 7, all clusters merge into one single cluster at distance 4.5. The main clusters separate at distance 3.9 and the sub-clusters separate at 3 and then 1. In relation to morphosyntactic variables, the {*can*$_{\text{il}}$ *can*$_{\text{native}}$} sub-cluster is the most strongly supported by the data (AU *p*-value of 100-89=11%), followed by the {*may*$_{\text{il}}${*can*$_{\text{il}}$ *can*$_{\text{native}}$}} sub-cluster (AU *p*-value of 100-81=19%) and in third place the {*may* $_{\text{native}}${*may*$_{\text{il}}${*can*$_{\text{il}}$ *can*$_{\text{native}}$}}} sub-cluster (AU *p*-value of 100-71=29%). With regard to Figure 8, all clusters are amalgamated in one overreaching cluster at distance 22. The main clusters separate at distance 18.1, 16.5 and 15.5. In relation to semantic variables, the {*can*$_{\text{il}}$ *can*$_{\text{native}}$} sub-cluster is, again, the most strongly supported sub-cluster by the data, (AU *p*-value of 72%), followed by {*may*$_{\text{il}}$ *may*$_{\text{native}}$} (AU *p*-value of 100-70=30%) and then {*pouvoir*{*can*$_{\text{il}}$ *can*$_{\text{native}}$}} (AU *p*-value of 100-66=34%).

Interestingly, the cluster results in Figure 7 and Figure 8 show that the intuitively reasonable dendrogram in Figure 6 is not replicated when considering morphosyntax or semantics alone. This contrasts to some extent with Gries and Otani's (2010) results, where their results did not differ very much between the three clusterings. The reasonable similarities of *may* and *can* do not emerge well when only the syntactic variables are considered. In particular, in both Figure 7 and Figure 8 *can*$_{\text{il}}$ and *can*$_{\text{native}}$ are grouped together, but the remaining forms are grouped differently. In the morphosyntactic dendrogram, the two kinds of *may* are successively amalgamated (but not grouped together first), and the French *pouvoir* is only added after all English forms have been clustered again supporting Salkie's study, though not as visually convincingly as Figure 7. In other words, morphosyntactically, we find a clear English-French divide, but interlanguage *may* is too different from native *may* to be grouped with it. To identify the source of this difference, I 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*$_{\text{il}}$ and *may*$_{\text{native}}$ (see Divjak and Gries 2009 or Gries and Otani 2010 for more examples). As indicated in Figure 9, learners deviate from native speakers in their morphosyntactic use of the modals mainly in that learners underuse *may* in subordinate clauses and in negated clauses. This is interesting because above we have seen, in individual tables that combined L1 and L2 data, that *may* is used more in more complex contexts than *can*. Here, however, we can see that this complexity-based approach also allows us to distinguish within uses of *may*, namely between native speakers and learners: 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 high processing load (as discussed with reference to Rohdenburg's (1996) Complexity Principle in Chapter 3).

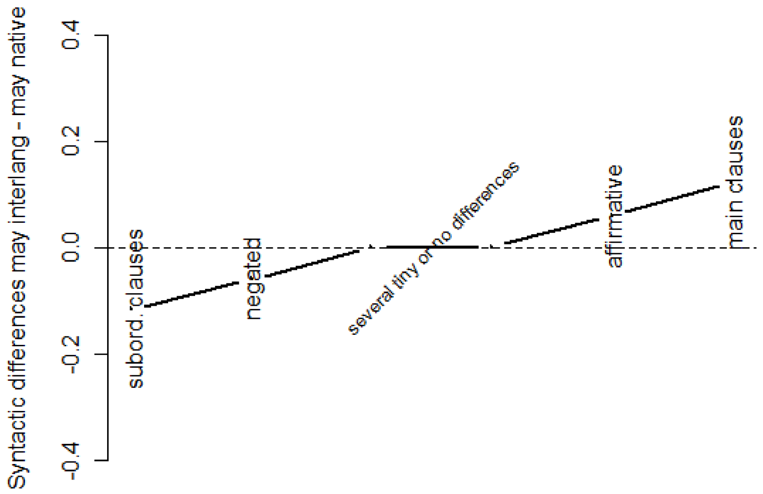


Figure 9 Snakeplot for the most extreme differences between syntactic ID tag levels of *may*

The semantic dendrogram in Figure 8 shows a different pattern. Semantically, *can*$_{\text{il}}$ and *can*$_{\text{native}}$ are again very similar and grouped together early, but the next clustering step groups the two forms of *may* together. However, it is not the English forms that are then all grouped together. In an interesting twist, *pouvoir* is not merged with one cluster of English-only forms which together with *pouvoir*'s being semantically more similar to *can* than *may* is, may seem to run counter to Salkie's (2004) earlier analysis. On the other hand, *pouvoir* is again only merged near the end of all amalgamations.

Overall the HAC has proved to be a useful technique to explore the extent to which the uses of *may* and *can* in French-English interlanguage are influenced by the uses of their

French equivalent *pouvoir* across the semantic, morphological and syntactic linguistic levels. The cluster results have made it possible to highlight the clear and pervasive contrasts between the uses of *pouvoir* on the one hand and the uses of native *may* and *can* on the other hand, at all three linguistic levels. More concretely, the cluster results have shown that:

- morphosyntactically, *pouvoir* contrasts most sharply with the uses of *can* (i.e., *can*$_{\text{il}}$ and *can*$_{\text{native}}$, with an AU *p*-value of 100-89=11%);
- semantically, although *pouvoir* yields enough similarity with *can* to be grouped together, the {*pouvoir*{*can*$_{\text{il}}$ *can*$_{\text{native}}$}} sub-cluster is not strongly supported by the data (AU *p*-value 100-66=34%);
- in contrast with the above two points, IL *may* and *can* are clustered together with their native equivalents across linguistic levels with AU *p*-values ranging from 100-89=11% in the case of the morphosyntactic uses of *can* to 100-70=30% in the case of the semantic uses of *may.*

Overall, those results indicate that French speakers use *may* and *can* in ways that are more similar to their uses by native speakers than the uses of *pouvoir* by native French speakers. In turn, this suggests the limited influence of *pouvoir* in the uses of *may* and *can* by French English learners. In fact, the corpus-based cluster results suggest that while using *may* and *can*, (advanced) French English learners rely more on their L2 English knowledge of the two modals rather than their knowledge of the uses of *pouvoir.* Following these results, it is reasonable to believe that focussing the analysis solely on the uses of *may* and *can* in native and learner English (rather than including *pouvoir)* and carrying out a fine-grained and systematic comparison of their uses in both English varieties may provide a more fruitful direction to further our understanding of how French speakers use the two modals (than studying *pouvoir* would do). At this point of the analysis, the binary logistic regression provides the necessary sophisticated statistical tool to explore in greater depth the (dis)similarities between *may* and *can* in both English varieties. In Section 6.2.2, I present the outcome of the binary logistic regression.

6.2.2 Binary logistic regression

As stated in the introduction to this chapter, the binary logistic regression provide the most fruitful results of the current study. The computation of the regression has led to an unprecedented systematic comparison of the uses of *may* and *can* in two varieties of English (i.e., native and learner English), in relation to a large number of predictors (see Table 14 for a detailed overview of all the predictors included in the study) across linguistic levels (i.e., semantic, morphological and syntactic) and on the basis of 3444 occurrences of the two modals across the two sub-corpora ICLE-FR and LOCNESS. The logistic regression analysis has not only helped to pinpoint the linguistic levels in which the uses of the two modals differ but also, and crucially, the specific variables that cause learners to use *may* and *can* in non-native ways. In more technical terms, the logistic regression has helped to identify which predictors interact with the dependent variable FORM as well as the predictor CORPUS. In what follows, I present the overall results of my final regression model. Throughout this section, I mainly focus on the interactions yielded by the model.

The model selection process involved thirteen steps during which insignificant predictors were discarded. The final and minimally adequate model includes 22 predictors – 16 significant variables and six significant interactions and returned a highly significant correlation: loglikelihood chi-square = 3296.47; *df*=60; *p*<0.001; the correlation between the 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%. The same classification accuracy was obtained with a leave-one-out classification approach. Table 42 summarises all the significant variables and interactions yielded in the final model.

Table 42 Overview of the final GLM model

Predictor	Chi-square (*df*): sign.		Predictor	Chi-square (*df*): sign.
Corpus	24.9 (1) ***		AnimType	98.2 (11) ***
GramAcc	13.8 (1) ***		Voice	55 (1) ***
Use	67.9 (1) ***		SentType	47.2 (1) ***
Elliptic	100 (2) ***		Neg	87.2 (1) ***
ClType	10.9 (1) ***		Corpus:ClType	60 (2) ***
VerbType	97.4 (2) ***		Corpus:VerbSemantics	32.2 (6) ***
VerbSemantics	384.9 (6) ***		Corpus:SubjNumber	37.4 (1) ***
SubjPerson	26.6 (2) ***		Corpus:RefAnim	122.2 (1) ***
SubjNumber	1.3 (1) ns		Corpus:AnimType	118.2 (11) ***
SubjMorph	49.1 (4) ***		Corpus : Neg	12 (1) ***
RefAnim	59.2 (1) ***			

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:Neg.

As previously noted in Section 2.2.2, existing literature concerned with native use of the modals recognises negation as an important aspect of modal meaning (Hermerén 1978). The current study not only confirms the need to include negation in an investigation of the uses of the modals (see the above discussion of the monofactorial results, e.g., Figure 5) but also recognises its significance as a morphological criterion to assess interlanguage (dis)similarity. Consider Figure 10 for the interaction Corpus:Neg.

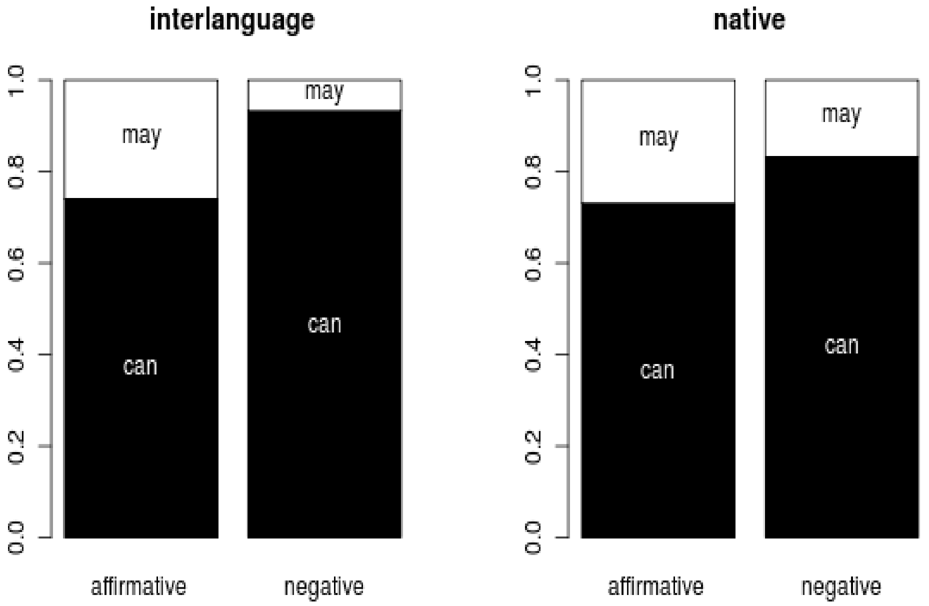


Figure 10 Bar plots of relative frequencies of CORPUS:NEG

Figure 10 shows that while all speakers prefer to use *can* in negated clauses, the interlanguage speakers do so more strongly. This result is interesting for two reasons. First, it generally reinforces Hawkins and Buttery's (to appear) idea of a correlation between structural complexity and frequency since negated clauses are more complex and preferred with the more frequent modal and secondly, this results suggests that this correlation applies to native and learner English in similar ways. It is worth noting, however, that where epistemic *may not* would be used in English, French speakers tend to use a lexical verb along with the adverb *peut-être* to indicate the speaker's uncertainty, as illustrated in the made-up example in (69).

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

The CORPUS:CLTYPE interaction indicates that the frequencies of *may* and *can* differ with regard to the types of clause they occur in in native and learner English (see Figure 14 in the appendix). The (weak) effect is that *can* is more strongly preferred over *may* in subordinate clauses in interlanguage English than it is in native English.

As for the interaction Corpus:SubjNumber, the results indicate that while native speakers use *can* more often with singular subjects than with plural subjects, it is the other way round for the learners, again a result compatible with the complexity principle, if we consider the plural form to involve added complexity (see Figure 15 in the appendix).

With regard to the interaction Corpus:RefAnim, 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 (see Figure 16 in the appendix).

Considering the interaction Corpus:VerbSemantics and given the larger number of ID tag levels involved in the variable VerbSemantics compared to the variables previously dealt with, I present the interaction graphically in Figure 11. The upper panel of Figure 11 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).

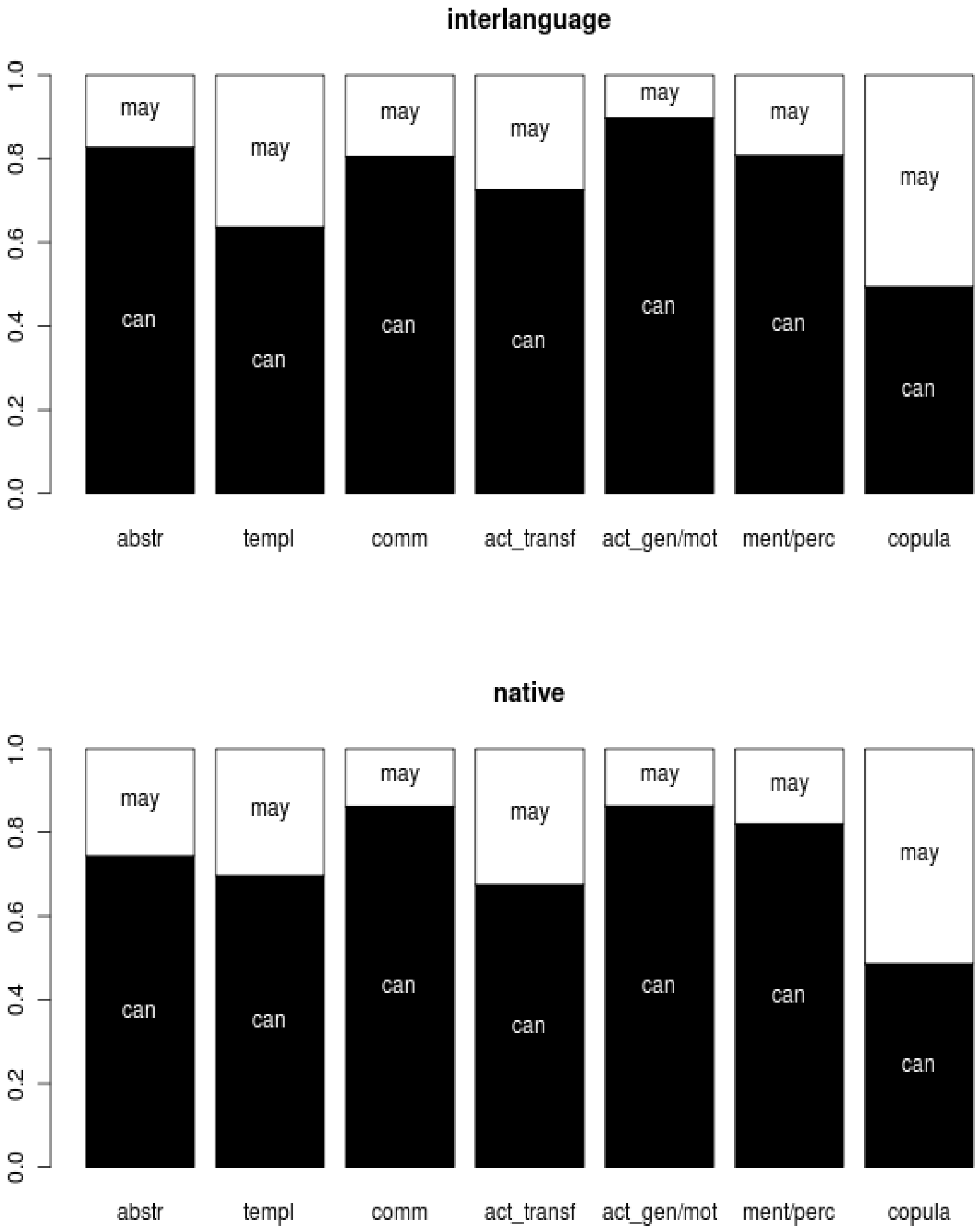


Figure 11 Bar plots of relative frequencies of Corpus:VerbSemantics

The learners and the native speakers differ most strongly with more abstract verbs and temporal verbs such as *achieve*, *cause*, *deprive* or *lead to* in the case of abstract verbs, and *end up*, *spend* or *begin* in the case of temporal verbs. 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. For instance, the learners prefer *may* with communication verbs and *can* with action-transformation verbs. Virtually no difference across corpora is found with copulas.

The final interaction, Corpus:AnimType, is not represented 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 means that this variable does not yield much in terms of interesting findings.

As for the main effects of the logistic regression, they are not discussed here in detail because these main effects by definition do not tell us anything about *can* and *may* across languages (since these variables do not interact with Corpus). However, they *do* tell us something about which modal verb is preferred by both native speakers and learners, so I have summarised them here visually in Figure 12. The *x*-axis lists the main effects, the *y*-axis shows 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.

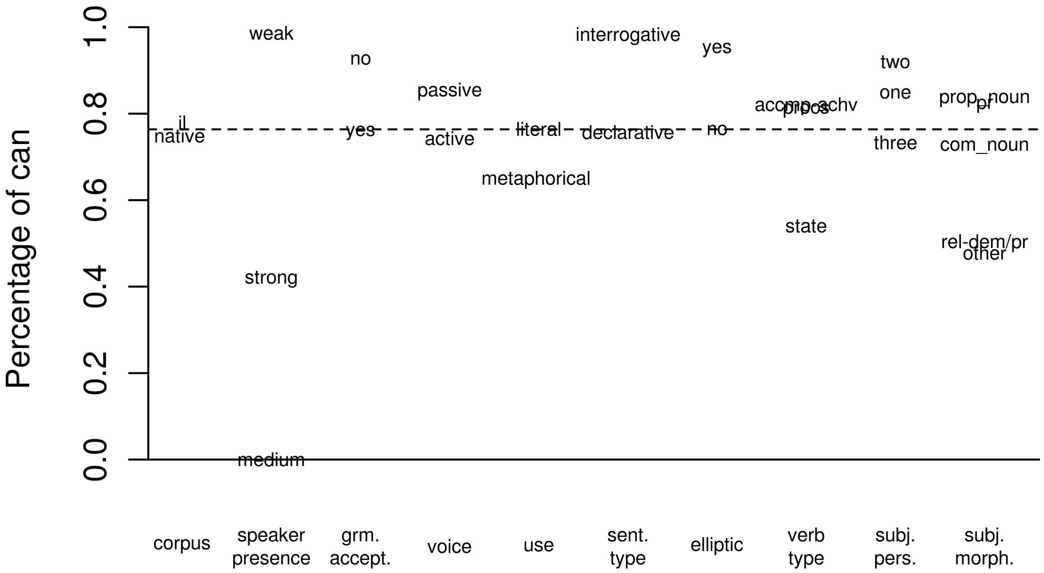


Figure 12 Main effects of the logistic regression

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 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 (70) and (71) below. While the ones in (70) illustrate the current point, (71) 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*.

(70) a. So we *may* say that …(ICLE-FR-UCL-0032.2)
b. To conclude: we *may* say that …(ICLE-FR-UCL-0040.2)
c. As a conclusion, we *may* say that …(ICLE-FR-UCL-0022.2)
d. This is why we *may* now speak of the stupefying effect …(ICLE-FR-ULG-0017.2)
e. That is the reason why we *may* say that …(ICLE-FR-UCL-0032.2)
(71) "Dresden is an old town", we *may* read of its history (ICLE-FR-UCL-0071.2)

6.2.3 Collexeme analysis

The distinctive collexeme analysis (see Section 4.3.4 for a detailed description of the method) of *may* and *can* in native English and French-English interlanguage presented in this subsection supports Gabrielatos and Sarmento's (2006) recommendation to focus on the collocation patterns of the modals in quantitative studies (see Section 2.4.2 for presentation and discussion of Gabrielatos and Sarmento 2006). Table 43 below presents the collostructional results for the first twenty most distinctive collexemes in the native data and Table 44 presents the results yielded by the learner data. Collexemes in bold feature in both the native and the learners' sets of results. For instance, considering the collexeme *be* in both tables, we can see that it is very strongly associated (p<0.001) with the *may* construction in both English varieties. In fact, in both cases, it yields the strongest association with that particular construction. The figures within parentheses alongside the collexemes refer to their observed frequencies in relation to *can* and *may.* In the case of native *can* in Table 43, for instance, *see* occurs 80 times with *can* and 4 times with *may*, and in the case of native *may*, *be* occurs 91 times with *may* and 79 times with *can*. Finally, the table includes the *p*-values of the statistically significant collexemes. Both Table 43 and Table 44 only include the top twenty most significant collexemes.

Table 43 Collexemes distinguishing between *can* and *may* in native English[61]

CAN (N=338)		*MAY* (N=155)	
Collexeme (*can*:*may*)	*Distinctiveness*	*Collexeme* (*may*:*can*)	*Distinctiveness*
see (80 : 4)	6.16 (*p*<0.001)	**be (91 : 79)**	10.28 (*p*<0.001)
do (43 : 3)	2.9 (*p*<0.01)	lead to (0 : 8)	4.87 (*p*<0.001)
afford (20 : 0)	2.49 (*p*<0.01)	want (1 : 7)	3.46 (*p*<0.001)
understand (19 : 0)	2.36 (*p*<0.01)	feel (8 : 12)	3.1 (*p*<0.001)
say (19 : 1)	1.6 (*p*<0.05)	**arise (0 : 4)**	2.43 (*p*<0.01)
learn (12 : 0)	1,48	**sound (0 : 4)**	2.43 (*p*<0.01)
go (11 : 0)	1,36	grow up (0 : 3)	1.82 (*p*<0.05)
expect (10 : 0)	1,23	need (0 : 3)	1.82 (*p*<0.05)
sympathise (9 : 0)	1,11	**seem (0 : 3)**	1.82 (*p*<0.05)
blame (7 : 0)	0,86	suffer (0 : 3)	1.82 (*p*<0.05)
relate (7 : 0)	0,86	have (27 : 17)	1.56 (*p*<0.05)
show (7 : 0)	0,86	**think (4 : 5)**	1.33 (*p*<0.05)
use (27 : 5)	0,8	**appear (1 : 3)**	1.3 (*p*<0.05)
achieve (6 : 0)	0,74	be able to (0 : 2)	1,21
buy (6 : 0)	0,74	deprive (0 : 2)	1,21
compete (6 : 0)	0,74	discover (0 : 2)	1,21
contract (6 : 0)	0,74	establish (0 : 2)	1,21
play (6 : 0)	0,74	face (0 : 2)	1,21
prove (6 : 0)	0,74	harm (0 : 2)	1,21
speak (6 : 0)	0,74	practise (0 : 2)	1,21
Total number of other attested collexemes	**318 types/1322 tokens**	**Total number of other attested collexemes**	**135 types/ 466 tokens**

[61] Collostruction strength values larger than 3, 2, and 1.3 are highly, very, and just significant respectively.

Table 44 Collexemes distinguishing between *can* and *may* in **IL** English

CAN (N=287)		MAY (N=101)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
see (75 : 4)	4.63 (*p*<0.001)	**be (88 : 71)**	10.86 (*p*<0.001)
do (41 : 0)	4.5 (*p*<0.001)	**seem (1 : 11)**	6.29 (*p*<0.001)
deny (22 : 0)	2.39 (*p*<0.01)	wonder (10 : 14)	3.95 (*p*<0.001)
live (15 : 0)	1.63 (*p*<0.05)	**think (7 : 12)**	3.9 (*p*<0.001)
afford (14 : 0)	1.52 (*p*<0.05)	**sound (0 : 5)**	3.29 (*p*<0.001)
compare (14 : 0)	1.52 (*p*<0.05)	**appear (1 : 6)**	3.19 (*p*<0.001)
find (37 : 4)	1.47 (*p*<0.05)	argue (1 : 6)	3.19 (*p*<0.001)
change (12 : 0)	1.3 (*p*<0.05)	lead (10 : 10)	2.28 (*p*<0.01)
use (12 : 0)	1.3 (*p*<0.05)	dream (1 : 4)	2.02 (*p*<0.01)
prevent (11 : 0)	1,19	represent (0 : 3)	1.97 (*p*<0.05)
understand (16 : 1)	1,08	ask (3 : 5)	1.8 (*p*<0.05)
imagine (9 : 0)	0,97	justify (1 : 3)	1.44 (*p*<0.05)
give (14 : 1)	0,9	turn out (1 : 3)	1.44 (*p*<0.05)
make (14 : 1)	0,9	**arise (0 : 2)**	1.31 (*p*<0.05)
mention (7 : 0)	0,75	commit (0 : 2)	1.31 (*p*<0.05)
realize (7 : 0)	0,75	disappear (0 : 2)	1.31 (*p*<0.05)
say (86 : 19)	0,72	exist (0 : 2)	1.31 (*p*<0.05)
conclude (6 : 0)	0,65	fear (0 : 2)	1.31 (*p*<0.05)
play (6 : 0)	0,65	look (0 : 2)	1.31 (*p*<0.05)
predict (6 : 0)	0,65	mean (0 : 2)	1.31 (*p*<0.05)
Total number of other attested collexemes	**267 types / 1142 tokens**	**Total number of other attested collexemes**	**81 types / 366 tokens**

Based on the comparison of Table 43 and Table 44, learners share with native speakers approximately one quarter of the twenty most significant collexemes, that is 25% of the most significant collexemes. This means that learners have internalised only a limited amount of the native-like patterns of verbal complementation for *can* and *may* and that their verbal preferences do not convincingly reflect those of native speakers. It is important to note, however, that it is with the *may* + copula construction that learners adopt their most native-like patterns (see Section 6.1 for monofactorial results in relation to the predictor VERBSEMANTICS).

It is interesting to note that a small set of verbs that are not distinctive; that is, they occur in relatively free variation with either *may* or *can*. Non-distinctive collexemes are more characteristic of the *may* construction: in the case of native *can*, for instance, there is no occurrence of a non-distinctive collexeme and only one features in the learner *can* data. *May*, on the other hand, yields a somewhat different pattern with learner *may* showing eight non-distinctive collexemes (*be* (88 : 71); *wonder* (10 : 14); *think* (7 : 12); *lead* (10 : 10); *dream* (1 : 4); *justify* (1 : 3); *ask* (3 : 5); *turn out* (1 : 3)) and native *may* showing 5 (i.e., *be* (91 : 79); *feel* (8 : 12); *have* (27 : 17); *appear* (1 : 3), *think* (4 : 5)). This result suggests a possible difference in the degrees of acquisition of L2 *may* and *can* by advanced learners. In other words, it is possible to envisage learners' lower level of acquisition of *may* in comparison to *can*, which, could explain their slightly greater use of *may* and *can* in free variation compared to their uses by native speakers.

Below, I represent graphically in Figure 13 the verb-specific preferences of *may* and *can* in (i) native English, (ii) native and (French) learner English, and (iii) (French) learner English only.

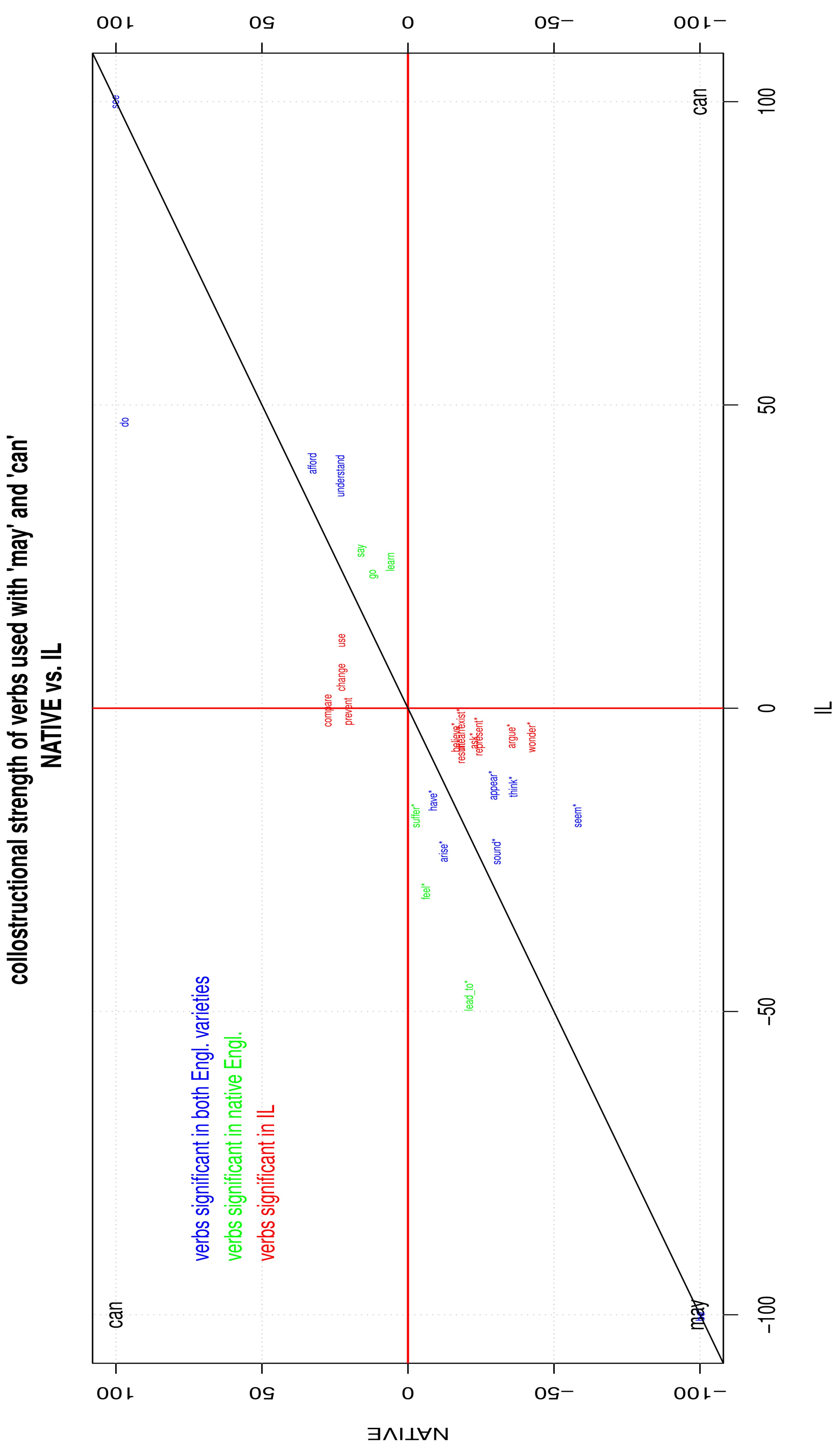


Figure 13 Distinctive Collexeme Analysis: *may* and *can*'s verb-specific preferences

Figure 13 provides a graphic indication of how, generally, learners' choices of lexical verbs with *may* and *can* reflect (or not) native-like patterns. The diagonal line on the graph represents the locations where lexical verbs chosen by non-native speakers would ideally be expected to be plotted. One first pattern yielded by the data is that learners unanimously recognise the strong associations of the lexical verbs *see* and *be* with *can* and *may*, respectively. Learners do indeed consistently use those two constructions in a native-like fashion. This, in turn, confirms even more clearly learners' sensitivity to, on the one hand, *may*'s preference for copula verbs and, on the other hand, *can*'s preference for verbs of perception (See Section 6.1 for each form's preferences in relation to VerbSemantics). It is worth noting, however, that the semantic and morphological proximity of the English adverb *maybe* as well as the semantic and morphological proximity of the French adverb *peut-être* (*maybe*) to the epistemic modal construction *peut être* (*may be*) may play an influencing role in learners' association of *be* with *may*. Furthermore, with regard to *can see*, it is also worth keeping in mind that the entrenchment of the construction in learners' mind may be influenced by rule-based instruction strategies that tend to single out perception verbs as characteristic of the use of *can*. Still, considering the entire range of lexical verbs featured on the graph, learners' choices are in general in line with those of native speakers: verbs solely selected by learners are not plotted at extreme points on the graph and neither do their plotting location show any particular significant deviation from those of verbs selected by either native speakers only (i.e., verbs in green) or by both native and learners (i.e., verbs in blue).

As a final point, Figure 13 suggests that learners may treat *may* and *can* differently. This view is based on the facts that

- in the case of *may*, collexemes in both the 'IL' and 'both Engl. varieties' sets are plotted close to each other and close to the diagonal line; and
- in the case of *can*, collexemes in both sets are plotted further apart from each other and from the diagonal line.

Those two observations suggest that learners would tend to use *may* in a more native-like way in comparison to *can*. It follows that out of *may* and *can*, *can* emerges as the most potentially linguistically interesting form for the identification of patterns of use in interlanguage. In fact, it is reasonable to hypothesise at this point *can*'s greater contribution (in comparison to *may*'s) to the emergence of non-native linguistic patterns. I develop this notion in Section 7.6.

6.2.4 Concluding remarks

In summary, the BP approach and the subsequent HAC and 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 of one modal over the other. On the whole, the combination of statistical techniques used in the current study has led to a number of discoveries. The HAC analysis, for instance, has shown that:

- native French speakers use *pouvoir* in ways that contrast with those of *may/can*'s native uses across linguistic levels and particularly at the morphological and syntactic levels;
- there is much more similarity between the IL and the native uses of *may/can* than between the uses of *pouvoir* and IL *may* and *can*, thus suggesting that native French speakers do not tend to map *pouvoir*'s patterns of use onto those of *may* and *can*.

The regression has shown that:

- six grammatical components trigger non-native use of *may* and *can* by native French speakers, namely: clause type, semantics of the modalised lexical verb, number of the grammatical subject, animacy of the referent, type of animacy of the referent and negation;
- in the context of the above variables, non-native use of *can* is consistently triggered in more complex grammatical environments such as subordinate

clauses, negated clauses, clauses including a plural referent number or clauses including an abstract referent.

Overall, in the cluster analysis, 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 indication of why that is so. The way learners choose one of the two verbs is often compatible with a processing-based account: they choose the more basic and frequent *can* over *may* when the environment is complex. But modal choice 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*, the results confirm that linguistic context plays an influential role in the uses of *may* and *can*. Indeed, the main effects in the final logistic regression model support studies that have identified morphosyntactic components such as VOICE and SENTTYPE as particularly influential categories (Leech 1969, 2004; Huddleston and Pullum 2002; Collins 2009), but the results also reveal the necessity of taking 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 (Bates and MacWhinney 1982, 1989), the cluster analysis and the high classification accuracy of the regression suggest that, on the whole, 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 (as compared to native English) and sometimes trigger a verb choice that is not in line with native speaker choices. It also shows that even this kind of incorrect choice is largely predictable (because the regression can still make the correct classifications). Although this is the first BP 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 therefore provide 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, that is where the categories of the learner are still substantially different from the native speaker. Needless to say, more and more rigorous testing is necessary (e.g., in particular for interactions of variables with CORPUS). However, this kind of approach generally demonstrates how the multifactorial approaches can go beyond the previous less fine-grained and less statistically comprehensive analyses and how learners' 'non-nativeness' tends to manifest itself at all linguistic levels simultaneously.

Chapter 7 General discussion

7.1 *Introduction*

The previous chapters have shed light on many aspects of the uses of *may* and *can* by French English learners as well as native English speakers. The modals have been subjected to a set of new methods of investigation which has led to (i) a description of their uses at a degree of granularity not attained in any previous corpus study of the modals, and (ii) the identification of linguistic variables that influence French English speakers' choices of *may* and *can* in systematic ways. The benefits of applying two multidimensional approaches, the multidimensional approach of behavioural profiles and the multifactorial regression approach, to learner language have been shown to reach beyond the benefits of the largely exclusively monofactorial results presented in previous studies.

First, the BP approach has provided a way to successfully reconcile a predictive methodological approach with a contrastive theoretical framework. This is important because while Gass (1996: 324) recognises that "[a] theory of language transfer requires that we have some ability to predict where the phenomena in question will and will not occur", she further adds that "[i]n this regard, contrastive analysis alone falls short" and "it is simply not predictive". Being able to predict the grammatical shape of learner language through a statistical modelling process and on the basis of a large scale data set represents a major breakthrough in the quest to understanding the mechanisms of interlanguage as a system in its own right. In addition, with its fine-grained quantitative approach to learner language, the multidimensional method has offered a new perspective for the analysis of interlanguage, namely one that combines psycholinguistics with the study of morphosyntactic and lexical semantic characteristics of language in use.

This chapter has two aims. The first is to evaluate to what extent the current study has improved existing work on *may* and *can* in L1 and L2. The second aim is to demonstrate and discuss how adopting a usage-based model of language suggests fruitful directions for future research in the field. Section 7.2 provides an overview of the types of approaches and results that are available in existing corpus-based studies of the modals and in Section 7.3 I present the major shortcoming of previous corpus-based quantitative studies of English modals. In Section 7.4, I relate the current work to existing work. In Section 7.5, I evaluate the success of the BP approach for the study of *may* and *can* in L1 and L2, and in Section 7.6 I assess to what extent the BP approach provides a helpful method for understanding the acquisition and the processing of L2 *may* and *can*. In the final part of this chapter, Section 7.7, I present recommendations for future research and conclude the study. Throughout that section, there is an emphasis on the notions that learners' choices of the two forms are cognitively motivated and that grammatical contexts constrain the acquisition of *may* and *can* in L2.

7.2 *Characteristics of previous studies on* may *and* can

As summarised at the start of Chapter 4, the existing literature on the English modals points towards a number of theoretical and methodological desiderata that a study on *may* and *can* in L1 and L2 should address. Across the literature, modal verbs have been approached both descriptively and quantitatively and generally, both types of approach tend to share the implicit assumption that uses of the modals are mainly determined by their senses. This explains why descriptive studies of *may* and *can* have mainly focused on how the forms relate to epistemic, deontic and dynamic and/or root meanings. However, at least some of the relevant literature has already strongly suggested the need to adopt a grammatically-grounded perspective to investigate modal verbs and by integrating simultaneously the semantic, syntactic and morphological linguistic levels. This view was convincingly argued both from a descriptive perspective (see Hermerén 1978) and from a theoretical perspective (Klinge and Müller 2005).

Quantitatively, studies such as Gabrielatos and Sarmento 2006 have provided empirical data indicating a correlation between the distribution of modal verbs and their syntactic contexts of utterance. Beyond native English, and in the field of EFL/ESL, Deuber 2010 has recognised that a number of grammatical features in Creole influence the behaviour of Trinidadian English. Interestingly, and unlike scholars such as Collins (2009) in the field of native English, Deuber (2010) has tried to demonstrate such influence quantitatively. Generally, studies like this that assume the existence of interactions between L1 and L2 grammars stand in sharp contrast with the work of scholars such as Adjemian (1976), for instance, who have argued for a fixed native-speaker speaker grammar and the permeability of learner grammar. Also in contrast with Adjemian's view is Jarvis (2000) who argues that both the effects from L1 and TL can be observed in L2. It is interesting to note that both Deuber and Jarvis have in common their methodological probabilistic approach to investigating cross-linguistic interactions.

7.3 *The limitation of previous corpus-based quantitative studies of the English modals*

In the field of psycholinguistics, Bates and MacWhinney (1982, 1989) have demonstrated the relevance of probabilistic approaches to language acquisition and use as well as processing-based models. Assuming such a theoretical framework enables the researcher to adopt a fine-grained quantitative methodological approach and to bridge descriptive and quantitative approaches. Ultimately, this helps the researcher to reach a high degree of granularity in the analysis as well as to provide an empirically reliable analysis. However, in spite of these pointers, there are, to my knowledge, no studies that try to bring all these desiderata together. Instead, studies tend to incorporate merely one of these desiderata by, for instance, including syntactic contexts in their studies of senses, or by adopting a cognitively-informed perspective. In other words, there is no work incorporating more of these aspects. For instance, to date, existing quantitative studies of English modals in L1 and L2 are yet to provide grammatically-grounded accounts of the uses of modal verbs. This, as Collins (2009) illustrates in L1 and Aijmer (2002) in L2, is often also due to limiting choices in terms of statistical methods to investigate the data. Generally, the results of such studies consist of raw and/or normalised frequency counts and ratios of the modal forms' occurrences, which means that quantitative studies tend to be limited to identifying and contrasting over- and under-use of individual modals and their related senses. In other words, quantitative results do not tend to be subjected to any statistical tests such as for instance a chi-square test to check for the statistical significance of the findings.

7.4 *Characteristics of the current study*

While the present study is far from being all-inclusive, it nevertheless begins to address many of these desiderata at the same time. In terms of approach, the current work contrasts with previous studies in several ways:

- it bridges the existing gap between descriptive and quantitative approaches to the English modals and provides a fine grained quantitative description of the behaviour of *may* and *can* that is based on more comprehensively annotated data sets than those used in corpus-informed descriptive studies;
- it accounts for the results of existing descriptive and quantitative studies by integrating the semantic, syntactic and morphological linguistic levels into the analysis of *may* and *can* (as recommended in Hermerén 1978) and by assessing quantitatively whether all three linguistic levels affect the uses of *may* and *can* in similar ways;
- it furthers Deuber's (2010) study by investigating simultaneously not only two different varieties of English (e.g., native language and learner language) but also two different native languages (e.g., English and French) whilst subjecting the data to a rigorous statistical treatment (as recommended in Jarvis 2000);
- in line with Bates and MacWhinney (1982, 1989) processing-based model of (second) language acquisition, it shows how a psychologically-informed corpus-based analysis of learner language and second language acquisition can be pursued.

In terms of method, the current work also contrasts with existing quantitative work in that

- it is not limited to frequency counts but is more versatile (as needed) and can, therefore address phenomena and distributions that are gradient in nature (e.g., Cramer' *V*, chi-square and *p*-values, Fisher-Yates exact test, percentages);

- it provides multidimensional results, that is, results that account for the behaviour of *may* and *can* on the basis of several predictors (see Gries 2010c for more details on this type of result);
- it provides results that not only yield a high degree of descriptive power and classification, but that can also be interpreted conceptually.

A major benefit of adopting a fine-grained quantitative approach is that it has made it possible to de-focus the scope of investigation of the modals and to enlarge it to include a precise characterisation of their linguistic uses and contexts. As a result, it has been possible to demonstrate empirically the validity of Hermerén's (1978) grammatically-grounded theoretical approach to modal verbs. More concretely, it was found that

- the senses of the forms do not influence speakers' choices of *may* and *can*; and that
- other grammatical features such as the semantics of modalised lexical verb or the type of referent's animacy influence learners' modal choices.

The results obtained in the present study have helped to provide some insight into the acquisition of *may* and *can* in L2 and, to some extent, their processing in L2 language production. However, in order to explain the relevance of corpus results for L2 acquisition, it is necessary to adopt a theoretical approach to language acquisition and use that is probabilistic in nature and thereby compatible with the BP method adopted in the current study. Such a suitable approach is the usage-/exemplar-based approach, described in Section 3.6.3. In what follows, I return to the BP approach and show from a descriptive perspective how beneficial it has proved to be in the present study.

7.5 *Evaluating the success of the BP approach and the use of sophisticated statistical techniques for the present study*

Throughout this study the BP approach has emerged as a fruitful method to investigate *may* and *can*. Generally, the application of the method has led to a number of discoveries not only with regard to the two modals and what characterises or distinguishes them as lexical items, but also with regard to how differently they are used by native and non-native English speakers. In the two following subsections, I explain in what ways the BP approach has shed light on the uses of (L2) *may* and *can*. I first cover the two modals as lexical items and then focus on their uses by native and non-native English speakers.

7.5.1 The BP approach and *may* and *can* as lexical items

The main point of focus throughout the present work has been to study *may* and *can* from the perspective of their co-occurrence patterns and in that regard, I have investigated the extent to which those patterns characterise *may* and *can*, individually. Methodologically, in order to carry out my research goal, I deviated from the traditional corpus-based methods currently applied in studies of the modals and instead I applied the BP approach which ultimately allowed me to adopt a multifactorial outlook on the uses of *may* and *can*. Overall, applying the BP approach helped me to establish that the uses of *may* and *can* are characterised by the distribution of their co-occurrence patterns across the semantic, syntactic and morphological linguistic levels. As anticipated in the introduction to the present work, this result is in line with Kennedy's (2002) claim that an analysis of the modals need to take into consideration their distribution throughout the data. Crucially, however, the present study furthers Kennedy's claim by indicating that it is the distribution of the forms' co-occurrence patterns (rather than the distribution of the lexical forms) that is most useful to the analyst.

Indeed, such distributional differences have revealed that *may* and *can* contrast sharply in the ways they behave in relation to other grammatical features. Consequently distinguishing between the two modals does not just involve accounting for the distribution of their respective senses across a given dataset, as carried out in Collins

2009, but rather involves identifying co-occurrence patterns across several linguistic levels. In that respect, the present study also furthers the work of Gabrielatos and Sarmento (2006) who, as we saw in Section 2.4.2, demonstrate the adequacy of a probabilistic outlook on the English modals in relation to syntactic features (see Section 2.4.2 for the complete list of investigated syntactic contexts). In a similar way, the present study provides empirical evidence that the semantic and morphological levels of analysis equally contribute to the characterisation of *may* and *can*. This is an important finding as it provides reliable empirical evidence supporting Hermerén's (1978) theoretical claim that the grammatical contexts of the modals contribute to their semantic import. In other words, the present study concludes that an adequate semantic account of *may* and *can* should be grammatically-grounded and should reflect the behaviour of the two forms within their sentential contexts.

Finally, while Perkins (1983) claims that the modals are most fully integrated within the structure of the clause, the present study is not only sympathetic towards this line of argumentation but, similarly to the case of Hermerén's study, it provides the necessary empirical evidence for the validation of the author's claim.

Given the result that the distribution of the co-occurrence patterns of *may* and *can* characterises the two modals, I briefly return to Coates' (1983) quantitative study on the semantics of the English modals presented in Section 2.4.1. There I expressed the concern that, methodologically, Coates's choice to dissociate her semantic clustering analysis of the modals from their subsequent form-syntax analysis prevents her from identifying potential interactions between the semantics of the modal forms and their co-occurring morphosyntactic features. In light of my results, I reiterate my concern about Coates's method and the extent to which it may have affected her results overall.

7.5.2 The BP approach and the (cross-linguistic) influence of grammatical contexts on IL *may* and *can*

The observed influence of *may* and *can*'s grammatical contexts on their uses in L1 (see the monofactorial results in Section 6.1) raises the question whether (and if so to what extent) grammatical contexts also influence the use of *may* and *can* cross-linguistically and across English varieties (i.e., from native English to learner English). Generally, the BP approach has allowed me to establish that grammatical contexts do not play a significant influencing role on modal selection cross-linguistically at any linguistic level. In other words, learners do not tend to apply the co-occurrence patterns of a given modal *a* in their native language *y* to *a*'s translational equivalent in second language *x*. This is an important result because it suggests the non-existence of cases of linguistic transfer at the co-occurrence level, at least from native French into French-English IL in the case of modal verbs. Indeed, and on the basis of Gilquin's (2008) previously mentioned claim (Section 3.4.1) that cases of transfer can only be identified as transfer if similarities between the learners' behaviour in interlanguage and in his/her native language can be established, my HAC analysis does not reveal cases of similarity between the co-occurrence patterns of *pouvoir* and those of IL *may* and *can* across the morphological and syntactic linguistic levels. As for the semantic level, even though some degree of resemblance is identified, such similarity is not convincingly supported by the data (see Figure 8).

With regard to native and non-native English varieties, I noted at the outset of this study that existing corpus-based work on L2 uses of the English modals hardly takes into consideration the interactions between investigated L2 lexical forms and their linguistic contexts (see Section 3.4.2). In contrast with such traditional work, the BP approach (followed by logistic regression) has led to the discovery that grammatical contexts influence the uses of English modals in L2 and constrain learners' modal choices.

This influential role of grammatical contexts in L2 was specifically observed in more complex grammatical environments. In such linguistic contexts, the differences between the uses of *may* and *can* in native and learner English were observed in relation to six grammatical features which were ultimately identified as causing non-native use of *may* and *can*: clause type, semantics of the modalised verb, subject number, animacy of the referent, type of animacy of the referent and negation. In addition, it was subsequently found that for each of those variables, learners prefer to use *can* rather than *may* with more complex features such as subordinate clauses, negated clauses, abstract lexical verbs, etc. (see Section 6.2.2 for the detailed results for the logistic regression). Overall, with regard to the two investigated English varieties, the multidimensional approaches emerged as highly successful to improve on existing studies on L2 modals as it helped to establish (i) the significant role of co-occurrence patterns as factors of linguistic variation and, as a result, (ii) the need to account for co-occurrence patterns in quantitative contrastive-linguistic studies.

More broadly, the BP method, combined with the use of sophisticated statistical confirmatory techniques, has proved to be a powerful approach to facilitate the formulation of corpus-informed and psychologically-motivated hypotheses on the emergence of co-occurrence patterns in IL grammars, their processing and their acquisition by learners. I illustrate this point below in Section 7.6.1 by proposing, for instance, a possible explanation for why IL *can*, rather than IL *may*, emerges as the preferred lexical variant in more complex grammatical environments. From a SLA perspective, and given the high degree of reliability of the logistic regression results (i.e., 99% classification accuracy), the formulation of such corpus-informed and psychologically-motivated hypotheses emerges as a useful outcome for the experimental assessment of the extent to which co-occurrence patterns contribute to learners' acquisition of *may* and *can* and their on-line processing. In the following section, I develop the notion that multifactorial corpus-based analyses facilitate the formulation of such hypotheses by (i) exploring the notion of *can* as a default modal term and (ii) by exploring the acquisition of L2 *may* and *can* from the perspective of prototypicality.

7.6 *A corpus-based exploration of the processing and acquisition of IL* may *and* can

7.6.1 *May*, *can* and the notion of a default term

Overall, the BP approach has emerged as a pivotal method in the investigation of learner language, in that it has proved powerful enough to capture emerging structure in IL. More precisely, fitting a binary logistic regression to a richly annotated dataset has led to the hypothesis that French English learners tend to use *can* as a default term in more complex grammatical environments. In this section, I demonstrate how methodologically, the use of the logistic regression contribute to the formulation of this hypothesis.

Overall, the logistic regression indicates that an important difference between *may* and *can* is the complexity of their linguistic context. In particular, the regression indicates that several variables that have something to do with complexity distinguish significantly between learners' and native speakers' uses of *may* and *can*. Within a usage-based and processing-based model, Rohdenburg (1996) argues that, when linguistic environments become too complex, speakers resort to a 'default' choice or construction. Despite its theoretical insight, Rohdenburg's claim calls for corpus-based validation. In what follows, I show how logistic regression can provide a useful statistical technique to explore Rohdenburg's claim more empirically and on the basis of native and learner corpus data.

Within the usage-based theoretical approach outlined above in section 3.6.3, three different perspectives can be adopted to explain the choices speakers make in the context of *may* and *can*.[62] So one question that emerges from this situation is how to predict which of the two variants, *may* or *can* is the default term. In what follows I present and briefly discuss three models that can potentially provide a way to identify a default item on the basis of a usage-/exemplar based theoretical approach.

Overall, the three perspectives I present below vary in that they make different assumptions about which kind of frequency motivates the emergence of a default form and consequently, each perspective leads to the identification of a different default item and, in the case of one particular perspective, the underlying assumptions make it impossible to decide at all whether *may* or *can* functions as the default.

The underlying assumption of the first perspective is that the default item is the form that is most frequently used throughout the data. In other words, this approach assumes that token frequency alone determines the emergence of a default item. On that basis,

[62] I remind the reader that *token frequency* refers to the actual number of occurrences of an exemplar and *type frequency* refers to "the number of distinct lexical items that can be substituted in a given slot in a construction" (Ellis and Collins 2009: 330). The use of token frequency as a predictor implies the assumption that the higher the token frequency of an exemplar, the more strongly it is represented in multidimensional knowledge system. In contrast with token frequency, the use of type frequency as a predictor implies the assumption that the larger the number of types that an element x co-occurs with in some pattern, the more diversely x is represented along all the dimensions in the knowledge space, and the more likely x will give rise to a schema. It is in this way that type frequency reflects productivity as well as likelihood of schematisation.

and in the current context, it is *can* that emerges as the default item with a total of 1290 raw occurrences throughout the learner data, in contrast with 366 raw occurrences for *may*.

The second possible perspective assumes that type frequency motivates the emergence of a default item. In other words, this approach relies on the notion that a more diversely used form is more likely to function as a default. At the core of this approach is the idea that a higher type frequency correlates with greater linguistic productivity and consequently, the more versatile a form is, the more likely it is to emerge as a default item. In contrast with the previous model, and in the specific case of *may* and *can*, this perspective does not allow for the identification of one of the two forms as a default item. This is because although the occurrences of the two modal verbs differ in terms of token frequency (i.e., the total number of cases where *can* is preferred is consistently higher than the total number of cases where *may* is preferred), they do not differ in terms of type frequency. Based on the raw BP data, both forms share the same type frequency. In fact, both *may* and *can* were attested for the same total number of ID tags (i.e., 76). Overall, this results indicates that the second approach is not reliable to predict speakers' modal choices.

The third perspective includes both type and token frequency and in this case, the identification of the default item is based on *may* and *can*'s type/token ratio. In the current context, this means that *may* would be identified as the default term due to the fact that although it is less frequent than *can* overall, its type frequency is the same as *can*'s. This is an important finding because if *may* can reach the same type frequency of ID tags with a lower token frequency, then this suggests that compared to *can*, *may* is a more versatile lexical form. Ultimately, this also suggests that by nature a default term exhibits a higher degree of variability.

The current work demonstrates that the logistic regression provides a clear-cut way to assess the above-described approaches and to identify the first one (i.e., based on token frequency) as the most reliable one. This result is based on the observation that while the first model identifies *can* as the default term, the logistic regression reveals that learners prefer *can* in grammatically complex environments. For example, in the cases of CLTYPE and VERBSEMANTICS while *can* is preferred in subordinate clauses, *may* is preferred in main and coordinate clauses and in the case of VERBSEMANTICS, while *may* is preferred in copula verbs, *can* is preferred with verbs denoting abstract processes. Overall, the results of the logistic regression show that

- it is possible to predict learner language on the basis of a corpus-based investigation and a usage-based theoretical framework;
- token frequency overrides type frequency in cases of emerging default items; and
- prototypicality can be seen to be a characteristic of default terms.

From a methodological perspective, the above results strengthen the case for the use of multifactorial statistical techniques in IL corpus research. From a more theoretical perspective, however, the above three-way approach presents a caveat, namely that, overall, the approach is somewhat simplistic considering the multidimensional nature of language advocated throughout the current work. Although, in the context of this study, the three-way approach serves the purpose of demonstrating the potential fruitfulness of a corpus-based study of default terms in L2, in order to validate *can*'s status as such a term, it would be necessary to adopt methods that can take the multidimensional nature of the data into consideration by, for instance, considering the entropy of the frequencies of ID tags and their intercorrelations.

7.6.2 Further arguments in favour of a default *can*

The higher frequency of *can* in the native sub-corpus in comparison to *may* explains *can*'s deeper entrenchment through both input and learners' output. On that basis, and in combination with the previously mentioned notion of semantic schematicity as a result of deeper entrenchment, it is possible to assume that *can* makes smaller cognitive demands in comparison to *may*. This characteristic of *can* provides an explanation with regard to observed distributional differences between *may* and *can* in more complex grammatical environments. More concretely,

- the way *may* and *can* interact with their grammatical contexts reflects the forms' degree of schematic specificity; and that
- a less schematically specified form such as *can* is more likely to be found in more complex grammatical environments as the form itself incurs less cognitive effort; and
- as such, *can* is faster to process than *may.*

Relating the notions of structural complexity with *can*'s lesser processing demands on learners leads me to revisit Hawkins' (2004) notion of *efficiency*, which I first briefly introduced in Section 6.1.3 while discussing monofactorial results in relation to the morphological variables. According to Hawkins (2009: 1), efficiency is concerned with a speaker's intended message and, according to the author, "[c]ommunication is efficient when the message intended by S [speaker] is delivered to H [hearer] in rapid time and with minimal processing effort". For Hawkins, grammars reflect both structural complexity and speakers' efficiency:

> [e]ven highly abstract and fundamental properties of syntax [are] derivable from simple principles of processing efficiency and complexity that are needed anyway in order to explain how language is used. As I see it, the emerging correlation between performance and grammars exist because grammars have conventionalized the preferences of performance, in proportion to their strength and in proportion to their number, as they apply to the relevant structures in the relevant language types. (Hawkins 2004: 2)

According to Hawkins (2009: 2), investigating grammars from the perspectives of structural complexity and efficiency "gives us a more complete picture of the forces that have shaped grammars and the resulting variations". More precisely,

> it puts structural complexity in its proper context, and it helps us understand the trade-offs better: preferred structures can be simpler in one respect, more complex in another; and the trade-off may involve simplicity competing with some other efficiency factor, e.g. speed of on-line property assignments in processing (Hawkins 2009: 14)

Following Hawkins' line of approach and in light of the logistic regression results presented in Section 6.2.2, one can say that *can* yields a higher degree of speaker's efficiency in comparison to *may*. It is indeed reasonable to believe that the six non-native-like co-occurrence patterns identified with the logistic regression (CORPUS:CLTYPE, CORPUS:VERBSEMANTICS, CORPUS:SUBJNUMBER, CORPUS:REFANIM, CORPUS:ANIMTYPE and CORPUS:NEG) emerge from on-line communication constraints imposed upon speakers and that, in the case of *can*, in order to "provide the earliest possible access to much of the ultimate syntactic and semantic representation as possible" (Hawkins 2004: 9), learners deviate from native-like co-occurrence patterns in more pronounced ways than they do with *may* and in that respect *can*'s efficiency is greater than *may*'s. To relate the current discussion to the notion of *can* as a default modal term, *can*'s greater efficiency suggests that default terms in IL emerge not only because they combine more flexibly with other grammatical features than non-default terms, but, crucially, because they facilitate learners' on-line selection and combination of grammatical features. Recognising *can* as a default term as well as *may* and *can*'s different degrees of efficiency is important to bear in mind for the purpose of investigating variability in IL varieties as Hawkins (2009: 13) notes that "the more efficiency there is in a given structure, the more grammars incorporate it as a convention". Such recognition should indeed be accounted for in order to adequately explain the emergence of non-native linguistic patterns.

Beyond *can*'s processing load, *can* is favoured over *may* as a default term because its behaviour in relation to semantic variables is more similar to *pouvoir*'s than to *may*'s in L2 English as shown by the cluster analysis. To assess *can* as a default term, it is necessary to take into consideration the extent to which the similarities between *can* and *pouvoir*'s co-occurrence patterns may explain learners' preference for *can* over *may*. In other words, to assess *can* as an L2 default term, it is necessary to compare *can*'s contextual behaviour in L1, in L2 and in contrast with that of native French *pouvoir*. This line of approach provides a way to establish:

– whether the use of *can* as a default term in French-English interlanguage is primed by the already established L1 *pouvoir*; and if so

- whether there is a particular grammatical level (i.e., semantic, syntactic or morphological) in which such priming is more noticeable.

In the field of L2 learning, Achard and Niemeier (2004) generally claim that learners undergo a mental "retraining" process in learning new sets of symbolic units that differ from sets they already know. In the context of the current study, this retraining process

- raises the question of the possible existence of cross-linguistic interferences between L1 *pouvoir* and L2 *may*/*can* at co-occurrence level;
- implies that greater cognitive effort is required from learners in comparison to native speakers in producing a particular L2; and
- suggests the possibility that to reduce the amount of cognitive effort during language production learners may select a linguistic form (or variant, in the case of *may* and *can*) whose co-occurrence patterns are the most similar to those of *pouvoir* in L1.

With regard to the retraining process and its above-stated implications, the cluster analyses have provided fruitful results. As a reminder, four cluster analyses were carried out in order to identify behavioural (dis)similarities between *can*$_{\text{native}}$, *can*$_{\text{il}}$, *may*$_{\text{native}}$, *may*$_{\text{il}}$ and *pouvoir*. For each analysis, the grammatical behaviours of the five modal items were compared on the basis of different groups of predictors: the first analysis included all predictors regardless of their linguistic level (i.e., semantic syntactic or morphological) and the other three analyses involved, in turn, all semantic, morphological and syntactic predictors. Cross-linguistically, the cluster analysis indicated that while *may* behaves differently from *pouvoir* morphologically and syntactically, *can* on the other hand, behaves similarly to *pouvoir* semantically. That is to say that from the standpoint of their semantic co-occurrence patterns *can* and *pouvoir* yield similar results. Arguably, the similarity between *can* and *pouvoir*'s semantic co-occurrence patterns may contribute to the emergence of *can* (rather than *may*) as a default term in French-English interlanguage because speakers are arguably more concerned with picking up semantic rather than morphosyntactic occurrence patterns during their learning process. In other words, on the basis of the results of the cluster analysis and on the basis that more complex environments are more cognitively challenging for both native and learners but even more so for learners who are subjected to additional retraining effort, one can hypothesise that in complex linguistic contexts learners select *can* as the lexical item that yields the highest degree of semantic, and hence more pertinent, similarity with L1 *pouvoir.*

To summarise, I have shown in this section that two factors support *can*'s status as a default term in French-English interlanguage. First, the fact that there are corpus-based reasons to believe that *can* incurs less processing effort than *may* and second, cross-linguistically and semantically, *can* behaves similarly to *pouvoir* which facilitates learners' retreat to *can* rather than to *may.* Despite the fruitfulness of the above-discussed results, it is crucial to keep in mind that the benefit of multidimensional approaches in corpus-based investigations of learner language lies in the fact that (i) they facilitate the identification of linguistic phenomena of potential interest, (ii) they help the researcher to relate such observed phenomena with cognitive mechanisms and (iii) (as a result) they facilitate the formulation of cognitively-inspired hypotheses on the processing and/or the acquisition of linguistic items in L2. Ultimately, while such hypotheses call for experimental validation, the BP approach provides a way to bridge both types of approach successfully (Divjak and Gries 2009). In the next section, I focus on the acquisition of *may* and *can* in L2 and I present a hypothesis that has emerged from my corpus-based analysis.

7.6.3 *May, can* and their acquisition in L2

In this section, I show how the above techniques and their application to learner corpus data allow the analyst to gain insights into the acquisition of *may* and *can* in L2. I consider the two modal forms both as lexical items and constructions such as *may*/*can* + lexical verb. The overall structure of this sub-section is based on the theoretical assumptions mentioned above, in particular the assumption that context plays an influential role in learners' choices of *may* or *can* and, in turn, it contributes to the acquisition of non-native patterns of use.

With regard to *may* and *can* as lexical forms, frequency analyses provide an exploratory way to

i. assess the proximate and relative degrees of acquisition of *may* and *can*; and also
ii. to (tentatively) establish whether a possible order in the acquisition of the two forms can be considered and if so, which one.

In both cases, the same assumption of the impact of token frequency discussed above prevails: a more frequent linguistic item is submitted to an entrenchment process at a faster pace and this process contributes to earlier acquisition. In light of the higher raw frequency of *can* and the lower raw frequency of *may*, one could then hypothesise English learners as first acquiring *can*, followed by *may*. An important implication of this claim is that in the acquisition process of *may* and *can*, their frequency of occurrence with particular ID tags overrides their dispersion across the range of ID tags they occur with. This means that if lexical item *x* is more frequent that lexical item *y*, learners are more likely to acquire it first even though *y* is used in a wider range of grammatical environments. With regard to the order of acquisition of *may* and *can*, the usage-/exemplar-based approach therefore predicts that *can* is acquired earlier: *can* is not only much more frequent than *may*, it also has a much lower type-token ratio of ID tags, which means it can be characterized as the low-variance sample that exemplar-based and construction-grammar approaches (Goldberg 2006) have demonstrated facilitates category acquisition. The higher type-token ratio of ID tags of *may* and its at the same time lower frequency makes it harder for learners to notice a form's co-occurrence patterns in proportions that are sufficient enough for those patterns to become path-breaking contexts.

From a second language acquisition perspective, the correlation between the degree of grammatical complexity of a linguistic context and the degree of schematicity of a given lexical item within that context is crucial, as is the perceived semantic similarity of items across languages. This is because such a correlation is, to a degree, responsible for the emergence of 'new' (i.e., non-native) co-occurrence patterns (such as for instance a significantly more frequent use of *can* by learners than by native speakers), which learners subsequently consolidate through entrenchment. From that perspective, one can say that while the acquisition of *may* and *can* is strongly influenced by the nature of their linguistic environments, it is negatively affected by it in contexts of grammatical complexity. This finding generally motivates the adoption of a constructionist outlook on L2 acquisition of *may* and *can*.

The relevance of a constructionist approach for L2 *may* and *can* is shown in two studies already cited above for their coverage of exemplar-based approaches in SLA: Ellis and Collins (2009) and Ellis and Ferreira-Junior (2009). Ellis and Collins claim, for instance, that, much like L1 acquisition and much as argued above, "L2 learning is driven by the frequency and frequency distribution of exemplars within construction". In the context of the current work, Ellis and Collins 2009 and Ellis and Ferreira-Junior (2009) raise the questions whether

i. English learners distinguish *may* and *can* as constructions;[63]
ii. L2 *may* and *can* may have lexically-specific preferences based on their co-occurrence patterns;
iii. those preferences may be identifiable when compared to those of *may* and *can* in L1;
iv. learners and native speakers share the same perception of which construction is the most prototypical (for example, which lexical verbs go with which modal) and which lexical verbs are most prototypically associated with *may* and *can* within each construction.

With regard to the first issue, the collexeme analysis in the current work indicates that English learners do distinguish *may* and *can* as constructions since they clearly exhibit very distinct lexicogrammatical preferences, which in Gries and Wulff (2005, 2009) is argued to be not only an indicator of constructionhood, but also, more pertinently, constitute part of constructional knowledge in L2. In the case of *can*, and recognising all due caveats given the register studied here, the identified prototypical construction is *can see* and in the case of *may*, the prototypical exemplar is found to be *may be.*[64]

[63] Here I refer particularly to the *may*/*can* + lexical verb construction.

[64] In the cases of *can see* and *may be*, the high degree of prototypicality I arrive at is based on collexeme

Interestingly, the degrees of association of *see* and *be* with their related modals not only represent the strongest of all associations in the data, but crucially, do so more radically than all other significant collexemes (recall Figure 13). So on the basis of the determinant of *prototypicality of meaning*, it is reasonable to hypothesise that *can see* (followed by *can do*) and *may be* (followed by *may seem*) are the most representative exemplars of their respective categories and that, more speculatively, the dynamic and epistemic senses of *can* and *may* are the two modal forms' main senses respectively (which is supported by inspecting the sense frequencies in those in which the logistic regression was most certain about its prediction). Those hypotheses come with all due caveats given the limitation of the corpus data analysed for the current work. Even more interesting is the fact that both collexemes *see* and *be* yield the same results in both native and learner English. So while this result is in line with Ellis and Ferreira-Junior (2009) who claim that "learners use first the most frequent, prototypical and generic exemplars" (abstract), it also provides further support for the claim that *can see* and *may be* can be considered as prototypical '*may*/*can* + lexical verb' constructions. Ultimately, this result also indicates that learners demonstrate that they have acquired considerable knowledge about both constructions' major usage patterns.

Adopting a constructionist approach to explore the acquisition of L2 *may* and *can* on the basis of corpus data provides a way to address the issue of speaker proficiency. As shown in Ellis and Ferreira-Junior (2009), construction learning involves the interaction between different determinants (*salience and perception*, *prototypicality*, etc.) whose individual contribution to the overall learning process is difficult to pinpoint. In line with Ellis and Ferreira-Junior's work, the collexeme analysis in the current study not only recognises the existence of interactions between determinants (*salience and perception*, *prototypicality*, etc.) but also reveals that determinants combine in ways that can reflect learners' level of proficiency in L2. This claim is based on the comparison of two sets of collexemes: on the one hand the collexemes that are jointly significant in native and learner English and on the other hand the collexemes that are significant in

strengths. As we saw in Section 3.6.5, however, according to Ellis and Ferreira-Junior (2009), the degree of prototypicality of a construction is also based on its frequency of occurrence: "[t]he greater the token frequency of an exemplar, the more it contributes to defining the category, and the greater the likelihood that it will be considered the prototype" (p. 371). The present study shows that, in addition to token frequency, collexeme strength provide a way to identify degrees of construction prototypicality and to assess the relative degrees of prototypicality of near-synonymous constructions.

learner English only. The comparison of those two sets of collexemes is interesting because if Ellis and Ferreira-Junior's thesis is correct, then collexemes that are significant in native/L2 would be expected to be more frequent, prototypical and generic than those that are only significant in the learner data. Indeed, as the first exemplars to be used by the learners, the uses of these collexemes with *may* and *can* would be expected to be more entrenched, more prototypical and schematically more generic. That is precisely the tendency emerging from the data. More concretely, while significant collexemes for *may* in native/L2 are *have*, *appear*, *think*, *sound* and *seem*, those found to be significant (also for *may*) in L2 only are *believe*, *represent*, *argue*, *wonder*. In both cases, all the collexeme verbs denote a highly generic process.[65] However, with regard to *Salience* and *Perception*, the two sets of results differ in that native/L2 collexemes denote both processes of perception (e.g., *appear*, *sound* and *seem*) and experiential processes (e.g., *have* and *think*), that is to say processes that involve the notions of possession and of intellectual activity.[66] The relation made here between experiencing an event and its salience is that the experience of an event makes that event more salient to a speaker. In contrast, L2 collexemes exclude perceptual processes altogether. However, like *think*, *wonder* and *believe*, they can be associated with experiential processes. Finally, *argue* and *represent* can also be said to refer to the experiential domain. I recognise that the above proposed collexeme analysis can only be limited and tentative given the small sample of lexical verbs that are being discussed in this section. However, generally, I hope to have demonstrated the potential usefulness of collexeme analyses for modal-related L2 acquisition research.

An additional implication emerging from the results of the collexeme analysis is concerned with the possible distinction between *salience* and *perception* as two separate determinants within *form*. Indeed, while Ellis and Ferreira-Junior approach the two

[65] By “highly generic process” I refer to courses of action that do not require any specific situational settings in order to be carried out.

[66] My categorisation of the verb *appear* may be controversial. It is motivated by the fact that the majority of occurrences where *appear* is used in the construction *may/can* + lexical verb, *appear* can be replaced with *seem* and generally refers to a perceptual process of a mental nature, as illustrated in the following examples extracted from the data: "This may *appear* to evade elitism", "as it may at first *appear*", "loss of sovereignty of the countries in favour of a central power can *appear* as a danger for some people", "may also *appear* as problems", "this new nation may also *appear* to some countries as", "this unification may *appear* as a drawback", “such sentence may *appear* as an old-fashion and caricatural statement" or "the first category may *appear* as the lesser of the two evils".

notions as a pair, the present data show that with regard to *may* and *can* at least, learners are sensitive to both notions in different ways: *perception* seems to promote a faster construction-learning process whereas *Salience* does not.

Emerging from this discussion is the notion that, with regard to L2 acquisition, synchronic corpus data present a much richer research source than the existing literature suggests. With particular regard to the L2 acquisition of *may* and *can*, the above discussion has shown that the frequencies of occurrence of *may* and *can* can help to establish a preliminary assessment of the two forms' relative degree and order of acquisition. In addition, we have seen that on the basis of synchronic corpus data *can*'s emerging 'new' patterns of use can be identified. We have also seen that, as part of constructions of the type *may/can* + lexical verb, the semantics of the lexical verbs modalised by *may* or *can* seems to interfere, to a certain extent, with *may* and *can*'s acquisition process. Given the richness and the multidimensional nature of (learner) corpus data, I hope to have demonstrated in the above discussion not only the suitability of the behavioural profile approach for corpus-based second language acquisition research but also the need to apply such an approach in order to identify fruitful directions to investigate the acquisition of interlanguage grammars.

Although I have shown that learner corpus data have a lot more to offer as traditionally assumed, I do recognise the limitations they present in that they capture a knowledge system that is specific to one moment in time. In other words, one may question how representative of the learning population the results of this corpus-based study actually are across all levels of acquisition. It would be fruitful to complement the current study by investigating whether the interactions observed in the logistic regression (i.e., CORPUS:NEG, CORPUS:CLTYPE, etc.) are also observed in a corpus of, say, intermediate or beginning learners. Furthermore, given the current compilation of longitudinal databases such as the *Longitudinal Database of Learner English* (LONGDALE), it is now possible to envisage studying, for instance, the order of acquisition hypothesis proposed in Section 7.6.3 and, particularly, the question of the order of acquisition of the two modals and how their diverse and low-variance aspects affect their acquisition by learners over time.[67]

7.7 *Future work and concluding remarks*

As the first multifactorial study of the modals in learner language, the current work has tried to advance our understanding of how *can* and *may* are used in L1 and L2 English, in what regards the two varieties differ (e.g., syntactically, semantically, etc.), and how different factors (e.g., the complexity principle) appear to explain at least several of the results (e.g., some of the interactions in the regression approach). However, given that works on the modals *can* and *may* that are corpus-based, contrastive with regard to varieties, multifactorial and predictive, and that adopt a cognitive-linguistic approach are scarce, much more exploration and testing is required before we can lay claim to a better understanding of the modals. This section provides an overview of the limitations relevant to a broader discussion of the work, with specific regard for areas of improvement for future studies.

One of the most obvious extensions is that the current work calls for further similar analyses involving a wider range of IL varieties. Such studies will not only allow researchers to assess the degree of validity of my results across a variety of IL varieties but also to address, on the basis of hopefully larger corpora, higher-level interactions that were not possible to explore in the current work. A similarly obvious extension is to broaden the scope of the modals studied. In addition to *can* and *may* as studied here, the obvious next step would be to include their closest relevant alternatives, *could*, *be able to* and *might*.

With regard to the modal verbs, much of the existing literature in L1 and L2 has mainly been concerned with their senses rather than with the forms as lexical items, broadly speaking. Given the benefits of the present multidimensional approach for the investigation of both the modals and learner language in general, the current work calls for further application of the approach to study to what degree the senses of the modals

[67] LONGDALE was initiated in January 2008 by the Centre for English Corpus Linguistics at the University of Louvain (UCL), Belgium.

yield particular distributional characteristics. Statistically, such studies could be carried out through the computing of a logistic regression involving Senses as the dependent variable (instead of Form, as is the case in the present work) and crucially involving, again, Corpus as a variable with which the other predictors can interact. From a second language use perspective, such a study would not only reveal the extent to which English learners use the senses of the modals in native-like ways but, crucially, such study would also shed light on the extent to which grammatical contexts determine the use of a modal in a particular sense.

From a cross-linguistic perspective, and given the six grammatical features identified in the current work as causing *may* and *can* to behave in non-native ways in French-English interlanguage, it would be useful to study how the grammatical behaviour of *pouvoir* compares to that of *may* and *can* in French-English interlanguage. Put differently, can the usage patterns of native *pouvoir* explain why the predictors that interacted significantly with Corpus exhibit the patterns they do? Can the usage patterns of native *pouvoir* explain the interactions that a regression on Senses would reveal? The above study of the (few) cases that the regression misclassified was a first similar step but more detailed study along these lines could be extremely interesting and would provide solidly-grounded empirical data that could potentially contribute significantly to on-going research in the field of cross-linguistic transfer.

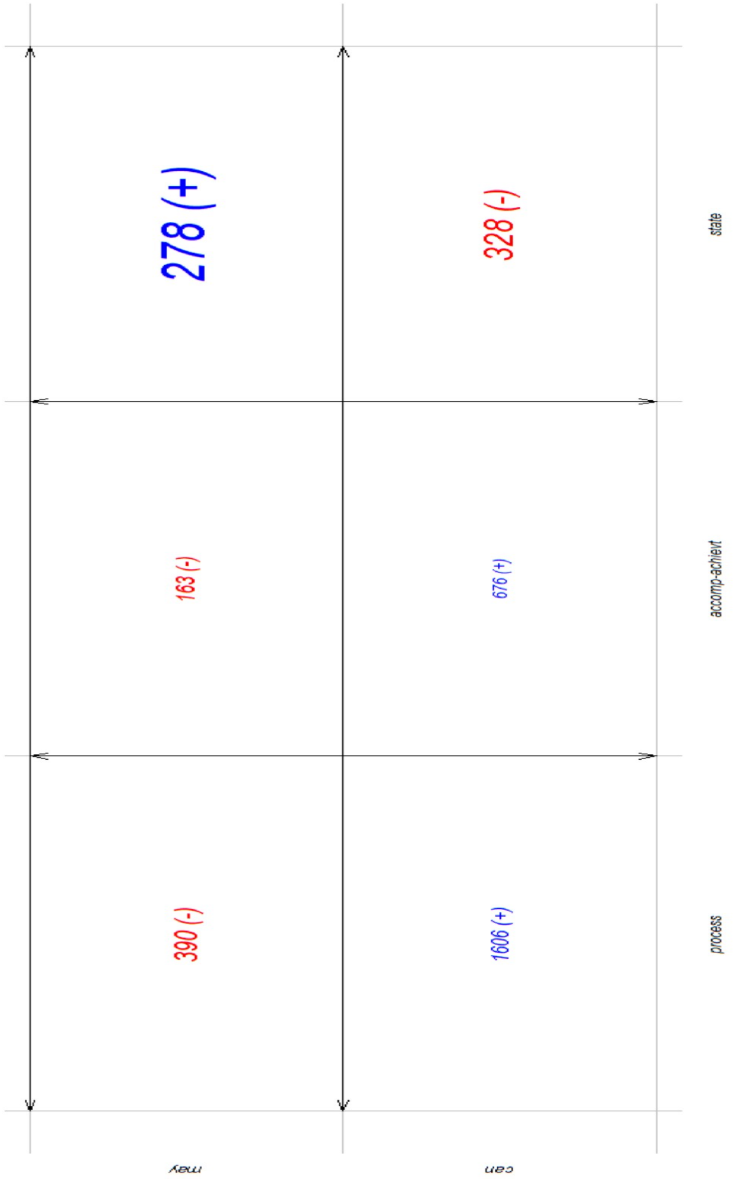
With regard to the collexeme analysis, follow-up study involving more modals would allow exploration of their contrastive patterns in more detail. That is, instead of separate distinctive collexeme analyses of *may* vs. *can* in each variety, one could either perform separate multiple distinctive collexeme analyses (of *may* vs. *might* vs. *can* vs. *could* vs. *be able to*) in each variety or even compute a hierarchical configural frequency analysis or a similar approach that would include Corpus in the analysis (cf. Stefanowitsch and Gries 2005 for an application of a conceptually similar approach).

Finally, the current work encourages interdisciplinary research enterprises. While, in this chapter, I have identified a possible connection between the schematic specificity of *may* and *can* and their co-occurrence patterns, this observation calls for further experimental research. Psycholinguistic studies on the co-occurrence patterns of *may* and *can* would allow investigation of the grammatical behaviour of the two forms in the context of their on-line production. This type of approach would allow the analyst to include in the analysis the factor 'processing speed' and work on the basis that a linguistic form that takes speakers less time to process is likely to be schematically less specified. Conversely, a linguistic form that takes speakers more time to process is likely to be schematically more specified. In the context of investigating L2 uses of *may* and *can,* such an approach would provide an experimental way to validate the proposed hypothesis that (i) there exists a connection between the degrees of schematic specificity of the modal forms and their co-occurrence patterns in learner language and that (ii) the form that is processed the fastest is the one consistently chosen in grammatically more complex environments

As an overall summary, I hope to have shown throughout this study that although learner corpus research has developed at a fast pace over the past fifteen years, the study of variation phenomena in interlanguage is still at a stage of infancy. The time has come to re-think the way to approach learner language both theoretically and methodologically. By adopting adequate methodological strategies that combine cognitive notions with sophisticated statistical methods, learner corpus research will be able to contribute significantly to the larger issue of the interplay of linguistics and cognitive aspects of language.

Appendix

Table 45 Cross-tabulation plot for the semantic variable VERBTYPE after conflation



	state	accomp-achievt	process
may	278 (+)	163 (-)	390 (-)
can	328 (-)	676 (+)	1606 (+)

Table 46 Cross-tabulation plot for the semantic variable REFANIM (no conflation)

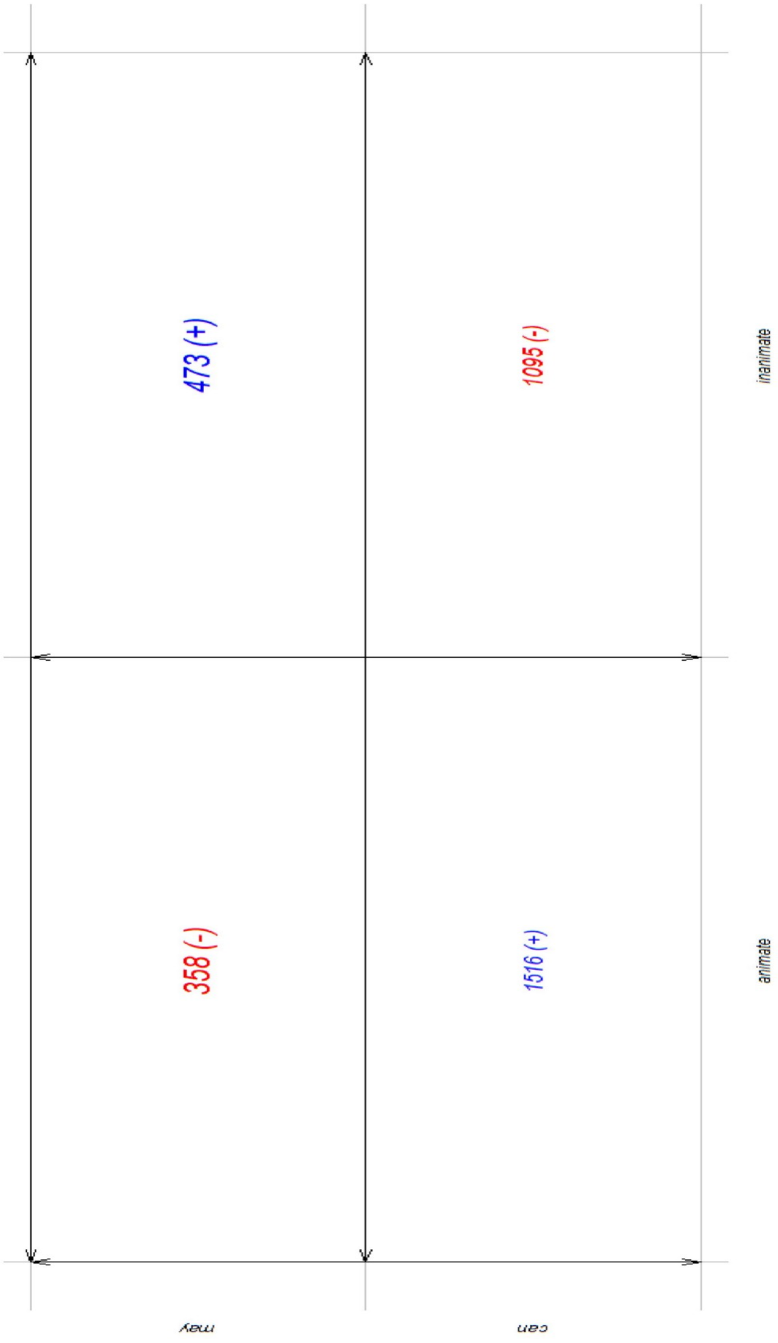
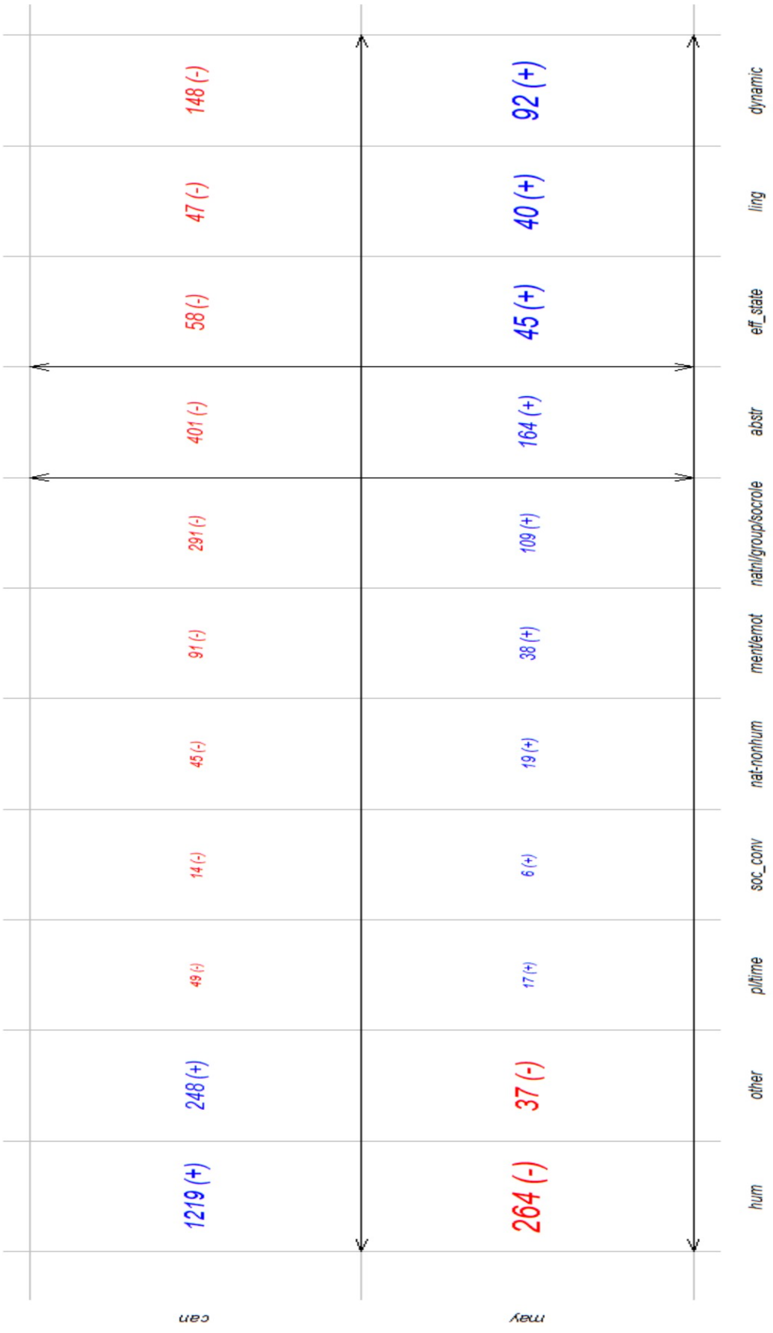


Table 47 Cross-tabulation plot for the semantic variable AnimType after conflation



	can	may
dynamic	148 (-)	92 (+)
ling	47 (-)	40 (+)
eff_state	58 (-)	45 (+)
abstr	401 (-)	164 (+)
natnl/group/socrole	291 (-)	109 (+)
ment/emot	91 (-)	38 (+)
nat-nonhum	45 (-)	19 (+)
soc_conv	14 (-)	6 (+)
pl/time	49 (-)	17 (+)
other	248 (+)	37 (-)
hum	1219 (+)	264 (-)

Table 48 Cross-tabulation plot for the semantic variable SPEAKPRESENCE

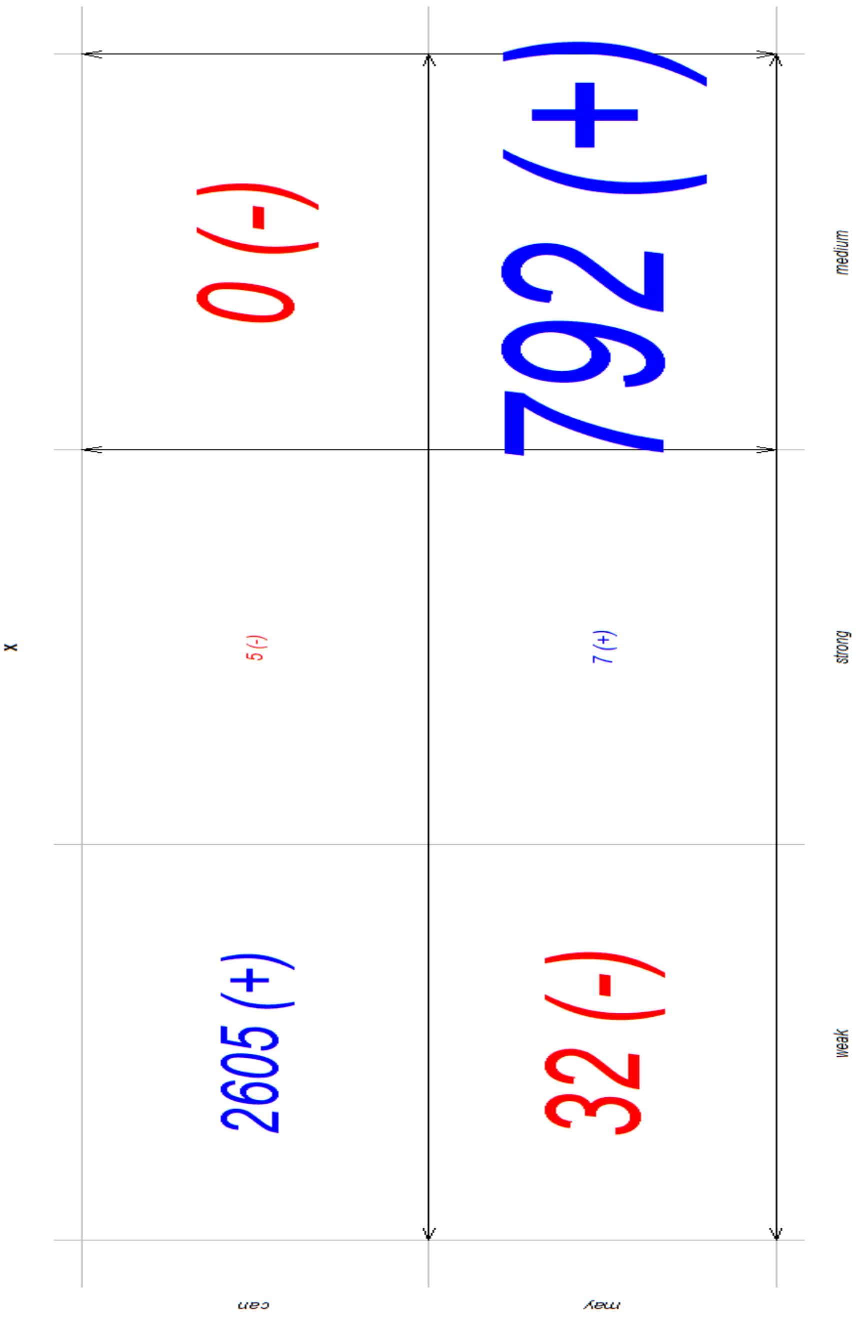
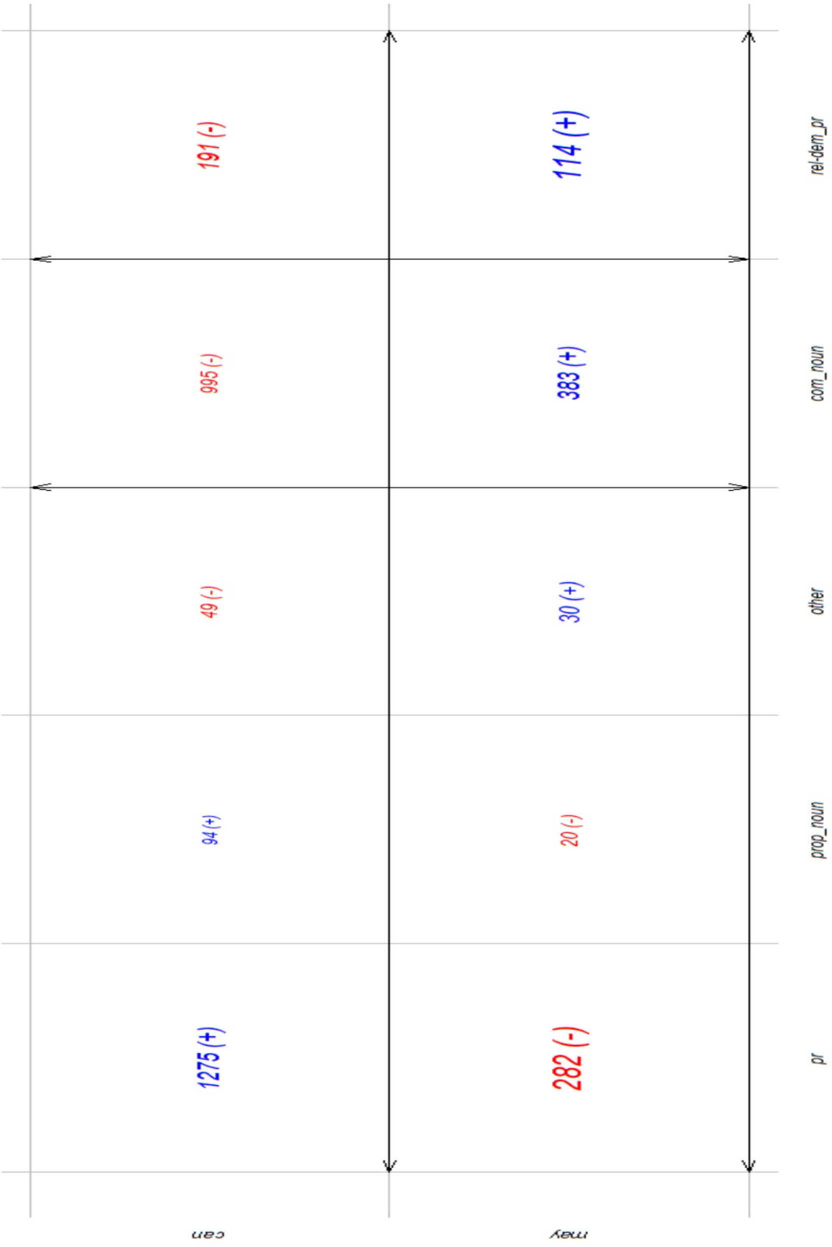
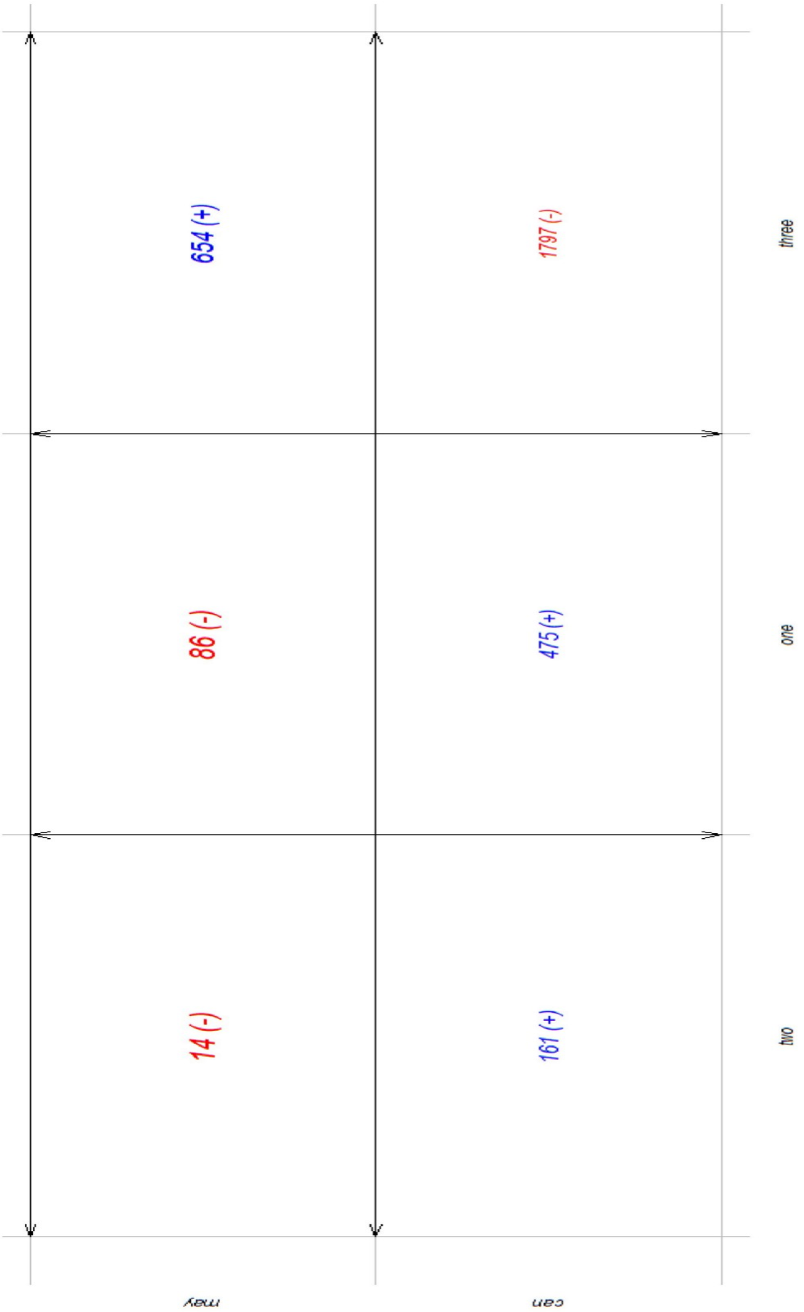


Table 49 Cross-tabulation plot for the morphological variable SUBJMORPH (after conflation)



	rel-dem_pr	com_noun	other	prop_noun	pr
can	191 (-)	995 (-)	49 (-)	94 (+)	1275 (+)
may	114 (+)	383 (+)	30 (+)	20 (-)	282 (-)

Table 50 Cross-tabulation plot for the morphological variable SUBJPERSON (no conflation)



	three	one	two
may	654 (+)	86 (-)	14 (-)
can	1797 (-)	475 (+)	161 (+)

Table 51 Cross-tabulation plot for the morphological variable Voice (no conflation)

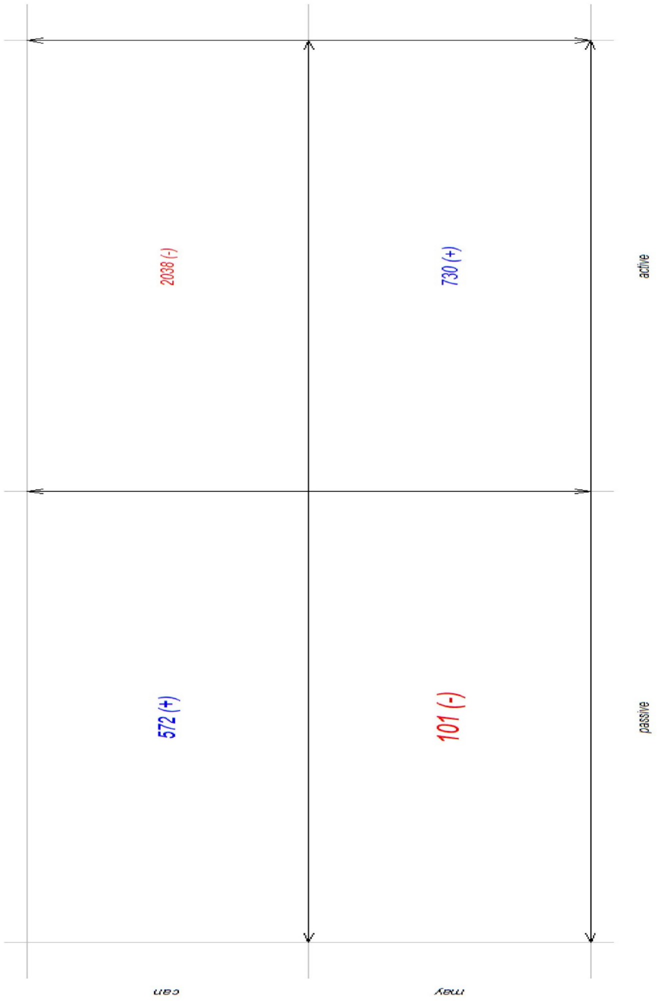


Table 52 Cross-tabulation plot for the morphological variable ASPECT (no conflation)

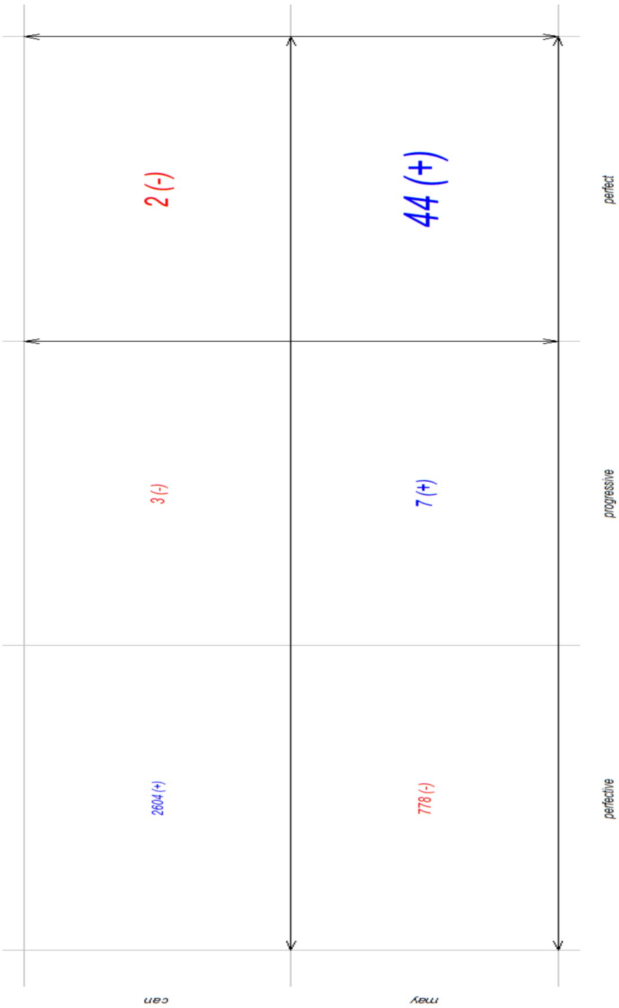
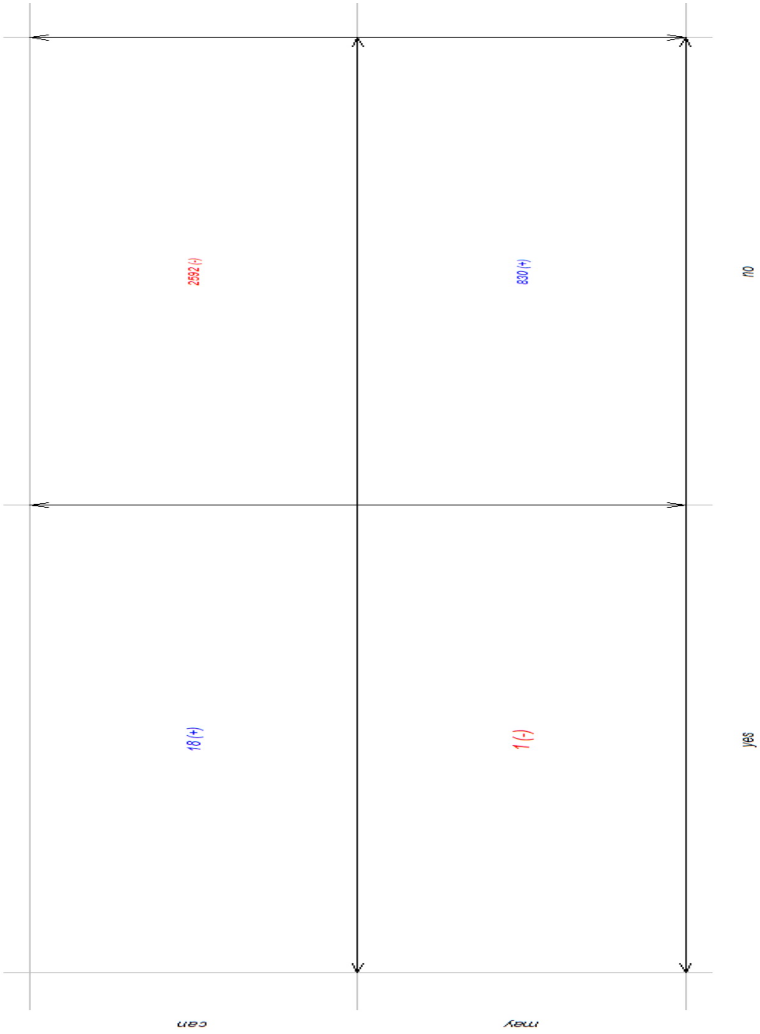


Table 53 Cross-tabulation plot for the morphological variable ELLIPTIC (no conflation)



	no	yes
can	2592 (-)	18 (+)
may	830 (+)	1 (-)

Table 54 Cross-tabulation plot for the syntactic variable SentType (no conflation)

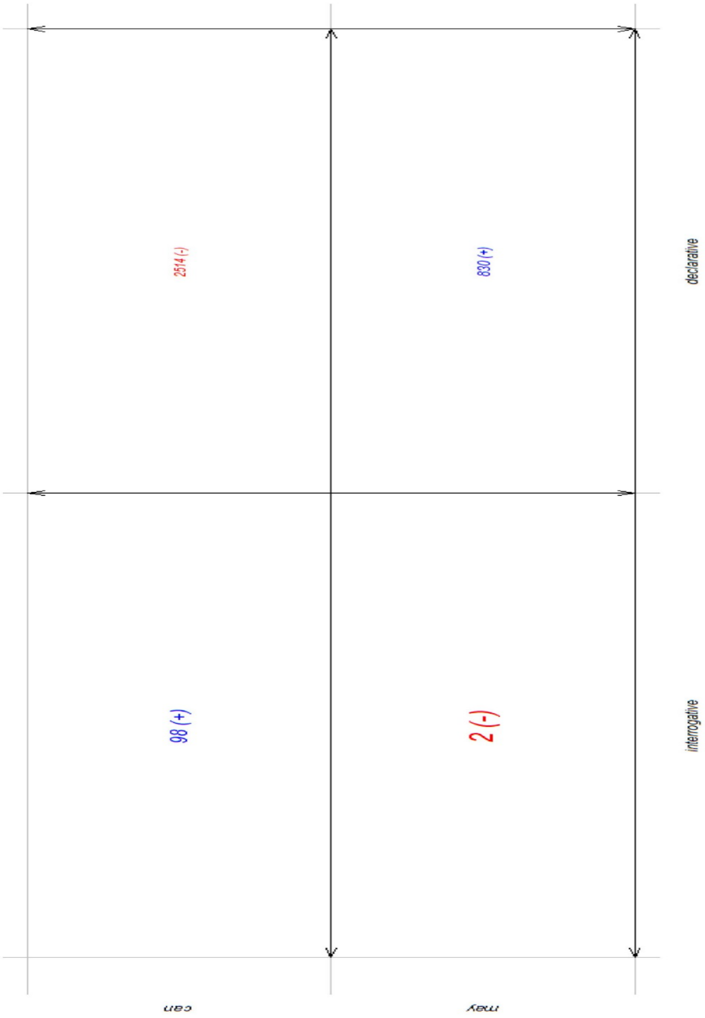


Table 55 Cross-tabulation plot for the syntactic variable ClType (no conflation)

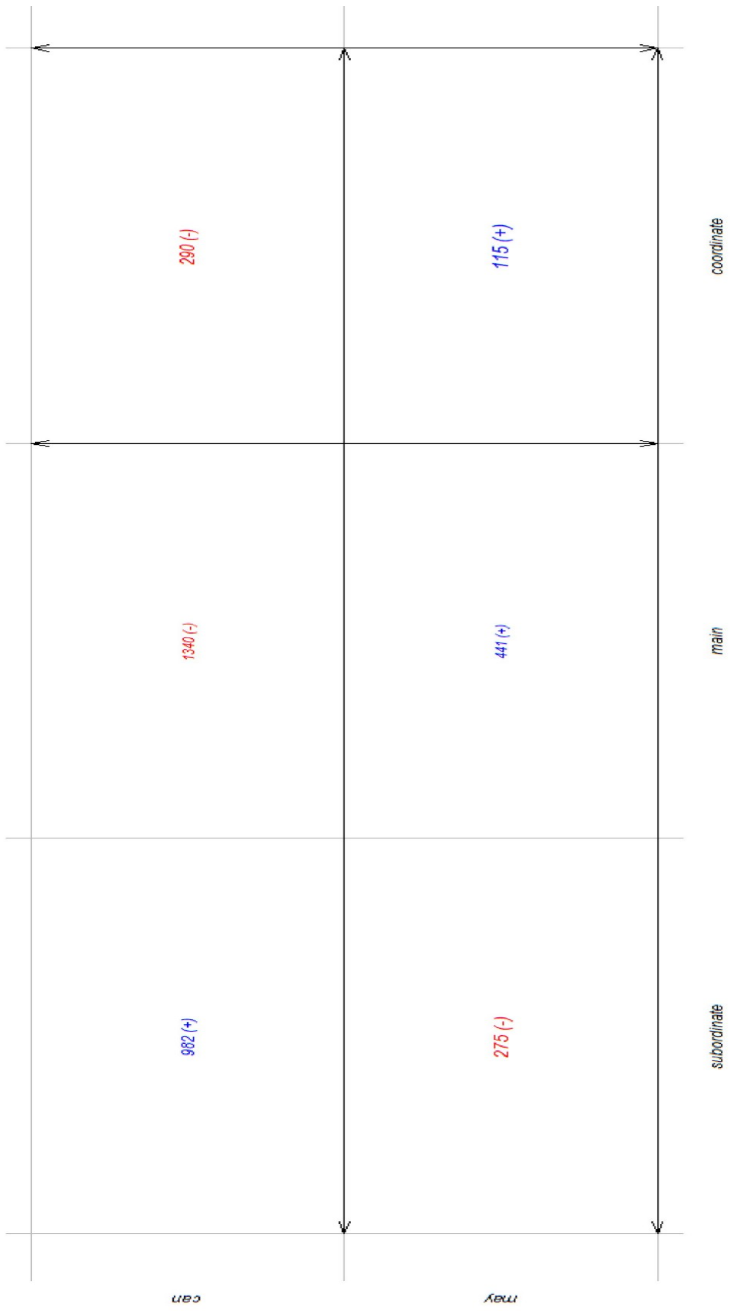


Table 56 Cross-tabulation plot for the syntactic variable NEG (no conflation)

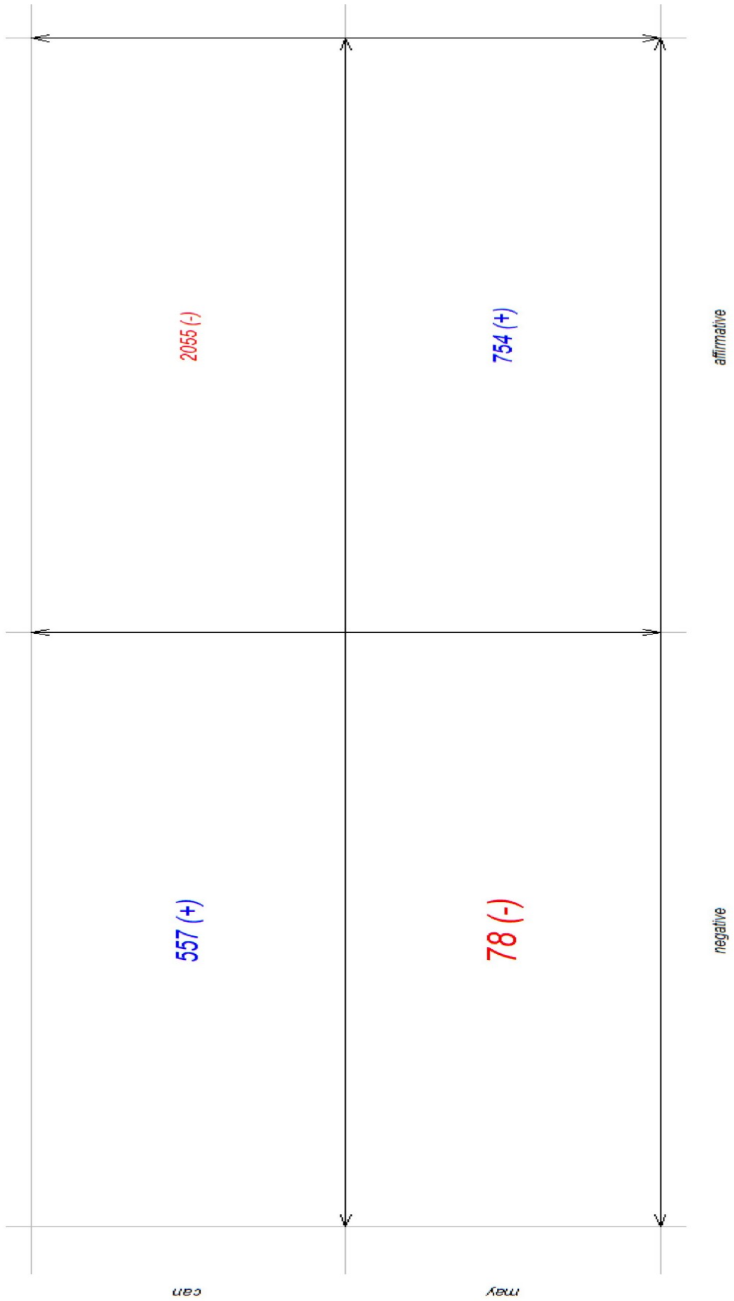


Table 57 Cross-tabulation plot for the data variable GramAcc

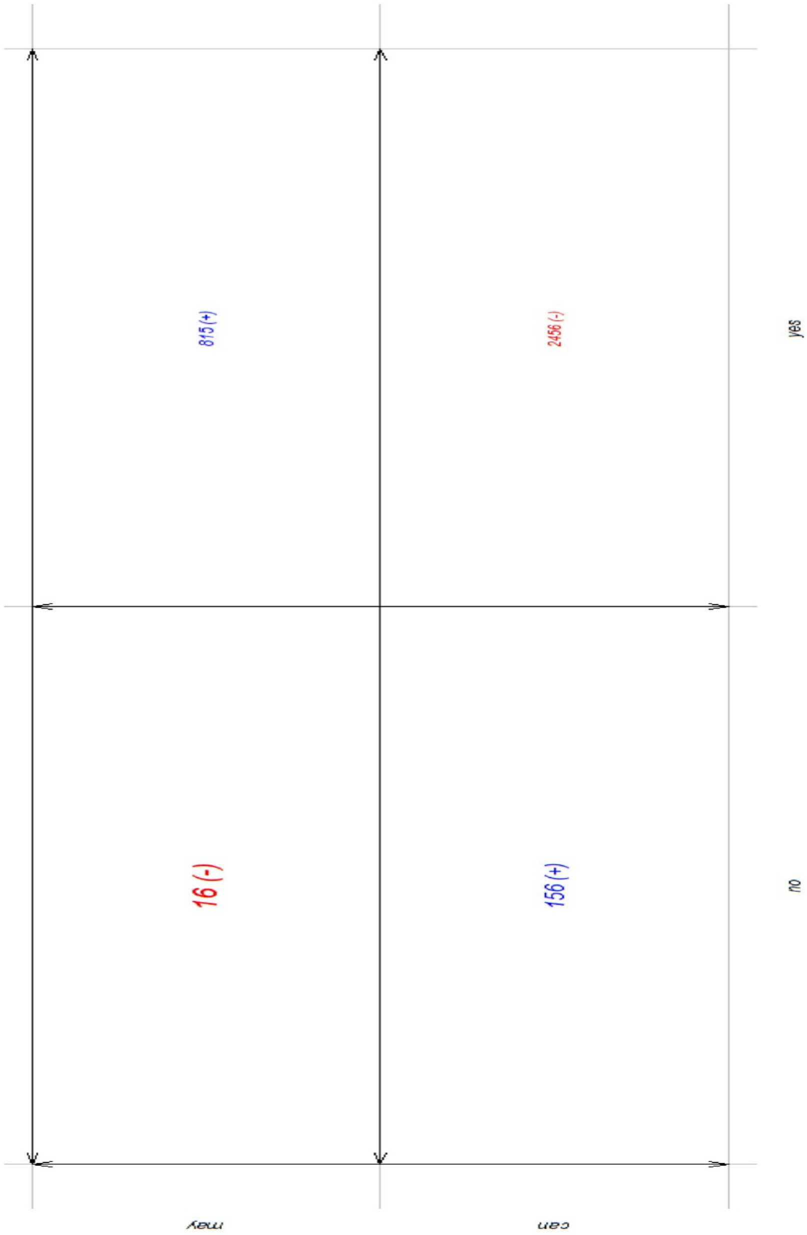


Table 58 Cross-tabulation plot for the data variable Corpus

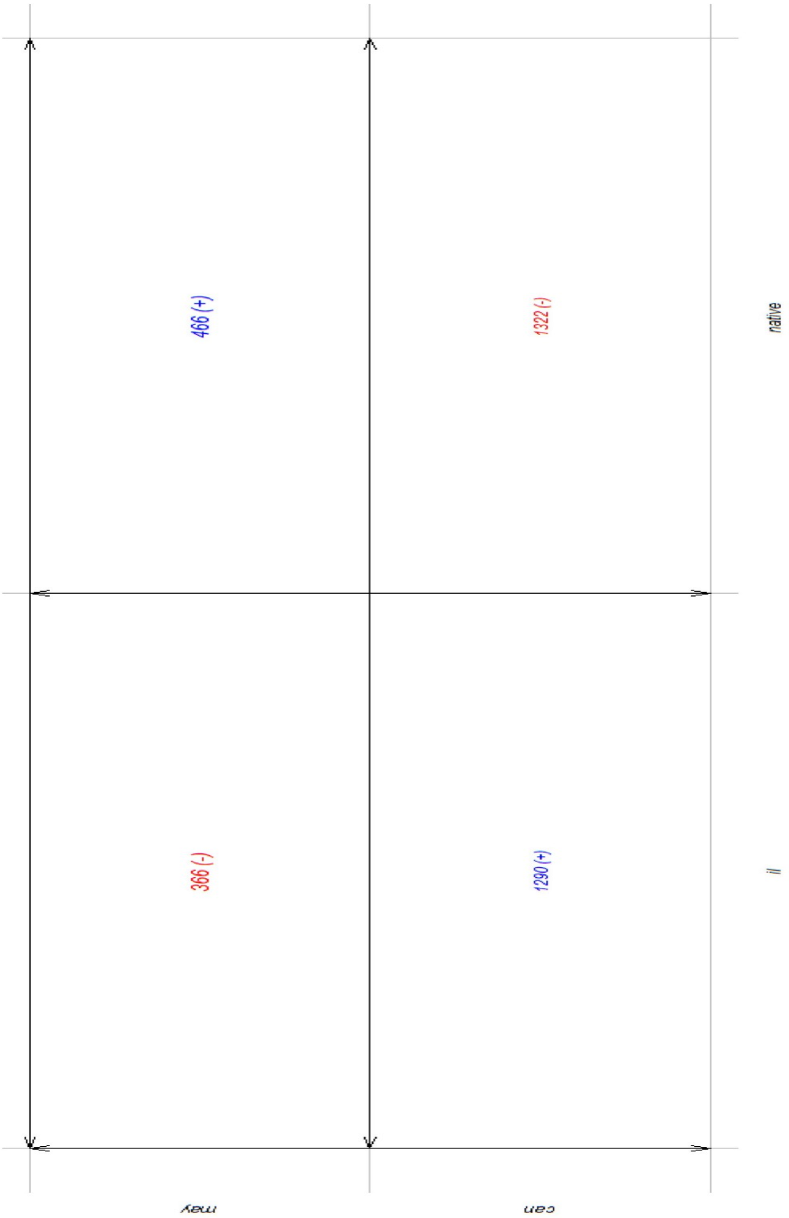


Figure 14 Bar plots of relative frequencies of CORPUS:CLTYPE

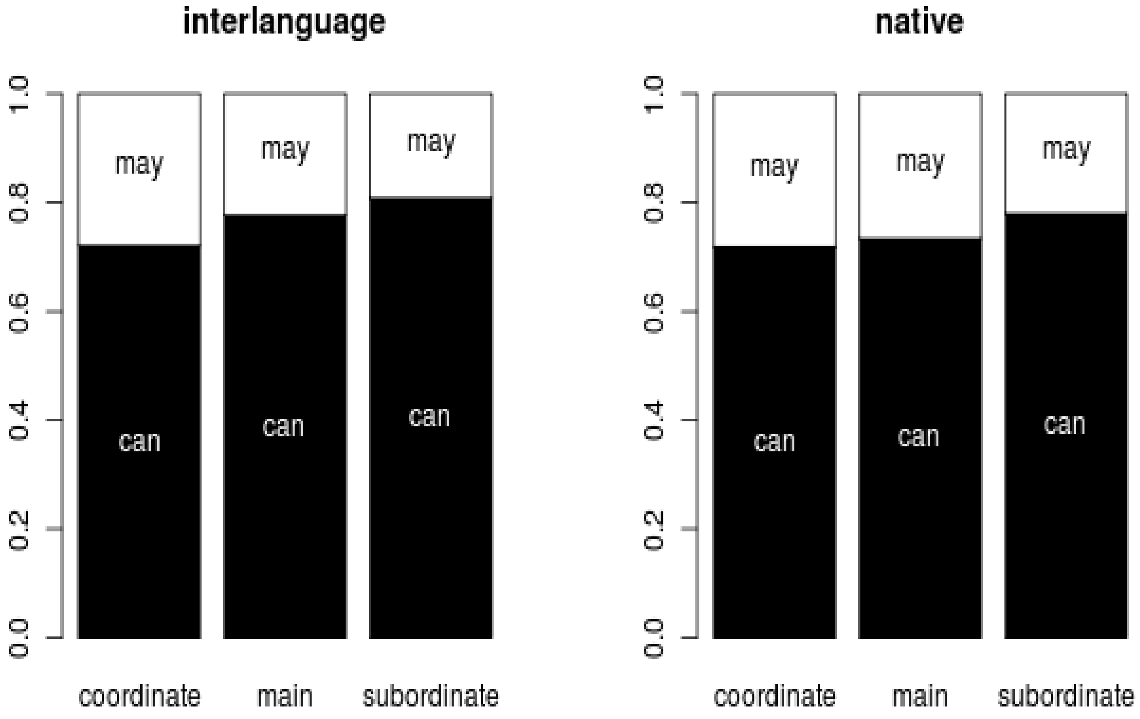


Figure 15 Bar plots of relative frequencies of CORPUS:SUBJNUMBER

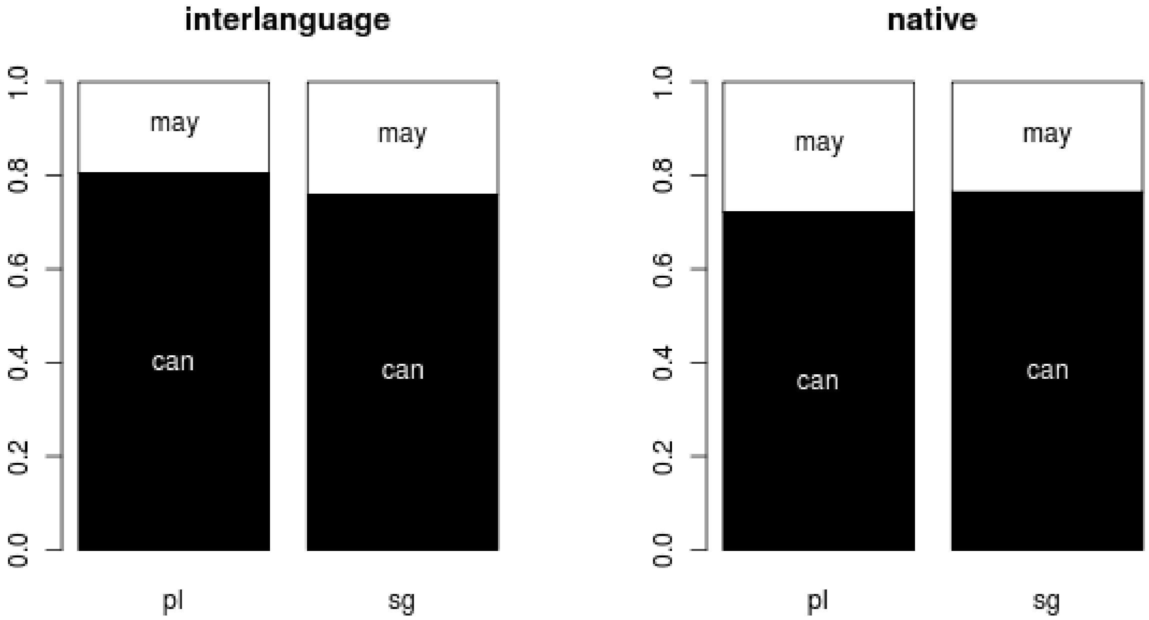


Figure 16 Bar plots of relative frequencies of Corpus:RefAnim

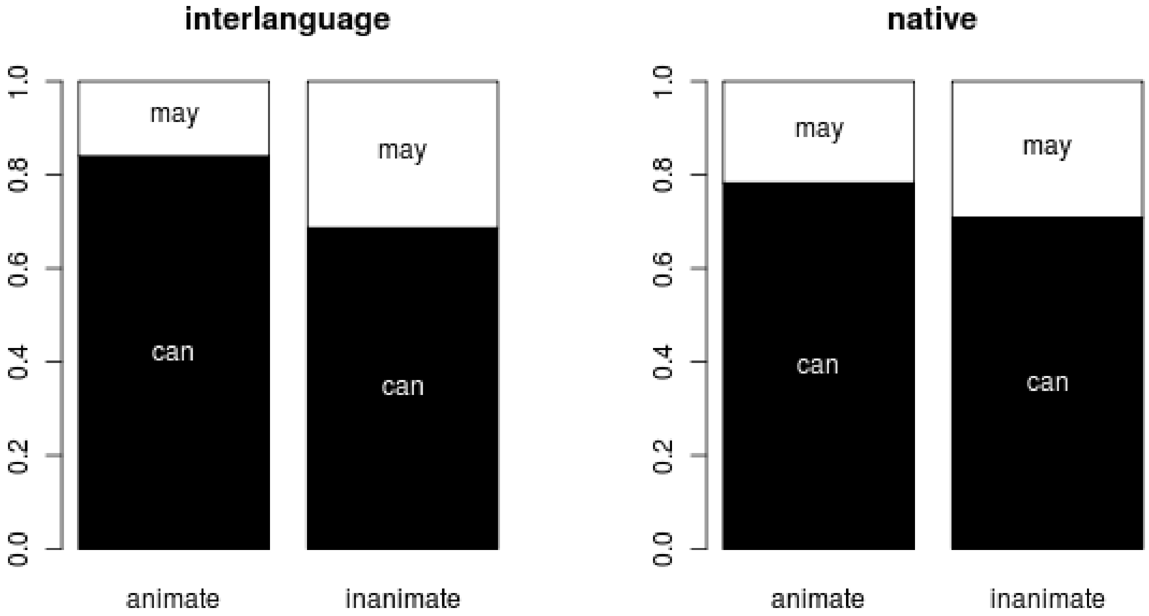


Table 59 Collexemes distinguishing between *can* and *may* in **native** English

CAN (N=338)		MAY (N=156)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
see (80:4)	6,16	be (91:79)	10,28
do (43:3)	2,9	lead to (0:8)	4,87
afford (20:0)	2,49	want (1:7)	3,46
understand (19:0)	2,36	feel (8:12)	3,1
say (19:1)	1,6	arise, sound (0:4)	2,43
learn (12:0)	1,48	grow up, need, seem, suffer (0:3)	1,82
go (11:0)	1,36	have (27:17)	1,56
expect (10:0)	1,23	think (4:5)	1,33
sympathise (9:0)	1,11	appear (1:3)	1,3
blame, relate, show (7:0)	0,86	be able to, deprive, discover, establish, face, harm, practise, prefer, require (0:2)	1,21
use (27:5)	0,8	continue, leave (2:3)	0,99
achieve, buy, compete, contract, play, prove, speak (6:0)	0,74	become (10:7)	0,99
lead (14:2)	0,69	cause (7:5)	0,81
Control, enjoy, teach, tell, transmit, travel (5:0)	0,61	contain, end_up, know, occur, persuade, receive, result	0,81
change (12:2)	0,54	affect, happen (3:3)	0,78
apply, conclude, create, deal with, do, escape, identify, imagine, justify, sell, solve. trust (4:0)	0,49	lose (4:3)	0,62
make (34:9)	0,45	find (14:7)	0,6
get (10:2)	0,4	abort, adopt, advise, ask, bring back up, centralise, cling on, close, combine, conceive, concern, consume, cost, deter, deteriorate, dictate, die, disagree, distort, eliminate, encounter, enter, evoke, exclude, fall, flare, force, go out of, handle, have, hinder, hit, impair, include, lead, lead away, lie, like, make up, marry, miss out, model, notice,	0,6

		perceive, pollute, put off, question, represent, seize back, send down, share, show_up, slave_away, slide, state, stunt, suggest, sway, take on, take back, throw, train, turn off, utilize, weed out, will, wish (0:1)	
accept, communicate, comprehend, condone, cope, determine, draw, gain, move, pass on, place, predict, produce, provide, read, recycle, reflect, sit, take on, watch, wear (3:0)	0,37	believe, kill, mean, spend, wonder (2:2)	0,58
give, out, start (6:1)	0,34	argue (7:4)	0,55
help, take (11:3)	0,27	come, live (5:3)	0,5
view, work (5:1)	0,26	try (3:2)	0,44
accomplish, admit, allow, answer, appeal, attribute, benefit, bind, break, build, calculate, categorize, come up, compare, condemn, correct, damage, define, demonstrate, depend on, describe, discuss, divide, draft, drink, exercise, express, fight, fit, forget, free, fulfil, function, get away, grow, hold, influence, inform, interfere, label, meet, modify, obtain, offer, overestimate, pass, pay, perform, put, refute, regain, replace, retire, select, separate, stay, succeed, survive, take, take up to, turn into, verify, win (2:0)	0,24	add, associate, assume, bring, choose, conform, deny, diversify, end, expand, improve, increase, last, lower, manufacture, overlook, recite, refer, refuse, reject, rule (1:1)	0,36
begin (7:2)	0,21	admire, decide (4:2)	0,33
look, prevent, remember, stop (4:1)	0,19	judge (5:2)	0,26
abandon, accrue, acknowledge, act, adapt, amend, analyse, appropriate, ask for, assert, attract, avoid, back up, break down, bring, broaden, by pass,	0,12	acquire, catch, contribute, destroy, exist, experience, push (2:1)	0,24

call, chart, chase, chose, claim, clarify, classify, clear, coexist, come about, come up, come up with, comfort, comment, commute, compromise, construct, contact, contrast, count, count on, counteract, cram, cut down, delay, derive, discriminate, devalue, differ, direct, discard, dispel, display, dissolve, distinguish, dream, drive, drive around, earn, ease, eat, embark, embody, encourage, endure, enforce, enrol, ensure, envisage, epitomise, erase, export, fall upon, fend, fertilise, fill, find out, fix, flood, focus, follow, foresee, gage, gather, get out, go, go back, go on, graduate, grasp, guarantee, halt, hear, hide, hope, illustrate, impact, implement, infect, enhance, inspire, instruct, integrate, interact, interpret, issue, jump, laugh, lay, lead into, legislate, legitimize, let, look around, look at, look down, look up, make up for, manipulate, measure, misuse, mix, mount, name, negotiate, observe, offset, open up, operate, overcome, parent, partake, participate, pass on, pick up, pinpoint, possess, pressure, print out, process, project, promote, protect, purchase, pursue, put in, range, rationalise, realise, redeem, reduce, reexamine, repeal, report, request, respect, restrict, retain, reverse, revolt, reward, rewrite, rival, ruin, screen, search, secure, shape,			

shift, shout, spare, spot, stand, stay on, store, subdue, subject, substitute, sue, sustain, tackle, take part, take away, talk, tap, tap, tolerate, torment, transfer, transform, translate, transport, turn around, turn out, turn up, underestimate, undo, veto, visualise, wait, walk in, wash, whip up, withhold (1:0)			
		alter, carry_out, consider, develop, explain, ignore, run, serve, support (3:1)	0,1

Table 60 Collexemes distinguishing between *can* and *may* in **IL** English

CAN (N=287)		MAY (N=101)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
see (75:4)	4,63	be (88:71)	10,86
do (41:0)	4,5	seem (1:11)	6,29
deny (22:0)	2,39	wonder (10:14)	3,95
live (15:0)	1,63	think (7:12)	3,9
afford, compare (14:0)	1,52	sound (0:5)	3,29
find (37:4)	1,47	appear, argue (1:6)	3,19
change, use (12:0)	1,3	lead (10:10)	2,28
prevent (11:0)	1,19	dream (1:4)	2,02
understand (16:1)	1,08	represent (0:3)	1,97
imagine (9:0)	0,97	ask (3:5)	1,8
give, make (14:1)	0,9	justify, turn out (1:3)	1,44
mention, realize (7:0)	0,75	arise, commit, disappear, exist, fear,look, mean, provide, result, stand for (0:2)	1,31
say (86:19)	0,72	believe (2:3)	1,12
conclude, play, predict(6:0)	0,65	regard (4:4)	1,11
reach (10:1)	0,57	have (18:9)	0,92
achieve, buy, escape, go, replace, see through, win (5:0)	0,54	account_for, cause, characterize (1:2)	0,9
help, take (13:2)	0,48	agree, forget, hope, try (2:2)	0,67
communicate, develop, divide, impose, inform, put, show, state, stop, view, watch (4:0)	0,43	feel (10:5)	0,65
explain (12:2)	0,42	accelerate, adore, articulate, confront, connect, deserve, devote, die, drop out, exacerbate, frighten, incite, induce, lack, lose, object, present, produce, push, reject, ruin, see, seek, sentence, shape, spoil (0:1)	0,65
work (8:1)	0,42	assert, last, refer (3:2)	0,51
adapt, benefit, defend, express, get, keep,	0,32	become (9:4)	0,49

misuse, observe, offer, perceive, remember, stand, survive, teach, tell (3:0)			
notice (10:2)	0,31	act, encourage, experience, function, give, let, meet, miss, neglect, punish, question, reconcile, reinforce, study, tempt, turn into (1:1)	0,4
draw, learn (6:1)	0,28	create (4:2)	0,4
consider (16:4)	0,26	claim (5:2)	0,31
call (9:2)	0,26	accept, come, doubt, grow, happen, influence, interpret, react, relate, reveal, save, suffer, summarize (2:1)	0,27
enjoy, remain, solve (5:1)	0,21	speak (13:4)	0,27
allow, bear, blame, break, combine, compete, conceive, continue, control, count on, criticize, deal with, decide, determine, discover, distinguish, exert, face, follow, free, get rid of, go on, guess, hear, ignore, illustrate, kill, maintain, manipulate, measure, move, obtain, perform, qualify, recreate, remark, stay, steal, succeed, suppress, talk, travel (2:0)	0,21	answer, apply, know, spend (3:1)	0,19
add, appreciate, assume, bring, read (4:1)	0,15		
put forward, acquire, address, admire, admit, affect, affirm, analyse, annihilate, approve of, arrest, ask for, assimilate, attend, attend to, avoid, base, be, be born, behave, bite off, bloom, blossom out, blot out, breed, bring, bring in, build, carry out, censure, challenge, cheat, check, chew choose, classify, come	0,1		

into, come to an end, comment, compare, complete, constitute, content, contest, convince, cope, cope_with, curse, dance, deal, define, denote, destroy, disagree, disapprove, dissociate, disturb, do away with, drive, earn, enable, enhance, ensure, enter, extract, extend, face up, fall, favour, fight, finish, foresee, foreshadow, forge, forgive, get off, get rid of, go away, guarantee, hand down, help, hope, identify, imply, improve, infer, introduce, invade, join, judge, keep alive, let out, liberate, lie, link, listen, make up, manage, muzzle, nest, occur, open out, organize, overcome, participate, penetrate, plead, point, praise, preserve, pretend, promote, prompt, prove, provoke, pull, push into, put asleep, put down to, quote, rearrange, recapture, reconstruct, reduce, reflect, regenerate, rely, rely on, remedy, reorganize, reply, rescue, resolve, rest, restore, satisfy, scrap, search, send out, separate, serve, shelter, sink, sit down, slow, spare, standardize, stick, strike, structure, sum up, supply, surround, take, take part, throw light, touch, trace back, transfer, transpose, treat, trust, turn, turn back, uncover, underestimate, undo, unsettle, visualise, wipe off, wipe out, witness, write down (1:0)			

Table 61 Behavioural Profile vectors for *can*$_{\text{il}}$, *can*$_{\text{native}}$, *may*$_{\text{il}}$, *may*$_{\text{native}}$ and *pouvoir* for all semantic predictors

ID tag	ID tag level	*can*$_{\text{il}}$	*can*$_{\text{native}}$	*may*$_{\text{il}}$	*may*$_{\text{native}}$	*pouvoir*
Senses	deontic	0,0007	0,0030	0,0219	0,0021	0,0000
Senses	dynamic	0,9977	0,9970	0,0847	0,0021	0,9925
Senses	epistemic	0,0000	0,0000	0,8934	0,9936	0,0075
Senses	NA	0,0016	0,0000	0,0000	0,0021	0,0000
Use	literal	0,9922	0,9970	0,9945	0,9893	0,9925
Use	metaphorical	0,0062	0,0030	0,0055	0,0086	0,0000
Use	NA	0,0016	0,0000	0,0000	0,0021	0,0075
VerbType	accomp/achvt	0,2566	0,2610	0,1776	0,2103	0,0868
VerbType	process	0,6310	0,5991	0,5055	0,4399	0,7811
VerbType	state	0,1109	0,1399	0,3169	0,3476	0,1245
VerbType	NA	0,0016	0,0000	0,0000	0,0021	0,0075
VerbSemantics	abstract	0,3504	0,3555	0,2623	0,3777	0,4302
VerbSemantics	act_gen/mot	0,1217	0,2020	0,0464	0,0901	0,0792
VerbSemantics	act_transf	0,0349	0,0514	0,0464	0,0644	0,0151
VerbSemantics	communication	0,1326	0,0545	0,1120	0,0279	0,0868
VerbSemantics	copula	0,0806	0,0870	0,2842	0,2639	0,1019
VerbSemantics	ment/perception	0,2729	0,2315	0,2377	0,1481	0,2755
VerbSemantics	temporal	0,0054	0,0182	0,0109	0,0258	0,0038
VerbSemantics	NA	0,0016	0,0000	0,0000	0,0021	0,0075
RefNumb	plural	0,4419	0,3593	0,3880	0,4034	0,2453
RefNumb	singular	0,5519	0,6172	0,5984	0,5708	0,7283
RefNumb	NA	0,0062	0,0234	0,0137	0,0258	0,0264
RefNumb	animate	0,6271	0,5348	0,4290	0,4313	0,5623
RefNumb	inanimate	0,3721	0,4652	0,5683	0,5687	0,4377
RefNumb	NA	0,0008	0,0000	0,0027	0,0000	0,0000
AnimType	abstract	0,1357	0,1710	0,2077	0,1888	0,2415
AnimType	dynamic	0,0434	0,0696	0,1038	0,1159	0,0528
AnimType	effect/state	0,0132	0,0310	0,0492	0,0579	0,0038
AnimType	human	0,5605	0,3752	0,3689	0,2768	0,5094
AnimType	linguistic	0,0147	0,0212	0,0601	0,0386	0,0113
AnimType	ment/emotional	0,0450	0,0250	0,0492	0,0429	0,0264
AnimType	natural/nonhum	0,0093	0,0250	0,0082	0,0343	0,0226
AnimType	national/group/ social role	0,0721	0,1498	0,0820	0,1695	0,0906
AnimType	other	0,0837	0,1059	0,0492	0,0408	0,0264

AnimType	place/time	0,0202	0,0174	0,0137	0,0258	0,0113
AnimType	social convention	0,0016	0,0091	0,0055	0,0086	0,0038
AnimType	NA	0,0008	0,0000	0,0027	0,0000	0,0000
Neg	affirmative	0,7620	0,8109	0,9399	0,8798	0,7547
Neg	negative	0,2380	0,1891	0,0601	0,1202	0,2415
Neg	NA	0,0000	0,0000	0,0000	0,0000	0,0038
SpeakPresence	medium	0,0000	0,0000	0,8962	0,9957	0,0075
SpeakPresence	strong	0,0008	0,0030	0,0191	0,0000	0,0000
SpeakPresence	weak	0,9977	0,9970	0,0847	0,0021	0,9925
SpeakPresence	NA	0,0016	0,0000	0,0000	0,0021	0,0000

Table 62 Behavioural Profile vectors for *can*$_{\text{il}}$, *can*$_{\text{native}}$, *may*$_{\text{il}}$, *may*$_{\text{native}}$ and *pouvoir* for all syntactic predictors

ID tag	ID tag level	*can*$_{\text{il}}$	*can*$_{\text{native}}$	*may*$_{\text{il}}$	*may*$_{\text{native}}$	*pouvoir*
NEG	affirmative	0,7620	0,8109	0,9399	0,8798	0,7547
NEG	negative	0,2380	0,1891	0,0601	0,1202	0,2415
NEG	NA	0,0000	0,0000	0,0000	0,0000	0,0038
SENTTYPE	declarative	0,9558	0,9690	0,9945	1,0000	0,9925
SENTTYPE	interrogative	0,0442	0,0310	0,0055	0,0000	0,0038
SENTTYPE	NA	0,0000	0,0000	0,0000	0,0000	0,0038
CLTYPE	coordinate	0,0992	0,1225	0,1366	0,1395	0,1132
CLTYPE	main	0,5760	0,4516	0,5984	0,4764	0,5925
CLTYPE	subordinate	0,3248	0,4259	0,2650	0,3820	0,2943
CLTYPE	NA	0,0000	0,0000	0,0000	0,0021	0,0000

Table 63 Behavioural Profile vectors for *can*$_{\text{il}}$, *can*$_{\text{native}}$, *may*$_{\text{il}}$, *may*$_{\text{native}}$ and *pouvoir* for all morphological predictors

ID tag	ID tag level	*can*$_{\text{il}}$	*can*$_{\text{native}}$	*may*$_{\text{il}}$	*may*$_{\text{native}}$	*pouvoir*
SubjMorph	common noun	0,3178	0,4425	0,3852	0,5193	0,3585
SubjMorph	other	0,0101	0,0272	0,0328	0,0386	0,0075
SubjMorph	pronoun	0,5767	0,4017	0,4426	0,2575	0,5358
SubjMorph	proper noun	0,0279	0,0439	0,0246	0,0236	0,0226
SubjMorph	relative/dem pronoun	0,0651	0,0809	0,1120	0,1567	0,0679
SubjMorph	NA	0,0023	0,0038	0,0027	0,0043	0,0075
SubjPerson	one	0,2643	0,1014	0,1803	0,0429	0,1245
SubjPerson	three	0,6171	0,7572	0,7350	0,8262	0,8151
SubjPerson	two	0,0628	0,0605	0,0246	0,0107	0,0000
SubjPerson	NA	0,0558	0,0809	0,0601	0,1202	0,0604
SubjNumber	plural	0,4419	0,3593	0,3880	0,4013	0,2453
SubjNumber	singular	0,5512	0,6157	0,5956	0,5730	0,7283
SubjNumber	NA	0,0070	0,0250	0,0164	0,0258	0,0264
Elliptic	no	0,9938	0,9909	1,0000	0,9957	0,9849
Elliptic	yes	0,0047	0,0091	0,0000	0,0021	0,0075
Elliptic	NA	0,0016	0,0000	0,0000	0,0021	0,0075
Voice	active	0,8140	0,7474	0,8552	0,8948	0,9208
Voice	passive	0,1853	0,2519	0,1448	0,1030	0,0566
Voice	NA	0,0008	0,0008	0,0000	0,0021	0,0226
Aspect	perfect	0,0016	0,0000	0,0164	0,0815	0,0000
Aspect	perfective	0,9953	0,9985	0,9836	0,8970	0,9774
Aspect	progressive	0,0008	0,0015	0,0000	0,0150	0,0000
Aspect	NA	0,0023	0,0000	0,0000	0,0064	0,0226

REFERENCES

Achard, Michel and Susanne Niemeier. 2004. Cognitive Linguistics, Language acquisition, and Pedagogy. In *Cognitive Linguistics, Second Language Acquisition and Foreign Language Teaching*, ed. by Michel Achard and Susanne Niemeier, 1-11. Berlin: Mouton de Gruyter.

Adjemian, Christian. 1976. On the nature of interlanguage systems. *Language Learning* 26(2):297-320.

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, ed. by Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, 55-76. Amsterdam: John Benjamins.

Altman, Roann. 1984. Assessing modal proficiency in English as a second language. Doctoral dissertation, University of Southern California.

Arppe, Antti. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonymy. Doctoral dissertation. Publications of the Department of General Linguistics, University of Helsinki.

Bartning, Inge. 2009. The advanced learner variety: 10 years later. In *The advanced learner variety: the case of French*, ed. by Emmanuelle Labeau and Florence Myles, 11-40. Oxford: Peter Lang.

Bates, Elizabeth and Brian MacWhinney. 1982. Functionalist approaches to grammar. In *Language acquisition: the state of the art*, ed. by Eric Wanner and Lila R. Gleitman, 173-218. Cambridge University Press.

Bates, Elizabeth and Brian MacWhinney. 1989. Functionalism and the competition model. In *The cross-linguistic study of sentence processing*, ed. by Brian MacWhinney and Elizabeth Bates, 3-73. Cambridge University Press.

Bialystok, Ellen and Michael Sharwood Smith. 1985. Interlanguage is not a state of mind: an evaluation of the construct for second language acquisition. *Applied Linguistics* 6(2):101-117.

Biber, Douglas. 1999. *Longman grammar of spoken and written English*. London: Longman.

Bolinger, Dwight. 1989. Extrinsic possibility and intrinsic potentiality: 7 on *may* and *can* +1. *Journal of Pragmatics*. 13:1-23.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive Foundation of Interpretation*, ed. by Gerolf Bouma, Irene Kraemer and Joost Zwarts, 69-94. Royal Netherlands Academy of Science, Amsterdam.

Bybee, Joan and Suzanne Fleischman. 1995. Modality in grammar and discourse, ed. by Joan Bybee and Suzanne Fleischman. Amsterdam: John Benjamins.

Byloo, Pieter. 2009. Modality and Negation: A Corpus-based Study. Doctoral dissertation. University of Antwerp.

Coates, Jennifer. 1980. On the non-equivalence of *may* and *can*. *Lingua* 50(3):209-220.

Coates, Jennifer. 1983. *The semantics of the modal auxiliaries.* London: Croom Helm.

Collins, Peter. 2009. *Modals and quasi modals in English*. Amsterdam: Rodopi.

Crawley, Michael J. 2007. *The R book*. Chichester, England: Wiley.

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

Deuber, Dagmar. 2010. Modal verb usage at the interface of English and a related Creole: a corpus-based study of *can/could* and *will/would* in Trinidadian English. *Journal of English Linguistics* 38(2):105-142.

Depraetere, Ilse and Susan Reed. 2006. Mood and modality in English. In *The Handbook of English Linguistics*, ed. by Bas Aarts and April MacMahon, 268-287. Oxford: Blackwell.

Divjak, Dagmar S. 2009. Mapping between domains. The aspect-modality interaction in Russian. *Russian Linguistics* 33(3):249-269.

Divjak, Dagmar S. 2010. *Structuring the lexicon: a clustered model for near-synonymy.* Berlin: Mouton de Gruyter.

Divjak, Dagmar S. and Stefan Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1):23-60.

Divjak, Dagmar S. and Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon* 3(2):188-213.

Divjak, Dagmar S. and Stefan Th. Gries. 2009. Corpus-based cognitive semantics: a contrastive study of phasal verbs in English and Russian. In *Studies in cognitive corpus linguistics*, ed. by Katarzyna Dziwirek and Barbara Lewandowska-Tomaszczyk, 273-296. Frankfurt am Main: Peter Lang.

Ellis, Nick C. and Laura Collins. 2009. Input and second language acquisition: the roles of frequency, form and function. *The Modern Language Journal* 93:329-335.

Ellis, Nick C. and Fernando Ferreira-Junior. 2009. Constructions and their acquisitions. *Annual Review of Cognitive Linguistics* 7:187–220.

Ellis, Rod. 1985. Sources of variability in interlanguage. Applied Linguistics 6(2):118-131.

Fries, Charles C. 1945. *Teaching and learning English as a foreign language*. University of Michigan Press.

Gabrielatos, Costas and Simone Sarmento. 2006. Central modals in an aviation corpus: frequency and distribution. *Letras de Hoje* 41(2):215-240.

Gass, Susan. 1996. Second language acquisition and linguistic theory: the role of language transfer. In *Handbook of second language acquisition*, ed. by William C. Ritchie and Tej K. Bhatia, 317-340. San Diego: Academic Press.

Geeraerts, Dirk. 2006. Introduction: a rough guide to Cognitive Linguistics. In *Cognitive Linguistics: basic readings*, ed. by Dirk Geeraerts, 1-28. Berlin: Mouton de Gruyter.

Gilquin, Gaëtanelle. 2008. Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In *Linking up Contrastive and Learner Corpus Research*, ed. by Gaëtanelle Gilquin, Szilvia Papp and María Belén Díez-Bedmar, 3-33. Amsterdam, Atlanta: Rodopi.

Goldberg, Adele. 1995. *Constructions: a Construction Grammar approach to argument structure*. The University of Chicago Press.

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goossens, Louis. 1992. *Cunnan*, *conne*(*n*), *can*: the development of a radial category. In *Diachrony within synchrony – language history and cognition: papers from the international symposium at the University of Duisburg*, 26-28 March 1990, 377-394. Frankfurt am Main: Peter Lang.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In *Languages in Contrast. Text-based cross-linguistic studies*, ed. by Karin Aijmer, Bengt Altenberg and Mats Johansson, 37-51. Lund University Press.

Granger, Sylviane. 1998. The computer learner corpus: a versatile new source of data for SLA research. In *Learner English on computer*, ed. by Sylviane Granger, 3-18. London: Longman.

Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In *Computer learner corpora, second language acquisition and foreign language teaching*, ed. by Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, 3-33. Amsterdam: John Benjamins.

Granger, Sylviane. 2003a. The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, ed. by Sylviane Granger, Jacques. Lerot and Stéphanie Petch-Tyson, 17-29. Amsterdam: Rodopi.

Granger, Sylviane. 2003b. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37(3):538-546.

Granger, Sylviane. 2004. Computer learner corpus research: current status and future prospects. In *Applied Corpus Linguistics: A Multidimensional Perspective,* ed. by Connor Ulla. and Upton Thomas, 123-145. Amsterdam: Rodopi.

Granger, Sylviane, Estelle Dagneaux, and Fanny Meunier. 2002. The International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylviane and Paul Rayson. 1998. Automatic lexical profiling of learner texts. In *Learner English on Computer*, ed. by Sylviane Granger, 119-131. London: Longman.

Gries, Stefan Th. 2003a. *Multifactorial Analysis in Corpus Linguistics: a study of Particle Placement*. London: Continuum Press.

Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1:1-27.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, ed. by Stefan Th. Gries and Anatol Stefanowitsch, 57-99. Berlin: Mouton de Gruyter.

Gries, Stefan Th. 2007. Coll. Analysis 3.2a. A program for R 2.1.7 and higher. http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/index.html [18/03/2011]

Gries, Stefan Th. 2008. Corpus-based methods in analyses of SLA data. In *Handbook of cognitive linguistics and second language acquisition*, ed. by Peter Robinson and Nick C. Ellis, 406-431. New York: Routledge, Taylor and Francis.

Gries, Stefan Th. 2009. *Statistics for linguists with R: a practical introduction*. Berlin, New York: Mouton de Gruyter.

Gries, Stefan Th. 2010a. Behavioural Profiles 1.01. A program for R 2.1.7 and higher. http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html [18/03/2011]

Gries, Stefan Th. 2010b. Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily …. *International Journal of Corpus Linguistics* 15(3):321-337.

Gries, Stefan Th. 2010c. Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5(3):323-346.

Gries, Stefan Th. To appear. Frequency tables, effect sizes and explorations. In *Polysemy and synonymy: corpus methods and applications in Cognitive Linguistics*, ed. by Dylan Glynn and Justyna Robinson. Amsterdam: John Benjamins.

Gries, Stefan Th. and Dagmar S. Divjak. 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In *New directions in cognitive linguistics*, ed. by Vyvyan Evans and Stephanie S. Pourcel, 57-75. Amsterdam: John Benjamins.

Gries, Stefan Th. and Dagmar S. Divjak. 2010. Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal. In, *Quantitative methods in cognitive semantics: corpus-driven approaches*, ed. by Dylan Glynn and Kerstin Fischer, 333-354. Berlin: Mouton de Gruyter.

Gries, Stefan Th. and Naoki Otani. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34:121-150.

Gries, Stefan Th. and Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1):97-129.

Gries, Stefan Th. and Stefanie Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3:182-200.

Gries, Stefan Th. and Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7:163-186.

Halliday, Michael A. K. 1970. Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of Language* 6(3):322-361.

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities*. 34(1/2):205-215.

Hawkins, John A. 1999. Processing complexity and filler-gap dependencies across grammars. *Language* 75(2):244-285.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford University Press.

Hawkins, John A. 2009. An efficiency theory of complexity and related phenomena. In *Complexity as an evolving variable*, ed. by David Gill, Geoffrey Sampson and Peter Trudgill, 252-268. Oxford University Press.

Hawkins, John A. To appear. Processing efficiency and complexity in typological patterns. In *Oxford Handbook of Language Typology*, ed. by Jae Jung Song. Oxford University Press.

Hawkins, John and Paula Buttery. To appear. Using learner language from corpora to profile levels of proficiency: insights from the English Profile Programme. In Proceedings of the 3rd ALTE Conference 2008. Cambridge University Press.

Heltoft, Lars. 2005. Modality and subjectivity. In *Modality studies in form and function*, ed. by Alex Klinge and Henrik Müller, 81-102. London: Equinox.

Hermerén, Lars. 1978. *On modality in English: a study of the semantics of the modals*. Lund: LiberLäromedel/Gleerups.

Herslund, Michael. 2005. Subjective and objective modality. In *Modality studies in form and function*, ed. by Alex Klinge and Henrik Müller, 39-48. London: Equinox.

Hoffmann, Thomas. 2006. Corpora and introspection as corroborating evidence: the case of preposition placement in English relative clauses. *Corpus Linguistics and Linguistic Theory* 2(2):165-195.

Huddleston. Rodney D. and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English Language*. Cambridge University Press.

Janda, Laura A. 2009. What is the role of semantic maps in cognitive linguistics? In *Cognitive approaches to language and linguistic data. Studies in honor of Barbara Lewandowska-Tomaszczyk*, ed. by Piotr Stalmaszczyk and Wieslaw Oleksy, 105-124. Frankfurt am Main: Peter Lang.

Jarvis, Scott. 2000. Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning*. 50(2):245-309.

Jordens, Peter. 1997. Introducing the basic variety. *Second Language Research* 13(4):289-300.

Kemmer, Suzanne and Michael Barlow. 1999. Introduction: a usage-based conception of language. In *Usage-based models of language*, ed. by Michael Barlow and Suzanne Kemmer, vii-xxviii. Stanford, CA: Center for the Study of Language and Information.

Kennedy, Graeme. 2002. Variation in the distribution of modal verbs in the British National Corpus. In *Using corpora to explore linguistic variation*, ed. by Randi Reppen, Susan M. Fitzmaurice and Douglas Biber, 73-90.Amsterdam: John Benjamins.

Kilborn Kerry and Takehiko Ito. 1989. Sentence processing strategies in adult bilinguals. In *The cross-linguistic study of sentence processing*, ed. by Brian MacWhinney and Elizabeth Bates, 257-291. Cambridge University Press.

Klinge, Alex and Henrik Høeg Müller. 2005. Modality: Intrigue and Inspiration. In *Modality studies in form and function*, ed. by Alex Klinge and Henrik Müller, 1-4. London: Equinox.

Lado, Robert. 1957. *Linguistics across cultures: applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.

Lakoff, George. 1990. The Invariance Hypothesis: is abstract reason based on image-schemas? *Cognitive Linguistics* 1(1):39-74.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford University Press.

Langacker, Ronald W. 1990. Subjectification. *Cognitive Linguistics* 1:5-38.

Langacker, Ronald W. 1999. A dynamic usage-based model. In *Usage-based models of language*, ed. by Michael Barlow and Suzanne Kemmer, 1-63. Stanford, CA: CSLI.

Larreya, Paul and Claude Rivière. 1999. *Grammaire explicative de l'anglais.* Nouvelle édition. London: Longman.

Leech, Geoffrey. 1969. *Towards a semantic description of English*. Bloomington: Indiana University Press.

Leech, Geoffrey. 2004. *Meaning and the English Verb*. 3rd ed. Longman.

Lenneberg, Eric H. 1967. *Biological foundations of language*. John Wiley and Sons Inc.

Le Querler, Nicole. 1996. *Typologie des modalités.* Presse Universitaire de Caen.

Lyons, John. 1977. *Semantics*. Cambridge University Press.

Meunier, Fanny. 1998. Computer tools for the analysis of learner corpora. In *Learner English on computer*, ed. by Sylviane Granger, 19-37. London: Longman.

Neff, JoAnne, Emma Dafouz, Honesto Herrera, Francisco Martínez and Juan Pedro Rica. 2003. Contrasting the use of learner corpora: the use of modal and reporting verbs in the expression of writer stance. In *Extending the scope of corpus-based research. New applications. New challenges*, ed. by Sylviane Granger and Stéphanie Petch-Tyson, 211-230. Amsterdam: Rodopi.

Nuyts, Jan. 2001. Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics* 33(3):383–400.

Nuyts, Jan. 2005. The modal confusion: on terminology and the concepts behind it. In *Modality studies in form and function*, ed. by Alex Klinge and Henrik Müller, 5-38. London: Equinox.

Nuyts, Jan. 2006. Modality: overview and linguistic issues. In *The expression of modality*, ed. by William Frawley, 1-26. Berlin: Mouton de Gruyter.

Palmer, Frank. 1979. *Modality and the English modals*. London: Longman.

Palmer, Frank. 1986. *Mood and modality.* Cambridge University Press.

Palmer, Frank. 1990. *Modality and the English modals.* 2nd ed. Longman.

Palmer, Frank. 1995. Negation and the modals of possibility and necessity. In *Modality in grammar and discourse*, ed. by Joan Bybee and Suzanne Fleischman, 453-472. Amsterdam: Benjamins.

Perkins, Michael. 1983. *Modal expressions in English*. London: Pinter.

Radden, Günter. 2007. Interaction of modality and negation. In *Cognition in Language: Volume in Honour of Professor Elżbieta Tabakowska*, ed. by Władysław Chłopicki, Andrzej Pawelec, Agnieszka Pokojska, 224-254. Kraków: Tertium.

Rayson, Paul and Andrew Wilson. 1996. The ACAMRIT semantic tagging system: progress report. In *Language engineering for document analysis and recognition*, LEDAR, AISB96 workshop proceedings, 13-20. Brighton, England.

R Development Core Team. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.

Reppen, Randi, Susan M. Fitzmaurice and Douglas Biber. 2002. Introduction. In *Using corpora to explore linguistic variation*, ed. by Randi Reppen, Susan M. Fitzmaurice and Douglas Biber, vii-xii. Amsterdam: John Benjamins.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2):149-182.

Salkie, Raphael. 1996. Modality in English and French: a corpus-based approach. *Language Sciences* 9(1-2):381-392.

Salkie, Raphael. 2001. Probability and necessity in English and German. In *Proceedings of the symposium "Information structure in a cross-linguistic perspective",* held at the Norwegian Academy of Science and Letters, November 30- December 2 2000.

Salkie, Raphael. 2004. Towards a non-unitary analysis of modality. In *Contrates: mélanges offerts à Jacqueline Guillemin-Flescher*, ed. by Lucie Gournay and Jean-Marie Merle, 169-182. Paris: Ophrys.

Salkie, Raphael. 2009. Degrees of modality. In *Modality in English: theory and description*, ed. by Raphael Salkie, Pierre Busuttil and Johan van der Auwera, 79-103. Berlin: Mouton de Gruyter.

Selinker, Larry. 1969. Language transfer. *General Linguistics* 9(2):67-92.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10(3):209-231.

Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2):209-43.

Stefanowitsch, Anatol and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1):1-43.

Sweetser, Eve. 1990. *From etymology to pragmatics: metaphorical and cultural aspects of semantic structure*. Cambridge University Press.

Talmy, Leonard. 2000. Towards a cognitive semantics. Cambridge, MA: MIT Press.

Tomasello, Michael. 2000. Do young children have adult syntactic competence? *Cognition* 74:209-253.

van der Auwera, Johan.1996. Modality: the three-layered scalar square. *Journal of Semantics* 13:181-195.

Vanparys, Jan. 1987. Toward a pragmatic approach to modality: the case of permissive *can* and *may*. In *The pragmatic perspective. Selected papers from the 1985 International Pragmatics Conference*, ed. by J. Verschueren, 229-238. Amsterdam: John Benjamins.

Vendler, Zeno. 1957. Verbs and times. *Linguistics in Philosophy* 66(2):143-160.

Verstraete, Jean-Christophe. 2001. Subjective and objective modality: interpersonal and ideational functions in the English modal auxiliary system. *Journal of Pragmatics* 33:1505-1528.

Verstraete, Jean-Christophe. 2005. Scalar quantity implicatures and the interpretation of modality problems in the deontic domain. *Journal of Pragmatics* 37:1401-1418.