http://dx.doi.org/10.1590/0102-445071548936077492

How Brazilian learners express modality through verbs and adverbs in their writing: a corpus-based study on n-grams

Como alunos brasileiros expressam modalidade por meio de verbos e advérbios em sua escrita: um estudo de corpus *baseado em* n-grams

Adriana Maria Tenuta (Universidade Federal de Minas Gerais)
Ana Larissa A M Oliveira (Universidade Federal de Minas Gerias)
Bárbara Malveira Orfanó (Universidade Federal de São João del Rei)

RESUMO

Com base na visão de modalidade do arcabouço teórico da sintaxe descritiva, este trabalho analisou um corpus de aprendizes em comparação com um corpus de falantes nativos do inglês com o objetivo de identificar padrões diferenciados na expressão de valores modais. Para tanto, o estudo concentrou sua análise em n-grams contendo verbos modais e advérbios que expressam modalidade. Essa análise revelou a prevalência dos valores modais epistêmicos em ambos os corpora, bem como a existência de padrões distintos na expressão desse tipo de modalidade. No corpus de não-nativos, percebeu-se uma expressão mais restrita da modalidade quando comparada à do corpus de nativos. No corpus de nativos, houve uma prevalência de advérbios com sentido modalizador. Nessa comparação, percebeu-se também um uso diferenciado de alguns verbos modais. Este estudo pode contribuir para a área emergente dos estudos linguísticos de corpora e para a área da sintaxe, com possíveis implicações para o ensino de escrita acadêmica em inglês.

Palavras-chave: *corpora de aprendizes; modalidade; modais; advérbios.*

ABSTRACT

Based on the view of modality in the theoretical framework of descriptive syntax, this study examined a corpus of learners compared with a corpus of native speakers of English, aiming to identify different patterns of expression of modal meanings, particularly, adverbs and modal verbs. Therefore, the study focused its analysis on n-grams containing modal verbs and adverbs that express modality. This analysis revealed the prevalence of epistemic values in both corpora, and the existence of distinct patterns in the expression of this type of modality. In the non-native corpus, the expression of modality is restricted when compared to the native speakers'. In the corpus of native speakers, there was a prevalence of adverbs with modalizing meanings. In addition, learners tend to use some modal verbs differently. This study may contribute to the emerging field of corpora linguistic studies as well as to the area of syntax, with possible implications for the teaching of academic writing in English.

Key-words: *learners' corpora; modality; modal verbs; adverbs.*

1. Introduction

Corpus-based studies on learners' production of written discourse have caught the attention of many researchers from different domains. Despite the difficulties in compiling and analysing students' production, recent findings have contributed to the understanding of their interlanguage by identifying linguistic features that are prevalent in their discourse (BERBER-SARDINHA, 2001; DUTRA, 2009, in Brazil). In this paper, we analyze a learners' corpus aiming at identifying how Brazilian learners of English express stance and attitude by employing modality items in their academic writing. We shall compare their production to that of native speakers of English in the same setting, that is, in the academic writing scenario. In order to do so, two corpora were analyzed: our reference corpus, CABrI - Corpus of Brazilian English Learners, in construction (BERBER-SARDINHA, 2001; DUTRA, 2009), and LOCNESS - Louvain Corpus of Native English Essays (GRANGER; DAGNEAUX; MEUNIER; PAQUOT, 2009). We believe that such an approach to the study of modality in English can contribute to the emerging area of corpora and the area of syntax. In order to begin this task, let us first outline the theoretical

DELTA
31.2
2015

framework we use in this paper, based on descriptive syntax analysis and corpora studies.

Modality has been studied by many researchers from different approaches. Following the Hallidayan (2004) model, modality conveys stance and attitude of language user. In this sense, the system of modality constitutes a region of uncertainty in discourse, displaying several degrees of commitment to the statements being expressed. Interpersonal meaning, then, plays an important role in relation to the topic in this study, which proposes to investigate how Brazilian learners of English express modality in the production of academic essays.

From a pedagogical perspective, research like the one conducted by Holmes (1997) has shown that teaching materials, in general, under-represent the use of modality in English, since this grammatical category is often dealt with through the restricted presentation of modal verbs. The present study argues along the same lines, as it also shows that students seem to heavily rely on modal verbs instead of other forms of modality that are also present in the native speaker's corpus. In this paper, modal verbs and adverbs were identified as the most common forms of modality found in learners' essays. For this reason, we set off to investigate how learners express stance and attitude through these forms. The structure of this paper is as follows: introduction, literature review and theoretical framework, analysis, and conclusion.

2. Literature Review

2.1. Mood and Modality in the English language

According to Palmer (2001), modality refers to the expression of the speaker's attitude or opinion regarding the proposition of a clause (PALMER, 2001). It is a linguistic feature generated by a variety of linguistic phenomena, as described by Downing and Locke (2006), among which modal verbs play a central role. Modality is to be understood as a grammatical category that covers notions such as possibility, probability, necessity, volition, obligation and permission. Mood, in its turn, is the realization of modality by the verbal morphology.

Therefore, modality can be connected to basic logical meanings, generating a few types: (a) epistemic modality, (b) deontic modality and (c) dynamic modality (DOWNING; LOCKE, 2006; HUDDLESTON; PULLUM, 2005; CARTER; MCCARTHY, 2006). The first two meanings, epistemic and deontic, are the central ones.

Epistemic modality comprises the various degrees of certainty/uncertainty referent to facts and therefore concerns the limitations on the speaker's knowledge about these same facts. Consequently, epistemic modality expresses meanings related to inference, prediction, expectation and probability (BIBER et al, 1999; DOWNING; LOCKE, 2006). Examples of epistemic modality are illustrated below:

It might rain tomorrow.
I suppose she did it by herself.
It's very unlikely that they will accept our offer.

Deontic modality, on the other hand, expresses meanings related to permission and obligation of various kinds, ranging from very strong obligation to more mild ones. Deontic modality is, therefore, very often associated with authority and judgment, rather than with knowledge or prediction, as it happens with epistemic modality. For this reason, deontic modality is a language resource used to influence people to do (or not to do) things, whereas epistemic modality is used to predict what speakers think is likely to happen.

Epistemic and deontic meanings can be associated with the same modal expression. For example, a modal verb can express both deontic and epistemic meanings, depending on the context given.

It must have been him. (epistemic)
You must leave now. (deontic)

Additionally, on many occasions, modality meanings may be rather ambiguous, allowing either interpretation. This is the case of the example below, in which *must* can express epistemic meanings of prediction or possibility (contextualized as: *I assume you are patient/she likes it, given certain evidences*) or deontic meanings of obligation or necessity (contextualized as: *there is a need for you to be very patient/there is a need for her to like it, according to my understanding of the situation*).

DELTA
31.2
2015

You must be very patient.
She must like it.

Although modality is centrally related to epistemic or deontic meanings, as we have stated, there are also other kinds of meanings associated with modality, all of them, however, play a more peripheral role in syntax analysis and are grouped under the label: dynamic.

These dynamic meanings are described as ability and courage (DOWNING; LOCKE, 2006) and ability, volition and courage (HUDDLESTON; PULLUM, 2005). They are often expressed by modal verbs like *can* and *will* and by semi-modals like *dare*. In this regard, according to Downing and Locke (2006), *dare* is actually the only semi-modal that is used only in the dynamic expression of modality.

Some examples of dynamic modality are displayed below:

I can speak Spanish. (ability)
She wants me to go, but I won't. (volition)
I daren't say this. (courage)

In certain cases, we can interpret an occurrence both as dynamic and epistemic at the same time, since both types of meanings can be identified:

You can't be right. (probability and/or ability)
She can play the piano. (possibility and/or ability)
I can speak four languages (possibility and/or ability)

There are authors that group modality meanings differently. Biber et al (1999) identifies three categories of modal verbs: (a) permission/possibility/ability (*can, could, may, might*), (b) obligation/necessity (*must, should, (had) better, have (got) to, need to, ought to, be supposed to*) and (c) volition/prediction (*will, would, shall, be going to*). This categorization does not correspond exactly to the distinction deontic/epistemic adopted in this work.

Modality also conveys meanings related to the concept of remoteness, illustrated in the examples below, in which the various degrees of remoteness imply a distinction between two kinds of

conditional construction, *open vs. remote*, as stated by Huddleston and Pullum (2005).

I hope she recovers soon. (open)
If she liked the place, she would have stayed. (remote)

We now present a broad picture on the realizations of modality in the English language:

At this point, it is important to focus on the distinctions among the terms modal, mood and modality. Modality is the most general term for this grammatical category in focus. It refers to a basic distinction of *realis* and *irrealis* meanings (PALMER, 2001). Mood would refer to the traditional system composed of the categories indicative, subjuntive and imperative, characterized by morphological inflexion on the verbal group (VG). The mood system, therefore, expresses modality. Indicative is the mood of certainty, expressing, *realis* meaning. Subjunctive, contrarily, is a morphological indication of *irrealis*, expressing doubt, possibility, uncertainty. Imperative, considering its imposition of a demand on the listener, is also *irrealis* in essence, since there is no assertion on the part of the speaker of the realization or effective occurrence of the event expressed in the proposition (TENUTA, 1995; 2006). Modal, by its turn, is the term used to refer to a set of auxiliary verbs, very frequent in the English language, that also expresses, in its variety, a profusion of nuances of *irrealis* content. Therefore, mood and modals are linguistic resources, of different kinds, for the expression of all types of modality. And these two resources are not the only ones.

Modality meanings may be reached, through different forms composing the VG: mood inflexions, modal auxiliaries, semi-modals, lexical auxiliaries, phased structures, lexical verbs. Modality meanings may also be found elsewhere in the clause, in elements such as adverbials, predicate adjectives and certain nouns.[1] These expressions are presented below (DOWNING; LOCKE, 2006):

1. Broadly speaking, *irrealis* can also be expressed through still other resources, such as, negative structures, interrogative structures, discourse markers (especially in oral language) showing little commitment of the speaker in relation to the realization of the content of his/her speech (TENUTA, 1995; 2006).

DELTA
31.2
2015

When expressed in the VG, modality can be realized by:
(a) modal verbs: *may, might, should, must, can, would, will, ought to, shall, could, need*;
(b) semi-modals (modals in certain uses): *need, dare, wish*;
(c) lexical auxiliaries (chain-like structures with primary verbs *be* and *have*): *be able to, be apt to, be due to, be going to, be liable to, be likely to, be certain to, be sure to, be to, be unlikely to, be supposed to, have to, have got to, had better, would rather, would sooner*;
(d) phased structures composed of a catenative verb, such as *need, want, regret, try, manage, hesitate, happen, chance, tend, seem, appear, pretend,* followed by a nonfinite verb form;
(e) subjunctive forms in embedded clauses, introduced by verbs such as: *expect, suppose, recommend, require, request, suspect, intend, think, guess, assume*.
(f) lexical verbs such as *allow, beg, command, forbid, guarantee, guess, promise, suggest, warn*;
(g) imperative forms;
(h) past tense to indicate remoteness from reality, as in *I thought I'd go along with you*, *if you don't mind* and
(i) conditional structures, as in *If you went, I would go too*.

Modality expressed elsewhere in the clause, may be found in:
(j) adverb and sentence modifiers: *maybe, supposedly, perhaps, possibly*;
(k) predicate adjectives: *possible, impossible, likely, conceivable, doubtful, certain, sure, positive* and
(l) nouns such as *possibility, probability, chance, likelihood*.

Modality can also be expressed at different points throughout the clause, concomitantly. Downing and Locke (2006) refer to this realization as modal harmony. According to them, modal harmony can be illustrated with the following example:

I'm sure she couldn't possibly have said that.

The categories proposed by different authors in relation to morpho-syntactic realizations of modality, nevertheless, may not correspond exactly to those just related. From the perspective of the Longman Student Grammar of Spoken and Written English, a corpus-based grammar (BIBER et al, 1999), modals are divided into three groups,

namely, 'modals', 'marginal auxiliary verbs' and 'semi-modals'. The first group encompasses *can, could, may, might, shall, should, will, would* and *must*. These modals have a number of specific features such as (a) being invariant forms, (b) preceding the subject in yes-no questions and (c) being followed by a verb in the bare infinitive. Marginal auxiliary verbs correspond to *need (to)*, *ought to*, *dare (to)* and *used to*. According to Biber et al (1999), these kinds of verbs are rare and occur almost only in British English. Fixed idiomatic phrases as *(had) better*, *have to*, *(have) got to*, *be supposed to* and *be going to* are called semi-modals by Biber et al (1999). Semi-modals differ from central modals because they can be marked for both tense and person. Besides, they can also occur as non-finite forms.

Taking into consideration the main aims of this research, the descriptive framework developed by Downing and Locke (2006) underpins the analysis. Following a more empirical approach to data, we rely on Biber et al (1999) and, more specifically, on Almeida Andrade (2011), who have investigated modal meanings using data from Brazilian learners.

3. Data and Methodology

This study comprises two corpora, as already mentioned: a sub-corpus taken from Corpus of Brazilians Learners of English (CABrI) and another sub-corpus taken from Louvain Corpus of Native English Essays (LOCNESS).

CaBri comprises essays written by undergraduate students taking English classes at the Liberal Arts course at the Federal University of Minas Gerais. The essays were handwritten outside the classroom and dictionaries, grammars, and other reference books were used during the writing process. Learners were instructed to write an essay from 500 to 1,000 words length. At the present moment, CABrI contains around 36,187.

Louvain Corpus of Native English Essays presents essays written by American and British speakers, ranging from academic to literary texts. The texts chosen to compose the sub-corpus belong to the

American argumentative section. In total, the LOCNESS sub-corpus used in this study contains 149,627 words.

Both corpora present differences and similarities concerning the expression of modality that, with the aid of corpus methodological tools, were the core of our analysis.

For this analysis, first, word lists were generated and both modal verbs and modal adverbs were isolated for analysis. This procedure, according to O'Keeffe, McCarthy, and Carter (2007), proves to be essential to identify the core vocabulary of English for pedagogical purposes, which is one of the aims of this study. Comparing frequency lists is, then, an essential starting point; however, only relying on frequency lists would not be sufficient. For that reason, in order to get a better notion of the pragmatic function of modals in the essays under investigation, n-gram lists were analyzed and items occurring more than five times were selected for investigation. After identifying the most common items, concordance lines were checked so that the modals could be observed in their particular linguistic contexts. It is believed that such an optimal approach generates concise results.

4. Analysis

4.1. Modal Verbs

As a starting point, we looked at the most frequent modal verbs in both corpora. Table 1 below describes our findings.

Table 1 – Modal verbs found in the data per million words[2]

Modals	**Cabri**	**LOCNESS**
Can	346	**10.546**
Will	229	641
Should	114	489
could	78	447
Would	77	1.055

2. Contracted forms are not included as they tend not to appear in written academic discourse.

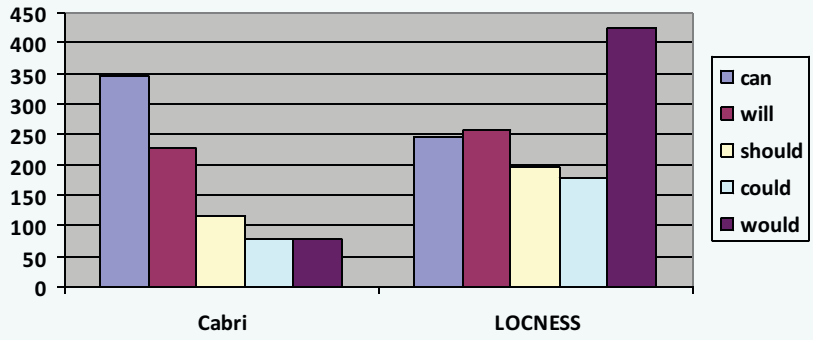

Figure 1 – Distribution of modal verbs in the data.

This preliminary analysis demonstrates that both corpora differ in terms of the frequency of these modal verbs. For example, the modal verb *would* is significantly more frequent in LOCNESS than in the learners' corpus. Actually, *would* is the modal verb least used by CABrI subjects. This finding drew our attention to the remarkable differences that can be found when we compare learners' to native speakers' production. In order to investigate this further, trying to explain these differences, we will rely on the n-grams analysis.

As highlighted by O'Keeffe, McCarthy, and Carter (2007), one of the benefits of looking at n-grams is the fact that they allow us a better view of the discourse under scrutiny. This means that, when looking at n-grams, one moves from an analysis on the level of the sentence to an investigation based on broader aspects of discourse, which, consequently, means an investigation of language in use. Thus, we concentrated on the n-grams containing modal verbs or adverbs with a modal meaning. In total, we found 22 n-grams in the learners' corpus. This amount represents the n-grams occurring more than 5 times in the entire corpus, which was the cut-off point established by the researchers, following Berber-Sardinha (2001). In addition, using the plot tool from the *Wordsmith tools software* one can observe that the item is evenly distributed in the entire *corpus*.

This present study, then, concentrated on the top-five 5 n-grams containing a modal verb and the top-five 5 n-grams containing a modal adverb, having in mind the main objective of this paper, which is to

DELTA
31.2
2015

describe the most frequent n-grams with modal verbs and adverbs in learners' production of academic English. In order to enhance the analysis, we compared data from the learners' corpus to data from the native speakers' corpus (LOCNESS). The items that are under analysis are bigrams and 3-grams, since they were the most frequent in the two corpora:

Table 2 – N-grams analyzed in both corpora

we cannot
can be seen
it would be
should be
this essay will

In this section, we present the n-grams that proved to be relevant in both corpora. The analysis will focus not only on frequency, but also on the prominent aspects of each n-gram, as they occur in the corpora.

4.1.1. N-grams with modal can

4.1.1.1. we cannot

In all occurrences, learners are using *we cannot* when they want to express a more general idea. The pronoun *we* refers to a group of people who share the same idea or perspective relative to a particular subject. In addition, this expression appears in the concluding section of the essay, which reinforces the claim that writers are using this n-gram to make an overall statement at the end of their writing.

4.1.1.2. can be seen

In contrast, concerning the use of *can*, writers in LOCNESS seem to prefer a more complex structure. Observing the n-grams list, we find that *can be seen* is the most prevalent n-gram in the native speakers' corpus. This confirms the hypothesis that native speakers tend to use more complex structures, such as the use of passive voice, which, many times, characterize the text genre 'academic essays'. Learners

seem to have difficulties in conforming to this genre characteristic; consequently, when comparing their essays to the ones written by native users of English, this relevant distinction can be observed. In Extracts 1 and 2 below, we can see the most frequent n-grams with the modal *can* in each corpus.

> Extract 1[3] (CABrI)
> Televisions are necessary, they provide an excellent form of leisure providing instant information and great accessibility beyond our imagination in the past; but we need to be conscious that it has intentions, it has interests and ***we cannot*** be their hostages. It is necessary to be more critic about TV shows, analyzing their point of view and discussing it, not accepting without reflection.

> Extract 2 (LOCNESS)
> The problem facing the Monarchy is how to adapt to modern life whilst still retaining a traditional role. The pressures **can be seen** with the growing question of whether the Prince of Wales may rule as king when being a divorcee. This brings up the issue of him being the Head of the Church of England and so the main problems lie in discussions about the future rather than the present. Also their role in society is a matter of debate: are they there in order to just open hospitals and wave to people? This is the image often seen by the normal subject, discounting of course the scandals within the newspapers.

These two extracts illustrate that Brazilian learners of English, on focus in this paper, use less complex structures and less vocabulary variation. For example, in Extract 2, written by a native speaker of English, the modal *can* is part of a passive voice structure, being passive constructions typical of the genre. Conversely, in the non-native speaker production (Extract 1), the use of the modal integrates a verbal group in a compound sentence. This structure is less complex than the one observed in the production of the native speaker. Academic essays demand more elaborated syntactic structures, due to the fact that they represent a more formal type of text register (HYLAND, 2004), and the non-native speakers' production analyzed in this paper shows less compliance with this requirement.

3. All extracts from the learners' corpus have been preserved the exact way they were written.

4.1.2. N-gram with modal would

4.1.2.1. it would be

As previously observed in Fig.1, the modal verb *would* is the least used by learners. When analyzing n-grams in the learners' corpus, we could not find a relevant number of this item. However, in both corpora, writers use n-grams with *would* when they wish to speculate about future possibilities in a more formal way. They express stance by using this modal verb in order to avoid committing themselves with the truth of their proposition. Learners mitigate their arguments using expressions that would save their faces (BROWN; LEVINSON, 1987). Extract 3 below illustrates the use of several modal expressions in combination (*would, certainly, will, probably* and *should*). This clustering of modals in one extract is representative of the way modality is used in English, including the use of modal harmony (DOWNING; LOCKE, 2006) to express attitude. Extracts 3 and 4 below illustrate the use of *it would be*.

> Extract 3 (CABrI)
> In conclusion, **it would be certainly** a lack of maturity to deny the importance of theory in academic courses as a basis for the formation of its professionals, since the university is definitely the place for them to have a contact with the references (authors, books, concepts, etc) that **will probably** help them throughout their careers whatever area they come to work in. However, in my opinion universities **should** adjust their curriculums in order to prepare well students to the market, moreover not to submit themselves anymore to study for long years only for the purpose of getting a degree – in the sense of a symbolic title required for the market – but for being eager to live somehow necessary experiences before graduating.

> Extract 4 (LOCNESS)
> **It would be** very hard to imagine a Britain without Beef but not so hard to imagine couses that may yet bring it about. For example there has been a constant stream of scares about BSE and CJD, (croytstelt Jacob Disease), so many in fact that much of the public has become blasé about such 'scientific' reports.

In both corpora the n-gram is used with the same syntactic and discursive function, that of allowing writers to speculate formally about

future possibilities and avoid committing themselves with the truth of their proposition, as stated previously.

4.1.3. N-gram with modal should

4.1.3.1. should

Hitherto, the most frequent use of modals had been to express epistemic modality, however, the observation of *should* led to two relevant findings. First, this modal verb was used in both corpora to express deontic modality, not epistemic. Secondly, when analyzing n-grams with *should* in LOCNESS, we found that they are usually followed by a passive construction, as can be seen in the following concordance lines. A certain occurrence of passive construction is indeed typical of the genre in focus.

N	Concordance
1	responsibility to act like a normal human being but should not be blamed for any misuse of thier final
2	Genetic manipulation should not be used to change normal humans into
3	believe that boxing is a blood-thirsty sport, and it should not be allowed in modern society.
4	or, the sex of a child that has been concieved should not be revealed to the parents expect in the
5	Another reason why it should not be banned is because foxes give chickens
6	had no child of their own but I personally feel that it should not be allowed as it does not only seem
7	Boxing, because it is a big money business, should not be banned because it would affect not

Figure 2 – Concordance lines for the n-gram *should* in LOCNESS.

The use of *should* (*not*) found in the concordance lines above, as already mentioned, belong to the deontic type of modality, therefore, they indicate necessity to change reality according to the writer's demands or expectations. From the point of view of syntactic complexity, *should* (*not*), in the native corpora, is followed by a passive voice form, as can be seen in the concordance lines just displayed. However, while observing concordance lines with the same n-gram in the learners' corpus, we found that the use of *should* associated with the passive voice is significantly lower: 22%, while in the native corpus this use corresponds to 82% of the total of occurrences.

Extracts 5 and 6 show examples of the use of *should* followed by passive voice in both corpora analyzed.

DELTA
31.2
2015

Extract 5 (CABrI)
Theory and practical **should** be connected, but, unfortunately, they are becoming more and more distant. Giving theoretical courses instead of theoretical and practical ones is more convenient for universities. It saves money, time and spares them the hard work. However, a practical formation would turn the students into qualified professionals by the moment they leave university before even having a first job.

Extract 6 (LOCNESS)
Technology has progressed quickly and in doing so ethics and practical guidelines have been left behind. I therefore think it is necessary to have certain regulations ie.1. Fertility treatment **should** not be given to post-menopausal women. The menopause is the body's way of telling you that you are to old and your body is no long capable of bearing a baby. Last year there was a case of a post menopausal woman who by lying about her age was given in vitro fertilisation (IVF). I don't think this is fair or morally correct to the child since her mother would be claiming her pension when she was at primary school and her mother would probably die while the child was in her teens....

4.1.4. N-gram with modal could

4.1.4.1. could

Although *could* appears to be a frequent modal verb in the production of both learners and native speakers, the analysis carried out in this paper showed that, when n-grams are investigated, *could* loses its importance, since we could not find more than 2-word n-grams in CaBrl. Thus, for the purpose of this paper, we do not take an in depth analysis of the modal *could.*

4.1.5. N-gram with modal will

4.1.5.1. will

When analyzing the n-grams, we found that, in LOCNESS, *there will be* is prevalent in the texts observed. Whereas, in the learners' corpus, the most common n-gram is *this essay will,* most of the times,

found in the introduction phase of the essays under investigation. This can be an evidence of the kind of instruction non-native speakers receive concerning essay structure in English, which leads to the presence, in their work, of more fixed discourse patterns. We can hypothesize that, when analyzing the production of learners, we might witness the emergence of new n-grams characteristic of learners' language that will not be found in native speakers' production, since these productions will not be influenced by formal language instruction in the same way. In order to have a better account of modality in English learners' academic writing, we set off to analyze the most frequent n-grams containing adverbs with modal values.

4.2. Adverbs

In this section, we start analyzing the most frequent adverbs found in the learners' corpus in order to verify their function in the data. These results are presented in the table below.

Table 3 – Most frequent adverbs found in CABrI

Item	Raw Freq.	Freq. per million words
probably	13	472
certainly	9	326
maybe	9	326
likely	9	326
simply	7	254
unfortunately	5	181
actually	5	181
Total	44	1,594

As the main aim of this paper is to compare learners' data with native speakers' production, we shall analyze LOCNESS for corresponding results. In table 4, then, we present the most frequent adverbs found in LOCNESS.

Table 4 – Most frequent adverbs found in LOCNESS

Items	Raw Freq.	Freq. per million words
likely	31	1.113
certainly	19	782
probably	42	701
perhaps	39	658
surely	1	534
possibly	20	429
maybe	31	425
unlikely	1	288
Total	184	4.930

Analyzing tables 3 and 4, we observe that native speakers seem to use more adverbs than non-native speakers. Also, when analyzing the frequency of adverbs in each group, we can observe that the native speakers' use of adverbs is more evenly distributed than that of learners. This first finding is in line with Holmes (1997), in that it provides evidence of a more balanced use of linguistic resources by native speakers to express modality. Throughout this paper, we make the case that learners tend to use a more fixed set of expressions to convey modality. In table 3, we contrast the adverbs found in both corpora, aiming to determine differences and/or similarities between them. Concerning the use of adverbs in the writing production of Brazilian learners of English, Almeida Andrade (2010), though working with a different set of adverbs, highlighted that the use of adverbs by Brazilian learners is more attached to the function of intensification than to the expression of epistemic meanings, as it happens with the writing production of native speakers.

Table 5 – Distribution of adverbs in the two corpora (raw results)

Items	LOCNESS	CABrI
Likely	31	9
Certainly	19	9
Probably	42	13
Perhaps	39	4
Surely	1	1
Possibly	20	0
Maybe	31	9
Unlikely	1	0

From this preliminary analysis, one can speculate that the use of adverbs to express modality is underpresented in learners' academic essays, as stated in a research carried out by Orfanó, Oliveira and Tenuta (2014), which has shown that Brazilian learners heavily rely on a rigid set of verbs to express modal values. However, the analysis proposed in this paper intends to go beyond stating quantitative differences. In fact, it aims at understanding the linguistic features that make up for these differences and their implications for the learners' written discourse production. In order to do so, at this point of the analysis, we shall concentrate on the most common n-grams containing an adverb in CABrI and LOCNESS. The following table shows this distribution in both corpora.

Table 6 – Distribution of n-grams containing an adverb in the data

N-grams	CABrI- raw frequency	Freq. per million words	LOCNESS-raw frequency	Freq. per million words
likely to	9	33	40	265
almost certainly	0	0	10	66
certainly not	0	0	10	66
will probably	9	32	0	0
would probably	0	0	15	99

After identifying the most common n-grams in both corpora, we set off to analyze each n-gram independently. In CABrI, there are only 2 n-grams being used, whereas, in LOCNESS, we found 4 n-grams. This fact reinforces the claim that native speakers express modality not only by using different adverbs, but also by combining them in different n-grams. The two n-grams used by learners in the corpus analyzed were: *likely to* and *will probably*.

4.2.1. N-gram with likely to

4.2.1.1. likely to

There are 33 occurrences[4] of the n-gram in CABrI and 265 in LOCNESS. This difference was expected, since, as it has been

4. The great majority of these occurrences involve the lexical auxiliary *be likeky to*, which contains the adverb *likely* in its formation.

DELTA
31.2
2015

mentioned previously, there is lower frequency of adverbs expressing modality in the learners' production. We consider important to speculate on this situation focusing on the learners' written discourse, bearing in mind that frequency differences between two datasets can indicate either overuse or underuse of linguistic features, which poses interesting pedagogical issues involving the teaching and learning of English.

> Extract 7 (LOCNESS)
> Now that rail privatisation has gone ahead, many people are likely to lose faith in trains, due to the percieved inefficiency of the operators (for example the time-table book full of errors or the recent survey in Which? magazine about overcharging). Fares are likely to increase, and many rural lines that used to be subsidised by the government face closure.

Figure 3 shows concordance lines for the n-gram *likely to* in CABrI. In all examples, it is possible to see that the epistemic use is prevalent in the data, which might be due to the text genre in focus. In texts of this genre, the writer, very frequently, has to commit him/herself, in different degrees, to the certainty of occurrence of a specific fact.

N	
1	ished severely and consequently they are more likely to commit crimes again.
2	country side towns around Brazil we are very likely to find a huge number of
3	by creating artificial dreams, graduates are likely to be shocked or unsure in
4	This kind of proficiency is more likely to be developed if one
5	This kind of proficiency is more likely to be developed if one
6	arget our limited resources for programs most likely to reduce recidivism and
7	Not likely to happen.
8	y and dreams be profitable, but they are also likely to be crucial ways to make
9	ot prepare graduate students to what they are likely to face in real life.
10	nt agencies show that such tragedies are more likely to occur to young adults

Figure 3 – Concordance lines for the n-grams *likely to* in CABrI.

The function of this n-gram is similar in both corpora; however, the frequency is significantly higher in the native speakers' corpus. *Likely to* is the n-gram containing an adverb preferred by native speakers to express possibility/probability, while English learners, in this study, expressed possibility/probability by using the n-gram *will probably*.

The low frequency of modal adverbs in the learners' academic writing (CABrI) was previously acknowledged by Orfanó, Oliveira and Tenuta (2014). The authors showed that learners seemed to heavily rely on modal verbs to express epistemic modality. However, in the corpus investigated, learners used a very narrow range of modal verbs with epistemic meaning, for example, mainly *can* and *will*. In this regard, the findings revealed that the distribution of modals in the native corpus was more varied, since native speakers employed, for example, *should*, *could* and *would*.

4.2.2. N-gram with will probably

4.2.2.1. will probably[5]

This n-gram follows the pattern *will probably* + verb.[6] We tend to conclude that the n-gram *will probably* might be more easily accessed by learners, becoming active in discourse through less mental effort (CHAFE, 1994), due to the fact that the pattern in which it occurs easily corresponds to *vai provavelmente* + verb, in the Portuguese language. In this study, learners used *will probably* three times more than native speakers did.

> Extract 8 (CABrI)
> There **will probably** be many reasons for dreaming and three possible - and believable – ones could be its profitability (for the enterprises which provide entertainment, for example), its help in making us stand and try to change our stressful reality, and the health benefits it provides us. Oddly enough, imagining can make a big profit from generating – and selling – brilliant ideas.

5. This is a case of modal harmony (modal verb and adverb).
6. We also observed that there was a high frequency in the use of *will probably* + a verb, since a third of the occurrences with the n-gram *will probably* followed this pattern. This finding can be further investigated.

N	Concordance
1	is. All you have to do is read and read. In fact, you **will probably** get all
2	e she does not recognize. As for Dee's sister, she **will probably** feel
3	he references (authors, books, concepts, etc) that **will probably** help
4	e way she could free herself from her family. They **will probably** go along
5	no chance to see in loco how things really work, it **will probably** be very
6	e inhabit a modern and industrialised world. There **will probably** be many
7	ity degrees, such as philosophy and anthropology, **will probably** have
8	ity degrees, such as philosophy and anthropology, **will probably** have
9	er mother, Dee is not the same anymore and they **will probably** be apart

Figure 4 – Concordance lines for the n-gram *will probably* in CABrI.

In the next section, we observe the n-grams found in LOCNESS.

4.2.3. N-gram with would probably

4.2.3.1. would probably[7]

The use of the n-gram w*ould probably* in LOCNESS might reflect a feature more closely related to formal register, which is considered adequate to the essay genre. We might speculate that learners do not use it because they might not be very proficient in the use of linguistic features that would enable them to express, together, a degree of both certainty and formality in written discourse, as can be seen in the concordance lines below. We could argue, then, that this n-gram has a twofold discursive function, which was employed by native speakers and ignored by learners in the corpora analyzed.

> Extract 9- LOCNESS
> If boxing was made illegal, then the sport would still take place, but it would not be under the control of doctors or referees and there would be no proper association to represent rules and regulations. It **would probably** become even more dangerous. Of course there are advantages banning the sport but they are very small compared to the disadvantages.

7. This is also a case of modal harmony (modal verb and adverb).

```
1       By banning fox hunting the law would probably drive the spor
2       ulations showed that charities would probably lose out, coll
3        primary school and her mother would probably die while the
4                                  It would probably become even mo
5       everal thousands of pounds and would probably have to sell u
6       were the highest bider as they would probably hide it as it
```

Figure 5 – Concordance lines for the n-gram *would probably* in LOCNESS.

5. Conclusion

In this study, we have concentrated on the most frequent n-grams expressing modality values found in the two corpora analyzed: CABrI and LOCNESS. First, we observed that, regarding the use of the modal *can*, in the learners' corpus, the most frequent n-gram found was *we cannot*. Conversely, the most frequent n-gram chosen by writers in LOCNESS is *can be seen*, which shows a higher level of complexity, due to the use of passive voice, in native speakers' production, which is expected in academic writing.

While analyzing n-grams with *would*, we found that the expression *it would be* plays an important role in LOCNESS, since it allows writers to avoid commitment with the truth of their statements. Learners seem unfamiliar with this pragmatic strategy, which is evident in their restrict choice of this modal verb.

The prevalent use of the fixed expression *this essay will* reinforces the claim that the learners' repertoire is relatively more rigid and is influenced by the input these speakers receive (HYLAND, 2004).

When observing native speakers of the English language, we found that they use a more varied expression of modality, including elements that do not appear, in the same way, in the repertoire of the learners' group we investigated. For example, passive voice combined with modal constructions is common in their essays, which is not the case with the learners' essays. We notice, then, that this combined use can be stimulated in the teaching of academic writing. Grammar topics such as this should be taught integrated, and not in isolation, as it seems to have been the case with the learning process gone through by the learners whose production we investigated. The integrated

DELTA
31.2
2015

pedagogical approach to grammar and academic writing we propose here could make learners become more aware of the requirements of the academic genre which include, as we have found in this study, the use of modal verbs and passive voice in combination.

Our findings also indicate that the expression of epistemic modality in the native speakers' written production of academic essays, referring to, specifically, the use of adverbs, seems to be more varied than in the learners' production. This can be evidenced, for example, by the non-use of adverbs, on the part of learners, and the use of expressions like *likely to*, which proved to be, on the other hand, widely used by native speakers.

By the same token, the data analyzed in this study show that the learners' preferred expression of epistemic modality including an adverb was *will probably*. This n-gram is frequently used by these speakers. Such use might be explained by the fact that *will probably* has a correspondent in the Portuguese language, *vai provavelmente*. These findings strengthen the argument that teachers and material developers should introduce students to a more varied set of vocabulary as well as structural choices, so that the learners' writing skills can be improved towards the production of texts which are better elaborated and more responsive not only to the grammatical and semantic, but also to the pragmatic demands of the genre in focus.

Another important remark about the learners' written production concerns the fact that, in the data analyzed, learners seemed to underuse the form *would probably*, which is highly used by native speakers as a way to express remoteness and little commitment to the certainty of facts. This finding also supports the claim that learners tend to rely on a very rigid set of structures to express modality. It also poses the issue of the role of language instruction and material design in raising students' awareness of how to structure their discourse in order to successfully meet the requirements of academic writing.

We would also like to comment on the importance of analyzing empirical language data for a broader understanding of how native speakers and learners can differ in their production and, if that is the case, decide on the best teaching strategies to help learners write more fluently and effectively.

Recebido em junho de 2012
Aprovado em abril de 2015
E-mails:
atenuta@gmail.com
adornomarciotto@gmail.com
bmalveira@yahoo.com.br

References

Almeida Andrade, M. I. 2010. *Prosa argumentativa em língua inglesa: um estudo contrastivo sobre advérbios em corpora digitais.* Dissertação de Mestrado. UERJ. [Orientadora Prof. Tânia M. G Shepherd]

Berber-Sardinha, A. P. 2001. O corpus de aprendizes Br-ICLE. Intercâmbio, v.10. p. 227-239.

_____. 2004. *Linguística de corpus*. Barueri, SP: Manole.

Berber-Sardinha, A. P.; Shepherd, T. 2008. An online system for error identification in Brazilian learner English. Anais do *8th Teaching and Language Corpora Conference.* Lisboa: Associação de Estudos e de Investigação Científica do ISLA-Lisboa. p. 257-262.

Berk, L. M. 1999. *English Syntax – From word to discourse*. Oxford: Oxford University Press. p. 317.

Biber, D. *et al.* 1999. *Longman Grammar of Spoken and Written English*. Longman. p. 1204.

Brown, P.; Levinson, S. 1987. *Politeness*: Some universals in language usage. Cambridge: Cambridge University Press. p. 347.

Carter, R.; Mccarthy, M. 2006. *Cambridge Grammar of English:* A Comprehensive Guide; Spoken and Written English Grammar and Usage. Cambridge: Cambridge. p. 973.

Chafe, W. 1994. The nature of consciousness. In: *Discourse, consciousness and time.* Chicago & London: The University of Chicago Press. p. 26-40.

Downing, A.; Locke, P. 2006. *English Grammar: a University Course*. New York: Routledge, Second edition, p. 610.

Dutra, D. P. 2009. Conscientização linguística através de corpora *online*. *Caderno de resumos do 17 INPLA*. São Paulo: PUC-SP. p. 62.

Granger, S. *et al.* 2009. International Corpus of Learner English-Version2. Louvain-La-Neuve: UCL Presses Universitaires de Louvain.

Halliday M. A. K. 2004. *An Introduction To Functional Grammar* London: Hodder Arnold Publication. Revised Edition, p. 689.

HOLMES, R. 1997. "Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines." *English for Specific Purposes*. 16, p. 321-338.

HUDDLESTON, R.; PULLUM, G. 2005. *A Student's Introduction to English Grammar.* London: Cambridge University Press, p. 312.

HYLAND, K. 2004. *Disciplinary discourse: social interactions in academic writing.* London: Longman, p. 211.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. 2007. *From Corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press, p. 313.

ORFANÓ, B. M; OLIVEIRA, A. L. A. M.; TENUTA, A. M. 2014. Epistemic modality through the use of adverbs: a corpus-based study on learners' academic written discourse. *Letras e Letras (online)* v. 30, p. 104-121.

PALMER, F. R. 2001. *Mood and Modality*. Cambridge: Cambridge University Press, p. 236.

TENUTA, A. M. 1995. Tempo, Modo e Aspecto verbal na estruturação do discurso narrativo. *Revista de Estudos da Linguagem*, Belo Horizonte, ano 4, v. 2, jul./dez. p. 179-195.

_____. 2006. *Estrutura narrativa e espaços mentais*. Belo Horizonte: Faculdade de Letras da UFMG, p. 249.