

**THE ROLE OF PRAGMATIC MARKERS IN ACADEMIC
SPOKEN INTERLANGUAGE:
A CORPUS-BASED STUDY OF A GROUP OF
BRAZILIAN EFL UNIVERSITY STUDENTS**
O PAPEL DOS MARCADORES PRAGMÁTICOS NA
INTERLÍNGUA FALADA ACADÊMICA:
UM ESTUDO BASEADO EM CORPUS DE UM GRUPO DE
ESTUDANTES UNIVERSITÁRIOS BRASILEIROS DE ILE

Bárbara Malveira Orfanò*
barbara.orfano@gmail.com

Ana Larissa Adorno Marciotto Oliveira*
adornomarciotto@gmail.com

Spencer Barbosa da Silva**
spencer.silva@ufop.edu.br

The present work addresses a group of university students of EFL (English as a Foreign Language) on how they use pragmatic markers in their oral productions. The initial hypothesis was that there would be differences both in usage and form in comparison to native speakers. In order to verify our claim, we set off to investigate two corpora: a learner oral corpus being compiled at the Federal University of Minas Gerais/Brazil and a sub-corpus from the British Academic Spoken English (BASE). While Brazilian students overuse items such as *maybe* and *just*, the data recorded in the UK displayed a more varied range of markers and multiword forms. Overall, the findings reinforce the importance of analyzing empirical data for a broader understanding of how native speakers and learners can differ in their oral academic production. The paper also sheds light on language teaching and learning in the academic setting from a pragmatic viewpoint.

Keywords: Learner *corpus*. Academic discourse. Politeness.

O presente trabalho aborda como um grupo de estudantes universitários de ILE (Inglês Língua Estrangeira) usam marcadores pragmáticos em suas produções

* Faculdade de Letras - Universidade Federal de Minas Gerais, Brasil.

** Departamento de Estatística - Universidade Federal de Ouro Preto, Brasil.

orais. O estudo consiste de dois corpora: um corpus de aprendiz sendo compilado na Universidade Federal de Minas Gerais-Brasil e um subcorpus do British Academic Spoken English (BASE). Enquanto aprendizes brasileiros sobreusam itens específicos como *just e maybe*, falantes nativos ou fluentes utilizam uma variedade maior de unidades multipalavra. Os resultados reforçam a importância da análise de dados empíricos no estudo da produção oral de aprendizes. Eles também lançam luz para o ensino e para a aprendizagem de inglês em contexto acadêmico de um ponto de vista pragmático.

Palavras-chave: *Corpus* de aprendiz. Discurso acadêmico. Polidez.



1. Introduction

Pragmatic markers, or discourse markers, allow for writers/speakers to communicate their stance or attitudes toward the information conveyed. They make room for negotiating the certainty of statements, functioning to ‘linguistically situate the intention of the writer, while priming the reader/listener to align with this intention’ (Ran 2003, p. 8).

Markers like these are commonly used to evaluate the certainty of a proposition, while concealing the author’s voice as the source of assessment and thus presenting the assertions as objective and impersonal (Biber 2006; Ran 2003). In the academic domain, pragmatic markers may also function as a strategy of negative politeness. They may contribute to soften the imposition of the research information (such as the hypotheses, the theoretical contributions, and the results) on the reader/listener, while they treat it as neutral or objective. In this sense, pragmatic markers operate as avoidance politeness strategies (Goffman 1967), in which the speaker/author prevents himself from invading the interlocutor/reader’s territory.

Politeness, taken in the comprehensive sense of speech oriented to an interactor’s public persona or ‘face’, is ubiquitous in language use (Oliveira, Cunha & Miranda 2017). It therefore meets the aim of the study we propose here, which focuses on the academic language domain. The reason for such a claim is associated with the fact that image projection is a hallmark of the academic domain. Likewise, the expression of stance in this field is also considered crucial. In this paper, we claim that the scope of pragmatic markers employed in the academic domain differ from native speakers and learners

in manifold ways. In order to verify this claim, we set off to investigate two corpora: a learner oral corpus being compiled at the Federal University of Minas Gerais/Brazil and a sub-corpus from the British Academic Spoken English (BASE).¹

In the past few years, pragmatic markers or metalinguistic monitors have been under scrutiny by different researchers. Erman (2001), Aijmer (2002; 2004), McCarthy and Carter (2006) and Fung and Carter (2007) have examined pragmatic markers in written and spoken discourse. However, studies concentrating on how Brazilian university students of English use such markers in spoken interlanguage are virtually non-existent. Considering that developing students pragmatic awareness is an essential part of their academic literacy, this paper aims to shed light upon how a group of Brazilian university students use pragmatic markers in their oral presentations. Upon the implications of this use, we will also focus on the way learners interact in English with their scientific community.

The learners in this study were undergraduate students taking the course English for Academic Purposes taught at a Federal University in Brazil. In order to better understand how Brazilian university students taking this course use pragmatic markers in their spoken language, we compared the results of our learner corpus with a native speaker corpus, focusing on underuse and overuse of the most significant patterns drawn from the data. In order to achieve this, the following research questions were addressed: What are the most commonly found patterns of pragmatic markers in the two corpora analysed? In case they are different, what is the possible impact of this discrepancy, considering the demands of the academic domain?

2. Pragmatic Markers and the interpersonal domain

The choice for the term 'pragmatic marker' is not a fortuitous one. In fact, it follows studies stating that it is impossible to ignore text type (written or spoken) context and the relationship between interlocutors. Arguing along the same lines, this paper follows the assumption that the term 'pragmatic

¹ The transcriptions used in this study come from the British Academic Spoken English (BASE) corpus project. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council.

marker' suggests a high degree of context sensitivity, as is also acknowledged by Andersen (2001).

An important issue concerning pragmatic markers is their multifunctionality. Their function and use vary depending on different issues ranging from discourse markers linking units of discourse, and then being responsible for coherence, to modal items within a more interpersonal dimension. This way, as pointed out by Fung and Carter (2007, p. 414), such markers are pragmatically significant and socially sensitive.

The literature in the field shows that pragmatic markers are not an easy term to define and definitions are usually associated with different approaches to their study, as well as with the functions related to them.

Fraser (1999) defined pragmatic markers as a pragmatic class, or as lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases which "signal a relationship between the segment they introduce, S2, and the prior segment, S1" (Fraser 1999, p. 63). In the example below, the pragmatic marker *In spite of*, relates the explicit interpretation of S2 to a non-explicit interpretation of S1. In S2, there is an implied proposition associated with S1, which is referenced by the use of *In spite of that*.

- (1) S1 We left late.
S2 In spite of that, we arrived on time.
(Fraser 1999, p. 64)

As we can see in (1), example taken from Fraser (1999), speakers can choose between hedges and approximators when they wish to minimize (or maximize) the effect of the message being communicated. Examples such as *I think*, *maybe* and *kind of* represent some of the most common face-saving markers (Goffman 1967) used by speakers in different contexts. Erman (2001) explains that, in the case of German studies on modality, interlocutors tend to concentrate more on the expressive attitude of the speaker towards the propositional contents of the utterance. This notion is closely related to Kriwonossow's (1977, p. 187) subjective modality and to Bublitz' (1978, p. 8) emotive modality. Both perspectives are oriented towards the speaker's attitude and also to the relationship between speaker and hearer.

In a similar vein, Aijmer (2004) acknowledges the importance of pragmatic markers in the study of learners' interlanguage. According to the author, these features need to be analyzed from the students' perspective, which means that looking at how learners use pragmatic markers in their

discourse might reveal important characteristics of their oral production. Aijmer (2013) also defines the role of pragmatic markers within a general pragmatic theory that concentrates on the language user and on the relationship between meaning and context. This is the position that is followed in this study, since our main goal is to determine how learners use such markers and the implications of this use in their academic discourse.

Along the same lines, Erman (2001) states that there are two well-established functions of pragmatic markers: they can be used as monitors of discourse and as interactional features. The author observes that, as a primary function, the markers fulfill the role of textual monitors, being responsible for turning fragmented pieces of discourse into a coherent text; however the secondary function markers operate as social monitors and their main role is to promote the negotiation of meaning and discourse management, ensuring that there is an open channel between interlocutors.

Advancing in his research, Erman (2001) proposes a third function labeled 'metalinguistic monitors' or, as he prefers to name it 'metalinguistic domain'. According to the author, markers within a metalinguistic domain are usually modal and speaker-oriented, having two main roles: to emphasize the speaker's authority as to the illocutionary force of an utterance and/or to serve as a *face-saving device* (Goffman 1967). This latter role, in particular, is the view adopted in this paper. In the next section, we will briefly address the foundation of this work, concerning Politeness Theory and the notion of face-work.

3. Politeness Strategies, face work and image projection

The notion of face-work, as was presented by Goffman (1967) refers to "the positive social value that a person effectively claims for himself by the line that others assume he has taken during a particular contact" (Goffman 1967, p. 223). More specifically, the term face-work refers to "the actions taken by a person to make whatever he is doing consistent with face. Face-work serves to counteract "incidents" – that is, events whose effective symbolic implications threaten face" (Goffman 1967, p. 12).

Also in the realm of image projection and face-work, Brown and Levinson (1987) have shown that certain lexical, grammatical, and prosodic phenomena can only be fully explained from the perspective of sociological factors (such as power relations) and pragmatic elements (such as the principle of politeness). From this viewpoint, elements of microlinguistic nature

(in this paper adverbs, prepositional phrases and conjunctions) may act as strategies of politeness, used to mitigate, or to intensify the virtual threats inherent to Face-Threatening Acts (FTA), such as criticisms, promises, compliments, among others (see Kerbrat-Orecchioni 2006; Cunha 2015; Oliveira, Cunha and Miranda, 2017).

In Brown and Levinson's Politeness Theory (1987, p. 61), the notions of face-work and territory, taken from Goffman (1976), are revisited in order to hold the concepts of 'positive face' and 'negative face', as in (a) and (b) below:

- a) negative face: the basic claim to territories, personal preserves, rights to non-distraction, to freedom of action and to freedom from imposition;
- b) positive face: the positive self-image or 'personality' claimed by interactants.

As Orfanò (2010) also postulates, in systematizing Goffman's approach to language studies, Brown and Levinson (1987) re-formulate the concept of facework. On the one hand, it becomes more restricted, since it only corresponds to the use of linguistic procedures (and not any procedures in general) that mitigate/intensify the threat of speech acts. On the other hand, the notion is deepened, in that it encompasses the strategies used to mitigate/intensify attacks on the negative face, and no longer only attacks on the positive face.

It is on the basis of this notion of face-work that, more recently, Brown (2015, p. 326) conceptualizes politeness in these terms: "Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors". Politeness strategies are, therefore, inherent to language use in general, which embraces the academic oral domain. Taking this claim into account, in the next section, we will present the methods of data collection and analysis of this study in an attempt to verify the potential similarities and differences between the two corpora in focus, in terms of the pragmatic markers students most frequently used.

4. Methodology of data collection

This study comprises two corpora. The main corpus, the Brazilian Academic Spoken English Corpus (BRASE), consists of 20-minute oral presentations given by students taking the course English for Academic Purposes at a

Federal University in Brazil. The undergraduate students are from different degree programs and their level of proficiency ranges from B1 to C1, following the Common European Framework of Reference for Languages (CERF). At the moment of data collection, this corpus had approximately 50,000 words. The reference corpus is a sub-corpus from the British Academic Spoken English (BASE), from the Humanities area, compiled for this specific study. BASE is a corpus designed by researchers from the Centre of Applied Linguistics of the University of Warwick-UK. In total the corpus has 1,644,942 words, encompassing the areas of Arts and Humanities, Life and Medical Sciences, Physical Sciences and Social Sciences.

After transcribing, organizing and including metadata, the main corpus was submitted for analysis using the software *Wordsmith Tools 5.0*. First, a frequency list was generated and items with the potential to function as metalinguistic monitors were isolated for analysis. In order to check if the items were in fact functioning as metalinguistic monitors, concordance lines were generated and items that were fulfilling different functions, such as circumstantial adjuncts or modifiers in noun phrases, were eliminated from the analysis and a list of metalinguistic monitors were designed.

Considering that metalinguistic monitors can also be composed of more than one word, cluster lists of 2, 3 and 4 words were generated and items fulfilling a metalinguistic monitor function were selected for analysis. However, the most fruitful list was the one containing 2 word-clusters, and for this reason only this list was included in the analysis. The same procedure was carried out for the reference corpus and lists of single and 2 word clusters were generated for both corpora to be compared.

Once the lists with the metalinguistic monitors were generated, the data was submitted to a set of statistical tests (see statistical analysis section). This procedure enables the researcher to be more accurate when analysing the data, thereby avoiding misinterpretation of the results from the corpora. As an example, the Log Likelihood² test allows the researcher to determine if the items under analysis when compared to another corpus present a significant difference. It is possible to check underuse and overuse of items in relation to the reference corpus and, in this way, obtain a better account of the main characteristics of the data being analysed.

2 <http://ucrel.lancs.ac.uk/llwizard.html>

5. Data analysis

The investigation begins with a single item frequency search isolating the items with the potential to function as metalinguistic monitors, following O’Keeffe, McCarthy and Carter’s (2007) framework. Searches for 2 word clusters were also carried out and again any forms likely to function as metalinguistic monitors were highlighted.

From these searches the highest number of potential items resulted from the 2 word cluster search, for example, *I mean*, *I think* and *I guess*, and for that reason the present paper concentrates on these examples. The following tables demonstrate the frequency of single words.

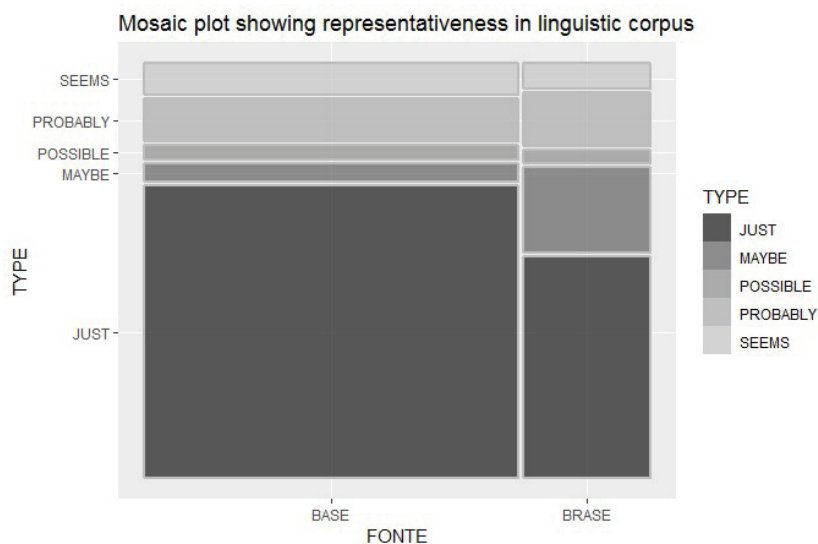
Table 1. Single words functioning as metalinguistic monitors in both corpora

BASE 150.000 words			BRASE 50.000 words		
Items	Raw freq.	Normalised Per 100.000	Item	Raw freq.	Normalised Per 100.000
JUST	401	267	JUST	103	206
ACTUALLY	316	210	MAYBE	40	80
RATHER	68	312	PROBABLY	25	50
PROBABLY	59	38	SEEMS	12	24
SUGGEST(s)	50	33	POSSIBLE	6	12
SEEM	26	46			
SEEMS	43				
MAYBE	24	16			
POSSIBLE	20	13			
POSSIBLY	13	9			
LIKELY	12	8			
SUPPOSED	11	7			
POSSIBILITY	10	6			
THINKS	10	6			
APPARENTLY	7	5			
SUPPOSEDLY	7	5			
TENDS	7	5			
INDICATIVE	6	4			
TOTAL	1.090	990	TOTAL	186	372

The first results from Table 1 indicate that there is variation in the forms used in each corpus and also in the number of forms. The frequency is higher in BASE than it is in the Brazilian group under investigation. In addition, in the British Corpus the use of items are evenly distributed, whereas in BRASE more than half of the uses are concentrated on two specific items. In order to analyze the features in more detail, we submitted the data to specific statistical tests.

5.1. Statistical Analysis for one-word elements

The contingency table that represents the qualitative observations extracted from the samples of the two corpora (BASE and BRASE) can be represented through the mosaic chart (Graph 1), in which each horizontally subdivided rectangle shows the proportionality of the results found in each corpus. Analysing Graph 1, we can observe a greater frequency of the words *just* and *probably* in BASE when compared to BRASE.



Graph 1. Mosaic display for sample of single words in the BASE and BRASE corpora

Table 2 below shows the results of the statistical tests carried out for the single word list. They were obtained from the comparison between the word frequency found in BRASE and BASE, via the application of Relative Risk (RR), Odds Ratio (OR) and P-value, when performing the chi-squared test in order to make proportions uniform. Besides, Log-likelihood test and P-value were also used to equalize proportions and the Confidence Interval at the 95% level was applied for the difference between proportions. These procedures were necessary due to the fact that the two corpora, BASE and BRASE, are unequal in size, each containing 150,000 words and 50,000 words respectively.

Concerning the results for Relative Risk (RR), generated by the ratio between the risk of occurrence of the word in BRASE and the risk of occurrence of that word in BASE, values higher than 1 were obtained in some situations, which means that the risk of *just* and *maybe* occurring in BRASE is higher than in BASE.

As for the Odds Ratio (OR), generated by the ratio between the odds (or “chance”) of occurrence of a word in BRASE and odds (or “chance”) of occurrence of that word in BASE, the word probably has a greater chance of being used in BRASE rather than in the BASE corpus.

The measures of the p-value generated from the chi-square test led to the finding that at the 5% significance level, a significant difference between the proportions relative to the words *just* and *maybe* can be found. In other words, these proportions differ statistically. Conversely, regarding the words *probably*, *seems*, and *possible*, the findings show that at the 5% level of significance the proportions do not differ statistically.

These same results are corroborated by the values obtained in the P-value through the test of equality of proportions, at the level of 5% of significance and by the Confidence Interval, at the level of 95% constructed for the difference in proportions (intervals containing zero). This proves that the proportions of these three words do not differ statistically. As for the values obtained from the Loglikelihood test, it is possible to conclude that the words *just* and *maybe* do not have the same proportion in BASE and in BRASE, suggesting their overuse in the Brazilian corpus. Table 2 presents the results from the tests described in this section.

Table 2. Statistical Results of occurrence and association of simple words

ITENS	RR	OR	P-VALUE χ^2	LL	P-VALUE = PROP	IC (0,95) for D
JUST	0,817	0,771	0,01811	5,85	0,02048	[-0,081; -0,010]
MAYBE	2,5	5	0	40,03	0	[0,255; 0,495]
PROBABLY	1,191	1,271	0,3136	0,98	0,38	[-0,051; 0,146]
SEEMS	0,873	0,837	0,586	-0,31	0,697	[-0,142; 0,078]
POSSIBLE	0,923	0,9	0,8209	-0,05	1	[- 0,843; 0,146]

In order to investigate the implications of these findings in the academic discourse of learners, the next sections will deal with these metalinguistic monitors in more detail.

5.2. The item *just* as a sample of overuse in BRASE

Following Aijmer's (2002, p. 158) claim that the pragmatic marker *just* has procedural meaning in that it functions as a signal to the hearer to interpret the speaker's utterance as an expression of an attitude. According to McCarthy and Carter (2006), the use of *just* in oral discourse has a number of functions: It can be used for emphasis, as a particularizer, temporal meaning, limiter and as a softener or downtoner (*idem*, p. 98). Significantly important for this study are Aijmer's (2002) observations, which assign *just* to a hedging function in the realms of both positive and negative politeness. From the perspective of negative politeness, *just* functions as a downtoning hedge, modifying the face threat carried by speech acts such as assertions, suggestions, criticisms or requests (*idem*, p. 169). The analysis demonstrates that *just* is overused in BRASE indicating that learners are familiar with the hedging role fulfilled by this marker in oral discourse. However, at this point, we can claim that overusing *just* as an epistemic³ marker in their oral presentations, learners are not complying with the characteristics of oral academic discourse, since they show a preference for a marker that belongs to the informal domain.

(2) BRASE IFA 2- B1+⁴

3 Epistemic modality refers to the degree of commitment one has in relation to what one says.

4 IFA stands for *Inglês para Fins Acadêmicos – English for Academic Purposes*. The example was taken from students taking this course. B1 corresponds to their proficiency (intermediate) following the Common European Framework of Reference for Languages.

*the volume is basically the volume of the cylinder and the cost is **just** the cost of all materials involved... here is the result ah all the solutions here and what was in the in this two genetic algorithm ...*

Examples like the one in (2) are common in BRASE. Even though this group of students are on an intermediate level, they seem to be unaware of other markers functioning as hedges. This overreliance on one marker tends to impoverish learners' oral academic performance. Advancing in our analysis, we decided that it would be important to look at multiword combinations considering that they are important in creating meaning and also responsible for fluency in oral discourse.

According to O'Keeffe *et al.* (2007, p. 60), "what corpora reveal is that much of our linguistic output consists of repeated multi-word units rather than just single words". The literature has dedicated considerable attention to this issue and the terminology varies depending on the researcher and the theoretical perspective adopted in each study. Biber (1999) calls the combination of repeated words lexical bundles, O'Keeffe *et al.* (2007) refer to them as cluster units and Orfanò (2010) names them as frequent items. Due to their importance they are also analysed in this paper. Table 3 accounts for the most frequent 2 word-clusters in the two datasets.

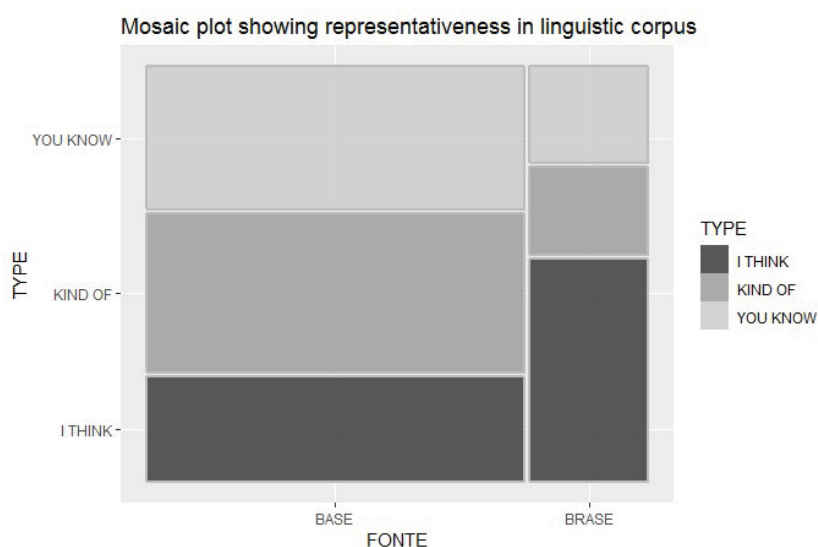
Table 3. Multiword units functioning as metalinguistic monitors in both corpora

BASE 150.000 words			BRASE 50.000 words		
Item	Raw freq.	Normalised Per 100.000	Item Freq. bruta	Raw freq.	Normalised Per 100.000
KIND OF A KIND OF	325	216	I THINK	143	286
YOU KNOW	293	195	YOU KNOW	62	124
I THINK	215	143	KIND OF	56	112
SORT OF	212	141	A LITTLE	25	50
I MEAN	73	48			
A SENSE	50				
TOTALS	2.340	861	TOTALS	286	572

Following Cortes (2002), a cut-off point of 20 occurrences per 100,000 words was established for the analysis of the multiwords units. A brief view of the list generated reveals interesting differences between the two datasets. The number of clusters in BASE outnumbers BRASE and the frequency in BASE is also higher than it is in the Brazilian data.

5.3. Statistical Analysis for 2-word clusters

When the use of multiword units is regarded, it is possible to verify, through the analysis of Graph 2, that there is a greater frequency of the expression *I think* in BRASE and of *kind of* in BASE, it is also possible to identify in the graph the size difference of the two corpora. That said, the frequency presented in the samples of these terms is displayed in a more homogeneous form in each corpus than it was in relation to the frequency of the words as shown in Graph 1.



Graph 2. Mosaic display for sample of multiword units in the BASE and BRASE corpora

With respect to the results found in the analysis of the contingency tables involving the multiword units displayed in Table 3, we can observe,

through the RR, that only the word combination *I think* presented a greater chance of occurring in BRASE in relation to BASE. Additionally, the same happened with respect to the result obtained from OR, that is, the word combination *I think* has a greater chance of being used in BRASE than in BASE (almost double the chance).

For the results obtained in comparison between the equality of the proportions of use for the words in BRASE and in BASE, in the four statistical measures found, p-value for chi-square test, p-value for equality of proportions and Confidence Interval confirmed that, at the level of 5% significance, the proportions differed. Likewise, the log-likelihood (LL) results were obtained, all of them were superior to the cut-off point of 3.8, associated with the significance level of 5%. These results confirm the overuse of *I think* in BRASE in relation to BASE, together with the underuse of the combinations *kind of* and *you know* in BRASE, when compared to BASE, as one can see in Table 4:

Table 4. Statistical Measures of occurrence and association of multiword terms

ITENS	RR	OR	P-VALUE χ^2	LL	P-VALUE = PROP	IC (0,95) for D
I THINK	1,598	1,995	0	38,47	0	[0,099; 0,200]
YOU KNOW	0,876	0,635	0,001	-11,62	0,001	[-0,075; -0,036]
KIND OF	0,588	0,517	0	-24,17	0	[-0,139;- 0,067]

The next sections will deal with the items that demonstrated a statistical significance for the purpose of this study.

5.4. The cluster *I think*: a sample of overuse

Holmes (1985; 1990) identifies two broad semantic categories of *I think*: deliberative and tentative. The former, according to Holmes (1985, p. 33), is used to express personal confidence in the proposition asserted and therefore adds weight to the speech act. The latter, is used to express uncertainty. In Table 4, we focused on the tentative function since the main aim of this paper is to concentrate on items functioning as face-saving devices and as negative politeness strategies (Brown & Levinson 1987; Goffman 1967). In

part, the overuse of *I think* can be attributed to first language interference. The expression *Eu acho* (Marcuschi 1989) in Portuguese is quite common when one is expressing his/her opinion about some issue.

However, it is important to state that the cluster *I think* might not be the most appropriate cluster to be used in an academic context as it is commonly associated with casual conversation (see McCarthy & Carter 2006). At this point of the analysis, it can be argued that this group of students are misusing the item *I think* in academic oral presentations. The examples show that they are borrowing an item from casual conversation and incorporating in their academic oral production in an excessive way.

(3) *I think that it might be a kind of revenge a kind of revenge. Revenge against the society against the society against the maybe against something that he has lift in the air ...*

Overall, the overuse of *I think* in learner's production may lead to an ineffective strategy of image projection and face-saving strategy, as it barely meets the interlocutor's expectations for exchanges held in the academic sphere.

In the British data, it can be noticed that the expression *I think* is commonly used clustering with other modal items reinforcing the preference for epistemic forms by native or near-native speakers interacting in an academic environment. In sum, our findings demonstrate that the use of *I think* is multifunctional. On the one hand, it displays uncertainty, while it also serves as an epistemic⁵ marker, influenced by the learners' mother tongue (Marcuschi, 1989).

5.5. The cluster *you know*: a sample of underuse

According to Östman (1981) and Holmes (1986), *you know* serves a variety of different, though closely related functions in discourse. Particularly important for this study are Holmes' (1986) observations on the marker, in particular, as a hedge device. She divides the functions of *you know* into two categories: Category I comprises instances of *you know* expressing speaker confidence or certainty (positive politeness) and category II involves the usage of *you know* to express uncertainty of various kinds (negative politeness). Once again, only examples of *you know* functioning as a hedge

5 Epistemic modality refers to the degree of commitment one has in relation to what one says.

were isolated for analysis. *You know* occurs 124 times in BRASE, and 195 in BASE (normalised occurrences).

The Log-likelihood test indicates that within a hedging function, learners underuse the item when compared to native speakers. A thorough analysis of the concordance lines for *you know* in BRASE demonstrates that the item is more commonly used as a marker of assertiveness. In the majority of the examples, learners were expressing their certainty on a proposition whereas in BASE the opposite is observed. Extract (4) below illustrates the discussion carried out in this section.

(4) BRASE IFA1-B1⁶

*now you've said something . you said something interesting because when Lula was campaigning for the first time there was the same euphoria . about him **you know** and remember when Lula was elected for president the parties around the country it was . in a way the same kind of feeling*

In the case of *you know*, it can be said that learners rely more often on the function of assertiveness and/or shared knowledge than on the role of a hedge. This finding reinforces the claim that students are not aware of the items that are more commonly associated with the academic domain, in the case here academic oral presentations. The underuse of the cluster *you know* shows that learners usually lack the linguistic repertoire needed to cope with negative politeness strategies. As a result, they tend to express themselves in an assertive way, which can be interpreted as an imposition, not complying with the norms of academic interaction.

6. Final remarks

This paper focused on the function and meaning of the metalinguistic monitors: *just, maybe, I think, you know* and *kind of* in two corpora: the Brazilian Academic Spoken English (BRASE) Corpus and the British Academic Spoken English (BASE) Corpus. These items were analysed according to their occurrence in the main corpus (BRASE) and then compared and contrasted with the reference corpus (BASE). The analysis was

⁶ IFA 1 stands for *Inglês para Fins Acadêmicos 1 – English for Academic Purposes Level 1*. The example sentence was produced by students taking this course. B1 corresponds to their proficiency level (intermediate) following the *Common European Framework of Reference for Languages* (CEFR).

organised following the results from the statistical tests RR, OR, P-value and Log-likelihood. These tests were instrumental in identifying the features that were responsible for the linguistic differences between the two corpora, indicating the items worth being analysed in more detail.

Hence, identifying the overuse and/or underuse of items in students' spoken interlanguage may provide language teachers a better account of learners' production, in the case of this study, in oral academic discourse. Once teachers understand learners' discourse in a more accurate way, they can design activities that will better suit their needs.

In this study, learners used *just* employing its hedging function. The overuse of *just* in BRASE confirms that although learners seem to be aware of face issues, their choice of items differ from that of BASE speakers. Its high frequency of occurrence in BRASE indicates that learners are employing a marker from informal oral discourse in their oral academic presentations. This fact suggests that this group of Brazilian learners is unaware of the norms guiding academic context and thus need to develop their oral academic literacy in English.

As for the items *I think* and *you know*, the first point worth mentioning is the total number of clusters in both corpora. There are more clusters in BASE than in BRASE. In addition, the overall occurrence also points to significant differences. There is more variation in BASE: different forms are evenly distributed in the British corpus, while in BRASE the occurrences are concentrated on a more restricted set of forms. These findings suggest implications to the discourse produced by learners. By relying on a limited set of clusters, learners are constrained to a rigid repertoire, which limits their discourse and their ability to communicate in a more effective way, especially in an academic environment. Failing to use these markers can hinder communication and pose difficulties for students willing to participate in a more globalized academic community.

Focusing on linguistic aspects like the ones presented in this paper, we can possibly contribute to improve learners' oral communicative skills. In addition, a refined description of learners' spoken interlanguage contributes to raise cultural and linguistic awareness, guiding material design and to the growing area of Brazilian curriculum concerning English for Academic Purposes.

Overall, the prevailing deployment of less pragmatically enriched pragmatic markers in the learners' production analysed here may reveal that explicit instruction is a crucial element to raise awareness of these items in learner's discourse. As a consequence, this guidance would help pave the

way for more effective interpersonal interactions in the academic scenario, particularly with regard to politeness strategies and face-work.

Finally, it is worth mentioning that the statistical methods employed in this paper allowed for the achievement of an optimal point of data cross-analysis, which we believe has led to more reliable and significant results. For this reason, we believe that these findings may, and should, be replicated and expanded to other populations and to different scenarios.

References

- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- Aijmer, K. (2002). *English discourse particles: Evidence from a corpus* (Vol. 10). Amsterdam: John Benjamins Publishing.
- Aijmer, K., & Simon-Vandenberg, A. M. (2004). A model and a methodology for the study of pragmatic markers: The semantic field of expectation. *Journal of Pragmatics*, 36(10), 1781–1805.
- Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, 3(1), 173–190.
- Aijmer, K. (2013). *Understanding pragmatic markers*. Edinburgh: Edinburgh University Press.
- Andersen, G. (2001). *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents* (Vol. 84). Amsterdam: John Benjamins.
- Biber, D., Johansson, S., Leech, G., & Conrad, S. E. Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2), 97–116.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge: Cambridge University Press.
- Brown, P. (2015). Politeness and language. In *The International Encyclopedia of the Social and Behavioural Sciences (IESBS)* (pp. 326–330). (2nd ed.) Elsevier.
- Bublitz, W. (1978). *Ausdrucksweisen der Sprechereinstellung im Deutschen und im Englischen*. Tübingen: Niemeyer.
- Cunha, G. X. (2015). As relações retóricas e a negociação de faces em debate eleitoral. *Confluência*, 1(47), 205–238.
- Cortes, V. (2002). Lexical bundles in freshman composition. *Using corpora to explore linguistic variation*, 9, 131–145.
- Downing, A., & Locke, P. (2006). *English grammar: A university course*. New York: Routledge.

- Erman, B. (2001). Pragmatic markers revisited with a focus on you know in adult and adolescent talk. *Journal of pragmatics*, 33(9), 1337–1359.
- Fraser, B. (1999). What are discourse markers? *Journal of pragmatics*, 31(7), 931–952.
- Fraser, B. (1990). An approach to discourse markers. *Journal of pragmatics*, 14(3), 383–398.
- Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied linguistics*, 28(3), 410–439.
- Goffman, E. (1967). *Interaction ritual: essays on face-to-face interaction*. New York: Anchor Books.
- Goffman, E. (1976). Replies and responses. *Language in society*, 5(3), 257–313.
- Halliday, M. A. K., Matthiessen, C., & Halliday, M. (2014). *An introduction to functional grammar*. London: Routledge.
- Holmes, J. (1985). Sex differences and miscommunication: Some data from New Zealand. *Cross-cultural Encounters: Communication and Miscommunication*, (pp. 24–43). Melbourne: River Seine.
- Holmes, J. (1986). Functions of you know in women's and men's speech. *Language in society*, 15(1), 1–21.
- Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication*, 10(3), 185–205.
- Haugh, M. (2013). Disentangling face, facework and in/politeness. *Pragmática sociocultural*, 1(1), 46.
- Kerbrat-Orecchioni, C. (2006). *Análise da conversação: Princípios e métodos*. São Paulo: Parábola Editorial.
- Kriwonossow, A. (1977). *Die modalen Partikeln in der deutschen Gegenwartssprache*. Vol. GAG 214. Göppingen: Kümmerle-Verlag.
- Marcuschi, L. A. (1989). *Marcadores conversacionais do português brasileiro: formas, posições e funções*. Campinas: Editora da UNICAMP.
- McCarthy, M. & Carter, R. (2006). This that and the other: *Multi-word clusters in spoken English* as visible patterns of interaction. *Explorations in corpus linguistics*, 7–26.
- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
- Oliveira, A. L. Adorno Marciotto, G. X. Cunha, & M. Vieira Miranda (2017). Nominalizations as complex strategies of politeness and face-work in scientific papers written in Brazilian Portuguese. *Cadernos de Estudos Lingüísticos*, 59(2), 361–374.
- Orfanò, B.M. (2010). *The representation of spoken language: a corpus-based study of sitcom discourse* [unpublished PhD dissertation], Limerick: Mary Immaculate 2College-University.
- Östman, J. O. (1981). *'You Know': A discourse-functional study*. Amsterdam: John Benjamins.
- Ran, Y. (2003). A pragmatic account of the discourse marker WELL. *Journal of Foreign Languages*, 3, 58–64.

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6) [held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1–8 October 2000], Hong Kong.

[submitted on March 31, 2018 and accepted for publication on September 17, 2018]