

Sense-related items and lexical patterning in English and Portuguese scientific abstracts

Carmen Dayrell

University of São Paulo

Introduction

Non-native speakers face major challenges when writing academic English. In addition to dealing with the various difficulties involved in writing in a foreign language, they also have to comply with the conventions and norms adopted by their academic discourse community, which may differ from one language to another.

Not surprisingly, within the specific context of English Language Teaching (ELT) and more specifically English for Academic Purposes (EAP), a number of scholars have focused on describing recurring phrases and lexical patterns which are frequently used in academic discourse (see, for instance, Swales 1990, 2004, Swales & Feak 2000 and Weissberg & Buker 1990). An invaluable contribution is also offered by corpus-based studies which use empirical data to explore the specific nature of academic texts (among others, Peacock 2006, Orasan 2001 and Thompson 2001). Most closely related to the present study are some recent corpus-based studies which place special emphasis on the language produced by non-native speakers and investigate similarities and differences between non-native and published scientific texts (for instance, Hyland 2008a, 2008b, Dayrell & Aluísio 2008, Cortes 2004, Hewings & Hewings 2002).

The present study pursues this last line of thought and examines the language produced

by Brazilian graduate students as opposed to published texts. The focus is on scientific abstracts and special emphasis is given to lexical choices and collocational patterns. As Genoves Jr. *et al.* (2007) explain, errors related to lexical use are by far the most frequent made by Brazilian students when writing scientific papers in English. The authors refer to the misuse of a word to express a particular meaning, which can be of three types: errors due to direct translations of a Portuguese item into a false cognate in English (*pretend* for *intend*), errors made in a common phrase (*as* for *such as*) and errors related to collocational patterns (*do contributions* for *make contributions*).

According to Granger (2002), when it comes to identifying the main difficulties faced by second-language learners with respect to the use of lexical items, phrases and structures, the data provided by native corpora will not suffice and needs to be complemented with information extracted from learner corpora. Thus, learner corpora – corpora containing texts produced by foreign or second language learners (Gilquin *et al.* 2007) – have opened up new perspectives in the field of language teaching and can provide useful insights and enhance our understanding of underlying regularities in the language produced by learners. As Gilquin *et al.* (2007:320) point out, second language “learners admittedly share a number of difficulties with novice native writers but they have also proven to have their own distinctive problems, which a careful corpus-based investigation can help uncover”.

This paper examines five sets of sense-related verbs which frequently occur in English academic discourse (see the Methodology section for details). The primary aim is to investigate potential differences in the frequency and lexical patterning of some specific English verbs in abstracts written by Brazilian graduate students in relation to published abstracts from the same disciplines. Two hypotheses are put forward in relation to a given set of verbs: (1) students may

show a more marked preference for a specific item; and (2) students may draw more heavily on specific lexical patterns. Thus, in addition to examining whether students misuse a given lexical item, this paper is particularly interested in exploring the question of naturalness, that is to say, whether the writer's lexical choice is the one most frequently used in a particular context. This hypothesis is based on the findings of previous work (Aluisio and Dayrell 2008) which revealed relevant differences between students' and published abstracts with respect to sense-related nouns: *work, paper, study, article* and *research*.

The results are validated by examining the frequency percentage of individual verbs and their respective lexical patterns in a reference corpus of English abstracts. The study then takes a step further and uses a reference corpus of Portuguese abstracts to investigate whether the lexical choices made by Brazilian graduate students can be said to have been influenced by the Portuguese language. The long-term objective is two-fold: to improve course materials and resources for academic English and to provide computer-aided writing tools with linguistic input so as to enable the automatic identification and correction of errors at the lexical, syntactical and rhetorical levels¹.

The Comparable Corpus of English Abstracts

The data analysed in this paper is drawn from a comparable corpus of English abstracts which consists of two separate subcorpora, containing 159 abstracts each: one made up of abstracts written by Brazilian graduate students and the other of abstracts extracted from published papers. They contain 33,836 and 28,117 words (tokens) respectively.

The corpus of abstracts written by students (hereafter EA-STs) contains abstracts collected from seven courses on academic writing offered to graduate students from various disciplines at two universities in Brazil between 2004 and 2008. Here, I examine the first version

of the abstracts, that is to say, abstracts handed in before the course starts. It should be mentioned that the students from these courses varied considerably with respect to their knowledge of English, ranging from lower to very advanced levels. As Genoves Jr. *et al.* (2007) explain, these abstracts display striking differences in terms of both quantitative (number of errors) and qualitative (types of errors) aspects. Table 1 summarises the current composition of the EA-STs.

	Discipline	Number of abstracts	Percentage of abstracts
1.	Physics	79	50%
2.	Pharmaceutical Sciences	38	24%
3.	Computer Science	26	16%
4.	Engineering	16	10%
	TOTAL	159	100%

Table 1: Composition of the EA-STs

Pharmaceutical sciences refers to students from pharmacology, chemistry and biology/genetics. Although the engineering departments at these universities do not offer courses on academic English, some students from these disciplines have attended the courses offered by the department of physics. It is therefore estimated that 10% of the abstracts included in the corpus are related to the field of engineering.

The corpus of abstracts from published papers (hereafter EA-PUB) was designed to match the specifications of the EA-STs so that the two collections were made comparable. All abstracts were extracted from published papers, which were randomly selected from various leading academic journals in the disciplines in question, such as *Physical Review Letters (A-D)*, *Science*, *Nature*, *Biotechnology Progress*, *ACM Transactions on Information Systems* and *International Journal of Mechanical Sciences*.

A relevant methodological point to make here is that by published abstracts I do not mean that they have necessarily been written by native speakers of English. What is assumed here is that, given that they have been published by recognised bodies of a given discipline, they are

presumably of acceptable quality and more likely to comply with the pre-established conventions adopted by the discourse community in question. Another difference between the two subcorpora is that most abstracts included in the EA-PUB come from papers by more than one author, which is not the case in the EA-STTS.

Reference corpora

Two reference corpora of abstracts are used in this study: one of English abstracts and the other of Brazilian Portuguese abstracts. These are independent abstracts, that is to say, they are not translations of one another. The selection of texts to be included in the reference corpus of English abstracts (hereafter REF-ENG) follows the criteria used to compile the EA-PUB. Thus, it includes abstracts of published papers, which have been published by major academic journals in the disciplines in question. It also follows the composition of the EA-PUB with respect to the percentage of abstracts from each discipline (Table 2). The corpus contains 1,170 abstracts (over 210,000 tokens)

	Discipline	Number of abstracts	Percentage of abstracts
1.	Physics	585	50%
2.	Pharmaceutical Sciences	281	24%
3.	Computer Science	187	16%
4.	Engineering	117	10%
	TOTAL	1,170	100%

Table 2: Composition of the reference corpus of English abstracts (REF-ENG)

The reference corpus of Portuguese abstracts (hereafter REF-PTG) was initially intended to be of similar size and have the same composition as the REF-ENG. However, for the disciplines under analysis, international journals vastly outnumber Brazilian journals and, to make matters worse, many Brazilian academic journals have English as their official language. This is why the REF-PTG is reduced in size (620 abstracts – over 100,000 tokens) and its

composition is slightly different from the REF-ENG (Table 3). Also, in addition to abstracts of papers published in major Brazilian academic journals, the REF-PTG also includes abstracts published in conference proceedings.

	Discipline	Number of abstracts	Percentage of abstracts
1.	Physics	200	32%
2.	Pharmaceutical Sciences	150	24%
3.	Computer Science	150	24%
4.	Engineering	120	19%
	TOTAL	620	100%

Table 3: Composition of the reference corpus of Portuguese abstracts (REF-PTG)

Methodology

This section explains the methodology adopted here in order to test the hypotheses I put forward. All procedures described below are carried out by means of the software package *WordSmith Tools*, version 4.0 (Scott 2004).

The first step is to identify five verbs which could serve as the starting point for the analysis. It is important to stress that the analysis takes into account lemmas, that is, all inflected forms of the verbs. For instance, the label STUDY² includes: *study*, *studies*, *studied* and *studying*. Thus, two basic criteria are adopted: (1) the frequency of the lemma in the EA-STTS; (2) the frequency of the lemma in academic discourse. The first criterion selects verbs (lemmas) with the highest number of occurrences in the EA-STTS. This is mainly because the focus is on the language produced by students and on identifying lexical items which frequently pose a challenge for Brazilian writers. The second criterion was established in order to select verbs which would typically occur in academic discourse. This is done by generating a list of all words from the REF-ENG whose frequency is unusually high in comparison with a reference corpus, that is, a keyword list³. Here, I have used the British National Corpus (BNC) as the reference

corpus⁴. Thus, in order to be considered for further analysis, the potential candidate should appear in the REF-ENG keyword list. For instance, MAKE is one of the most frequent verbs in the EA-STS but it was discarded because it does not appear in the keyword list.

Once a given verb has been selected, the next step is to search for its near-synonyms. By near-synonyms, I do not mean that they are interchangeable but instead that their meaning is related in one way or another. Near-synonyms are selected on the basis of the entries in the *Collins Thesaurus* (2002). In addition, the suggested verb should appear at least once in either the EA-STS or the EA-PUB. The verb STUDY can serve as an example to illustrate how the near-synonyms are selected. The *Collins Thesaurus* (2002) suggests the following as synonyms for STUDY: *analyse, examine, investigate, look into, peruse, research, scrutinize, survey* and *work over*. The verbs *peruse, scrutinize, look into* and *work over* are discarded because they occur neither in the EA-STS nor in the EA-PUB. Table 4 lists the five sets of verbs selected for analysis. SHOW and PRESENT were both included among the most frequent verbs in the EA-STS and were initially considered as separate entries. However, they are regarded by *Collins Thesaurus* (2002) as synonyms and this is why they have been grouped together.

Set	Verbs	Sense-Related Verbs
1.	USE	APPLY / EMPLOY / UTILIZE
2.	SHOW	PRESENT / DEMONSTRATE / EXHIBIT / DISPLAY
3.	OBTAIN	COLLECT / ACHIEVE / ATTAIN / ACQUIRE
4.	FIND	OBSERVE / DETECT / DISCOVER / EXPERIENCE / NOTE / NOTICE / PERCEIVE
5.	STUDY	ANALYSE / INVESTIGATE / EXAMINE / RESEARCH / SURVEY

Table 4: Verbs selected for analysis

The first hypothesis is tested by comparing the frequency percentages of each verb in the EA-STS and the EA-PUB. The results are validated by examining the frequency percentage of each verb in the REF-ENG. The study then investigates whether the lexical choices made by

Brazilian graduate students can be said to have been influenced by the Portuguese language. In order to do this, I look at the frequency percentage of cognate translations of the English lexical items under analysis in the REF-PTG. For instance, for set-1, the following Portuguese verbs are examined: USAR [USE], APLICAR [APPLY], EMPREGAR [EMPLOY] and UTILIZAR [UTILIZE].

The second hypothesis focuses on the recurring patterns with the highest number of instances in at least one subcorpus, EA-STS or EA-PUB, and discusses the main similarities and differences between students' and published abstracts. The comparison takes into consideration the percentages of each pattern in the two subcorpora, rather than the raw number of instances. Patterns are retrieved by sorting the concordance lines by different positions on the left and on the right of the verb and examining the surrounding context. In the specific case of this study, I examine all lexical items which occur in a span of five words to the right and five words to the left of the search-verb (5:5). The findings are then validated in the REF-ENG. Portuguese translations of individual lexical patterns are not examined in the REF-PTG because it would involve a more comprehensive analysis of the lexical and syntactical differences between the two language systems, which goes beyond the scope of this study.

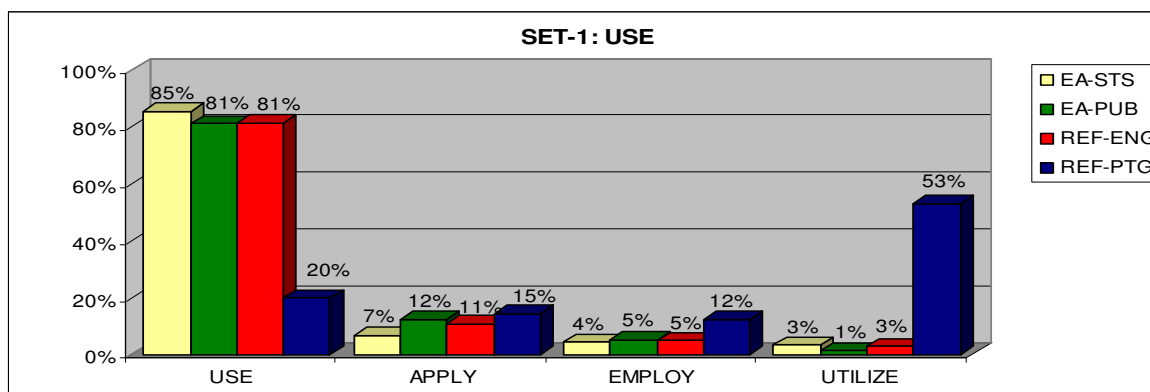
Data Analysis

This section describes the results of the data analysis. As we shall see, the findings related to the third hypothesis, which tests whether the lexical choices made by Brazilian students can be said to have been influenced by the Portuguese language, are presented together with the results of the first hypothesis. The data is represented in graphs for each set of verbs separately, ordered by the number of instances in the EA-STS subcorpus. The discussion focuses on those cases in which there is at least a five-percentage-point difference between the corpora, be it either between the EA-STS and the EA-PUB subcorpora or in relation to the reference corpora.

Hypothesis (1): Preference for specific items of a given set of verbs

Taking into consideration the verbs within set-1 – USE, APPLY, EMPLOY and UTILIZE –, we find that both the EA-STG and the EA-PUB show a strong preference for the verb USE, which accounts for more than 80% of instances in the two subcorpora (Graph 1). The REF-ENG confirms that USE is by far the most frequent verb within set-1. The percentage of APPLY is five percentage points (pp) higher in the EA-PUB in comparison with the EA-STG (12% and 7% respectively) and it accounts for 11% of instances in the REF-ENG. These figures suggest that students seem to draw more heavily on USE and employ it in cases where APPLY would also fit.

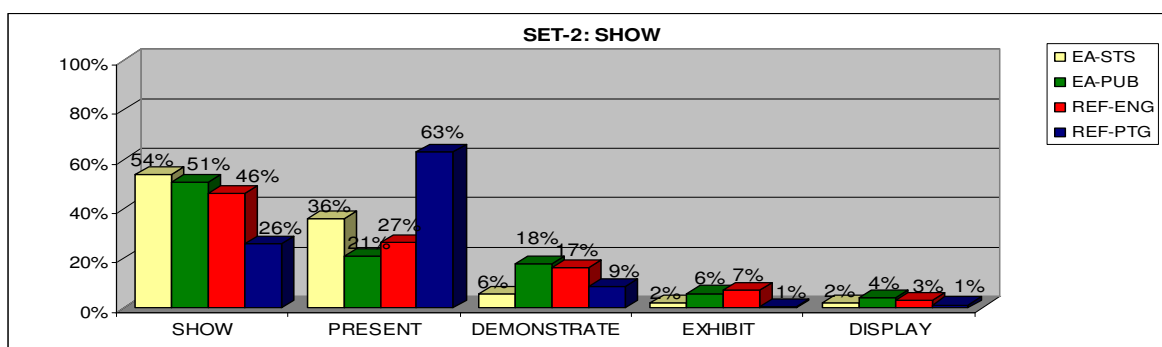
As for the cognate translations of the English verbs within set-1, the following Portuguese verbs are analysed in the REF-PTG: USAR [USE], APLICAR [APPLY], EMPREGAR [EMPLOY] and UTILIZAR [UTILIZE]. The findings do not confirm the hypothesis that the lexical choices made by Brazilian students were influenced by the Portuguese language. Portuguese abstracts show a strong preference for UTILIZAR (53%) while USAR occurs in only 20% of instances. The lower percentage of APPLY in the EA-STG is not explained either; APLICAR is relatively more frequent in the REF-PTG (15%) than in the REF-ENG (11%). The same can be said about EMPLOY. It is used in similar proportion in the EA-STG, EA-PUB and the REF-ENG (4%, 5% and 5% respectively) while it is much more frequent in the REF-PTG (12%).



Graph 1: Frequency percentages of verbs within set-1

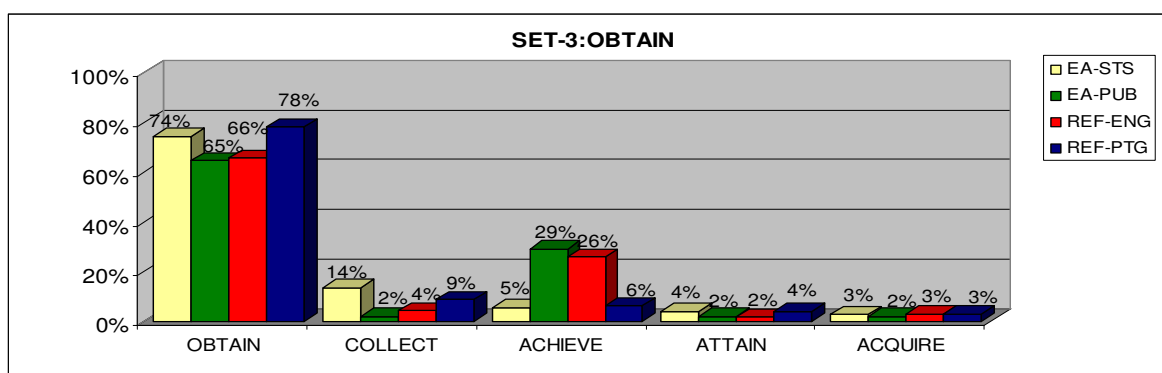
As for set-2 – SHOW, PRESENT, DEMONSTRATE, EXHIBIT and DISPLAY – more than 50% of instances in both the EA-STS and the EA-PUB refer to the verb SHOW (Graph 2). The EA-STS shows a higher percentage of instances for PRESENT (36%) in comparison with the EA-PUB (21%). DEMONSTRATE and EXHIBIT are more frequent in published abstracts than in students' writing: 12pp and 4pp higher respectively. The REF-ENG confirms that SHOW is the most frequent and PRESENT is the second most frequent verb within set-2. However, their relative frequencies of occurrence in the EA-STS exceed their frequencies in the REF-ENG: 8pp and 9pp higher respectively. The REF-ENG also suggests that DEMONSTRATE and EXHIBIT are underused by students.

The following Portuguese verbs have been analysed: MOSTRAR [SHOW], APRESENTAR [PRESENT], DEMONSTRAR [DEMONSTRATE], EXIBIR [EXHIBIT] and EXPOR [DISPLAY]. APRESENTAR is by far the most frequent verb in Portuguese abstracts, with 63% of occurrences, and MOSTRAR comes second with 26% of all instances. The clear preference of students for SHOW does not seem to reflect the influence of the Portuguese language. However, the higher percentage of PRESENT in the EA-STS in relation to English published abstracts may, in principle, be said to have been influenced by the marked preference of Portuguese abstracts for APRESENTAR. Similarly, the underuse of DEMONSTRATE and EXHIBIT by students may also be explained by the low frequency of DEMONSTRAR and EXIBIR in the REF-PTG.



Graph 2: Frequency percentages of verbs within set-2

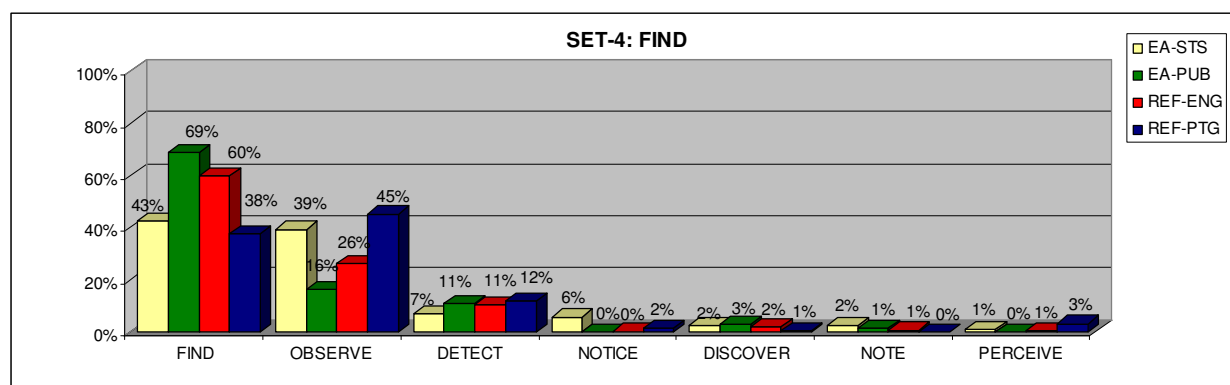
As regards set-3 – OBTAIN, COLLECT, ACHIEVE, ATTAIN and ACQUIRE –, all corpora display a strong preference for OBTAIN (Graph-3). This includes the frequency of its Portuguese translation (OBTER) in the REF-PTG. However, here again, the tendency is more marked in students' than in published abstracts: 9pp higher in relation to the EA-PUB and 8pp in relation to the REF-ENG. This may be explained by the influence of Portuguese: OBTER accounts for 78% of instances in the REF-PTG. COLLECT and ACHIEVE are used in similar proportion in the EA-PUB and the REF-ENG whereas their frequencies in the EA-STS seem to reflect the frequencies of their corresponding Portuguese translations (COLETAR and ALCANÇAR) in the REF-PTG. ATTAIN, ACQUIRE and their corresponding translations into Portuguese (ATINGIR and ADQUIRIR) are used with a similar low frequency in all corpora.



Graph 3: Frequency percentages of verbs within set-3

Taking into consideration the verbs within set-4 – FIND, OBSERVE, DETECT, DISCOVER, EXPERIENCE, NOTE, NOTICE and PERCEIVE –, we find that the EA-PUB shows a clear preference for one specific verb (FIND), which represents 69% of instances (Graph 4). Although slightly less marked, this tendency is confirmed by the REF-ENG (60%). By contrast, the EA-STS shows a lower percentage of FIND (43%), which is in fact closer to the percentage of its corresponding translations into Portuguese (38%)⁵. For OBSERVE, the percentage in the EA-STS is again closer to the percentage of its corresponding translation (OBSERVAR) in the REF-PTG: 39% and 45%

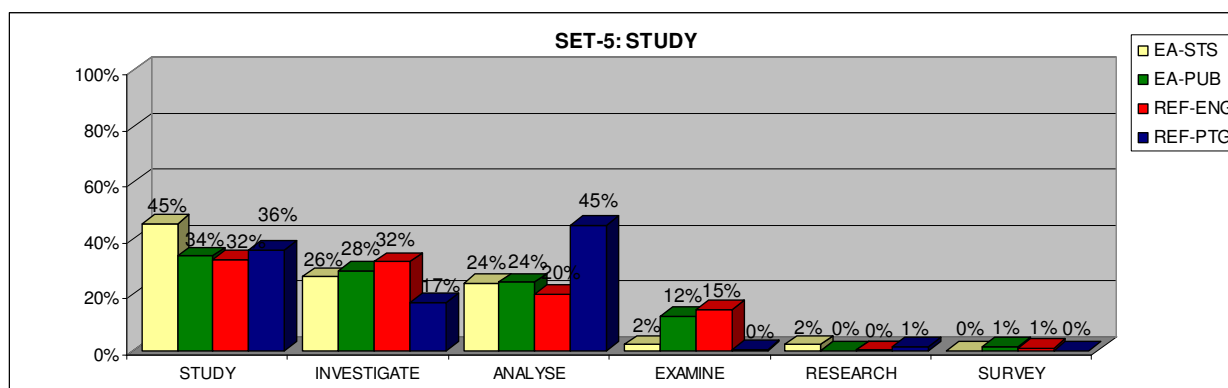
respectively. OBSERVE occurs in only 16% of instances in the EA-PUB and it is 10pp more frequent in the REF-ENG (26%). DETECT seems to be underused in the EA-STS (7%) in relation to published abstracts. This is not explained by the influence of the Portuguese language since its corresponding translation (DETECTAR) represents 12% of instances in the REF-PTG. NOTICE appears in 6% of occurrences in the EA-STS and shows no occurrences in the EA-PUB and only one instance in the REF-ENG. When it comes to Portuguese, both NOTICE and NOTE can be translated as NOTAR. Here, I have contrasted all three instances of NOTAR with NOTICE and this is why it shows 2% in the REF-PTG.



Graph 4: Frequency percentages of verbs within set-4

For set-5 – STUDY, ANALYSE, INVESTIGATE, EXAMINE, RESEARCH and SURVEY –, students again show a more marked preference for one specific verb: STUDY accounts for 45% of instances in the EA-STS (Graph 5). In the EA-PUB, the preference for STUDY is not as marked (34%) and it is slightly less frequent in the REF-ENG (32%). INVESTIGATE and ANALYSE are also frequent in the EA-STS, representing 26% and 24% of instances respectively and they appear in similar proportion in the EA-PUB, 28% and 24% respectively. In the REF-ENG, the percentage of INVESTIGATE is slightly higher (32%) and the percentage of ANALYSE is lower (20%). EXAMINE is much more frequent in the EA-PUB (12%) than in the EA-STS (2%) and its underuse by students is confirmed by the REF-ENG.

The following Portuguese verbs have been analysed: ESTUDAR [STUDY], INVESTIGAR [INVESTIGATE], ANALIZAR [ANALYSE], EXAMINAR [EXAMINE] and PESQUISAR [RESEARCH and SURVEY]. Portuguese abstracts display a strong preference for ANALIZAR, which accounts for 45% of instances. ESTUDAR is also frequent with 36% of occurrences while INVESTIGAR represents 17% of instances. For these three verbs, the figures indicate that the lexical choices made by Brazilian students do not seem to have been influenced by the Portuguese language. By contrast, the influence of Portuguese in the students' choices is clearly seen in the frequency of EXAMINE in the EA-STTS. EXAMINE and its cognate translation (EXAMINAR) appear with similar low frequencies in the EA-STTS (2%) and the REF-PTG (0%). In the EA-PUB and the REF-ENG, EXAMINE occurs in 12% and 15% of instances respectively.



Graph 5: Frequency percentages of verbs within set-5

Hypothesis (2): Preference for specific lexical patterns

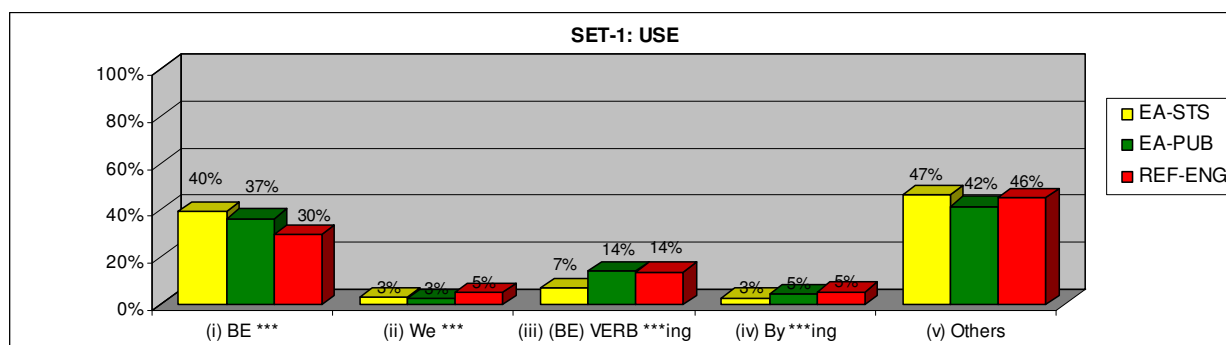
For the second hypothesis, the notation *** is used to indicate the position of the verb under analysis. For example, **we ***** stands for *we show*, *we investigate*, *we find*, *we use*, etc. PP and INF indicate any verb other than the one under analysis, in the participle and infinitive forms respectively. ADV refers to adverbs and semantic categories are represented by *SMALL CAPITALS* in italics. The category “others” is used to group together all instances which do not yield recurring lexico-grammatical patterns. The last column shows some examples, which are by no

means restricted to the ones presented.

For set-1 – USE, APPLY, EMPLOY and UTILIZE –, the following lexico-grammatical patterns have been identified:

	Patterns	Examples
i	(i) BE ***	<i>is used, were applied</i>
ii	(ii) We ***	<i>we used, we use</i>
iii	(iii) (BE) PP ***ing	<i>are treated using, were performed using</i>
iv	(iv) By ***ing	<i>by using</i>

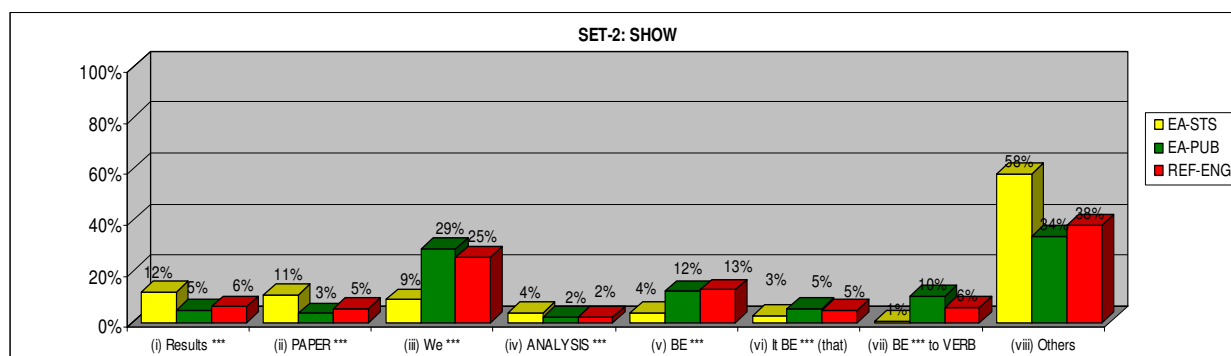
The analysis shows that both the EA-STs and the EA-PUB draw heavily on pattern (i), which occurs in similar proportion: 40% in the former and 37% in the latter (Graph 6). Nearly 50% of instances in both subcorpora fall under the category “others”. We also noticed that the frequency of pattern (iii) is 7pp higher in the EA-PUB in relation to the EA-STs. The REF-ENG confirms all tendencies revealed in the EA-PUB.



Graph 6: Lexico-grammatical patterns for set-1

Taking into consideration the verbs within set-2 – SHOW, PRESENT, DEMONSTRATE, EXHIBIT and DISPLAY –, the following recurring patterns were found, where *PAPER* refers to the lexical items: *paper, study* and *work*, and *ANALYSIS* refers to *analysis, experiment* and *tests*.

Patterns	Examples
i. results (BE) ***	<i>results demonstrated, results are presented</i>
ii. PAPER ***	<i>this paper shows, this work presents</i>
iii. we ***	<i>we show, we present</i>
iv. ANALYSIS ***	<i>analysis showed, tests show</i>
v. BE ***	<i>has been exhibited, is showed</i>
vi. it BE *** (that)	<i>it is showed that</i>
vii. BE *** to INF	<i>is shown to depend/be</i>

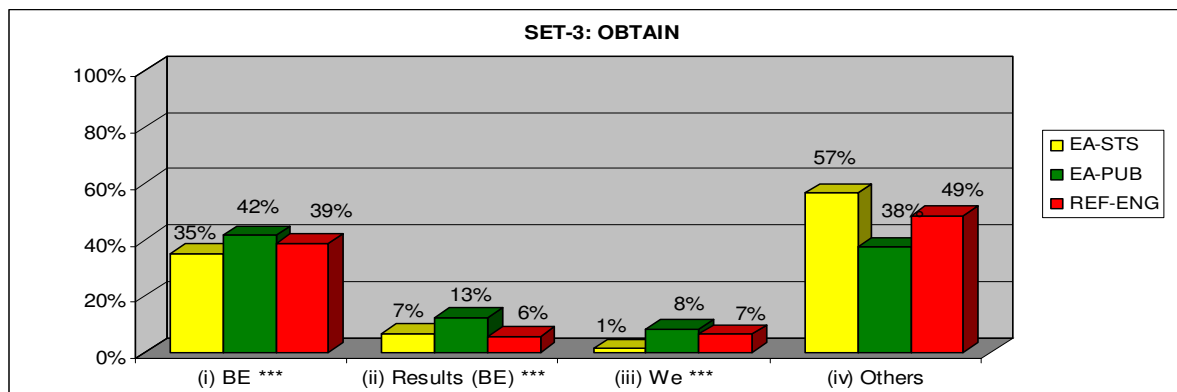


Graph 7: Lexico-grammatical patterns for set-2

We find that 58% of instances in the EA-STS subcorpus do not yield recurring lexical patterns as opposed to 34% in the EA-PUB collection (Graph 7). Patterns (i) and (ii) are more frequent in the EA-STS (12% and 11%) than in the EA-PUB (5% and 3% respectively) whereas patterns (iii), (v) and (vii) are far more frequent in the EA-PUB than in the EA-STS (20pp, 8pp and 9pp higher respectively). The only patterns which occur in similar proportion in both subcorpora are patterns (iv) and (vi). The REF-ENG confirms all tendencies displayed by the EA-PUB.

Set-3 – OBTAIN, COLLECT, ACHIEVE, ATTAIN and ACQUIRE – yields three recurring patterns:

Patterns	Examples
i. BE ***	<i>can be obtained, are collected</i>
ii. results (BE) ***	<i>results (were) obtained</i>
iii. we ***	<i>we obtained</i>

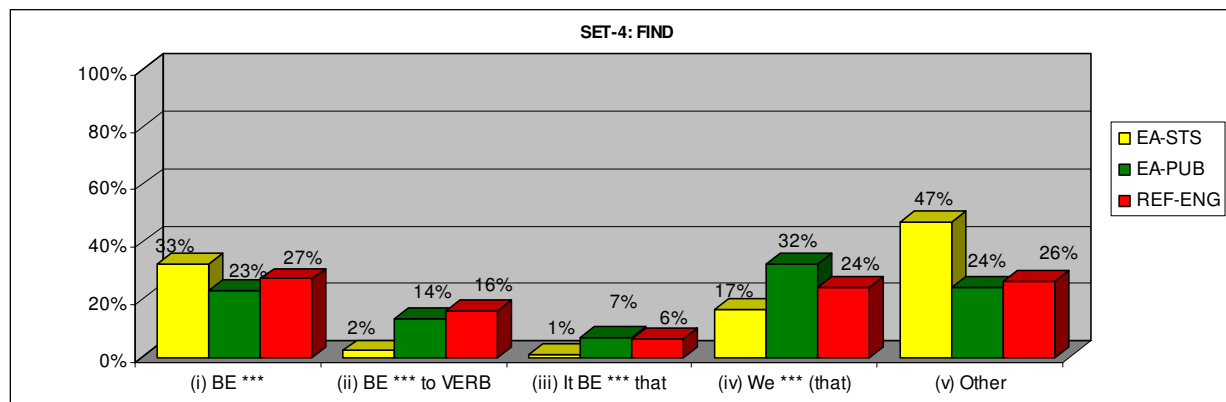


Graph 8: Lexico-grammatical patterns for set-3

The vast majority of instances in the EA-STS (57%) do not yield recurring patterns whereas, in the EA-PUB, the percentage is 38% (Graph 8). However, the REF-ENG suggests that the verbs within this set seem to be used more freely since 49% of instances fall under the category “others”. Taking into consideration the three recurring patterns mentioned above, we find that they are all less frequent in the EA-STS in comparison with the EA-PUB. The percentages of patterns (i) and (iii) are similar in the EA-PUB and the REF-ENG. However, for pattern (ii), the REF-ENG does not confirm the tendency displayed by the EA-PUB, since the percentage of instances is 7pp higher in the latter. In fact, the pattern occurs in similar proportion in the REF-ENG and the EA-STS.

As regards set-4 – FIND, OBSERVE, DETECT, DISCOVER, EXPERIENCE, NOTE, NOTICE and PERCEIVE, the following patterns have been analysed:

Patterns	Examples
i. BE (ADV) ***	<i>was found, was first observed</i>
ii. BE *** to INF	<i>were found to be</i>
iii. it BE *** that	<i>it was observed that</i>
iv. we *** (that)	<i>we find that, we have not detected</i>



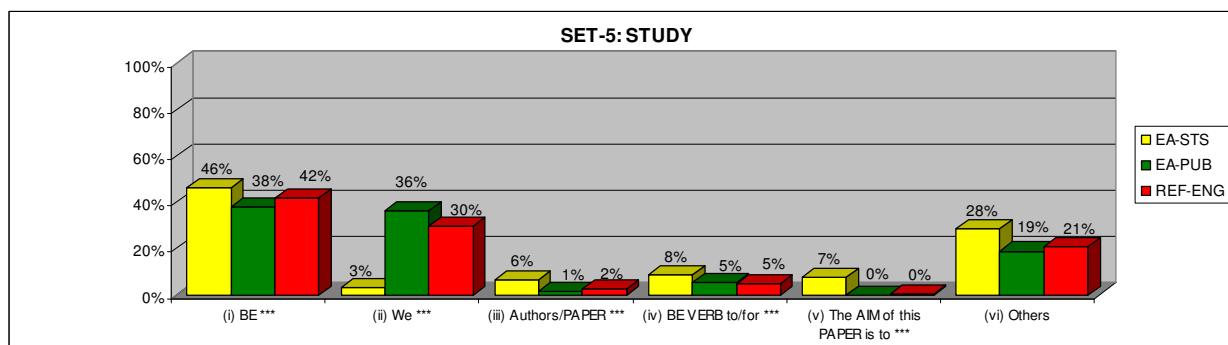
Graph 9: Lexico-grammatical patterns for set-4

Here again, most instances in the EA-STS (47%) do not yield recurring patterns (Graph 9). This is not the case in published abstracts for which the percentage of instances in the category “others” is 23pp lower in EA-PUB and 21 pp lower in the REF-ENG. Pattern (i) appears in 33% of instances in the EA-STS whereas patterns (ii) and (iii), which also refer to the passive voice, are hardly used. More importantly, for pattern (i), students rely heavily on the verb OBSERVE, which accounts for 53% of all instances; FIND comes second with 28% of instances. By contrast, published abstracts show a lower percentage of instances of pattern (i) (23%) and higher percentages of patterns (ii) and (iii), 14% and 7% respectively. These figures are similar to the ones found in the REF-ENG. For pattern (i), published abstracts use FIND and OBSERVE in similar proportion, 19% of instances in each. Patterns (ii) and (iii) are related to the verb FIND only. The overall percentage of the passive voice (patterns (i), (ii) and (iii)) is lower in the EA-STS (36%) in relation to the EA-PUB (44%) and the REF-ENG (49%). Pattern (iv) is much more frequent in published (32%) than in students’ abstracts (17%). The percentage of instances in the REF-ENG (24%) is not as high as in the EA-PUB, but it is 8pp higher than in the EA-STS. As a whole, these figures seem to suggest that students do not use the recurring patterns most commonly associated with the verbs within set-4.

Five recurring patterns are identified for the verbs within set-5 – STUDY, ANALYSE,

INVESTIGATE, EXAMINE, RESEARCH and SURVEY, where *PAPER* refers to the following lexical items: *paper*, *study*, *research* and *work* and *AIM* refers to *aim*, *objective* and *purpose*:

Patterns	Examples
i. BE ***	<i>were studied, have been investigated</i>
ii. we ***	<i>we analyzed</i>
iii. authors/ <i>PAPER</i> ***	<i>the authors investigate, this paper analyses</i>
iv. BE PP to/for ***	<i>were used to investigate, have been used for studying</i>
v. the <i>AIM</i> of this <i>PAPER</i> is to ***	<i>the aim of this study is to analyze</i>



Graph 10: Lexico-grammatical patterns for set-5

Both students' and published abstracts display a marked preference for pattern (i) (Graph 10); the percentage is nevertheless higher in the EA-STs (46%). The REF-ENG confirms this clear tendency towards the passive voice. Pattern (iii), which is also impersonal, represents 6% of occurrences in the EA-STs and no more than 2% in the other two corpora. A striking difference is seen between the percentages of pattern (ii) in the EA-STs and EA-PUB: it represents 36% of instances in the EA-PUB and only 3% of instances in the EA-STs. The REF-ENG also shows a high proportion of this pattern (30%). Pattern (v) appears in the EA-STs only, with 7% of instances. The category "others" is again more frequent in the EA-STs (28%) in relation to the other two corpora: 19% in the EA-PUB and 21% in the REF-ENG.

Discussion

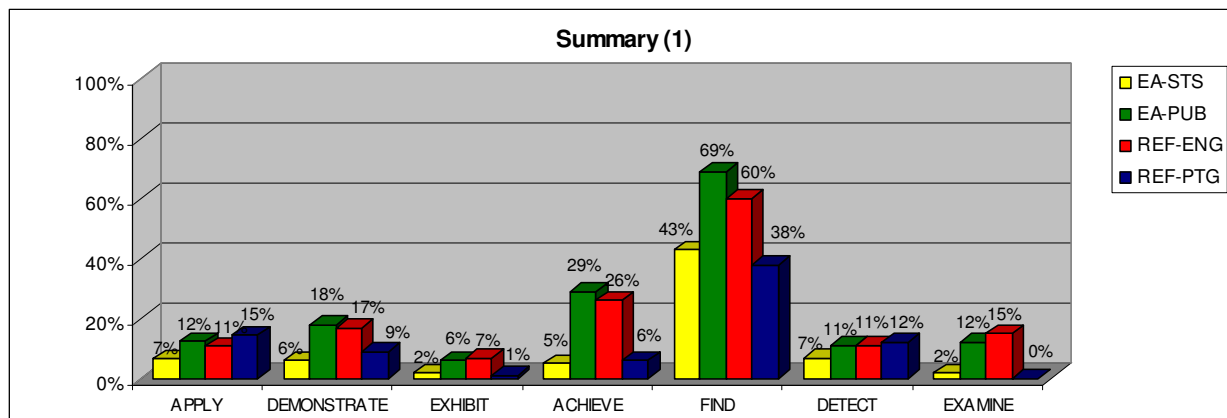
The analysis of the data indicates that students tend towards the most frequent verb within each set. With the exception of FIND, the percentages of instances are consistently higher in the EA-STS (Table 5) as opposed to the EA-PUB. This in other words means that students tend to draw more heavily on the most recurring verbs and use them in contexts where other sense-related verbs would apply.

Set	Verb	Percentage in the EA-STS	Percentage in the EA-PUB	Difference in pp	Subcorpus with higher percentage of instances
1	USE	85%	81%	4 pp	(STS)
2	SHOW	54%	51%	3 pp	(STS)
3	OBTAIN	74%	65%	9 pp	(STS)
4	FIND	43%	69%	26 pp	(PUB)
5	STUDY	45%	34%	11 pp	(STS)

Table 5: Percentages of the most frequent verbs of each set in the EA-STS and the EA-PUB

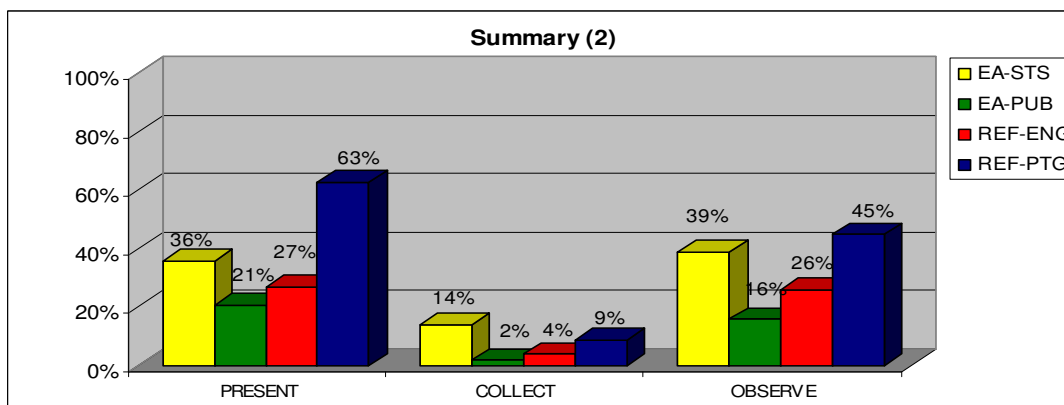
FIND is a special case since it accounts for 69% of instances in the EA-PUB whereas the EA-STS uses FIND and OBSERVE in similar proportion, 43% and 39% respectively (see Graph 4). The percentages of these two verbs in EA-STS seem to be justified by the interference of the Portuguese language; its corresponding translations represent 38% and 45% of instances in the REF-PTG respectively.

Seven verbs are clearly underused by students since they show a low percentage of instances in the EA-STS and a higher percentage in both the EA-PUB and the REF-ENG. These are: APPLY, DEMONSTRATE, EXHIBIT, ACHIEVE, FIND, DETECT and EXAMINE (Graph 11). With the exceptions of APPLY and DETECT, the lower percentages of instances in the EA-STS seem to be due to the interference of the Portuguese language for which the percentages are also low.



Graph 11: Verbs underused by students

Three verbs seem to be overused by students: PRESENT, COLLECT and OBSERVE (Graph 12). They are relatively frequent in the EA-STS and far less frequent in the EA-PUB and the REF-ENG. This also seems to happen because of the influence of the Portuguese language since all the verbs are much more frequent in the REF-PTG than in the REF-ENG.



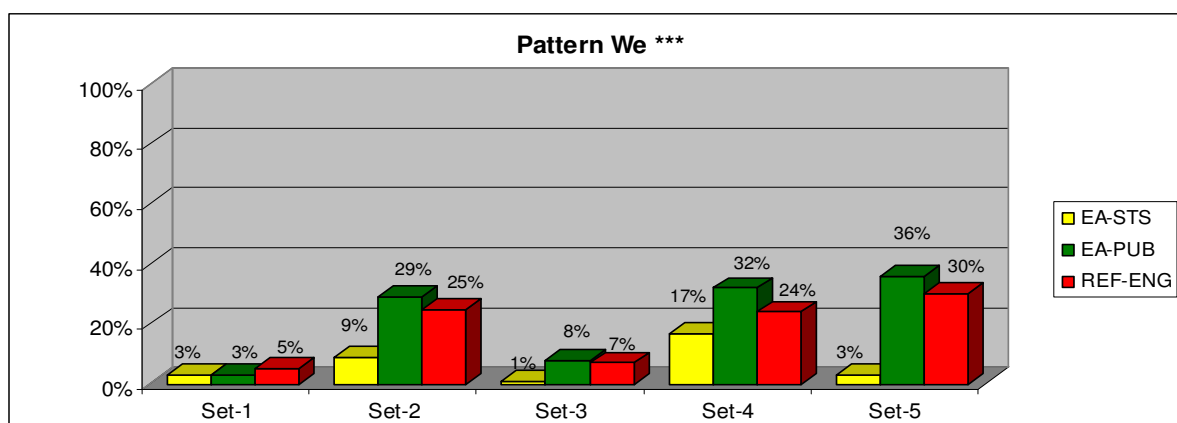
Graph 12: Verbs overused by students

In terms of recurring patterns, with the exception of set-2 in the EA-STS, we find a high number of instances with BE in the immediate positions on the left of the search-verb in both the EA-STS and the EA-PUB (Table 6). From a syntactical perspective, this means that the passive voice is widely used in both collections.

Set	Percentage in the EA-STs	Percentage in the EA-PUB
1	40%	37%
2	8%	27%
3	35%	42%
4	36%	44%
5	46%	38%

Table 6: Percentages of instances with BE in the immediate positions on the left of the search-verb

The active voice (**we *****), which clearly puts the researchers in the spotlight, is far more frequent in the EA-PUB than in the EA-STs (Graph 13). The only exception is for set-1, where both subcorpora show similar low percentages for this specific pattern. In the specific case of sets 2 and 5, other means of disguising authorship (patterns (i), (ii) and (iv) in set-2 and pattern (iii) in set-5, see Graphs 7 and 10) are shown to be more frequent in the EA-STs than in the EA-PUB.



Graph 13: Percentages of the pattern We * in all sets**

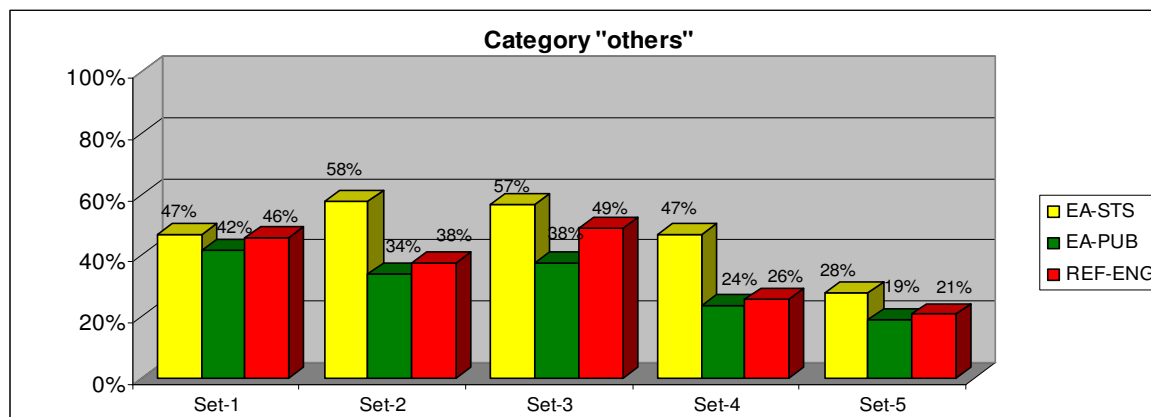
These findings seem to suggest that students display a clear preference for author anonymity. However, one cannot afford to ignore that this paper focuses on the lexical patterning of specific verbs. Generalisations on the use of the passive and active voice would require a more comprehensive syntactical analysis of a wider range of verbs. This is nevertheless an issue worth further investigation since it can offer useful insights for the development of pedagogic materials. These textual features can be used to investigate how the researcher places him/herself

towards his/her work, which may in turn mirror cultural practices and preferences. Unlike English, the use of first person pronouns is somewhat discouraged in academic Portuguese, which tends to favour impersonal writing and rely strongly on the passive voice. A contrastive study on the use of the passive and active voice in English and Portuguese abstracts would therefore be of special relevance.

For the time being, what is important to mention is that these findings are in line with Hyland (2002, 2008a, 2008b), who finds that master's dissertations, and to a lesser extent PhD theses, written by Hong Kong students exhibit a lower proportion of stance markers such as first person pronouns. Hyland (2008a) also finds a higher percentage of anticipatory-*it* phrases (*it can be seen, it should be noted that, etc.*) in students' writing in relation to published research papers. However, this is not the case in the abstracts written by Brazilian students. The pattern **it BE *** that** (pattern (vi) in set 2 and pattern (iii) in set-4, see Graphs 7 and 9) is more frequent in the EA-PUB than in the EA-STTS.

Another interesting point worth commenting on is that the percentages of the category "others" are consistently higher in the EA-STTS in comparison with the EA-PUB (Graph 14). For sets 2, 3 and 4, the difference in percentages is at least 19pp. This indicates that Brazilian students do not resort to recurring patterns as much as published abstracts do, which contradicts Hyland's (2008a, 2008b) suggestion that students tend to draw more heavily on pre-fabricated phrases. This may be explained by the composition of the EA-STTS. As explained earlier, the level of English of Brazilian students varies dramatically and many students may not be familiar with the most frequent lexical patterns in academic English. Hyland (2008a, 2008b), on the other hand, examines texts written by students from five Hong Kong universities who are taught by British and American instructors and hence would presumably have a better command of the

English language than most of the Brazilian students in question.



Graph 14: Percentages of the category “others” in all sets

Last but not least, it is worth commenting on the use of reference corpora in the present paper. As can be seen in the Data Analysis section, the REF-ENG confirms all tendencies revealed by the EA-PUB. The only exception is the pattern **results (BE) ***** within set-3 (OBTAIN, COLLECT, ACHIEVE, ATTAIN and ACQUIRE – see Graph 8), which occurs in similar proportion in the REF-ENG and the EA-STs (6% and 7% respectively) and is more frequent in the EA-PUB (13%). Thus, it turns out that, in the specific case of this study, the main contribution of the REF-ENG was to give the researcher confidence in pointing out the most fundamental differences between students and published abstracts. Similarly, the use of the REF-PTG to analyse the interference of Portuguese in the language produced by students has allowed the researcher to draw conclusions on the basis of empirical data rather than resorting to subjective judgement.

Final Remarks

This paper investigated the frequency and lexical patterning of five sets of sense-related verbs in English abstracts written by Brazilian graduate students in relation to published abstracts from

the same disciplines. Relevant differences were found between the two subcorpora in terms of both frequency percentages of individual verbs within each set and preference for specific recurring patterns. The analysis also revealed that students tend to underuse some verbs and overuse others, and this seems to be caused by the interference of the Portuguese language.

The long-term objective of this study is to incorporate these findings into course materials and computer-assisted writing tools and hence contribute to improving the quality of the academic English produced by Brazilian graduate students. By identifying differences in the lexico-grammatical patterning of abstracts written by students and published abstracts, we hope to be able to raise students' awareness of their most frequent errors as well as draw their attention to the use of chunks which are regularly used within their academic discourse community.

¹ Here, I specifically refer to *SciPo-Farmácia*, a corpus-based writing tool, developed by the Centre for Computational Linguistics (NILC) at the University of São Paulo, whose primary goal is to assist novice researchers to write scientific papers in English. This is done by providing users with extracts from authentic research papers retrieved from a reference corpus of the discipline in question. Further details at <http://www.nilc.icmc.usp.br/scipo-farmacia/>

² SMALL CAPITALS have traditionally been used to represent lemmas.

³ This is done by means of the *Keywords* feature in the software package *WordSmith Tools* (Scott 2004).

⁴ The BNC is a 100-million-word corpus of texts originally produced in English. Further information at <http://info.ox.ac.uk/bnc>.

⁵ Here, I have considered both ENCONTRAR and ACHAR as direct translations of FIND. However, it is worth mentioning that there are 71 instances of the former and only one instance of the latter.

Acknowledgements

Special thanks are due to FAPESP (Brazil) for providing the financial support needed in this research. I would also like to thank my supervisor Professor Stella Tagnin (University of São Paulo, Brazil) for her relevant comments on an earlier version of this paper.

References

Collins English Dictionary & Thesaurus (2002), Harper-Collins Publishers Ltd., Version 3.0

Software.

- Cortes, V. (2004) 'Lexical bundles in published and student disciplinary writing: examples from history and biology'. In *English for Specific Purposes*, 23:397-423.
- Dayrell, C. and S. Aluísio (2008) 'Using a comparable corpus to investigate lexical patterning in English abstracts written by non-native speakers'. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Workshop on Comparable Corpora*. Marrakech, Morocco, 31st May 2008. Available at: <http://www.lrec-conf.org/lrec2008/>
- Genoves Jr., L., R. Lizotte, E. Schuster, C. Dayrell, S. Aluísio (2007) 'A two-tiered approach to detecting English article usage: An application in scientific paper writing tools'. In *Proceedings of the International Conference RANLP'2007*. Borovetz, Bulgaria, 26th Sep 2007, pp. 225-239.
- Gilquin, G., S. Granger and M. Paquot (2007) 'Learner corpora: The missing link in EAP pedagogy'. In *Journal of English for Specific Purposes*, 6:319-335.
- Granger, S. (2002) 'A bird's-eye view of learner corpus research'. In S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam/Philadelphia: John Benjamins Publishing, pp. 3-33.
- Hewings, M. and A. Hewings (2002) 'It is interesting to note that ...': A comparative study of anticipatory 'it' in student and published writing'. In *English for specific Purposes*, 21:367-383.
- Hyland, K. (2002) 'Authority and invisibility; authorial identity in academic writing'. In *Journal of Pragmatics*, 34(8):1091-1112.
- Hyland, K. (2008a) 'Academic clusters: Text patterning in published and postgraduate writing'.

- In *International Journal of Applied Linguistics*, 18(1): 41-61.
- Hyland, K. (2008b) 'As can be seen: lexical bundles and disciplinary variation'. In *English for Specific Purposes*, 27: 4-21.
- Orasan, C. (2001) 'Patterns in scientific abstracts'. In *Proceedings of Corpus Linguistics 2001 Conference*. Lancaster: Lancaster University, pp. 433-443
- Scott, M. (2004). *WordSmith Tools* version 4. Oxford: Oxford University Press.
- Peacock, M. (2006) 'A cross-disciplinary comparison of boosting in research articles'. In *Corpora. Corpus-Based Language Learning, Language Processing and Linguistics*, 1(1): 61-84.
- Swales, J.M. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Swales, J.M. (2004) *Research Genres. Explorations and Applications*. Cambridge: Cambridge University Press.
- Swales, J.M. and C. B. Feak (2000) *English in Today's Research World: A Writing Guide*. Michigan: The University of Michigan Press.
- Thompson, P. (2001) 'Looking at citations: Using corpora in English for Academic Purposes'. In *Language Learning and Technology* 5(3):91-105.
- Weissberg, R. and S. Buker (1990). *Writing up Research: Experimental Research Report Writing for Students of English*. Englewood Cliffs (NJ): Prentice Hall Regents.

Author: Carmen Dayrell
 Department of Modern Languages - University of São Paulo (Brazil)
 Address: Rua Prof. Luciano Gualberto, 403/sala 14
 Cidade Universitária - São Paulo
 CEP: 05508-900 Brazil
 E-mail: dayrellc@gmail.com