New Language Technologies and Linguistic Research

New Language Technologies and Linguistic Research: A Two-Way Road

Edited by

Sandra Maria Aluisio and Stella E. O. Tagnin

CAMBRIDGE SCHOLARS PUBLISHING

New Language Technologies and Linguistic Research: A Two-Way Road,
Edited by Sandra Maria Aluisio and Stella E. O. Tagnin

This book first published 2014

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-5377-1, ISBN (13): 978-1-4438-5377-4

TABLE OF CONTENTS

Part V: Corpus Annotation

Part VI: Corpora and Multiple Documents

Part VII: Plenary Papers

INTRODUCTION*

This book features 12 papers on six topics addressed at the 11th Corpus Linguistics Symposium (ELC 2012), held at the *Instituto de Ciências Matemáticas e de Computação* (Institute of Mathematics and Computer Science) of the University of São Paulo, in São Carlos, state of São Paulo, Brazil. The topics are: *Corpus Linguistics and Language Description*; *Translation, Terminology and Corpora*; *Spoken Language and Corpora*; *Natural Language Processing and Corpora*; *Corpus Annotation*, and *Corpora and Multiple Documents*.

The overall theme of the conference was *Technological convergence for language processing and analysis: new technologies for linguistic research and linguistic research for new technologies*. Natural Language Processing (NLP), also known as Computational Linguistics, and Corpus Linguistics (CL) have experienced a significant development in the last decades, particularly in Europe and the United States, mainly for the English language. In Brazil, despite considerable advances in the last decade, such areas are not yet widespread and are restricted to just a few universities, often as subdivisions of broader areas such as Computing or Linguistics. With the proposed convergence we hope to call attention to the importance of activities that arise under the broader scope of NLP with CL. For this reason, ELC 2012 intended to bring together researchers in Linguistics, Computer Science, Historical Linguistics, Applied Linguistics, Cognitive Linguistics, Information Sciences, and other Corpus Linguistics-related fields to submit studies that have been completed or are in progress in that multidisciplinary area.

ELC 2012 took place right after the 6th Brazilian School of Computational Linguistics which opened with a plenary talk on speech synthesis and recognition, by Sara Candeias. Two more plenary sessions were held at the 11th ELC, given by Diana Santos and Nancy Ide, along with a roundtable on *Speech recognition and synthesis for Portuguese and interactions between linguists and engineers: experiences and*

* We would like to thank Danilo Murakami, from the Department of Modern Languages at the University of São Paulo, for helping us put this book together. He revised every single article to make sure they all conformed to the formatting guidelines.

opportunities. Two papers, described below, give an overview of the topics dealt with in these sessions.

The first paper, in this collection, on the topic *Corpus Linguistics and Language Description* is by Deise Prina Dutra, Barbara Malveira Orfano and Tony Berber Sardinha. The authors present the analysis of four-word bundles extracted from native and non-native corpora of argumentative essays. In addition, the study highlights the differences among the corpora as far as stance expressions are concerned and to detect if these differences are mainly structural or related to frequency within a specific function. As a final aim the paper discusses the use of stance expression bundles in academic writing. Such bundles are important as they carry the opinion or judgment of the writer or speaker.

The second paper is by Sandra Navarro and addresses the topic *Translation, Terminology and Corpora*. The author describes the process of building a bilingual English-Portuguese corpus for the hotel industry and extracting equivalents in both languages. She discusses four collocations with the word "room" and shows how collocational differences affect the translation of the terminology. These differences are highlighted in the entries presented to constitute the glossary.

The next two papers are by Maryualê Malvessi Mittmann, Tommaso Raso, Adriellen Arruda and by Élida Gama Chaves and Vera Lúcia Santiago Araújo. Both papers address *Spoken Language and Corpora*. The first paper presents a cross-linguistic study on the usage of dialogic units in spoken Brazilian Portuguese and Italian. It aims to show the distribution of information units in both languages and to discuss some interesting aspects regarding the usage of dialogic units in Brazilian and Italian. The second paper puts forward a proposal of customized tags to investigate the segmentation in intralingual Brazilian Portuguese subtitles for the deaf and the hard-of-hearing.

The fifth paper is by Leonardo Zilio, Adriano Zanette and Carolina Scarton on the topic *Natural Language Processing and Corpora*. The authors present a system to extract subcategorization frames (SCFs) from corpora written in Portuguese and describe how the system is used in a verb classification task and in the labeling of semantic roles.

The following three papers deal with *Corpus Annotation*. The first one, by Lianet Sepúlveda Torres, Roana Rodrigues and Sandra Maria Aluísio presents the compilation of a corpus composed of texts written in Portuguese by Spanish speakers enrolled in Brazilian graduate programs. The purpose is to formalize a typology of the main errors committed by these speakers when writing theses and dissertations in Portuguese. The second paper, by Rove Chishman, Anderson Bertoldi, João Gabriel

Padilha, and Diego Spader de Souza, discusses the challenges encountered in annotating a Brazilian Portuguese corpus of soccer language with semantic frames. The third paper in this group, by Cláudia Freitas, Eduardo Motta, Ruy Luiz Milidiú and Juliana César, describes the construction of ReLi – a corpus of book reviews manually annotated with respect to the expression of opinion. Specifically, it reports on the decisions made during the annotation process based on the results of an inter-annotator agreement study, and briefly explores the corpus created.

The following papers address *Corpora and Multiple Documents*. The first one, by Renata T. Camargo, Verônica Agostini, Ariani Di Felippo, and Thiago A. S. Pardo, discusses Multi-Document Summarization (MDS), which involves the production of a single summary from a group of texts that deal with the same subject. The paper mainly addresses the alignment of source texts to their human (manually produced) multi-document summaries, which allows a linguistic analysis of human summarization strategies that may subsidize the creation of rules and models for MDS methods that are more linguistically motivated. The second one, by Di Felippo et al., aims to contribute to the linguistic characterization of summaries. For that purpose, the group developed a corpus-based analysis of aspects conveyed by human multi-document summaries. The corpus used was the CSTNews corpus with texts from various online Brazilian news agencies. Each human summary of the corpus was manually annotated with 17 aspects resulting from the refinement of a previous set of aspects.

And last, but by no means least, there are two papers by our invited speakers Diana Santos and Sara Candeias. Santos presents a personal view of quantitative, and especially statistical, approaches to language study, filling a pedagogical gap in this area. In addition, she introduces a new grammar of Portuguese, which is being developed at Linguateca, based on the AC/DC (Access to corpora /Availability of corpora) project. Candeias discusses the role of language theory for designing and developing applications in speech technology, providing an overview of the typical architecture of speech technology systems (recognizers and synthesizers). The author also introduces the core application areas in speech technology, which are raising new challenges for linguistic research.

We hope this collection will show that New Language Technologies and Linguistic Research really constitute a two-way road.

Sandra M. Aluísio and Stella E. O. Tagnin
Organizers

PART I

CORPUS LINGUISTICS AND LANGUAGE DESCRIPTION

CHAPTER ONE

STANCE BUNDLES IN LEARNER CORPORA

DEISE PRINA DUTRA[†],
BARBARA MALVEIRA ORFANO[‡]
AND TONY BERBER SARDINHA[§]

1. Introduction

This article is part of a broader research project, which aims at analyzing in detail the many functions and uses of lexical bundles (Biber et al. 1999 et seq.) from quantitative and qualitative views. "Lexical bundles are combinations of three or more words which are identified empirically in a corpus of natural language" (Cortes 2006, p. 392). The identification of words that 'go together', in a corpus linguistic perspective, has been addressed under different terms, such as fixed collocations, extended collocations, cue phrases, clusters, and ngrams, to name a few. As Cortes (2006) notes, lexical bundles might be related to proficient and fluent language production (e.g. Hyland 2008a,b).

As our understanding of lexical bundles increases, so does the need for more research on how they operate in different discourse contexts, such as in second language (L2) writing. In this article, we look at stance lexical bundles in learner corpora, and contrast their frequency and functional associations with native speaker writing. It is organized as follows. First, we review some of the studies in the next paragraphs (Section 1.1). Secondly, we introduce the corpora and methods used, then we present the findings, and draw the paper to a close with considerations on the role of lexical bundles in shaping L2 writing.

[†] UFMG – Federal University of Minas Gerais.
[‡] UFSJ – Federal University of São João Del Rey.
[§] PUC-SP – Catholic University at São Paulo.

1.1. Lexical bundles in corpus research

Lexical bundles are defined as "simply sequences of word forms that commonly go together in natural discourse" (Biber et al. 1999, p. 990). They are generally not structurally complete[1], and perform different discourse functions, the main of which are referential, stance and discourse organizing (Biber et al. 2004). Above all, they act as building blocks of discourse (Biber et al. 2004), and, therefore, "serve the most important communicative needs of a register" (Biber 2009, p. 285).

Hyland (2008a) explores the structure and function of 4-word bundles in a corpus of academic discourse. The data for his study consisted of three corpora (research articles, doctoral dissertations and master's theses) comprising 3.5 million words. The results show significant differences in frequencies across registers, for instance fewer lexical bundles in doctoral theses than in articles; surprisingly, there are more bundles in master's theses, which might be explained by a number of possible reasons, such as the possibility that novice writers might rely on formulaic expressions more than more experienced ones due to a restricted vocabulary, or that these novice writers might incorporate more fixed expressions with the intent of being more readily accepted in the community. In a similar direction and adopting a frequency-driven approach, Chen and Baker (2010) identify the most frequent lexical bundles in three written corpora: 1) a sub-corpus from FLOB (academic prose section), 2) BAWE-CH (Chinese students of English) and 3) BAWE-EN (English students). Their comparative study shows both differences and similarities between native and learner academic writing. The use of lexical bundles in non-native and native student essays, for example, is very similar from a structural point of view. They both have more VP-based bundles and discourse organizers than native expert writing, whereas native professional writers exhibit a wider range of NP-based bundles and referential markers.

Following a pragmatic-functional approach Simpson-Vlach and Ellis (2010) also looked at the most common lexical bundles in academic discourse in both oral and written corpora. Their data comprised the Michigan Corpus of Academic Spoken English (MICASE), the oral academic part of the British National Corpus (BNC), the Hyland 2004

[1] This lexical bundles characteristic has been recently challenged in Cortes (2013: 38) as she found bundles that are "complete structures, complete clauses, and sometimes even sentences. That is why the following new category was created:
d. Lexical bundles that include noun phrases and verb phrases (fragments or whole phrases or clauses) in bundles such as *the rest of the paper is organized as follows*, and *the objective of this study was to evaluate*."

corpus and the written BNC files of various academic subjects. They extracted 3 and 4-word *n-grams* and had ESP instructors judge if those were chunks, meaning in that study that the n-gram was associated with a particular function and was worth teaching. As a result, they proposed the Academic Formulas List (AFL) with 435 lexical bundles distributed in 18 subcategories (see Table 1 with its subcategories and examples). Since it attempts to cover the most frequent lexical bundles in academic written discourse, the Simpson-Vlach and Ellis (2010) list has contributed significantly to our research, serving as a coding tool and as a basis for the analysis.

Table 1: Functional classifications of lexical bundles based on Simpson-Vlach and Ellis (2010)

Group A – Referential Expressions	***Examples***
1. specification of attributes	
1.a intangible framing attributes	*form of the; in terms of a*
1.b tangible framing attributes	*a list of; both of these*
1.c quantity specification	*a high degree; a large number of*
2. identification and focus	*[an/the] example of (a); does not have; there has been*
3. contrast and comparison	*be related to; (on) the other (hand) (the)*
4. deictics and locatives	*at this point; at the time of*
5. vagueness markers	*and so on*
Group B – Stance expressions	***Examples***
1. hedges	*are likely to; it appears that*
2. epistemic stance	*we can see; assumed to be*
3. obligation and directive	*(it should) be noted; take into account (the)*
4. expressions of ability and possibility	*allows us to; can also be, most likely to*
5. evaluation	*important role in; it is necessary (to)*
6. intention/volition, prediction	*to do so; we do not*
Group C - Discourse organizing expressions	***Examples***
1. metadiscourse and textual reference	*as shown in; in the present study*
2. topic introduction and focus	*for example (if/in/the); what are the*

3. topic elaboration	
3.a non-causal	*are as follows; see for example*
3.b cause and effect	*as a consequence; for this reason*
4. discourse markers	*in other words; even though the*

In Dutra & Berber-Sardinha (2013: 121) we presented the general counts of the three broad categories (referential, stance and discourse organizing expressions) across three corpora (ICLE – the International Corpus of Learner English, Br-ICLE – the Brazilian subcorpus of ICLE, and LOCNESS - Louvain Corpus of Native English Essays), which is reproduced below (Figure 1), and a summary of the statistical analysis (Chi-square) (Table 2). These results came up after the application of the frequency cut-off point of 20 times per million words (pmw), which yielded a total of 676 four-word bundle types which remained in the data, across the three corpora. The category with the highest count of bundles for all three corpora are referential expressions, followed by stance expressions, and discourse organizers (Figure 1). There are statistical differences across the corpora (Chi-square 17.126, df=4, p=0.002).

Similar results have been found in Biber et al. (2004), but in Chen & Baker (2010) the discourse organizing bundles were the second highest group of frequent bundles. This difference may be due to the fact that Biber et al. (2004) used a native speaker corpus, the T2K-SWAL, containing over 2 million words from oral and written registers (classroom teaching and textbooks). On the other hand, Chen & Baker (2010)'s corpus included a mixture of expert native speaker, native speaker student, and non-native learner data. Another reason that might explain this difference in stance expression frequency is the fact that argumentative essays may require more expressions of opinion (stance bundles) while expert academic texts, such as the ones in the academic prose section of the Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus, might rely less on stance bundles and more on referential ones. Other differences have been found in Hyland (2008b), in which stance bundles were the least frequent across the four disciplines corpora investigated (Biology, Electrical Engineering, Applied Linguistics and Business Studies)[2]. These studies suggest that the relative frequency of functional associations is sensitive to the genre of texts included in the corpora, and consequently the rank of

[2] This paper, which investigates theses, dissertations and research articles, uses a different terminology from the one adopted in our paper and stance bundles are called participant-oriented bundles.

functional categories in a particular study cannot be predicted ahead of time.

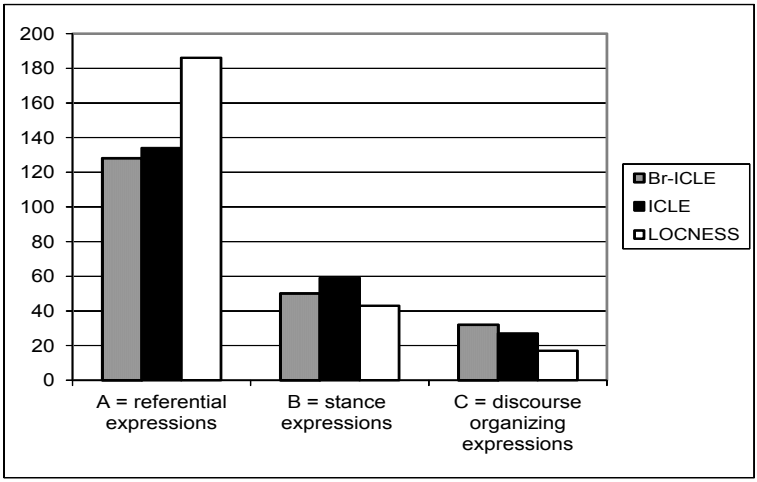

Fig. 1. Lexical bundle category count by corpus.

Table 2: Chi-square test – Category count by corpus

	Value	Df	Asymp.Sig. (2-sided)
Pearson Chi-Square	17.126	4	0.002
Likelihood Ratio	17.508	4	0.002
N of Valid Cases	676		

1.2. Research goals

The main goal of this paper is to show the relevance of analyzing and contrasting types of stance bundles produced by native and non-native speakers in argumentative essays. It is believed that the discussion encompassing the frequency of lexical bundle types that appear under the category "stance expressions" can bring interesting insights to what lies behind quantitative and qualitative corpus studies. In addition, as a secondary goal, we investigate if the similarities or differences in the frequency of stance bundles are mainly structural or functional. The final goal of this study is to discuss the role of stance expression bundles in L2 academic writing.

2. Materials and Methods

This study comprises three corpora of argumentative essays: a) LOCNESS (Louvain Corpus of Native English Essays), which contains 324,006 words written by American and British university students; b) ICLE, the International Corpus of Learner English containing 3.7 million words – this corpus is comprised of 16 subcorpora of written essays by international students from many different countries (Japan, China, Italy, Finland and etc.; for a full description, see Granger et al. (2009)); and c) Br-ICLE, which has approximately 159,000 words of academic argumentative essays written by Brazilian university students, and which does not form part of the consolidated ICLE distribution yet.

Having described the data used in this research, we proceed to the methodological framework. First, sequences of four words were extracted from each corpus with scripts specially developed for our research project. Only bundles that occurred in five different essays were included in our list. These sequences were categorized both manually and semi-automatically according to the AFL framework, which includes its three major categories, namely referential expressions, stance expressions and discourse organizing functions, and also according to its 18 different subcategories. Secondly, the sequences were counted, their frequencies normed, and those sequences that occurred at least 20 times pmw were selected for further inspection. Finally, frequency comparisons across corpora and categories were carried out to determine which categories were predominant in the data. As part of our broader investigation we have extracted lexical bundles of 3, 4 and 5 words. In this paper we focus on the analysis of 4-word bundles, as a follow-up on our previous work (Dutra & Berber Sardinha 2013). In this research, we use a conservative cut-off point and focus on bundles that occurred at least 20 times pmw (see Cortes 2008).

3. Results

In a previous paper (Dutra & Berber Sardinha 2013) we argued that referential expression bundles have a crucial role in how learners present their argumentation. In this article, we turn to stance expression bundles since it is also relevant to investigate how judgments and opinions are conveyed in the three corpora, contributing to a better understanding of how learners express themselves in the academic register. Sample 1 shows an example of the presence of stance expressions in Br-ICLE (all samples

are shown exactly as they were produced by the students, without proof-reading).

Sample 1

Unfortunatelly, we live in a society where a lot of people still think that the certificate is more important than all others things in the life. I don't think that the study is not important but I think that it is not vital.
Today, there are a lot of Universities graduating people, but there are not enough work to every these people. The graduation will not help people to get a work. There are a lot of people that know how to embroider, to sew, people that work in the field planting our food. There are millions of people that know to do many kinds of works that are very important to our life even not being in the universities' curriculum. <ICLE-BR-FMG-0012.1>

In sample 1, the learner uses *I + think*, a very common construction employed by users of English to express opinion[3]. We shall return to this construction in more detail later in this section. After running Fisher's Exact test on the count of different bundles for the stance expression subcategories: hedges, epistemic stance, obligation/directives, ability/possibility, evaluation, intention/prediction/volition , no statistically significant difference was found (Data from Table 2 and p = 0.09365). Yet, it seemed to us that it was worth to closely investigate the choices made by the users (native and non-native speakers) as the bundles classified in the stance subcategories were not always the same in structure which affect the meaning they convey. Table 3 shows a breakdown of the frequencies by stance type within each corpus.

Table 3: Stance bundle type normed frequency[4]

	LOCNESS	**ICLE**	**Br-ICLE**
Hedges	9.26	0.26	6.28
Epistemic stance	33.95	3.98	75.38
Obligation and directives	27.77	2.38	87.94
Expressions of ability and possibility	18.52	2.65	62.82

[3] Simpson-Vlach and Ellis (2010) included the bundle *I think this is* in their Academic Formulas List as it appeared as a frequent epistemic stance bundle in primarily academic spoken discourse.
[4] The frequency counts refer to bundle types in each subcategory.

Evaluation	37.04	5.57	75.38
Intention, volition and prediction	6.17	0.79	31.41

Despite the lack of significant statistical difference across corpora in terms of the stance bundle types, there may be relevant differences if these bundles are qualitatively analyzed. Due to space restrictions we will focus our discussion on three stance sub-categories, namely hedge, epistemic stance, and obligation & directives. The following subsections present, first, a description of the stance bundle structure in the data. Secondly, we discuss register adequacy in the learner corpora (mainly in Br-ICLE) from the point of view of argumentation structure. Finally, we argue that analyses that consider bundles from a qualitative standpoint as well have advantages over studies that are limited to a purely quantitive approach to bundle use.

3.1. Stance Bundle Structure

We analyzed all instances of the stance lexical bundles in all three corpora and established a cut-off point of at least 20 times pmw words for a bundle to be on our shortlist for further consideration. We categorized these shortlisted bundles structurally as shown in Table 4. These patterns are neither complete structural units nor idiomatic expressions (Biber et al., 1999). In five out of the six bundle types there is a verb, mainly a VP with a modal (e.g. *would have to be*). Two of these verbal patterns include a pronoun: antecipatory *it* + VP/AdjP (*it should not be*) and (NP)+ VP + (*that*- or *to*-clause) (e.g. *think that it is*, *we need to be*). Prepositional phrases and noun phrases (e.g. *to a certain extent*, *my point of view*) also appear, even though less frequently.

Table 4: Bundle structure

Bundle structure	Example
(Preposition) + NP	*to a certain extent*
Passive	*can be seen to*
(NP)+ VP + (*that*- or *to*-clause)	*think that it is*, *we need to be*
VP (Modal + V)	*would have to be*
Copula *be* + NP or AdjP	*is a kind of*
Antecipatory *it* + VP/AdjP	*it should not be*

3.2. Pragmatic functional subcategories

To fully understand the choice of bundles to express stance meanings, we turn now to how the structures presented in Table 4 are distributed among the following subcategories: hedge, epistemic stance, and obligation & directives, across the three corpora (Table 5).

Table 5: Bundle examples (subcategories by corpus)

	LOCNESS	ICLE	Br-ICLE
Hedge	*to a certain extent* *could be used to* *can be seen to*	*is a kind of*	*is a kind of*
Epistemic stance	*is shown to be* *I think that the* *I feel that the* *can be seen as* *is seen to be*	*I think it is* *I do not think* *I think that the* *my point of view* *seems to be a*	*it has been argued that some people think that think that it is* *my point of view*
Obligation and directives	*would have to be* *it should not be* *should be able to* *should not be allowed* *should be allowed to*	*do not want to* *they do not have to* *think that it is* *do not have to* *should be able to*	*what they want to* *you do not have* *we need to be* *do not need to*

The results in Table 5 show that learners (non-native speakers –NNSs) typically choose the bundle *is a kind of* as a hedging device, which indicates that their repertoire for this function might be very limited, compared to the range of bundles put in play by the native speaker students (LOCNESS). This bundle is formed by one of the most common stance adverbials in English (Biber et al., 1999), *a kind of*. Biber et al. (1991) note that this stance adverbial is more frequent in conversation, being a marker of imprecision mainly preferred in American English.[5] The overuse of the lexical bundle *is a kind of* both in ICLE and Br-ICLE might signal some form of 'mode crossover' in learner discourse, that is, the use

[5] "… BrE conversation shows a preference for *sort of."* (Biber et al., 1999: 867)

in writing of bundles most commonly associated with speech. Sample 2 illustrates this use in a text written by a Brazilian student.

Sample 2

Beliefs like these are part of a great range of concepts which comes in and goes out throughout minds that constitute our society. Sexism ***is a kind of*** *prejudice which is neither too weak to be ignored nor too strong to complaint about it. Prejudice does not favor women or men, both just have to deal with the limitations it puts in their lives.* <ICLE-BR-CAT-0059.1>

While modal verbs such as *can* and *could* are chosen by native speakers to convey a hedging function, non-native speakers' overuse of the bundle *is a kind of* is an example of learners' restricted linguistic hedging choices. The LOCNESS results, however, indicate that native speakers use three different hedging bundles with high frequency (*to a certain extent*, *could be used to, can be seen to*) (Table 5). This choice of bundle gives native speakers the possibility of varying the way they present an opinion or judgment.

Sample 3, from LOCNESS, gives an example of how native speakers (NSs) hedge their arguments with a modal (*could*) while making a claim about a particular topic.

Sample 3

This treatment is costly. On the NHS, treatment costs ***could be used to*** *fund more life-saving treatments rather than to bring more children into an overpopulated world. There are many children who are unwanted and who need adopting.* <ICLE-ALEV-0030.8>

In LOCNESS, another hedging lexical bundle is *to a certain extent*, which combines an adjective expressing likelihood (*certain*) with a stance noun (*extent*) (see sample 4). A similar lexical bundle in form and function (*to some extent*) has been reported by Simpson-Vlach and Ellis (2010) as being frequent in both written and oral academic discourse, and is therefore included in the Academic Formulas List (AFL).

Sample 4

With health experts, however, our marvellous machine provokes problems to our health which must not be ignored. Apart from being a necessity to some, it makes the human being lazy ***to a certain extent*** *and thus, allows us to use our brains even less and less. It is true to say that a computer is*

able to perform many hundreds of useful tasks which our brains are completely indifferent to. <ICLE-ALEV-0013.9>

While hedging lexical bundles are more frequent in LOCNESS, expressions that convey an obligation or an intention of being more assertive are more frequent in Br-ICLE and ICLE (Figure 2 and Tables 3 and 5). Assertiveness can be seen as the opposite in meaning to hedgings; the use of assertive bundles such as *they do not have to*, *we do not need* is greater than that of the hedge *is a kind of*, a frequent choice in the NNS corpora. The difference in the use of hedges in the three corpora suggests that native speakers might make use of more 'inexplicit language', here characterized by hedges, which, in turn, lends more modalization to their assertions. Even when NSs choose to be more assertive, they do so by using bundles such as *would have to be*, *should not be*, *should be able to* and *should not be allowed,* all of which incorporate a modal (*should*, *would*), which in turn are less assertive than the periphrastic modals (*have* to, *need* to) preferred by the NNSs. In addition, the passive construction is common in LOCNESS as a means of adding impersonality to the obligation and directive bundles. Conversely, NNSs seem to lack this ability and express their arguments in different ways. The results in Table 3 show that the Br-ICLE students use obligation markers more than twice as often as NSs. This overuse of directives by learners shapes the way their writings are interpreted by readers in general. The apparent excess of obligation might be seen as odd in written academic registers, since those are usually associated with a less assertive style.

3.3. Going beyond frequency

The discussion presented in the previous section shows how a corpus can reveal the different ways in which particular groups present arguments: NSs seem to have more types of hedges at their disposal, and they also use them more frequently than learners. The stance lexical bundles in LOCNESS are more polite and impersonal (e.g. with anticipatory *it* + passive). On the other hand, the Brazilian students display an overuse of obligation and directive bundles. The native speakers use bundles in similar functions, yet the proportion is half of that of the Brazilians, and the type structures are again more impersonal (passive + anticipatory *it*).

Looking closely at all three corpora, personal pronouns (*I,* we) appear in lexical bundles in all three corpora analysed. Both personal and

possessive pronouns (*I*, *my*) are characteristic features of spoken discourse, revealing personal engagement of the participants[6]. Personal pronouns such as *I* and *you* are usually in the first top ten words in most spoken corpora. Combined with the choice of using the periphrastic modal verb *have to*, this demonstrates the inclusion of oral features in literate discourse, especially in the writing of Brazilian students. This is due to the fact that spoken communication implies a more dyadic interaction between participants which necessitates frequent use of first and second person pronouns. It seems that learners attempt to replicate this engagement in writing by making use of characteristics from oral communication, such as through the bundle *I think that it is*. The choice of this bundle to express epistemic stance instead of bundles that include the passive voice (e.g. *can be seen as*) might suggest that learners do not possess a wide range of forms to perform this function, or that hedging is not common in academic writing in their mother tongue, among other reasons.

4. Final remarks

The analysis of the types of stance bundles across the three corpora shows that the Brazilian student corpus presents a less diverse use of these bundles than ICLE or LOCNESS, especially with hedge bundles. The qualitative analysis reveals that Br-ICLE and ICLE bundles are more personal and are formed less frequently with antecipatory *it* and passive structures. There seems to be an overuse of directive and obligation bundles in the Br-ICLE corpus, which may mark the texts as peremptory.

We would like to see more qualitative analyses in lexical bundle research to complement the more mainstream quantitative analyses. If we had only considered the statistically significant differences across the corpora, we would have probably missed out on both the influence of oral discourse in written stance bundles, and the choice of less complex yet more personal bundles in the learner corpora.

[6] Although we claim that there might be influence of oral features into the Brazilian learners' written discourse, there may be other explanations for their choices. For instance, there is a tendency, recently, for scientific article writers to stand up for their results and, therefore, use personal pronouns (e.g. *we*) (Dayrell and Candido Jr, 2013).

References

Biber , D. 2009, A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3), 275-311.

Biber, D., Conrad, S. & Cortes, V. 2004, *If you look at...* Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999, *Longman Grammar of Spoken and Written English*. Essex: Longman.

Chen, Y. & Baker, P. 2010, Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, v.14, n.2, 30-49.

Cortes, V. 2004, Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, v. 23, 397-423.

—. 2006, Teaching lexical bundles in the disciplines: An example from a writing intensive history class. Linguistics and Education, v. 17, 391–406.

—. 2008, A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1), 43-57.

—. 2013, The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, v. 12, 33–43.

Dayrell, C.; Candido JR., A. 2013, Textual patterns and rethorical moves in English scientific abstracts: comparing student and published writings. Paper presented at Learner Corpus Research 2013. Bergen, Norway.

Dutra, D. P. & Berber Sardinha, T. 2013, Referential expressions in English learner argumentative writing. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 117-127.

Granger, S. et al. 2009, *International Corpus of Learner English*: *Version 2*. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Hyland, K. 2008a, Academic clusters: text patterning in published and Postgraduate writing. *International Journal of Applied Linguistics*, v.18, 41-62.

—. 2008b, As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes* v.27, p. 4-21.

O'Keeffe, A., McCarthy, M. & Carter, R. 2007, *From Corpus to Classroom*: *Language Use and Language Teaching*. Cambridge: Cambridge.

Simpson-Vlach, R. & Ellis, N. 2010, An academic formulas list: New methods in phraseology research. *Applied Linguistics*, v. 31, n.4, p. 487–51.

PART II

TRANSLATION, TERMINOLOGY AND CORPORA

CHAPTER TWO

A BILINGUAL GLOSSARY OF COLLOCATIONS TYPICAL OF THE HOTEL INDUSTRY: A MODEL IN LIGHT OF CORPUS LINGUISTICS

SANDRA NAVARRO*

1. Introduction

Tourism is a major economic and cultural sector around the world. Fueled largely by international travels, this field has great impact on several other sectors, especially the hotel industry. If back in the early days people stopped by on roadside homes seeking shelter, today the purpose of many trips is to stay in luxury hotels, which bring together all sorts of entertainment. Therefore, the hotel industry has evolved considerably over the years and has become increasingly multicultural. To mediate the relations of individuals of different nationalities and cultures, communication, usually in English, plays a crucial role and translation is a constant need.

One of the major sources of demand for translation in this area is the Internet. In order to cater to a larger audience, many hotel websites, booking websites, travel guides, among others, choose to translate their content into different languages. At this point, the translator working in this specialized field faces numerous challenges.

One of them concerns the close relationship the hotel industry holds with cultural aspects of each country and region. In the words of experts, "Tourism, besides being an economic activity, is culture in its essence. Hotels lie within this context" (Gregson, 2009: IX). Indeed, hotel descriptions encompass topics as varied as cuisine, architecture, décor, sports, entertainment, different types of establishments and services and even historical and geographic features. Therefore, translating hotel texts

* University of São Paulo.

is an attempt to bring closer together two distinct realities, two culturally different worlds.

Despite this difficulty and the great demand for translations, the area lacks terminological materials, especially bilingual English-Portuguese publications. This fact could be observed over two years of my personal experience as a translator of hotel websites. During that period, my team and I did not make use of any bilingual dictionary simply because the publications available did not meet our needs and were completely outdated. In fact, none of the publications available is targeted exclusively at translators and consequently does not help this professional in his fundamental task: to produce a technically accurate and natural translation that can be read as an original text.

Hence, in order to meet the needs of a text producer, reference materials should go beyond a list of terms and their equivalents. For the professional translator, it is important to know how terms and surrounding words function in the target language. Therefore, translators would benefit from examples of usage, suggestions of equivalents and translation solutions, information about the typical ways of expression in a given context, the frequency of equivalents and cultural information, to name a few options. This scenario gave rise to my desire to narrow the gap between the high demand for translations in this field and the shortage of specialized reference materials to support the challenging task of translators. To this end, I have conducted a master's research (Navarro, 2011) whose objective was to propose a model for a bilingual unidirectional glossary (English - Portuguese) of collocations typical of the hotel industry, aimed at the translator. This paper presents some of the results from this recently completed research. First, in section 2, we briefly outline the theoretical background that guided the research. Section 3 presents the study corpus and summarizes the procedures to identify the collocations and their equivalents. Section 4 discusses the results and presents glossary entries for some collocations. The paper concludes with final remarks, in section 5, about the main findings of this study.

2. Theory

The theoretical framework of this study comprises three related fields: Corpus Linguistics (Halliday, 1991; Sinclair, 1991), Terminology (Hoffmann, 1999; Krieger and Finatto, 2004) and Translation (Chesterman, 1998). These areas provided subsidies for a linguistic research with an empirical approach, focused on the observation of language in use.

The empiricist approach underpins Corpus Linguistics (CL), a method of linguistic investigation based on the analysis of a collection of authentic texts in a corpus. CL considers language a probabilistic system, whereby although there are a number of possible lexical choices and combinations, they do not occur with the same frequency. In other words, some lexical combinations are more recurrent than others. Thus we can say that language is conventionalized, following a phraseological tendency, or the "idiom principle" as put forward by Sinclair (1991), and this view of language gives rise to collocations.

For the purpose of this study, a collocation is the recurrent association of lexical items (Sinclair, 1991). Even though collocations do not necessarily represent a comprehension problem, they can certainly pose a challenge to translators. Take the case of "complimentary breakfast", which could be literally translated into Portuguese as "*café da manhã de cortesia*", but according to our corpus research, "*café da manhã incluído na diária*" [breakfast included in the daily rate] is the most common equivalent. Moreover, studies demonstrate that collocations constitute approximately 70% of terminological occurrences in specialized languages (Krieger and Finatto, 2004: 81); hence the importance of developing terminology reference materials with a special focus on collocations.

Regarding Terminology, this study adopts a descriptive approach, in line with the principles of socio-communicative theories (Krieger and Finatto, 2004: 78), such as Textual Terminology (Hoffmann, 1988). This *in vivo* approach acknowledges the fundamental role of context in terminological research, represented by the texts that make up the corpus.

Corpus-based translation studies have influenced the notion of equivalence, a crucial element for this study. In this sense, we adopt the notion of context equivalence put forward by Chesterman (1998: 31). Under this view, equivalents must fulfill the same role, within the same context, in both source and target languages. This perspective expands the concept of equivalence beyond the formula "word x = word y", reflecting more accurately the complexity of terminological correspondence. It is worth mentioning again the example of "complimentary breakfast", whose functional equivalent in Portuguese is "*café da manhã incluído na diária*" [breakfast included in the daily rate] instead of the literal translation "*café da manhã de cortesia*".

3. Methodology

To carry out this study, I have compiled a comparable corpus (texts written originally in English and in Portuguese) of texts taken from