

**Language and learner corpora in ELT: reviewing the impact of Corpus Linguistics  
in the teaching of English**

**Corpora linguísticos e de aprendizes no ensino da língua inglesa: uma discussão do  
impacto da Linguística de Corpus no ensino de inglês**

Elisa Mattos de Sá<sup>1</sup>

Leonardo Pereira Nunes<sup>2</sup>

***Abstract:** This paper discusses the impact language and learner corpora have had in English Language Teaching (ELT). Corpus Linguistics can be defined as the study of language through compilation of authentically produced texts which undergo automatic or semi-automatic analyses. Over the past decade Corpus Linguistics studies have been increasingly adopted in ELT, ranging from applications in teaching materials (O’Keeffe, McCarthy and Carter, 2007) to understanding how learners and non-native speakers use English (Granger, 2003, 1998). Language and learner corpora, both spoken and written, have shed light into how the English language is actually used around the world, and are consequently changing the way we teach English. In this paper we will show how such corpora are influencing ELT, in general, and more specifically in Brazil, as this is our research and teaching context.*

***Keywords:** Language Corpora; Learner Corpora; English Language Teaching.*

***Resumo:** Este artigo discute o impacto dos corpora linguísticos e de aprendizes no ensino de inglês. A Linguística de Corpus pode ser definida como o estudo da linguagem por meio da compilação de textos produzidos autenticamente e analisados automática ou semi-automaticamente. Na última década, os estudos provenientes da Linguística de Corpus têm sido cada vez mais adotados no ensino de inglês, indo das aplicações nos materiais didáticos (O’Keeffe, McCarthy and Carter, 2007) ao entendimento sobre como aprendizes e falantes não nativos usam a língua (Granger, 2003, 1998). Corpora linguísticos e de aprendizes, falados e escritos, têm revelado como a língua inglesa é usada espontaneamente no mundo, e estão, desse modo, mudando a forma como ensinamos inglês. Neste artigo, mostraremos como esses corpora estão influenciando o ensino de inglês, em geral, e mais especificamente no Brasil, nosso contexto de ensino e de pesquisa.*

***Palavras-chave:** Corpora Linguísticos; Corpora de Aprendizes; Ensino de Inglês.*

---

<sup>1</sup> Especialista em Ensino de Inglês pela Universidade Federal de Minas Gerais (UFMG) e colaboradora do projeto LINDSEI-BR na mesma instituição. Em preparação para o mestrado.

<sup>2</sup> Doutorando em Estudos Linguísticos na Universidade Federal de Minas Gerais (UFMG) e mestre em Linguística Aplicada pela mesma instituição.

## 1 Introduction

Over the last ten years, the interface between corpora-based studies and English Language Teaching (ELT) has been a growing trend. Teachers and ELT scholars have acknowledged the importance of deploying didactic materials entailing spontaneously produced language by native speakers of English (language corpora). Also, such authentic language produced by foreign and/or second language learners of English (learner corpora) around the globe has been likewise under scrutiny so that a more comprehensive grasp of linguistic features can be reached and in turn be used for teaching and research purposes.

Considering such scientific and pedagogical rationale, this paper addresses how language and learner corpora have impacted ELT both in the global and local (Brazilian) teaching and research contexts.

A background on Corpus Linguistics as a methodological tool will be firstly presented, and will be followed by an overall picture of the impact of language and learner corpora in ELT in the middle sections. Finally, remarks will be made reinforcing the importance of such studies in English teaching and learning.

## 2 Corpus Linguistics

Corpus Linguistics can be defined as a methodological framework which involves the investigation of spontaneously produced texts based on a specific set of principles. It aims at exploring linguistic elements of a certain language or language variety through empirical evidence found in a *corpus* (or *corpora* in the Latin plural), a group of texts compiled on the grounds of well-established linguistic criteria (McEnery & Wilson, 2001; Sardinha, 2004). As to modern Linguistics, a corpus is only deemed as such if linguistic data in it can be both retrieved and manipulated by electronic means, i.e. in machine-readable format (McEnery & Wilson, 2001).

Corpora can be of numerous types and their categorization depends on compilation criteria, which vary according to the linguistic features to be queried. Sardinha (2004, p. 20) states that such features include: *mode*, *time dimension*, *selection*, *content*, *authorship*, *internal layout* and *purpose*. Concerning *mode*, corpora may represent either written or spoken language; as to *time*, they can be synchronic (covering a certain period in time), diachronic (dealing with a wider range of periods), contemporary (representing present times) or historic (addressing past times); regarding *selection*, they may be either static (containing a limited set of textual portions), dynamic (allowing for size expansion or reduction) or balanced (in which texts are equally distributed according to genre or type, for instance); as to *content*, they may be specialized (covering specific genres and registers), regional or dialectal (pertaining to one or

more sociolinguistic varieties) or multilingual (covering two or more languages); *authorship* entails either language produced by native speakers or by learners of a foreign and/or second language; as to *internal layout*, corpora can be comparable and aligned (original and translated texts); lastly, when it comes to *purpose*, corpora may either be for study (of a specific feature), for contrast (as a reference to another corpus) and for training (testing tools and applications for linguistic analyses).

In order to be a representative body of language (in which data concerning a specific element can be generalized), a corpus is to be designed and compiled based on parameters regarding balance, sampling and size.

A balanced corpus should cover a broad scope of text categories supposedly representative of a language or language variety. Such categories may be based on text typology, in which texts are classified and characterized according to standards such as subject field, period of production, medium of production and level of formality<sup>3</sup> (McEnery, Xiao and Tono, 2006, p. 17). However, it can be stated that other grounds may apply and that clear research questions on the linguistic object of study are necessary in order to build a balanced corpus (McEnery, Xiao and Tono, 2006, p. 18).

Sampling can be seen as representative if it is a scale-down version of a population of texts. Such population corresponds to the assembly of sampling units (e.g. books, journals, newspapers), which can be extracted either through *simple random sampling* (by numbering such units and choosing them randomly) or through *stratified random sampling* (by dividing the population into homogeneous groups to be chosen randomly) (McEnery, Xiao and Tono, 2006, p. 20). Either one or the other type of sampling will depend on which linguistic aspect to be investigated.

Even though a corpus is only a very limited portion of all the writing and speech produced in a certain language or language variety, the total number of words (tokens) and different words (types) as well as the number of categories and samples are also aspects to be taken into account when size is concerned (Kennedy, 1998, p. 66). Biber (1990; 1993b), for instance, argues that samples ranging from 2000 to 5000 words are enough to represent text categories, although the author posits that qualitative pilot studies based on theory should be carried out in order to test such design. Thus, likewise balance and sampling, a clear question as to which specific elements are to be scrutinized is of utmost importance so that a certain amount of texts can be accordingly integrated into the corpus.

As previously stated, considering the influence of modern technology on linguistic studies, texts in a corpus are to be in electronic format so that quantitative data can be automatically or semi-automatically obtained with the use of computational resources.

---

<sup>3</sup> These criteria have been used for the compilation of the British National Corpus (BNC). See on <http://www.natcorp.ox.ac.uk/corpus/creating.xml>

Softwares such as *TextSTAT*<sup>4</sup> and *Wordsmith Tools*© are able to produce statistical information about a corpus, such word frequency (total number of types and tokens) and the ratio between types and tokens. They can also display concordance lines (a specific word and its surrounding co-text of occurrence) and lists of collocates and clusters, among others.

Corpora may either be in raw format (plain text) or annotated, i.e. enhanced with linguistic information. To name but a few, added elements may include the disambiguation of word categories in different parts of speech (POS tagging<sup>5</sup>) or the encoding of texts through international standards, as in the *Text Encoding Initiative*<sup>6</sup> (TEI).

As far as such methodological assets are concerned, these have been employed in a wide range of fields within Applied Linguistics, including Translation Studies and ELT. As to the latter, such interface has been rather fruitful both in the realm of didactics and research, as shown in the following sections.

### 3 Language corpora and ELT

Corpus Linguistics has undoubtedly revolutionized ELT in a number of ways. As seen in O’Keeffe, McCarthy & Carter (2007), the application of corpora-based studies in ELT is quite extensive, ranging from highly positive contributions for material designers to understanding how non-natives speakers use the English language. In this section, we will look more closely at how corpora research has influenced teaching materials.

Corpus Linguistics studies provide empirical evidence as to how language is used in authentic contexts and can therefore guide material developers into adopting more natural dialogs and sillabi, as opposed to highly scripted language and unrealistic study topics (O’Keeffe, McCarthy and Carter, 2007).

One example of how corpora findings have influenced material design is the *Business Result* series (Baade, 2009; Duckworth, Turner and Lane, 2009; Hughes, Naunton and Lane, 2009; Grant and Hudson, 2009; Grant, Hughes and Turner, 2009), whose audios not only cover several varieties of the English language, but also include hesitations, interruptions and discourse markers, all of which are naturally found in real-life business meetings, teleconferences, presentations and conversation in general. The series has been used at Authentic English, in Belo Horizonte, by Sá (2012), along with the Global English Corporate Learning Service™.

Another example of how coursebook developers have been adopting corpora findings in ELT material is the *Touchstone* series (McCarthy, McCarten and Sandiford, 2005a, 2005b,

---

<sup>4</sup> This open source programme is available at <http://neon.niederlandistik.fu-berlin.de/en/textstat/>.

<sup>5</sup> Some POS taggers can be found at [http://cst.dk/online/pos\\_tagger/uk/](http://cst.dk/online/pos_tagger/uk/) and <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>.

<sup>6</sup> See <http://www.tei-c.org/index.xml> for further information.

2006a, 2006b), which presents discourse markers and wh-cleft sentences as part of conversation strategies and speaking skills. The series was largely adopted at the Center for Continuing Education (CENEX)<sup>7</sup> at the Faculty of Letters, in the Federal University of Minas Gerais, in Belo Horizonte, also as a result of the University's research in Corpus Linguistics and its associations to ELT (see, for example, Dutra, 2009, and Dutra and Silero, 2010).

In addition to coursebooks, dictionaries have been greatly impacted by Corpus Linguistics investigations. In fact, corpus-based dictionaries were the very first result of the impact of Corpus Linguistics in ELT back in the 1980's (Mukherjee, 2006) and now the most important English language dictionaries are entirely produced using corpus-driven information. Among many, the Collins Cobuild dictionaries are the forerunners when using corpus studies in the making of English dictionaries. The data for the Collins dictionaries is retrieved from the Collin Corpus<sup>8</sup>, which contains over 2 billion words<sup>9</sup>.

It is precisely the work in Corpus Linguistics that has paved the way for changes such as this: the acknowledgement that what was being taught in coursebooks and listed in dictionaries was quite different from the language produced by native speakers in actual interaction. It is crucial that learners be exposed to language as it is used in authentic situations, for it is those situations that learners will encounter as they engage in communication in the English language<sup>10</sup>.

One aspect that stands out in the *Touchstone* and *Business Result* series is the way spoken language is dealt with. It is clear that the development of speaking and listening skills is being reexamined in ELT. The content, the audios, and the type of activities, all of which include discourse markers and cleft sentences, among others, can be seen as a shift from the (somewhat) artificially produced recordings and the rigidly scripted use of language to a more natural and dynamic work on language skills.

In terms of studies on spoken corpora, it is worth mentioning that McCarthy (2006) has examined the notion of fluency in contrast to what speakers really produce in free flowing conversation. His research revealed that the so-called fluency patterns (rate of talk, lack of pauses, natural rhythm, stress and intonation), adopted as targets in ELT, are not always performed by native speakers, suggesting that a great deal of fluency also comes from the interactional support speakers give "to the flow of talk" (p.5).

---

<sup>7</sup> <http://www.letras.ufmg.br/cenex/>.

<sup>8</sup> <http://www.mycobuild.com/about-collins-corpus.aspx>.

<sup>9</sup> Of which 650 million words come from the Bank of English<sup>TM</sup>. New words are added to the Collin Corpus every month.

<sup>10</sup> Evidently, we are reflecting upon how corpora have influenced the teaching and the studies of the English language, as this is the language we have chosen to investigate. We acknowledge, however, that many other languages have been benefited by corpora studies. For Brazilian Portuguese, for instance, we refer to Raso & Mello (2009), who have created the C-ORAL-BRASIL (<http://www.c-oral-brasil.org/>), a corpus of spoken Brazilian Portuguese.

O’Keeffe, McCarthy & Carter (2007) develop this idea further, analyzing how short utterances and non-verbal signals (“right”, “yeah”, “good”, “umhum”, “is that true?”, among others) can guide speakers and listeners in interaction. These short utterances and non-verbal signals can perform several functions in conversation, acting as continuer and convergence response tokens and engagement tokens, forming a great part of spoken vocabulary (O’Keeffe, McCarthy & Carter, 2007).

Such observations should be taken into account when designing teaching materials and in teacher training and teacher development literature, as they represent what speakers actually do with the language and may thus require rethinking certain long-standing criteria and/or redesigning teaching and learning goals, especially when dealing with speaking and listening skills. Successful communicators are not only those who can use the language correctly, in its grammatical and lexical aspects, but are also those who engage bi-directionally in interaction through the use of linguistic and discourse mechanisms and strategies.

Corpus Linguistics has also produced abundant findings in lexis, as seen in McCarthy & Carter (2002), in which multi-word clusters in spoken British English are used as interactional patterns, showing that teaching should also be less focused on single, decontextualized words and more concerned with chunks of language as well as with how words are organized within specific contexts; and in Nelson (2006), whose analysis of semantic prosody in Business English has shown that Business English’s semantics are quite specific to the area, and that the meanings of words in Business English seem more limited, differently from its grammatical patterns, which are used with much more flexibility than in general English.

But these are just some examples. Virtually any language pattern can be examined through corpus-based analysis, from lexical and structural properties to phonological and discourse patterns (Krieger, 2003). Findings, some more comprehensive than others, include research on Early Modern English (Jacobsson, 2002) as well as to metaphorical language and English for specific purposes (Sardinha, 2011), among many others. When it comes to Corpus Linguistics, it all depends on the purpose of the research and, in ELT, how such research can be pedagogically applied.

Nelson (2000), for instance, has compiled an extensive corpus of Business English (spoken and written), analyzing highly frequent words, and cross-checking the corpus with the language used in teaching materials specifically designed for Business. In terms of the language compiled and examined in the corpus and the language of Business English materials, Nelson’s conclusion is that the two differ in many ways. The lexis in the corpus the researcher compiled is more diverse than that found in the materials, which tends to overemphasize politeness, social situations and travel, among others.

Nelson’s research goes to showing that although advances have been made, more corpora-based research is needed, especially considering that millions of people interact in

English every single day, in various contexts, contributing in one way or another to language modifications and novelties over time.

#### 4 Learner corpora and ELT

Learner corpora can assist academics and teachers in better understanding how second and/or foreign language learners use English. Defined by Granger (2003) as a collection of authentically produced texts by foreign or second language learners of English, learner corpora can help assemble objective information about learner language (Granger, 1998), which can then be described and analyzed by linguists, and can also be used as reference or consulting material for educators.

Pravec (2002) argues that learner corpora are important for researchers and language teachers because “they provide a deviation from the standard, i.e. language of the native speakers of a particular language” (p.81). We will once again mention the *Business Result* series and the many varieties of English covered in the series audios, including non-native speech, such as conversations carried out by Italians and Brazilians.

Seeing that English is now an international language with a significant number of non-native speakers around the world<sup>11</sup>, it seems paramount that non-native corpora be compiled, examined and described for research and pedagogical purposes, as it is such recorded language samples that, once examined, will show how the English language is being used around the world.

It is especially interesting to note how learner corpora can be of use in ELT, from the linguistic research that goes into compiling, tagging, coding and describing a learner corpus, to its potential applications in the language classroom. Compilation and description of learner corpora can cast light into how non-native speakers use the English language, especially when dealing with error-coded corpora. Through rigorous investigations of learner errors and overall use of language, one can comprehend language transfers, errors and other difficulties non-native speakers may have when learning English.

In Brazil, researchers have been making fruitful partnerships with universities and research projects from different parts of the world, and such partnerships have resulted in interesting learner corpora. The Br-ICLE Corpus<sup>12</sup> (Sardinha, 2001), for instance, as part of the ICLE (International Corpus of Learner English) project<sup>13</sup>, is a compilation of academic texts

---

<sup>11</sup> Graddol (1998) states that the number of non-native speakers of English has surpassed the number of native speakers around the world.

<sup>12</sup> Limited availability online: <http://www2.lael.pucsp.br/corpora/bricle/>. For more information about the project, refer to Tagnin (2003).

<sup>13</sup> <http://www.uclouvain.be/en-cecl-icle.html>. Project director: Sylviane Granger.

produced by Brazilian learners of English. The corpus has 200,000 words (Zara, 2011) and can be used for various types of research within written academic discourse.

Sardinha and Dutra (2010) have studied lexical bundles in the corpus<sup>14</sup>, resulting in the creation of a software specifically designed for identifying lexical bundles in electronic corpora. The authors argue that the software can help researchers in the investigation of lexical bundles in learner corpora, especially in academic discourse, seeing that such lexical patterns – as categorized in the research: referential expressions and discourse organizers, among others – are typical of argumentative texts.

Similar to the Br-ICLE Corpus, the CABrI Corpus (Corpus de Aprendizes Brasileiros do Inglês) is under compilation at the Federal University of Minas Gerais, and is part of Professor Deise Dutra's research on Corpus Linguistics and ELT. The CABrI is composed of essays written by Brazilian undergraduate students of English at the university and partially follows the ICLE guidelines for corpus compilation.

The LINDSEI-BR (Mello, 2011), a subcorpus of the LINDSEI (Louvain International Database of Spoken English Interlanguage) project<sup>15</sup>, is under compilation at the Federal University of Minas Gerais. Launched in association with the COBAI – Brazilian Oral Corpus of Learner English, the corpus contains fifty recordings by Brazilian high intermediate and advanced learners of English following the LINDSEI project guidelines, which includes a narrative, a free discussion and a description (Mello *et. al.*, 2012).

Another corpus of relevance in Brazil is the COMET<sup>16</sup> Corpus (Corpus Multilíngue para Ensino e Tradução), a project from the University of São Paulo under the coordination of Professor Stella Tagnin. The corpus has a subcorpus called CoMAprend Corpus (Corpus Multilíngue de Aprendizes) containing essays written by Brazilian undergraduate students of English, Spanish, French, Italian and German, and can be used for contrastive analysis, lexicographical analysis and teaching purposes.

Besides the CoMAprend, the COMET Corpus has two other subcorpora: the CorTrad<sup>17</sup> Corpus (Corpus de Tradução), which contains literary, technical and journalistic texts in English and has been POS-tagged, and the CorTec<sup>18</sup> Corpus (Corpus Técnico-Científico), made up of scientific and technical texts in English and Brazilian Portuguese. The CorTec Corpus has favored texts from the Information Technology, Commercial Law, Environmental and Dentistry areas.

---

<sup>14</sup> The post-doctorate research involved analyzing lexical bundles in the Louvain Corpus of Native English Essays – LOCNESS and the International Corpus of Learner English – ICLE, *as well as* in the Br-ICLE.

<sup>15</sup> <http://www.uclouvain.be/en-cecl-lindsei.html>. Project director: Sylviane Granger.

<sup>16</sup> <http://www.fflch.usp.br/dlm/comet/projeto.html>, <http://www.fflch.usp.br/dlm/comet/>.

<sup>17</sup> [http://www.fflch.usp.br/dlm/comet/consulta\\_cortrad.html](http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html).

<sup>18</sup> [http://www.fflch.usp.br/dlm/comet/consulta\\_cortec.html](http://www.fflch.usp.br/dlm/comet/consulta_cortec.html).



## 5 Final considerations

The studies and corpora reviewed in this paper are valuable resources for language professionals and educators, and constitute examples of how corpus linguistics studies have influenced ELT. Had not been for the advances in technology, coupled with the scientific status and development of Linguistics studies, investigations and applications such as the ones cited in this paper would probably not be a reality in the contemporary scenario of English language teaching and learning.

In regards to teaching materials, it is true that ELT pedagogy has been dealing with language study in more contextualized practice, prioritizing language chunks over single words and more authentic communication over highly scripted language. However, it is a very different thing reaching such conclusions as a result of corpora findings by examining how language is actually used in authentic contexts and not as a by-product of theoretical models or teaching approaches based on native speakers' intuitions or fabricated examples.

If we contrast the genres encompassed in language corpora with the ones covered in learner corpora we will see that there is still a large amount of research to be developed in the latter. While language corpora such as the Collins Corpus contains a wide range of spoken and written genres, learner corpora, at least in Brazil, is limited to essays, descriptions and interviews. It is a start, and one that should be praised. Nevertheless, learner corpora is an area in which much research remains to be done.

## References

BAADE, K. *Business Result: Advanced: Student's Book Pack*. Oxford: Oxford University Press, 2009.

BIBER, D. M. Methodological issues regarding corpus-based analyses of linguistic variation. In *Literary and Linguistic Computing*, Oxford: Oxford University Press, v.5, n.4, 1990. pp. 257-269.

\_\_\_\_\_. Representativeness in corpus design. In *Literary and Linguistic Computing*, Oxford: Oxford University Press, v.8, n.4, 1993. pp. 243-257.

DUCKWORTH, M., TURNER, R., and LANE, A. *Business Result: Upper Intermediate: Student's Book Pack*. Oxford: Oxford University Press, 2009.

DUTRA, D. P. Conscientização linguística com base em corpora online. In *Revista Intercâmbio*, volume XX: 79-98, 2009. São Paulo: LAEL/PUC-SP. Available at: <http://www4.pucsp.br/pos/lael/intercambio/pdf/XX/5%20Deise%20Dutra%20Intercambio%20v%20lume%2020.pdf>. Accessed on April 4, 2012.

DUTRA, D. P., SILEIRO, R. P. Descobertas linguísticas para pesquisadores e aprendizes: a Linguística de Corpus e o ensino de gramática. In *Revista Brasileira de Linguística Aplicada*. v. 10, n. 4, 2010. Available at: <http://www.periodicos.letras.ufmg.br/index.php/rbla/article/view/890>. Accessed on April 4, 2012.

GRADDOL, D. The decline of the native speaker. In *AILA Review 13: English in a changing world*. 1998. Available at: <http://www.ffri.uniri.hr/datoteke/Sirola/graddolnativespeaker.pdf>. Accessed on April 4, 2012

GRANGER, S. (ed.) *Learner English on Computer*. London & New York: Addison Wesley Longman, 1998.

\_\_\_\_\_. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. In *TESOL Quarterly*, v.37, n. 3, 2003. pp. 538-546.

GRANT, D., HUGHES, J., and TURNER, R. *Business Result: Elementary: Student's Book Pack*. Oxford: Oxford University Press, 2009.

GRANT, D. and HUDSON, J. *Business Result: Pre-Intermediate: Student's Book Pack*. Oxford: Oxford University Press, 2009.

HUNSTON, S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

HUGHES, J., NAUNTON, J., and LANE, A. *Business Result: Intermediate: Student's Book Pack*. Oxford: Oxford University Press, 2009.

JACOBSSON, M. Thank you and thanks in Early Modern English. In *ICAME Journal*. No. 26, April 2002. Available at: <http://icame.uib.no/ij26/jacobsson.pdf>. Accessed on April 04, 2012.

KENNEDY, G. *An Introduction to Corpus Linguistics*. Essex: Longman, 1998.

KRIEGER, D. Corpus Linguistics: What it is and how it can be applied to teaching. In *The Internet TESL Journal*, Vol. IX, No. 3, March 2003. Available at: <http://iteslj.org/Articles/Krieger-Corpus.html>. Accessed on April 04, 2012.

McCARTHY, M., and CARTER, R. This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. In: *Teanga: The Irish Yearbook of Applied Linguistics*, vol. 21, 2002, pp. 30-52.

McCARTHY, M., McCARTEN, J., and SANDIFORD, H. *Touchstone. Student's Book 1*. Cambridge: Cambridge University Press, 2005a.

\_\_\_\_\_. *Touchstone. Student's Book 2*. Cambridge: Cambridge University Press, 2005b.

\_\_\_\_\_. *Touchstone. Student's Book 3*. Cambridge: Cambridge University Press, 2006a.

\_\_\_\_\_. *Touchstone. Student's Book 4*. Cambridge: Cambridge University Press, 2006b.

McCARTHY, M. *Explorations in Corpus Linguistics*. New York: Cambridge University Press, 2006.

McKENERY, T.; WILSON, A. *Corpus Linguistics*. Edinburgh: Edinburgh UP, 2001.

McKENERY, T.; XIAO, R.; TONO, Y.; *Corpus-Based Language Studies: An Advanced Resource Book*. London & New York: Routledge, 2006.

MELLO, H. *LINDSEI-BR: corpus oral de aprendizes brasileiros de inglês*. Research project. Faculdade Letras. Universidade Federal de Minas Gerais, 2011.

MELLO, H. *et al.* LINDSEI-BR: an oral English interlanguage corpus. In *GSCP 2012 International Conference*. Belo Horizonte, 2012. Available at: <http://www.lettras.ufmg.br/gscp2012-eng/data1/arquivos/80.pdf>. Accessed on April 07, 2012.

MUKHERJEE, J. *Corpus linguistics and language pedagogy: The state of the art – and beyond*. Justus Liebig University. Gessen: Germany, 2006. Available at: <http://www.uni-giessen.de/anglistik/LING/Staff/mukherjee/pdfs/CorpusLinguisticsandLanguagePedagogy.pdf>. Accessed on April 8, 2012.

NELSON, M. *A corpus-based study of the lexis of business English and business English teaching materials*. Unpublished thesis. University of Manchester. Manchester. 2000. Available online: <http://users.utu.fi/micnel/thesis.html>. Access on April 04, 2012.

\_\_\_\_\_. Semantic associations in Business English: a corpus-based analysis. In: *English for Specific Purposes*. Volume 25, Issue 2, 2006, pp. 217–234.

O'KEEFFE, A., McCARTHY, M., CARTER, R. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press, 2007.

PRAVEC, N. A. Survey of learner corpora. In *ICAME Journal*. No. 26, April 2002. Available at: <http://icame.uib.no/ij26/pravec.pdf>. Accessed on April 04, 2012.

RASO, T.; and MELLO, H. The C-ORAL-BRASIL Project. In *AACL 2009 – American Association for Corpus Linguistics*. Canada, 2009. Available at: <http://www.ualberta.ca/~aac12009/PDFs/RasoMello2009AACL.pdf>. Accessed on April 08, 2012.

SÁ, E. M. The Global English Corporate Learning Service™ as a tool for developing learner autonomy in Business English Language Teaching and Learning. *INTED 2012, 6th International Technology, Education and Development Conference*. Valencia, Spain, 2012a.

SARDINHA, T.B. *Linguística de corpus*. Barueri, SP: Manole, 2004.

\_\_\_\_\_. Metáforas e Linguística de Corpus: metodologia de análise aplicada a um gênero de negócios. In *DELTA*, São Paulo, v.27, n.1, 2011. Available at: [http://www.scielo.br/scielo.php?pid=S0102-44502011000100001&script=sci\\_arttext](http://www.scielo.br/scielo.php?pid=S0102-44502011000100001&script=sci_arttext). Accessed on April 8, 2012.

\_\_\_\_\_. Br-ICLE - *The Brazilian Portuguese Sub-corpus of ICLE, The International Corpus of Learner English*, 2001. Available at: <http://www2.lael.pucsp.br/corpora/bricle/>. Accessed on April 04, 2012.

SARDINHA, T. B., and DUTRA, D. P. Pacotes lexicais em corpora de aprendizes. *ELC 2010 IX Encontro de Linguística de Corpus*. Porto Alegre, 2010. Available at: <http://corpuslg.org/gelc/elc2010.php/2010/11/05/baixar-apresentacoes>. Accessed on April 07, 2012.

TAGNIN, S. E. O. *A multilingual learner corpus in Brazil*. Lancaster: Corpus Linguistics – Learner Corpus Workshop, 2003. Available at:

<http://www.fflch.usp.br/dlm/comet/artigos/A%20multilingual%20learner%20corpus%20in%20Brazil.pdf>. Accessed on April 04, 2012.

ZARA, J. Sobre a expressão linguística de eventos de transferência de posse na interlíngua português-inglês. In *ELC 2011. X Encontro de Linguística de Corpus*. Belo Horizonte, 2011. Available at:

[http://www.letras.ufmg.br/linguisticacorpus2011/data1/arquivos/ta\\_juliazara.pdf](http://www.letras.ufmg.br/linguisticacorpus2011/data1/arquivos/ta_juliazara.pdf). Accessed on April 07, 2012.

Received: April 30, 2012

Accepted: September 15, 2012

E-mail:

[mattos.elisa@gmail.com](mailto:mattos.elisa@gmail.com)

[leopereiranunes@gmail.com](mailto:leopereiranunes@gmail.com)