




VI Escola Brasileira de Linguística Computacional (EBRALC 2012) - 11 e 12 de setembro de 2012
XI Encontro de Linguística de Corpus (ELC 2012) - 13, 14 e 15 de setembro de 2012
São Carlos - SP - Brasil

Bundles in learner corpora: what a type and token analysis can reveal

Deise Prina Dutra- UFMG
deisepdutra@gmail.com

Barbara Malveira Orfano-UFSJ
bmalveira@yahoo.com.br

Tony Berber–Sardinha-PUC-SP
tony@pucsp.br

Acknowledgment

- Faculdade de Letras da UFMG
  - LEEL
- Centro de Extensão da UFSJ
- PUC-SP
- FAPESP

Introduction

- Corpus Linguistics (CL) has valued the investigation of group of words rather then words in isolation
  - Collocations (Sinclair 1991)
- Studies have concentrated on lexical bundles in a variety of contexts
  - in business contexts – genre based analysis of business report (Berber Sardinha 2003);
  - in the university – oral and written discourse - (Biber et al. 2004; 2006; 2009);
  - in different disciplines– electric engineering, biology, administration, applied linguistics (Hyland 2008);
  - in academia, where Simpson-Vlach and Ellis (2010) propose a list of the most commonly used bundles in academic registers.

Lexical Bundles

- simply sequences of word forms that commonly go together in natural discourse (Biber et al. 1999: 990)
  - *in terms of the*
  - *a list of*
  - *the fact that*
  - *it has been argued that*
  - *to a certain extent*
  - *my point of view*

Research on lexical bundles

- Biber et al. (2004)
  - Frequency approach
  - Classroom teaching and textbooks
  - Structural patterns and function
    - Three major functional categories
      - » Referential expressions
      - » Stance expressions
      - » Discourse organizing functions
- Simpson-Vlach e Ellis (2010)
  - oral and written corpora
  - MICASE + BNC (oral academic part)
  - Hyland corpus (2004) + BNC files (various academic subjects)
  - Academic Formulas List (AFL)- 435 lexical bundles

Aims

- to discuss the relevance of analyzing and contrasting types and tokens of bundles produced by native and non-native speakers in argumentative essays;

- to highlight the differences among the corpora as far as stance expressions are concerned;

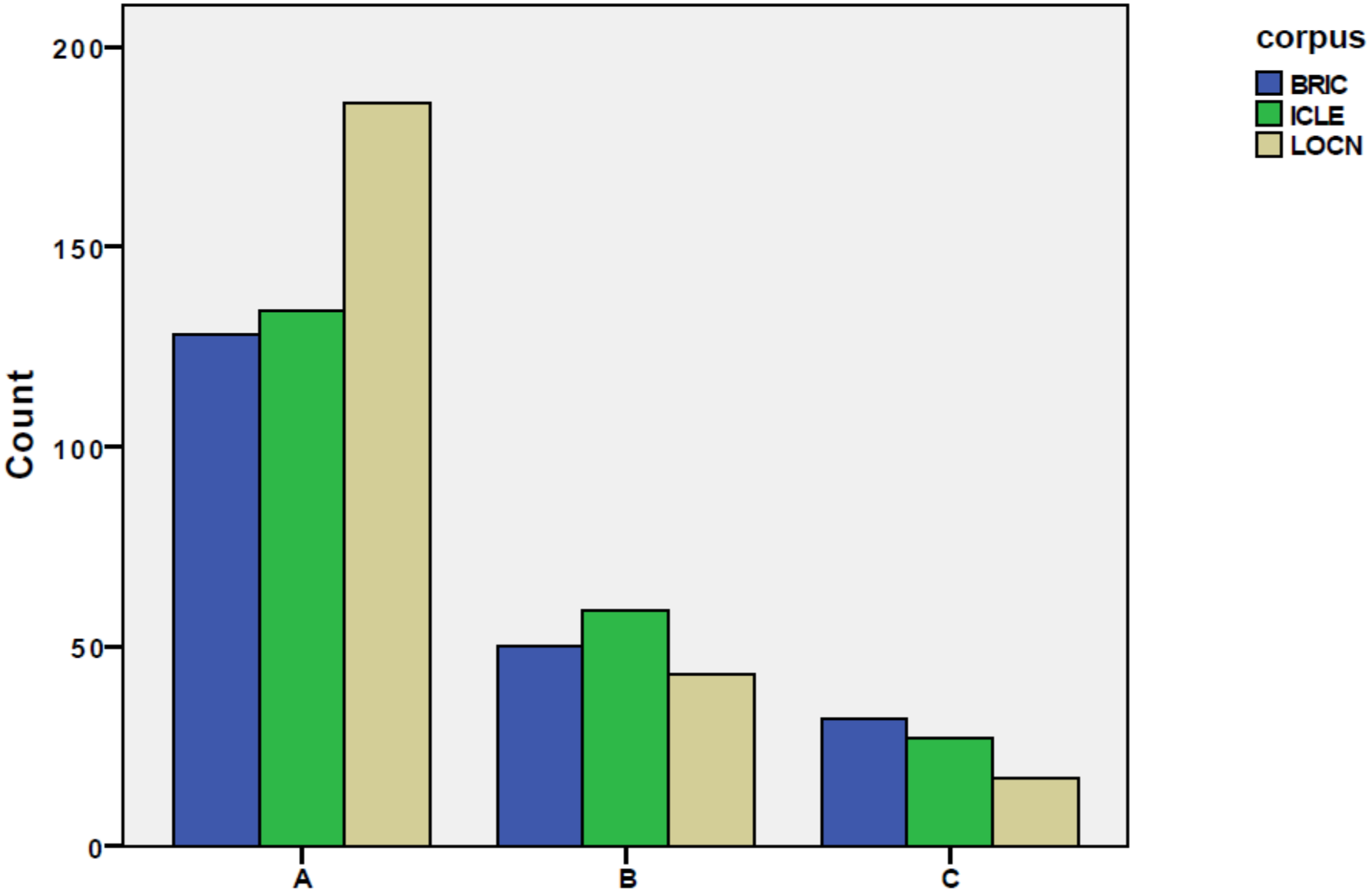
- to detect if these differences are mainly structural or related to frequency within a specific function.

Data - Essays

- LOCNESS (Louvain Corpus of Native English Essays)
  - 324,006 words
  - written language
  - American and British university students
- ICLE (International Corpus of Learner English)
  - 3.7 million words (Granger et al. 2009)
  - written language
  - 16 subcorpora (Japan, China, Italy, Finland ...)
- Br-ICLE (Berber Sardinha 2001)
  - In 2009-> 159,000 words (aim 200,000 words)
    - CABrI (Corpus de Aprendizes Brasileiros de Inglês – UFMG)
- Total – 4,251,714 words

Methodology

- bundles of 4 words were extracted from each corpus with scripts specially developed for our research project;
- the bundles were categorized manually and automatically according to the AFL framework
  - 3 major categories: referential expressions, stance expressions and discourse organizing functions - 18 specific subcategories
- the most frequent categories in each corpora were identified and isolated and we detected the differences in terms of types of bundles across the broad categories (>= 20 wpm);
- token frequency analysis was done to investigate the extent to which they could reveal significant differences among the subcategories;
- we ran statistical tests to identify differences within each category;
- concordance lines for the most frequent bundles in each corpora were generated in order to identify differences in use across the 3 datasets.



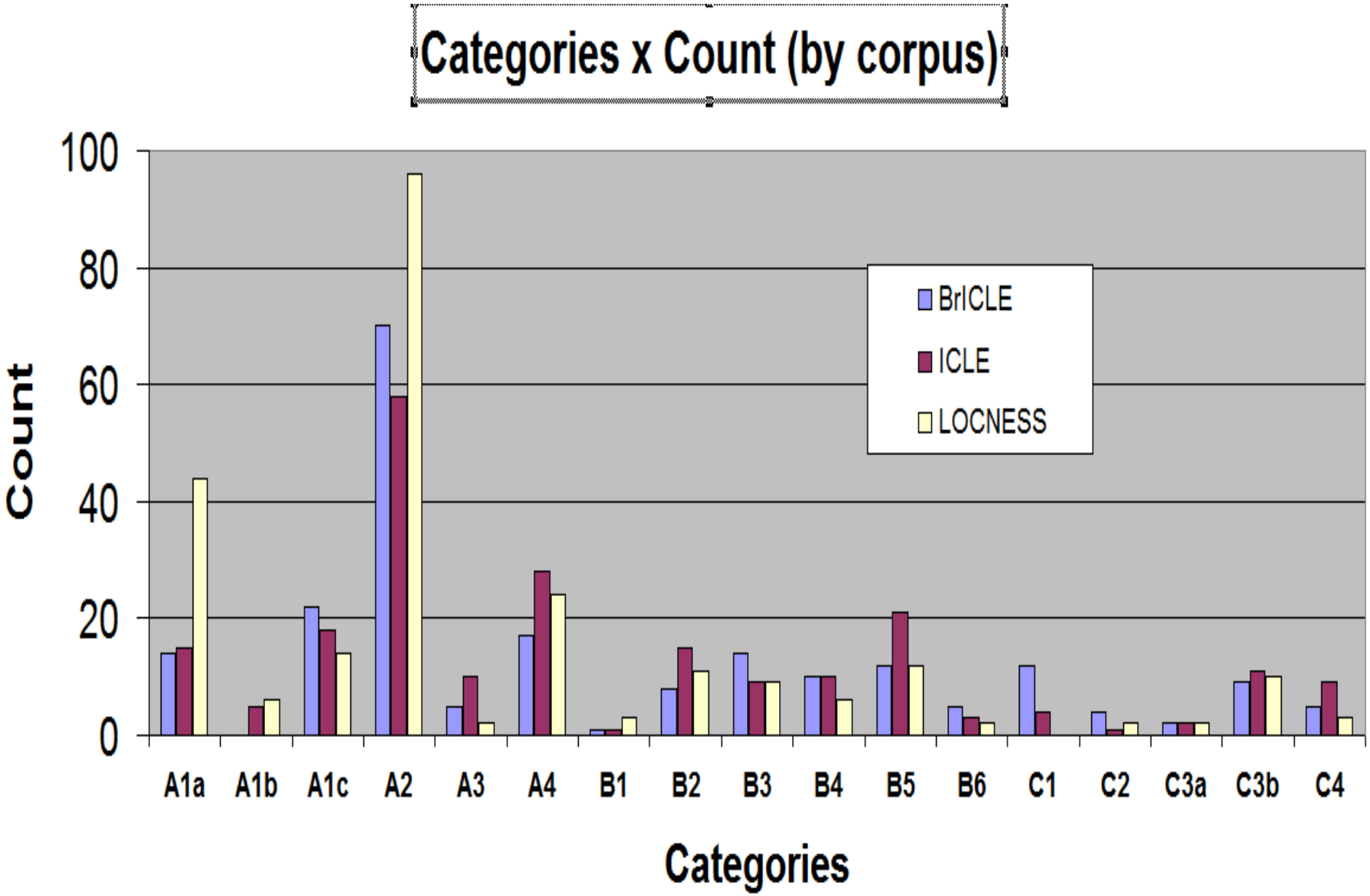
A = referential expressions

B = stance expressions

C= discourse organizing expressions

Chi-square Test

	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		


Categories x Count (by corpus)
100
80
60
40
20
0
Count
BrICLE
ICLE
LOCNESS
A1a
A1b
A1c
A2
A3
A4
B1
B2
B3
B4
B5
B6
C1
C2
C3a
C3b
C4
Categories

Chi-square test
all subcategories

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.624	34	0.000
Likelihood Ratio	23.112	34	0.000
N Valid Cases	676		

SUB-CATEGORY X CORPUS (TOKEN FREQUENCY)

	LOCNESS		ICLE		BRICLE	
	raw	wpm	raw	wpm	raw	wpm
B1 Hedges	33	101.851	104	27.597	12	75.385
B2 Epistemic stance	83	255.992	2128	564.678	23	144.488
B3 Obligation and directives	75	231.478	1485	394.054	71	477,443
B4 Expressions ability and possibility	97	299.379	1252	332.225	53	332.971
B5 Evaluation	129	370.364	2485	624.647	90	396.8
B6 Intention, volition and prediction	32	98.763	748	198.487	25	156.209

	LOCNESS	ICLE	BR-ICLE
B1	to a certain extent could be used to can be seen to	is a kind of	is a kind of
B2	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
B3	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

Bundle Structure

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

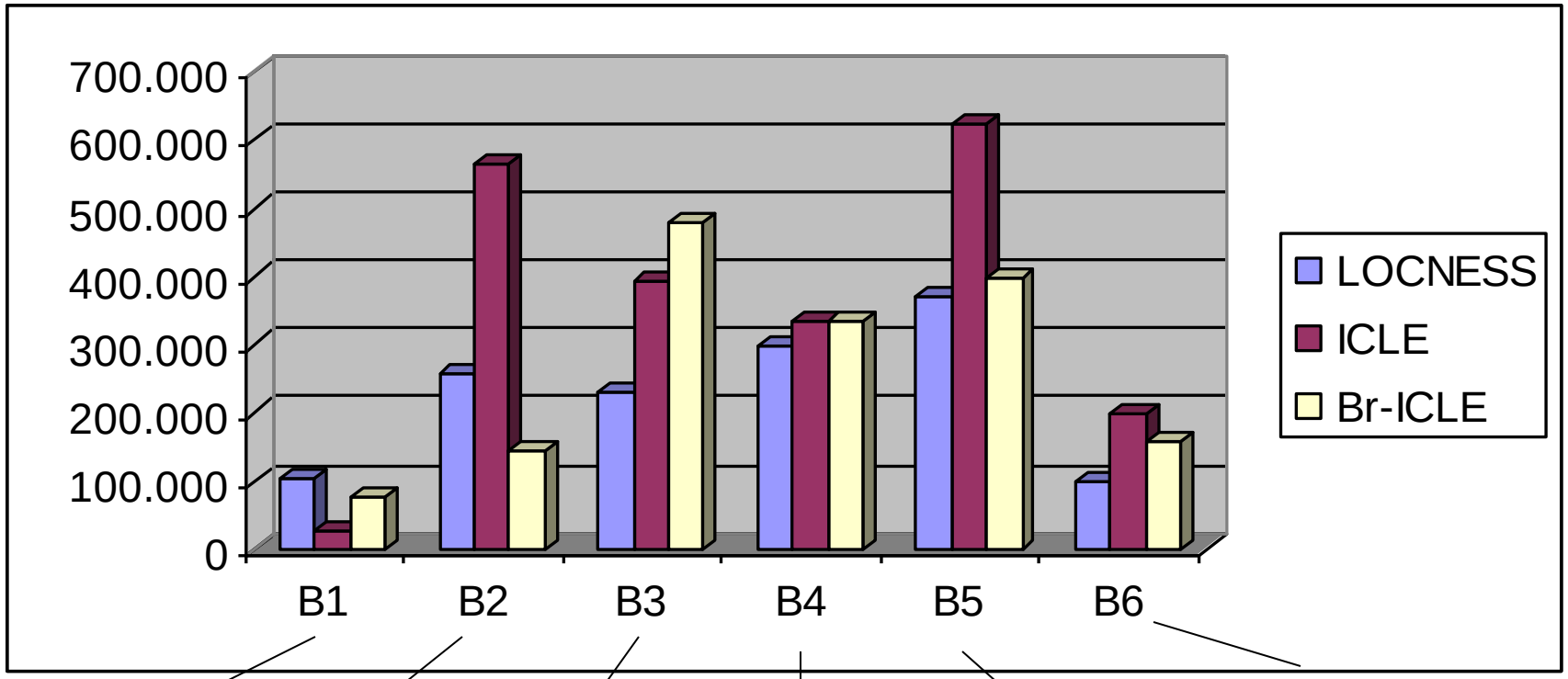
Register appropriateness

- Written vs Spoken
  - Hedging (cautious language)
    - LOCNESS
      - *to a certain extent / could be use to /can be used to*
    - ICLE and Br-ICLE
      - *is a kind of*
- Participant-oriented (reader or writer oriented)
  - Epistemic
    - LOCNESS
      - *is shown to be / can be seen as / is seen to be*
      - *I think that the / I feel that the*
    - ICLE and Br-ICLE
      - *I think it is / some people think that / my point of view*
      - *it has been argued that*

Chi-square test
Stance expressions

	Value	df	*p*-value
Pearson Chi-Square	8.742	10	0.557
Valid Cases	149		

Normalized token frequency



Hedging

Epistemic

Obligation
Directives

Ability
possibility

Evaluation

Intention

Volition

Prediction

Obligation and Directives

B3	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

Conclusion
Br-ICLE

- Types
  - Less diverse use of stance bundles
- Tokens
  - More personal
    - bundle structure
      - fewer antecipatory *it* and passive structures
  - Directive and obligation
    - Participant-oriented
      - fewer hedging bundles
      - instead there is overuse of bundles that carry an overstating tone
- Lexical bundle studies
  - Token analysis complements type analysis helping to describe different corpora even when there are no statistically significant differences.

Future actions

- Classify more bundles - >10 wpm
  - Improve automatic bundle classification
- Bundle analyzer
  - Make it available to
    - Teachers
    - Students
- Add to the bundle analysis
  - Readability measures

Bibliography

BERBER SARDINHA, T. O corpus de aprendiz Br-ICLE. *Intercâmbio*, v. 10, 2001, p. 227-39.

BIBER, D.; CONRAD, S.; CORTES, V. *If you look at*... Lexical bundles in university teaching and textbooks. Applied Linguistics, v. 25, n, 3. p. 371-405. 2004.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *Longman grammar of spoken and written English*. Essex:Longman. 1999.

CARTER, R.; MCCARTHY, M. *Cambridge Grammar of English*. Cambridge: Cambridge. 2006

CHEN, Y.; BAKER, P. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology.* June 2010, Volume 14, Number 2 pp. 30–49

CORTES, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397–423.

DE COCK, S. et al. An automated apporach to the phrasicon on EFL learners. In: GRANGER, S. (ed.) *Learner English on Computer*. London & New York: Addison Wesley Longman. 1998. p.67-80.

De COCK, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 51–68). Amsterdam: Rodopi.

DUTRA, D. P.; BERBER-SARDINHA, T. Pacotes lexicais em corpora de aprendizes. (in press)

MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in foreign language learning and teaching*. Cambridge: Cambridge. 2008.

NESSELHAULF, N. *Collocations in a learner* corpus. Amsterdam: John Benjamins. 2005.

NEKRASOVA, T. English L1 and L2 Speakers' Knowledge of Lexical Bundles. *Language* Learning v. 59, n. 3. p. 647, 486. 2009.

O´KEEFFE, A.; MCCARTHY, M.; CARTER, R. *From corpus to classroom*: language use and language teaching. Cambridge: CUP. 2007.

OLIVEIRA, M. ; DUTRA, D. Pacotes lexicais ou palavras isoladas? Organizadores discursivos em corpora de aprendizes e de falantes nativos. 2011

SHEPHERD, T. Corpora de aprendiz de língua estrangeira:um estudo contrastivo de n-gramas. *Veredas* n.2. p. 100-116. 2009.

SINCLAIR, J. M. *Corpus, concordance, collocation*. Oxford. Oxford University Press. 1991.

SIMPSON-VLACH, R; ELLIS, N. An Academic Formulas List: New Methods in Phraseology Research*Applied Linguistics*, p. 1-26. 2010.

Thank you!

Deise Prina Dutra- UFMG
deisepdutra@gmail.com

Barbara Malveira Orfano-UFSJ
bmalveira@yahoo.com.br

Tony Berber–Sardinha-PUC-SP
tony@pucsp.br