Diamesia-related bundles across native and non-native written and oral corpora

Deise P. Dutra - Heliana Mello
Bárbara Orfano - Carolina Bohórquez
UFMG – UFSJ

LCR 2013
Bergen, Norway
Sept, 2013

Acknowledgment

- *Faculdade de Letras* - UFMG
  - LEEL
- *PROPE*- UFSJ
- FAPEMIG

Lexical bundles in academic discourse

- Academic discourse has caught the attention of researchers and it is fair to say that many of them have focused on lexical bundles both in written and in oral discourse.
- Why are lexical bundles relevant for academic context?
  - they are building blocks of discourse (Biber 2004)
  - they "serve the most important communicative needs of a register" (Biber, 2009, p. 285)
  - important component of fluent linguistic production (Hyland, 2008)
- There is a paucity of research about the pervasiveness of oral bundles into written discourse

Functional pragmatic bundle classification

Biber et al. (2004)

- oral and written corpora
- Structural patterns and functional categories
  - Three major functional categories
    - » Referential expressions
    - » Stance expressions
    - » Discourse organizing functions

- Simpson-Vlach e Ellis (2010)
  - oral and written corpora
    - » They proposed the Academic Formulas List (AFL)

    435 lexical bundles – 18 subcategories

Definition of pragmatic functional categories

- Referential expressions
  - Express identification of entities or attributes in a text, which are essential in the presentation of ideas where arguments build up from
- Stance expressions
  - judgments and opinions conveyed by the writers or speakers
- Discourse organizing expressions
  - Explicit markers of text organization

Examples from written corpora

- Spanish-ICLE
  - *This person have to get by with the computer, and **in the case of** being male, have done the military service.* (referential expression – intangible framing attribute)
- Br-ICLE
  - *The power of imagination **is a kind of** gift that people are born with, but it seems that they don't remember this currently, like some years ago.* (stance expression– hedge)

Examples from oral corpora

- SBC

  he comes and says
  well
  he goes
  I don't know if *you've*
  if you've
  packed this or not (**epistemic stance**)

- LINDSEI-Br

  … I will do everything at the same time but ***I think this is not*** *the way you have to do in your time and I will let my life* (**epistemic stance**)

Research questions

- Is there a significant difference in the frequency of referential, stance and discourse organizing bundles across NN and N written and oral corpora?
- Does difference correlate proficiency levels (NN corpora)?
- To what extent do frequent 4-word oral bundles appear in written learner corpora?

Methodology

- 4 NN written corpora (argumentative essays)
  - Chinese – ICLE subcorpus (490,617 words)
  - Spanish - ICLE subcorpus (198,131 words)
  - Dutch - ICLE subcorpus (234,723 words)
    - Proficiency range from intermediate to advanced (B2 – C1/C2 -ICLEV2)
  - Br-ICLE, the Brazilian *subcorpus* (201,771 words) - not yet part of ICLE – overall group proficiency is not clear
- 1 NN oral corpus (quasi spontaneous production)
  - LINDSEI-BR (40,456 words) - under-construction *subcorpus* of the Louvain International Database of Spoken English Interlanguage (LINDSEI)
- 2 N corpora
  - Louvain Corpus of Native English Essays (LOCNESS) (322,985 words)
  - the Santa Barbara Corpus of Spoken English (SBC) - 200,000 words - subcorpus of 56,856 words (spontaneous conversation

Methodology

- Bundle lists: Collocate 1.0 (Barlow 2004)
- Conservative cut-off point: > 20 wpm

Methodology

- Topic-related and overlapping bundles were manually eliminated in all the corpora
- Eliminations are relevant as the maintenance of such bundles can bias bundle frequency results (Bohórquez et al. 2012; Staples et al. 2013)
- Analysis
  - Fisher's Exact Test
    - **R** scripts

Results
Essays – Academic Discourse

- Total of 1284 bundles
- Category with the highest count of bundles
  - referential expressions
  - Except for the Chinese ICLE subcorpus, LINDSEI and SBC
    - Stance expressions bundles present a higher count

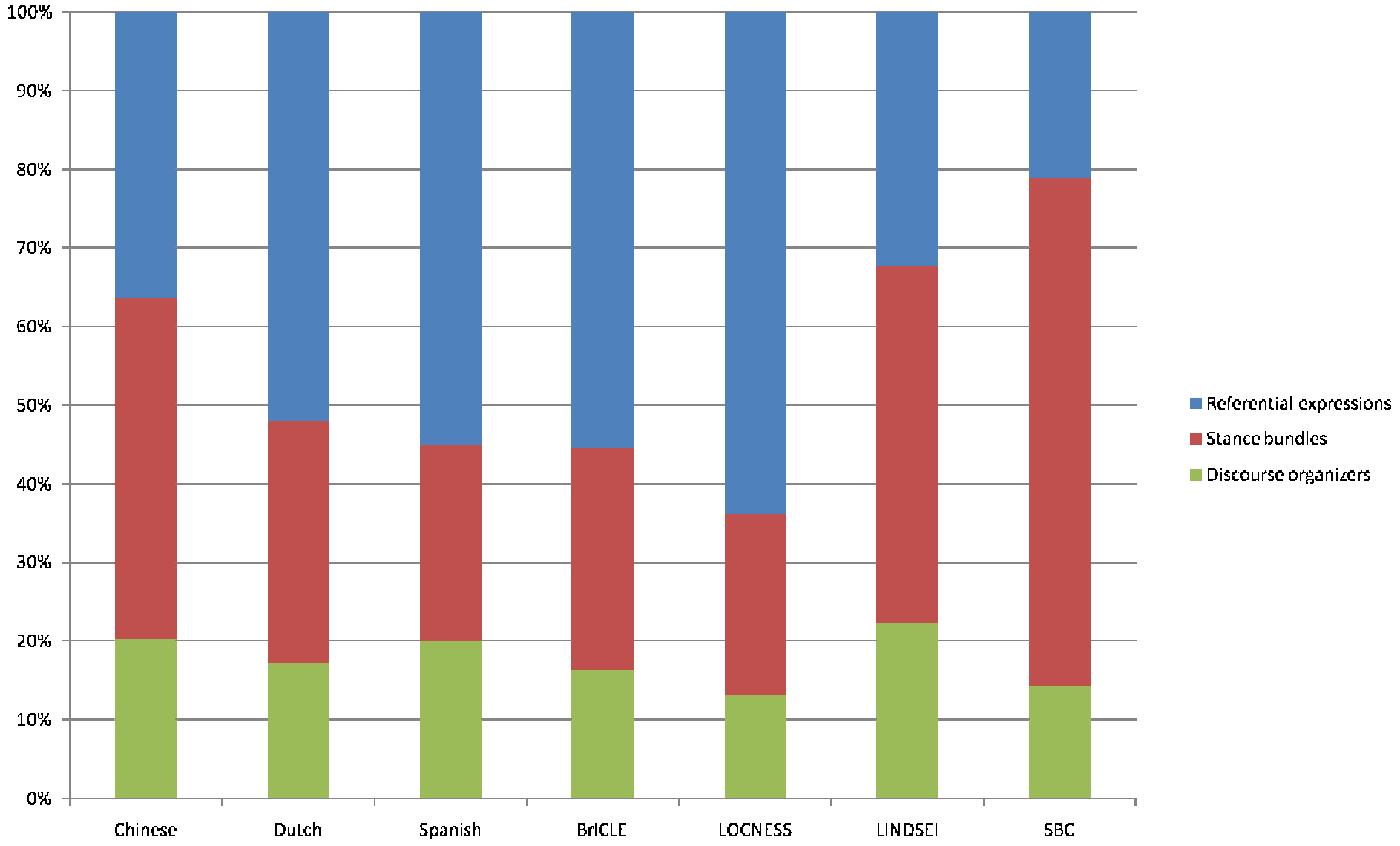
Bundle type results - raw and normalized frequency

Bundle / Corpora	Chinese	Dutch	Spanish	Br-ICLE	LOCNESS	LINDSEI	SBC
Referential	80	138	218	125	86	10	3
	163	588	1100	619	266	247	53
Stance	95	82	100	63	31	14	9
	193	349	505	312	96	346	158
Discourse Organizaing	44	44	79	37	17	7	2
	90	187	399	183	53	173	35

Results: 3 majors categories

- Fisher's Exact Test
  - *p*-value = 0.0001
    - two.sided
- Significant difference
  - Frequency of bundle types across all corpora (oral and written + N and NN)

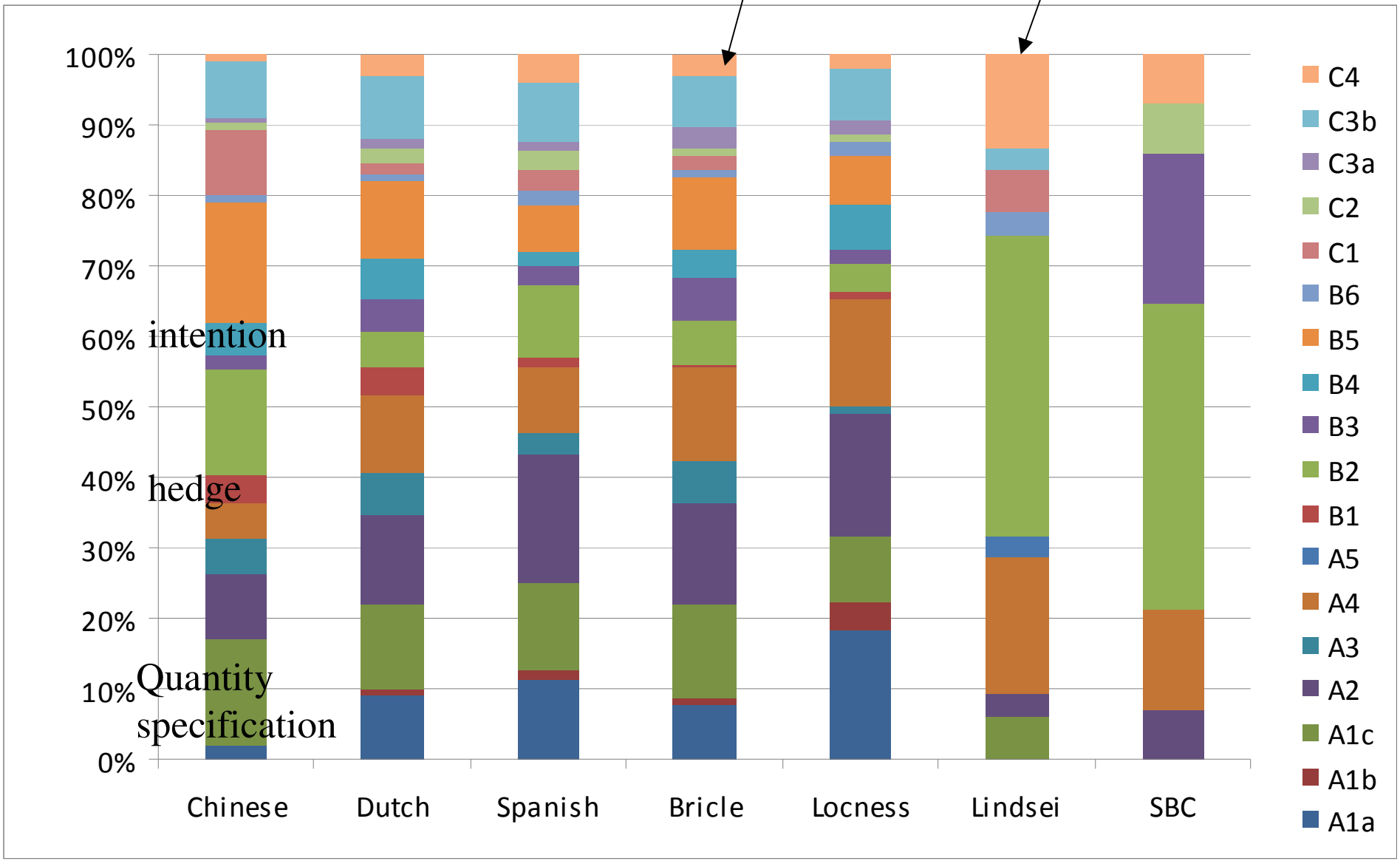
Functional distribution (types)



Results: 18 subcategories

- Fisher's Exact Test
  - *p*-value = 0.0001
    - two.sided
- Significant difference
  - Frequency of bundle types across all corpora (oral and written + N and NN)

Subcategory type functional distribution



Oral -> written

Bundle subcategory /Corpus	**Chinese**	**Dutch**	**Spanish**	**Br-ICLE**
Quantity specification (there are/is/ get /have) **a lot of** (money /people...)	X	X	X	X
Hedge **is a kind of**	X	as a sort of	X is a sort of	X

Additional comparison of oral and written bundles

- Another common stance bundle across all written corpora
  - *I would like to*
    - There is the use of the first person pronoun, this may not reveal an influence of oral discourse. It may be a characteristic of opinion essay “genre”

Our research questions

- Is there a significant difference in the frequency of referential, stance and discourse organizing bundles across NN and N written and oral corpora?
  - YES
- Does this difference correlate proficiency level ?
  - General categories: maybe
    - Chinese (B2 –CEF) – Stance bundles- most frequent category in SBC
  - Subcategories: ? (e.g. hedge – *likely / seem*)
    - Dutch: 10 types
    - Spanish: 6 types
    - Chinese: 9 types
    - Brazilian: 1type – Hedge – clearly an oral bundle (*is a kind of*)
- To what extent do frequent 4-word oral bundles appear in written learner corpora?
  - *is a kind of*
  - *a lot of / sort of* - > 2 or 3-word bundles that appear frequently with slot variation in written corpora

Conclusions

- At first our comparison focused on identical 4-word bundles
- LINDSEI-Br and Br-ICLE
- There is some kind of overlap between what students say orally and what they write (but it was lower than what we first hypothesized)
    - Very conservative cut-off point of 20 wpm
    - 4-word bundles (3, 2-word bundles)
- Some lack of genre adequacy

Future studies

- Look at smaller bundles
- Increase the oral corpora size
- Access other oral corpora to confirm the traces in found in other ICLE subcorpora
- Our research group has been developing an automatized bundle elimination
  - Topic-related
  - Overlapping bundle
    - This will allow us to do this study with all the other ICLE subcorpora

Thank you!

- Barbara Malveira - bmalveira@yahoo.com.br
- Carolina Bohorquez - carolinaboho@gmail.com
- Deise Dutra - dpdutra@ufmg.br
- Heliana Mello - hmello@ufmg.br

AFL list

A. Referential expressions	**B. Stance expressions**	**C. Discourse organizing functions**
1.Specifications of attributes a.Intangible framming attributes b.b. tangible framming attributtes c.Quantity specification	1.Hedges	1.Metadiscourse and textual reference
2. Identification and focus	2. Epistemic stance	2.Topic introduction and focus
3. Contrast and comparions	3. Obligation and directives	3.Topic elaboration a.non-causal b.Cause and effect
4. Deictics and locatives	4.Expressions of ability and possibility	4.Discouse markers
5. Vagueness markers (spoken discourse)	5.Evaluation	
	6.Intention/volition, prediction	

Bibliography

Biber, D. et al. *If you look at ...*: lexical bundles in university teaching and textbooks. *Applied Linguistics*, v.25, n.3, p. 371-405, 2004.

Biber, D. *University Language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins. 2006.

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

Bohórquez, C. et al. (2012) O impacto da eliminação de pacotes lexicais relacionados ao tópico e em contexto de sobreposição In: ENCONTRO DE LINGUÍSTICA DE CORPUS, 11, 2012, São Carlos. *Anais*...São Paulo: Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, 2012.

Chen, Y.; Baker, P. (2010) Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology,* vol. 14, nº 2. p 30-49.

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.

Simpson-Vlach, R; Ellis, N. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistic*, v.31, n.4, p. 487-512. 2010.

Staples,S., Egbert J.,Biber, D., McClair, A. (2013) Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, v. 12, p. 214–225.

Tannen, D. Oral and Literate Strategies in Spoken and Written Discourse. *Literacy for life: The demand for reading and writing*, Richard W. Bailey and Robin Melanie Fosheim (Ed.). NY: The Modern Language Association, 1983.