CLILLAC-ARP Workshop - Beyond CEFR level prediction of texts in learner corpora: Exploring feedback to learners and learning analytics - Paris October 30, 2019

Investigating criterial features of learner English : Microsystem features

•••

Thomas Gaillat Nicolas Ballier Andrew Simpkin Bernardo Stearns
Manon Bouyé Manel Zarrouk

PHC Ulysses 2019 France-Ireland

Acknowledgement

With the financial support of the French Ministry for Europe and Foreign Affairs (Ministères de l’Europe et des affaires étrangères, MEAE) and the French Ministry of Higher Education, Research and Innovation (Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, MESRI).

PHC Hubert Currien Ulysse 2019 (ref 43121RJ)

Research Questions

- What criterial features (Hawkins & Buttery 2010) can be identified as predictors for CEFR levels?
  - How to operationalise linguistic learner microsystem into metrics?
    - Are there variations within microsystems?

Previous work

Work on metrics

- Shared-tasks: Spoken CALL (Baur et al., 2017) & CAP18 (Arnold, et al. 2018)
- Automatic learner language analysis: criterial features (Crossley et al, 2011; Hawkins & Filipović, 2012)
- Complexity metrics (Lu 2010)

Corpus

EFCAMDAT @ Cambridge University (Geertzen, Alexopoulou, & Korhonen, 2013).

- French and Spanish L1 components
- 83 million word learner corpus
- Writing essays of different Englishtown levels mapped onto the six CEFR levels

Number of writings	A1	A2	B1	B2	C1	C2
French L1	17,605	11,584	8,105	3,514	742	76
Spanish L1	2,572	2,066	2,005	1,176	340	32

Annotation & metrics

Annotation and pattern frequency tools

- LCA (TreeTagger) - TAACO - TAALES - TAASC -TEXTSTAT - PYENCHANT
- Modified version of L2SCA New features based on paradigmatic microsystems L2SCA_MS

Metrics

- Syntactic e.g. amount of coordination, subordination, **microsystems**
- Semantic e.g. ambiguity
- Lexical e.g. density, sophistication
- Pragmatic e.g. cohesion

Linguistic microsystems in learner English

Instability in syntactic structures

- Paradigmatic confusions/interactions between words of the same syntactic function but of different semantic implications.
- The article microsystem: *a, the* or *0?*

"Ladies and Gentlemans, My flat was robbed the previous evening. In coming back at my home, I saw that ***the*** window was broken." (EFCAMDAT writing ID: 2498)

"What do you think about positive discrimination in ***the*** companies?" (EFCAMDAT writing ID: 569744)

"Why *the* gender's discrimination is still a problem in our society?" (EFCAMDAT writing ID: 579779)

And more microsystems

Microsystems	Components
Nominal density	determiner genitive; noun-of/for-noun constructions, compound nouns
Modals for possibility	may; can; might; could
Modals for obligation	must; have to
Proforms	it; this; that
Articles	a; the; 0
Relativisers	that; which; who
Complementizer vs relativizer	that
Duration/start/date	For; since; ago; from; during
Prepositional constructions	For; to
Quantifiers	Some vs any; many vs much vs most; few vs little

Feature Extraction

Modifying L2SCA (Lu 2010)

1. Identifying patterns
2. Counting patterns

Nominal construct extraction

- Determiner genitive
  - Determiner use of N's
    - NP <POS < (DT !< /[a|A]/)
      - gen_dt
- N prep N (with for and of)
  - Two nouns linked by a preposition
    - NP <(PP < (IN [<of | <for]) <NP) <NP
      - n_prep_n
- N2 N1
  - Compound nouns
    - NP < (NN $+/NN.*/)
      - n_n

Determiner extraction

- a
  - DT </^[A|a]$/ $++ /^NN$/
    - art_a
- the
  - DT </^[T|t]he$/ $++ /^NNS?$/
    - art_the
- zero
  - plural noun preceded by zero article
    - /^NN$/ >(NP !<DT !</PRP.*/ !<<POS) $- /^NN$/
      - Art_zero_sing
  - singular noun preceded by zero article
    - /^NNS$/ >(NP !<DT !</PRP.*/)
      - art_zero_plur

Computing metrics

Modified L2SCA

MS nominal constructs = ratios

- gen_dt/(gen_dt + n_prep_n + n_n)
  - ms_multi_noun_gen
- n_prep_n/(gen_dt + n_prep_n + n_n)
  - ms_multi_noun_prep
- n_n/(gen_dt + n_prep_n + n_n)
  - ms_multi_noun_0

Computing metrics

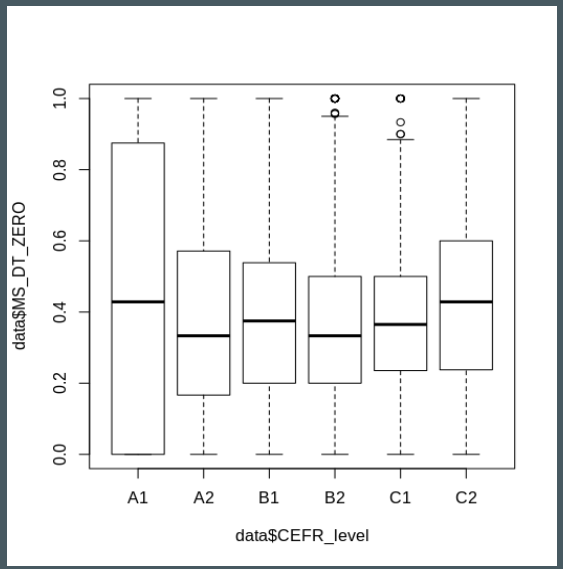
MS_determiner = ratios

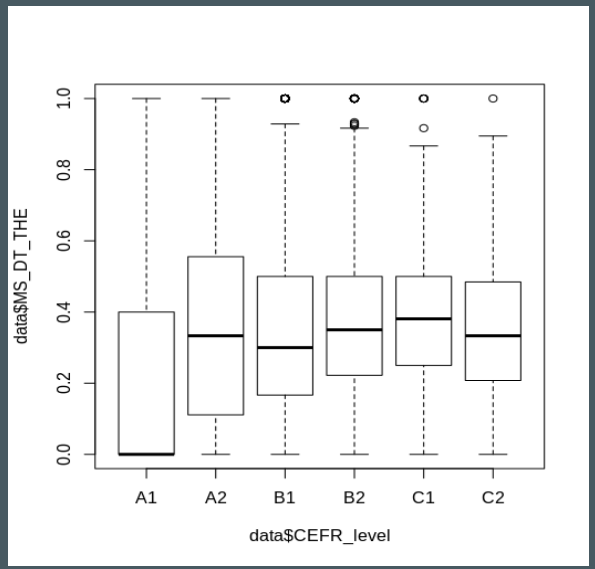
- MS_dt_a = art_a/(art_a+art_the+art_zero_sing+art_zero_plur)
- MS_dt_the = art_the/(art_a+art_the+art_zero_sing+art_zero_plur)
- MS_dt_zero = (art_zero_sing+art_zero_plur)/
  (art_a+art_the+art_zero_sing+art_zero_plur)

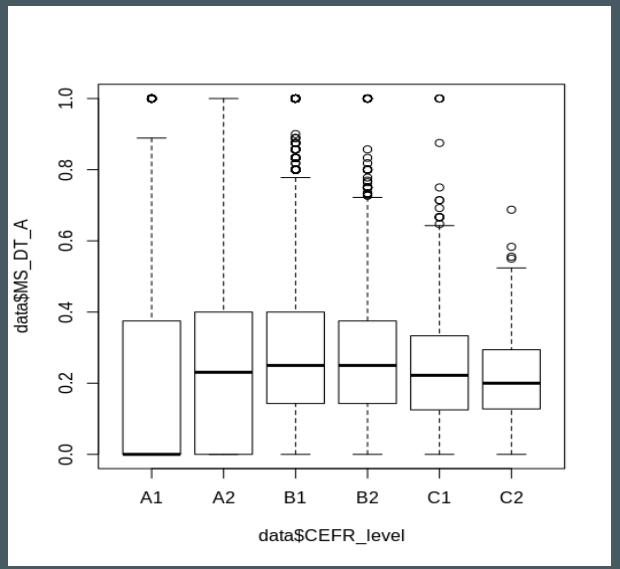
Variations in a microsystem

Proportions of one element in relation to the other of the same system

Assumption: Proportions vary according to individuals: Example of Determiner MS







Discussion

Objective linguistic features correlate with levels

- Insight in Interlanguage.
- Features per learning stage operationalised by CEFR levels

Learner language analysis requires features based on paradigmatic relationships

Next steps

- More L1s, more micro-systems (Tregex)
- More annotation (Dependency)
- Capture more than proportions.

References

Chen, Xiaobin, and Detmar Meurers. 2016. “Characterizing Text Difficulty with Word Frequencies.” Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, 84–94.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. Language Testing, 28(4), 561–580.

Crossley, Scott A., Tom Salsbury, Danielle S. McNamara, and Scott Jarvis. 2011. “Predicting Lexical Proficiency in Language Learner Texts Using Computational Indices.” Language Testing 28 (4): 561–80.

Díaz-Negrillo, Ana, Nicolas Ballier, and Paul Thompson, eds. 2013. Automatic treatment and analysis of learner corpus data. Studies in Corpus Linguistics 59. Amsterdam, Pays-Bas, Etats-Unis: John Benjamins Publishing Co.

Ellis, Rod. 1994. The Study of Second Language Acquisition. Oxford, United Kingdom: Oxford University Press.

Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen. 2013. “Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat).” In Proceedings of the 31st Second Language Research Forum, edited by R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Miguel, A. Tseng, A. Tuninetti, and D Walter. Carnegie Mellon: Cascadilla Press.

Granger, Sylviane, Gaëtanelle Gilquin, and Fanny Meunier, eds. 2015. The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press.

Hawkins, John A., and Luna Filipović. 2012. Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework. United Kingdom: Cambridge University Press.

Khushik, Ghulam Abbas, and Ari Huhta. 2019. “Investigating Syntactic Complexity in EFL Learners’ Writing across Common European Framework of Reference Levels A1, A2, and B1.” Applied Linguistics amy064.

Kim, Minkyung, and Scott A. Crossley. 2018. “Modeling Second Language Writing Quality: A Structural Equation Investigation of Lexical, Syntactic, and Cohesive Features in Source-Based and Independent Writing.” Assessing Writing 37: 39–56.

Kyle, Kristopher, Scott Crossley, and Cynthia Berger. 2018. “The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0.” Behavior Research Methods 50 (3): 1030–46.

Lu, Xiaofei. 2010. “Automatic Analysis of Syntactic Complexity in Second Language Writing.” International Journal of Corpus Linguistics 15 (4): 474–496.

———. 2012. “The Relationship of Lexical Richness to the Quality of ESL Learners’ Oral Narratives.” The Modern Language Journal 96 (2): 190–208.

———. 2014. Computational Methods for Corpus Annotation and Analysis. Dordrecht: Springer.

Pilán, Ildikó, and Elena Volodina. 2018. “Investigating the Importance of Linguistic Complexity Features across Different Datasets Related to Language Learning.” In Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing, 49–58. Santa Fe, New-Mexico: Association for Computational Linguistics.

Tono, Yukio. 2013. “Automatic Extraction of L2 Criterial Lexicogrammatical Features across Pseudo-Longitudinal Learner Corpora: Using Edit Distance and Variability-Based Neighbour Clustering.” In L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis, edited by Camilla Bardel, Christina Lindqvist, and Batia Laufer, 149–176. Eurosla Monographs Series 2. The European Second Language Association.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. Assessing Writing, 18(1), 85–99.

Many thanks to:

Xiaofei Lu

Detmar Meurers

Dora Alexopoulou

Kyle and Crossley

Schmid (Treetagger)

Alternate slides

Extra-details on DataViz

Customisable micro-systems with TregEx: L1-driven features ??

Tregex queries in L2SCA for microsystems

Expressions for nominal constructions:

- n_prep_n='NP <(PP < (IN [<of | <for]) <NP) <NP'
- n_n='NP < (NN $+/NN.*/)'
- #n_of_n='NP <(PP < (IN < of) <NP) <NP'
- #n_for_n='NP <(PP < (IN < for) <NP) <NP'

Code in L2SCA

- div = shortcut_to_count["gen_dt"] + shortcut_to_count["n_prep_n"] + shortcut_to_count["n_n"]
- shortcut_to_count["ms_multi_noun_gen"]= division(shortcut_to_count["gen_dt"],div)
- shortcut_to_count["ms_multi_noun_prep"]= division(shortcut_to_count["n_prep_n"],div)
- shortcut_to_count["ms_multi_noun_0"]= division(shortcut_to_count["n_n"],div)