EUROCALL 2019 Louvain-la-Neuve, Belgium · 28-31 August 2019

Investigating criterial features of learner English and predicting CEFR levels in French and Spanish learners of English

•••

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

PHC Ulysses 2019 France-Ireland

Problem statement

Learning a language (at universty level)

- For individuals > requires regular assessments for both learners and teachers
- For institutions > a growing demand to group learners homogeneously & fast
- CEFR framework

Solution: Automatic Scoring System (ASS) BUT labour intensive, short-context and rule-based exercises and error-focused

Our proposal: AI- based ASS

- Supervised learning approach
- Morphological, Syntactic, semantic and ‘pragmatic’ features of texts.

Research Question

What criterial features (Hawkins & Buttery 2010) can be identified as predictors for CEFR levels?

Outline

Previous studies

Corpus data

Microsystems (and metrics)

Workflow

Current performance of our system

Next steps

Previous work

Converging methods in automatic learner language analysis

- ASS: for learner language (Shermis et al., 2010; Weigle, 2013)

  Shared tasks: Spoken CALL (Baur et al., 2017) & CAp18 (Arnold *et al*. 2018)

- Automatic learner error analysis (Leacock, 2015)
- Automatic learner language analysis: criterial features (Crossley *et al*, 2011; Hawkins & Filipović, 2012)

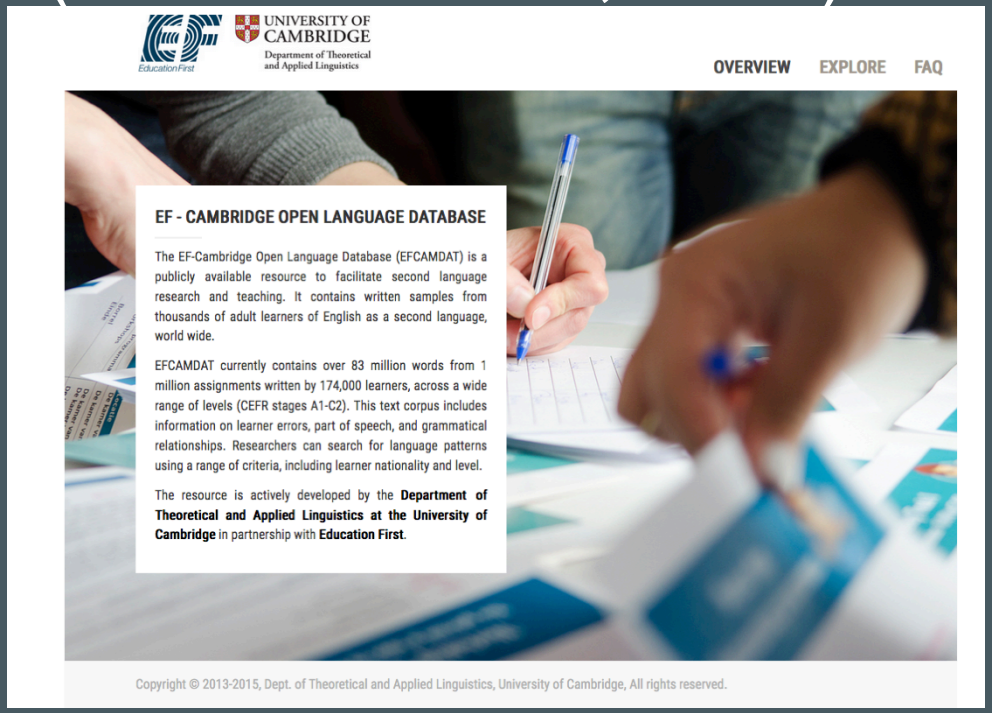
Corpus

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

EFCAMDAT (Geertzen *et al.*, 2013)



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

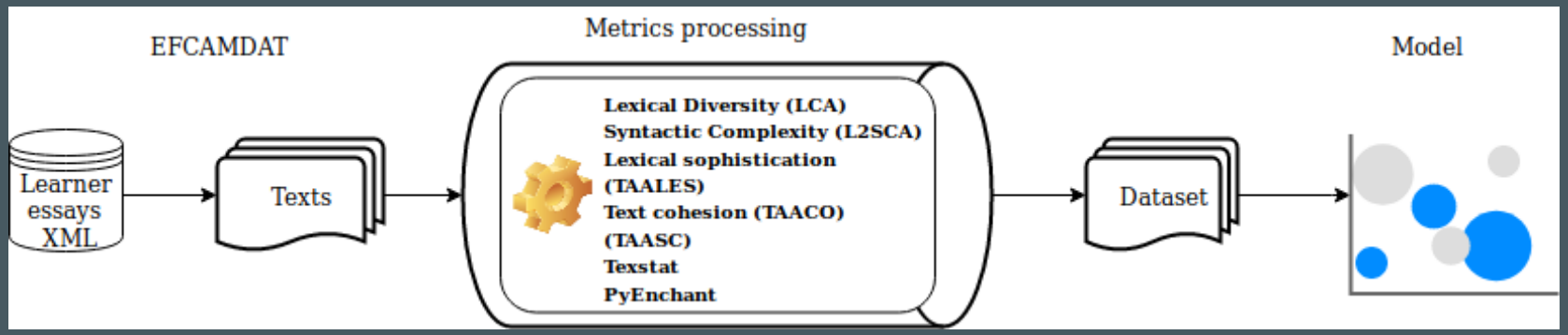
Microsystems	Components
Nominal density	determiner genitive; noun-of/for-noun constructions, compound nouns
Modals for possibility	may; can; might; could
Modals for obligation	must; have to
Proforms	it; this; that
Articles	a; the; 0
Relativisers	that; which; who
Complementizer vs relativizer	that
Duration/start/date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

Processing pipeline and Data set

Features

- 768 metrics for each text
- CEFR levels

Pipeline



Experimental setup

Supervised learning approach

- Dataset: Train (75%) and test set (25%)
- Method: multinomial logistic regression

Stage 1: Classification in CEFR levels without and with Microsystem features

Stage 2: Feature explanation per CEFR level

Results: Stage 1

6 CEFR class classification with MS features

- 80% to 81 % F1-Score accuracy improvement with MS features
- High predictive ability for A1, A2 and B1 learners but poor performance for B2, C1 and C2 learners

	A1	A2	B1	B2	C1	C2	
Pred A1	4506	343	76	28	6	3	
Pred A2	376	2717	326	37	8	0	Pred B1
115	267	2031	313	42	7		
Pred B2	40	27	122	793	120	19	
Pred C1	7	4	12	33	72	4	
Pred C2	0	0	0	0	0	0	

	precision	recall	f1-score	support
A1	0.91	0.89	0.90	5044
A2	0.78	0.81	0.79	3358
B1	0.73	0.79	0.76	2567
B2	0.71	0.66	0.68	1204
C1	0.55	0.29	0.38	248
C2	0.00	0.00	0.00	33
avg / total	0.81	0.81	0.81	12454

Results Stage 1

Binary classification in 2 levels: beginner (A1,A2,B1) and advanced (B2,C1,C2)

- 95% Average F1-Score

	Real Advanced	Real Beginner
Pred Advanced	1150	335
Pred Beginner	258	10711

	precision	recall	f1-score	support
advanced	0.82	0.77	0.80	1485
beginner	0.97	0.98	0.97	10969
avg / total	0.95	0.95	0.95	12454

Results Stage 2

3 Binary classification per level

i.e. A1 vs A2, B1 vs B2, C1 vs C2

-> which metrics come up as best predictors?

Results Stage 2

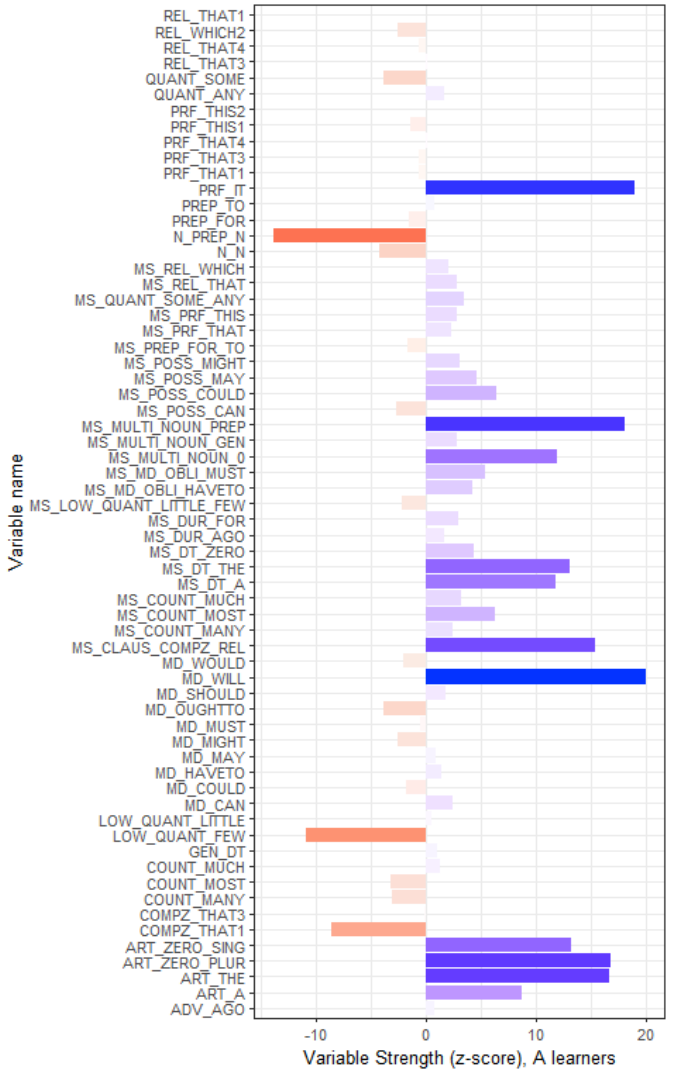
3 Binary classification per level

A1 vs A2

Features: MS nominal constructions and determination

predictor of A2

compared with A1



Results Stage 2

3 Binary classification per level

B1 vs B2

MS quantifiers Few vs little predictor of B1

MS Determination predictor of B1 compared with B2

Modal *should* predictor of B2

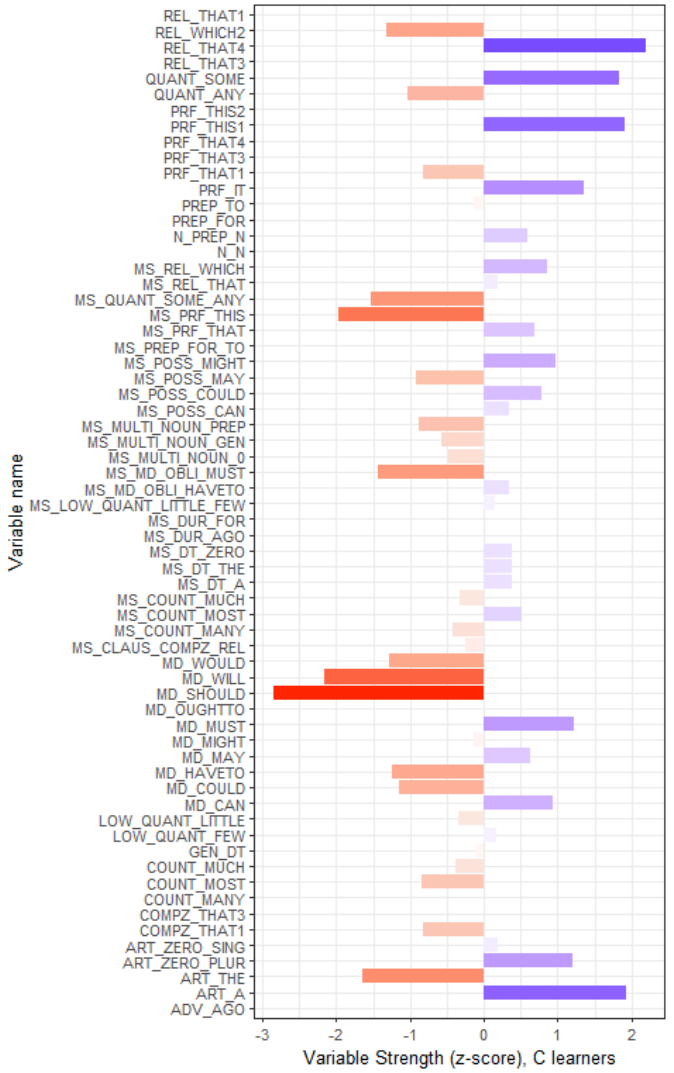
Results Stage 2

3 Binary classification per level

C1 vs C2

MS proforms predictor of C1 compared with C2

Modal should predictor of C1



Discussion

Objective linguistic features correlate with levels

- Insight for Interlanguage (customisable micro-systems)
- Features per learning stage operationalised by CEFR levels

Microsystem features are significant (but limited number of L1s)

Learner language analysis requires features based on paradigmatic relationships

Building pipeline requires CPU power

Next steps

- More L1s, more micro-systems (Tregex-based implementation)
- L1 parameter : NLI ; specific interlanguage settings for L1s reflected in our metrics?
- Reverse engineering : SP vs. FR : L1 prediction on the basis of features
- On-line prototype (Docker : python-based pipeline)

- Workshop in Paris : Oct 31st : website (typed text -> CEFR level prediction)

- From ensemble learning to detailed analysis of metrics
- > from error detection to CALL (feedback and metalinguistic knowledge)

Moodle implementation (and keylog capture) for DataViz

Discussing metrics variability

- Baliskas 2018 on discrepancies between scores for the same metrics according to the implementation (KoRpus versus textstats)

syllable counts versus hyphenation rules : textstats / quanteda / Kyle and Crosley

- - hyphenation (textstats: pyphen: hyphenation rules)
- - quanteda (CMU dictionary)

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:

Dora Alexopoulou (EFCAMDAT)

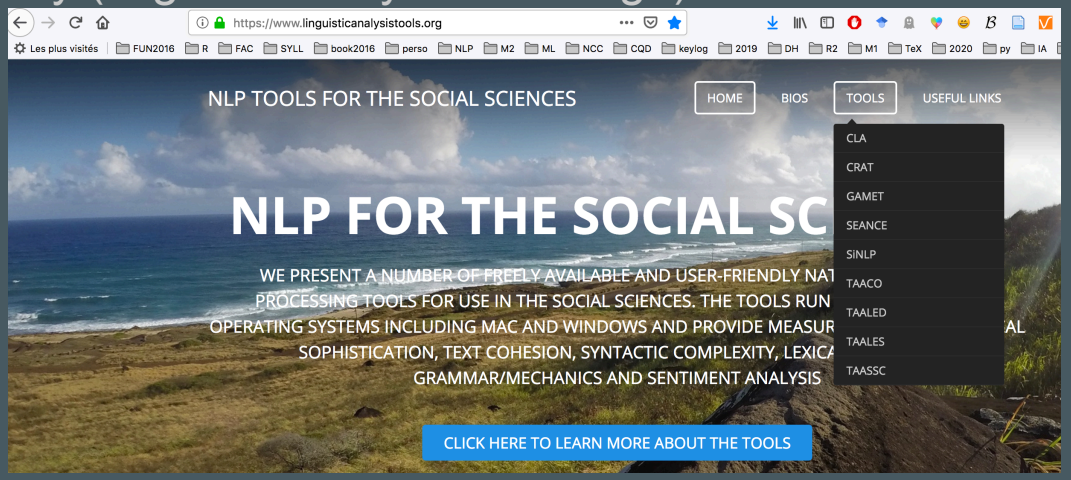
Stéphane Canu (Cap2018 main organisor)

Kristopher Kyle and Scott Crossley (linguisticanalysistools.org/)

Xiaofei Lu (L2SCA)

Detmar Meurers (CTAP)

Helmut Schmid (Treetagger)



Questions ?

- Thank you from the team (Galway)

PHC
Hubert Currien
Ulysse 2019
(ref 43121RJ)

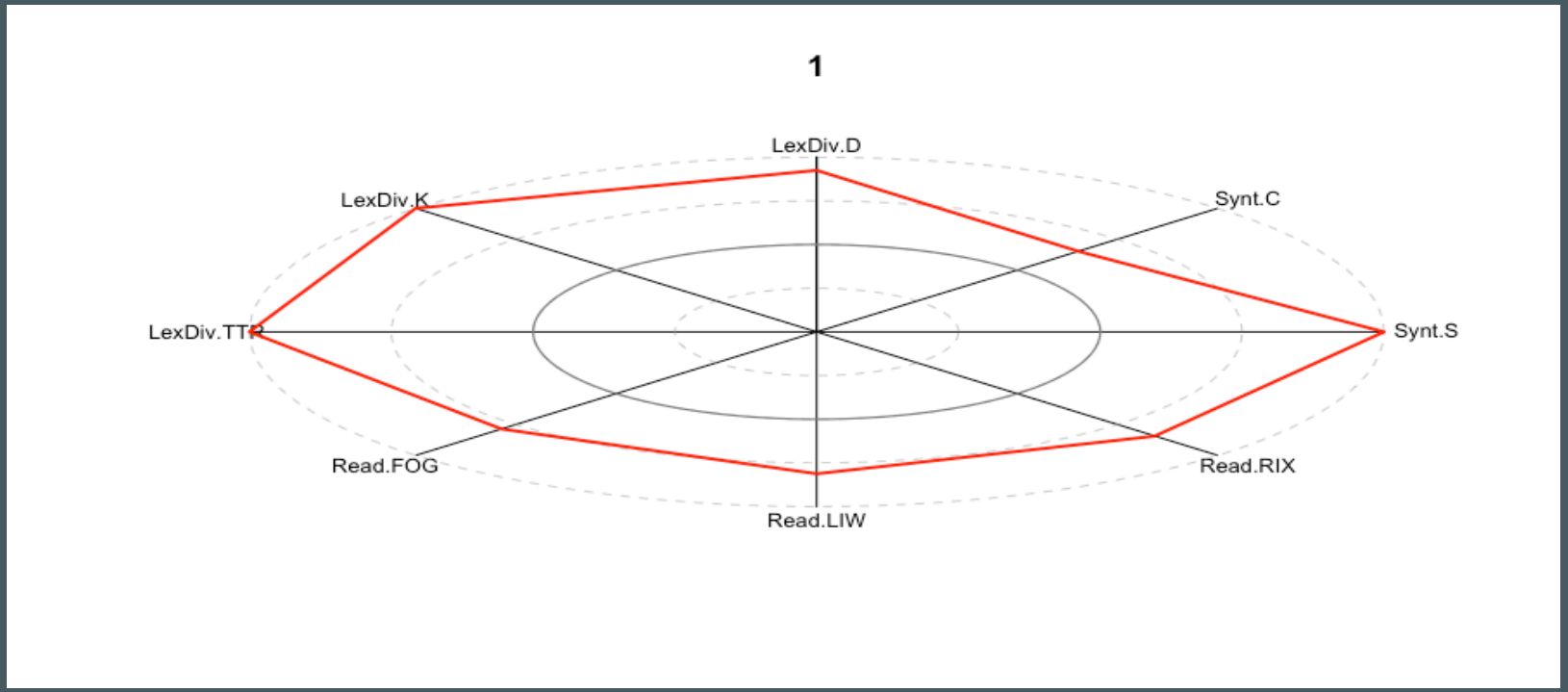
Alternate slides

Extra-details on DataViz

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

More on micro-systems

DASHBOARD : using metrics for feedback



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)

Micro-systems (articles)

(1) "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)

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

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

Relevance of micro-system features

Models	Without microsystem variables				Including microsystem variables			
Level	Precision	Recall	F1	Support	Precision	Recall	F1	Support
A1	0.89	0.90	0.89	5087	0.89	0.90	0.89	5087
A2	0.76	0.76	0.76	3346	0.77	0.77	0.77	3346
B1	0.70	0.78	0.74	2530	0.72	0.78	0.76	2530
B2	0.65	0.60	0.62	1184	0.68	0.63	0.66	1184
C1	0.57	0.17	0.26	273	0.58	0.27	0.37	273
C2	0.00	0.00	0.00	31	0.01	0.01	0.01	31
Mean	0.78	0.79	0.78	12451	0.78	0.79	0.78	12451