Vizualisations of linguistic profiles in learner written productions

Thomas Gaillat[1], Nicolas Ballier[2], Antoine Lafontaine[2] & Anas Knefati[3]
[1] Université Rennes 1&2 [2] Université de Paris (Paris Diderot) [3] ENSAI

Introduction

Motivation:
Language acquisition
- Provide regular proficiency assessment reports to teachers and learners.
- Institutions need to group students homogeneously for adapted teaching methods.

Task:
Describes learners' linguistic measurements according to CEFR levels.

Objective:
Offer immediate and individual feedback in terms of linguistic characteristics (lexical & syntactic complexity ...)

Data

Learners:
• 286 productions ESP learners collected at Université de Rennes 1, Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp)
• 62 724 tokens
• Texts classified according to CEFR classes - 2 grades :
- Student results from Written Dialang (Alderson & Huhta, 2005) & graded .
- Text graded by 2 expert teachers. (IR agreement 0.71 n =50)

	A1	A2	B1	B2	C1	C2
Informatique et électronique	6	12	46	12	4	1
Médecine	0	5	23	19	2	0
Pharmacie	2	20	49	10	13	2
Sciences de la vie et de l'environnement	20	29	9	2	0	0

	A1	A2	B1	B2	C1	C2
L1	12	8	3	1	0	0
L2	8	21	6	1	0	0
L3	8	17	69	31	6	1
M1	0	20	49	10	12	2
M2	0	0	0	0	1	0

Tableau 1. Distribution per domain (left), per university level (right).

Natives:
- 20 comparable written essays from Lancaster University students extracted from ICE-GB (Nelson, Wallis & Aarts, 2002)
- 43 000 tokens, ~ 2 100 tokens per text

Processing pipeline

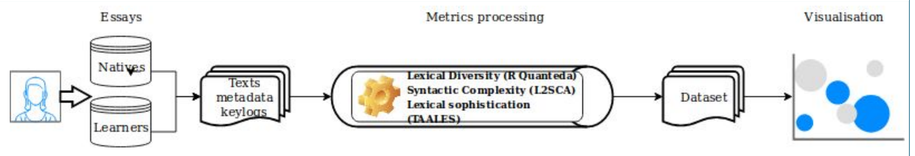

Fig.2 : Data colelcted in MOODLE and processed in R

1. Texts are typed in by student in MOODLE
2. Data extraction (texts and keystrokes)
3. Processing with R Quandeta (Benoit et al.,2018) , L2SCA (Lu, 2010), et TAALES (Kyle, Crossley & Berger, 2018)
4. Creation of visualisations

Method

Specific metrics selected to compare individuals with groups (learners of specfic level or natives (**Fig.2**):
- Lexical diversity: K indicating word repetition and Corrected Type Token Ratio (TTR), NDW.
- Readability: FOG, LIW et RIX selon Lissón & Ballier, 2018
- Syntactic complexity: CN/C, C/S, DC/C et MLC obtained with L2SCA (Lu, 2010).

Discussion

- Most significant metrics to be determined
  vs
- Most relevant in terms of feedback clarity

Associations between CEFR grades and metrics (p-values)

CTTR	0.0000 ***	NDW	0.0000 ***
W	0.0000 ***	MLC	0.6679
S.1	0.0000 ***	MLT	0.0020 *
T	0.0000 ***	CN.C	0.6172
FOG	0.0467 *	CN.T	0.0407 *
RIX	0.0112 *	CP.C	0.9942
K	0.0000 ***	CP.T	0.4660

Conclusion and future developments

This tools provides feedback automation & enhances thinking over language usage.

Future directions:
- Integrate level prediction and visualizations with recommender system

Results

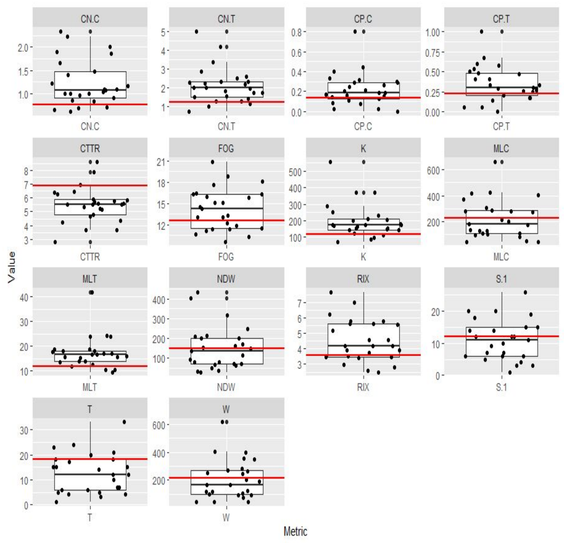

Fig. 2 : Visualization of a learner's scores in relation to 20 native scores

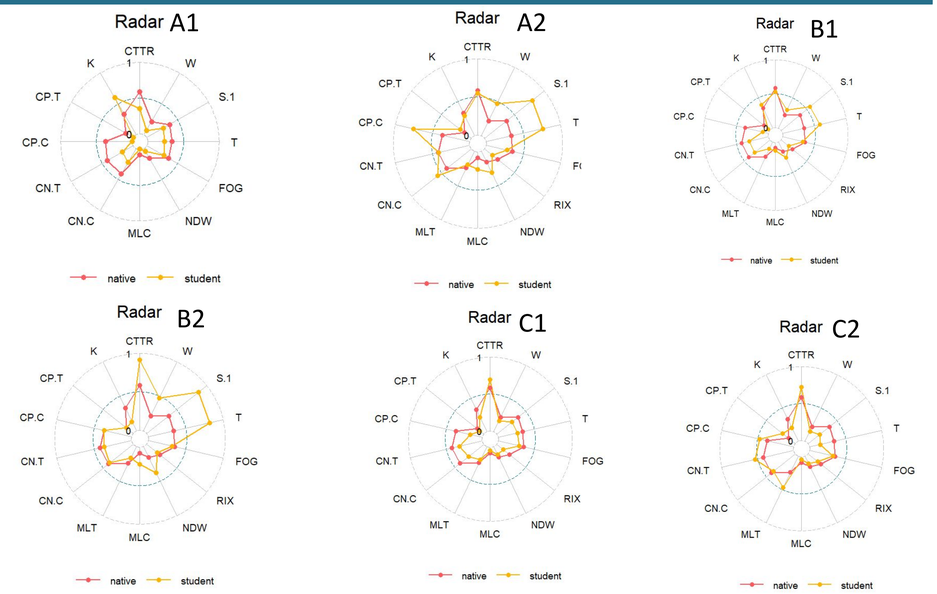

Fig. 2 : A learner's scores in a radar chart. The native line indicates an average over 20 natives..

Contact

nicolas.ballier@univ-paris-diderot.fr
thomas.gaillat@rennes1.fr

References

J. C. Alderson, and A. Huhta. "The Development of a Suite of Computer-Based Diagnostic Tests Based on the Common European Framework." Language Testing vol. 22, no. 3 (2005) 301–20
K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo, "Quanteda: An R Package for the Quantitative Analysis of Textual Data," Journal of Open Source Software, vol. 3, no.30, (2018), 774.
A. Díaz-Negrillo, M. Callies, C. Lozano, "Designing and compiling a learner corpus of written and spoken narratives: The "Corpus of English as a Foreign Language" (COREFL)". ARISLA workshop (Anaphora Resolution in Second Language Acquisition, Universidad de Granada (España), (2018)
K. Kyle, S. Crossley, and C. Berger, "The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0," Behavior research methods, vol. 50, no. 3, pp. 1030–1046, (2018).
P. Lissón et N. Ballier 'Investigating lexical progression through lexical diversity metrics in a corpus of French L3', *Discours,* 23, 2, December 2018, à paraître début juin 2019
X. Lu, "Automatic analysis of syntactic complexity in second language writing," International journal of corpus linguistics, vol. 15, no. 4, (2010), 474–496.
G. Nelson, S. Wallis, and B. Aarts, Exploring natural language: working with the British component of the International Corpus of English, vol. 29. John Benjamins Publishing, (2002)