On the link between L2 learner's vocabulary knowledge and pronunciation accuracy: a corpus-based study

Paolo Mairano, Fabian Santiago

Mairano, P. & Santiago, F. (2019) On the link between vocabulary knowledge and pronunciation accuracy: a corpus-based study. *Actes des 10èmes Journées Internationales de Linguistique de Corpus*, 26-28 novembre 2019, Grenoble.

On the link between L2 learner's vocabulary knowledge and pronunciation accuracy: a corpus-based study

Paolo Mairano [2], Fabian Santiago [2]
[1] STL UMR 8153, Université de Lille
[2] SFL UMR 7023, Université de Paris 8
paolo.mairano@univ-lille.fr, fabian.santiago-vargas@univ-paris8.fr

Introduction

In recent years there has been growing evidence that measures of vocabulary size are good predictors of L2 competence, and vocabulary tests are therefore often used as a quick evaluation of L2 proficiency level (Meara, 2010; Milton, 2013). In effect, vocabulary size has been shown to correlate strongly with reading, writing, and listening skills (Stæhr, 2008); this is of course grounded on the fact that knowledge of a higher number of words is likely to result in a better comprehension of text. However, vocabulary tests may give a smaller indication of learners' speaking skills, and in particular pronunciation accuracy. In effect, evidence about the correlation between vocabulary size and speaking skills is scant. Koizumi and In'nami (2013) report on 9 existing studies, most of which evaluate speaking skills in terms of fluency (speech rate, length of utterances, etc.), never in terms of pronunciation accuracy, foreign accentedness or intelligibility – something that may be due to the lack of standard metrics for assessing any aspect of L2 pronunciation. The only existing studies investigating the relation between vocabulary knowledge and L2 pronunciation are in fact very recent. Uchihara & Saito (2019) found significant correlations of productive vocabulary size with speech rate but not with ratings of accenteness and comprehensibility for Japanese learners of L2 English. Similarly, in our previous study (Mairano & Santiago, forthcoming), we found that a measure of receptive vocabulary size (Dialang vocabulary test) showed low to medium correlations with speech rate but not with ratings of foreign accentedness, nor with acoustic measures of vowels, for Italian learners of L2 French. Additionally, we computed metrics of lexical diversity (vocd-D, MTLD, MTLD-MA, cf. McCarthy & Jarvis, 2010) from learners' productions as an indication of their productive vocabulary size and found that they did not correlate significantly with any pronunciation measure.

The aim of this study is to expand the investigation reported by Mairano & Santiago (forthcoming), by computing learners' lexical profiles and verifying their correlation with various L2 pronunciation metrics. In our previous study, we used lexical diversity metrics to estimate learners' productive vocabulary size, as in many other studies (e.g., Arnold et al., 2018). However, low lexical diversity in learners' productions does not necessarily imply small vocabulary size, and some authors have suggested that lexical profiles may give a better indication of productive vocabulary size (Laufer & Nation, 1995: Edwards & Collins, 2011). This is because low proficiency learners tend to reuse frequent lexical items, while more proficient learners tend to use more infrequent lexical items. We therefore expand our previous analysis by computing learners' lexical profiles and verifying if they correlate with our L2 pronunciation metrics.

Corpus and methodology

Corpus

We used the Italian section of the *ProSeg* corpus (Delais-Roussarie et al., 2018), which includes recordings of 25 Italian learners of L2 French in a university setting. Students of L2 French (21 females and 4 males; B1 to C1 levels) at the University of Turin (Italy) were recorded in a sound-proof booth, thereby guaranteeing high-quality audio, apt for acoustic analysis. All participants signed a consent form and filled a questionnaire gathering information about their acquisition process and other useful sociolinguistic information. They were asked to perform the following tasks:

- a read-aloud task of 8 short passages in French (907 words in total)
- a read-aloud of a longer passage in French
- a picture description task
- a monologue (telling a film/book/holiday)
- Dialang vocabulary test
- a read-aloud task of 8 short passages in Italian

The audio was transcribed orthographically and an automatic transcription was forced-aligned to the signal via *EasyAlign* (Goldman, 2011) and subsequently manually checked on *Praat* (Boersma & Weenink, 2018) for all reading tasks and for the initial 5 minutes of the semi-spontaneous tasks (picture description and monologue), taking care to preserve transcriptions that were as close as possible to the target phonemes.

Learners' lexical profiles

In order to quantify learners' productions from the point of view of word frequency, morphological complexity and phonological complexity, we analysed the first 5 minutes of each learner's production for the picture description task. After lemmatisation of our orthographic transcription with *TreeTagger* (Schmid, 1995), we computed the following metrics on the list of words and lemmas used by every speaker (with reference to the *Lexique 383* corpus, cf. New et al., 2005):

- percent of words ranking >4000, >3000, >2000 and >1000 in the *Lexique* corpus;
- percent of lemmas ranking >3000, >2000 and >1000 in the *Lexique* corpus;
- percent of words with >2 and >1 morpheme (excl. inflectional suffixes);
- percent of words longer than 10, 8, 6 and 4 phonemes.

Measures of L2 pronunciation

We used various measures of L2 pronunciation, all presented in detail in our previous study (Mairano & Santiago, forthcoming):

- fluency was evaluated in terms of speech rate (SR, phon/sec incl. pauses), articulation rate (AR, phon/sec excl. pauses) and inverted number of pauses (NP);
- global foreign accent was evaluated via ratings of foreign accentedness (FA) on a 5-point Likert scale provided by 3 native French phoneticians (*ICC* = .89) based on 8 sentences extracted from every learner's productions;
- nasal vowels were evaluated via ratings of nasality for /ɛ̃, ɑ̃, ɔ̃/ on a 5-point Likert scale provided by 3 native French phoneticians (*ICC* = .88, .67, .85 respectively for each vowel) based on 9 words extracted from every learner's productions;
- the degree of distinctness of the problematic /y - u/, /ø - e/, /œ - ɛ/ vowel pairs was evaluated via acoustic distances (D) and Pillai scores (P) (Hall-Lew, 2010), following the approach proposed by Mairano et al. (2019) for L1 English.

Results and discussion

Firstly, we analysed the relation among our lexical variables by computing a correlation matrix and plotting it as a Fruchterman-Reingold graph via the *qgraph* package (Epscamp et al., 2012). As shown in figure 1, metrics relating to word frequencies (in green) and lemma frequencies (in blue) are all strongly correlated with each other. Additionally, they correlate (more moderately) with metrics of morphological (in yellow) and phonological (in rose) complexity, but not with measures of lexical diversity (in orange) nor with learners' scores for the Dialang test.

Finally, we calculated the correlation between pronunciation measures and lexical metrics computed on productions of our learners. As shown in figure 2, correlations are very low and mostly do not reach statistical significance: only measures of phonological complexity correlate at $r > .4$ and significantly with (some) pronunciation metrics, suggesting that good pronouncers tend to use longer words. Instead, correlations involving lexical frequency or morphological complexity are low and non-significant, reflecting our previous results obtained with Dialang scores and metrics of lexical diversity. This may be due to the limited dataset and other methodological limitations outlined in Mairano & Santiago (forthcoming); but, for the moment, we cannot provide any tangible proof of a relation between the acquisition of vocabulary and L1 phonology.

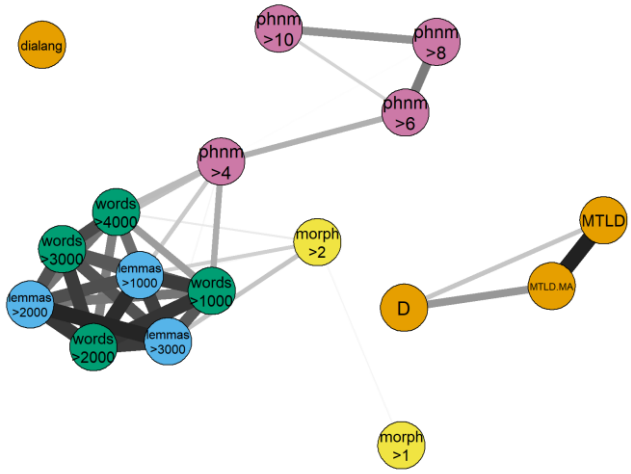

figure . 1 **Fruchterman-Reingold graph showing the relation among lexical variables computed on learners' productions. Lines are plotted only for significant correlations.**

	FA	D u-y	D e-ø	D ɛ-œ	P u-y	P e-ø	P ɛ-œ	ɛ̃	ɑ̃	ɔ̃	AR	NP	SR
% words with >10 phonemes	0.27	0.51	-0.15	-0.17	0.31	-0.00	-0.03	0.23	0.21	0.29	0.21	0.08	0.17
% words with >8 phonemes	0.09	0.42	-0.22	-0.22	0.28	0.07	-0.22	0.40	0.22	0.04	0.01	-0.13	-0.04
% words with >6 phonemes	0.10	0.33	-0.03	-0.09	0.17	0.32	-0.07	0.17	0.41	0.05	-0.06	-0.06	-0.12
% words with >4 phonemes	-0.36	-0.16	-0.15	-0.21	-0.18	0.03	-0.29	-0.05	0.17	-0.07	-0.28	-0.18	-0.28
% words with >2 morphemes	-0.25	0.04	-0.08	-0.11	0.03	0.26	-0.03	0.01	0.11	-0.05	-0.21	-0.39	-0.25
% words with >1 morphemes	-0.22	-0.18	-0.01	-0.06	-0.14	-0.02	-0.10	-0.01	0.02	-0.16	0.00	-0.31	-0.05
% words with rank > 4000	-0.09	0.01	0.08	-0.02	0.02	0.24	0.13	0.16	0.15	0.01	-0.20	-0.21	-0.18
% words with rank > 3000	-0.27	-0.12	0.20	0.04	-0.12	0.35	0.22	0.02	-0.04	0.03	-0.23	-0.09	-0.23
% words with rank > 2000	-0.14	-0.01	0.34	0.21	-0.00	0.30	0.32	0.06	-0.19	0.19	-0.23	-0.03	-0.21
% words with rank > 1000	-0.45	-0.22	0.14	0.01	-0.22	0.18	0.13	0.01	-0.27	0.02	-0.31	-0.08	-0.27
% lemmas with rank > 3000	-0.18	-0.07	0.24	0.08	-0.10	0.19	0.22	0.15	-0.03	0.29	-0.23	-0.11	-0.22
% lemmas with rank > 2000	-0.07	0.06	0.30	0.15	0.08	0.25	0.31	0.12	-0.21	0.21	-0.17	-0.09	-0.14
% lemmas with rank > 1000	-0.23	0.02	0.14	0.07	0.01	0.36	0.26	0.04	-0.05	0.16	-0.27	-0.21	-0.25

figure . 2 **Correlations (Pearson's *r*) between lexical variables (rows) and pronunciation measures (columns).**

References

Arnold, T., Ballier, N., Gaillat, T. and Lissón, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts. *Proc. of the CAP conference (Conférence sur l'Apprentissage Automatique)*, arXiv:1806.11099

Boersma, P. & Weenink, D. (2019). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.49, retrieved 2 March 2019 from http://www.praat.org/

Delais-Roussarie, E., Kupisch, T., Mairano, P., Santiago, F. & Splendido, F. (2018) ProSeg: a comparable corpus of spoken L2 French. Poster presented at *EuroSLA*, 5-8 September 2018, Münster (Germany).

Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's law. *Language Learning*, 61(1), 1-30.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1-18.

Goldman, J. Ph. (2011). EasyAlign: a friendly automatic phonetic alignment tool under Praat. *Proc. of the 12th INTERSPEECH 2011*, 3233-3236.

Hall-Lew, L. 2010. Improved representation of variance in measures of vowel merger. *Proc. of Meetings on Acoustics* (Vol. 9, No. 1).

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, *4*(5), 900-913.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, 16(3), 307-322

Mairano, P., Bouzon, C., Capliez, M. & De Iacovo, V. (2019). Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness. *Proceedings of ICPhS2019 (International Congress of Phonetic Sciences)* (pp. 1104-1108), Melbourne (Australia), 5-9 August 2019.

Mairano, P. & Santiago, F. (forthcoming) What vocabulary size tells us about pronunciation skills: Issues in assessing L2 learners. *Journal of French Language Studies*.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392.

Meara, P. (2010). *EFL vocabulary tests* (2nd ed.). ERIC Clearinghouse.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (eds.) *Eurosla Monographs Series*, 2, 57-78.

New, B., Pallier, C., & Ferrand, L. (2005). Manuel de Lexique 3. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524.

Richards, B. J. & Malvern, D. (1997). *Quantifying lexical diversity in the study of language development*. Reading: Faculty of Education and Community Studies.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Uchihara, T. & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, 47(1), 64-75.