L1 and L2 Typological Distance Effects on the Learnability of Articles in L2 English: A Large-Scale Learner Corpus Analysis

Doğuş Öksüz[(1)] Kate Derkach[(1)], Dora Alexopoulou[(1)]

[(1)] University of Cambridge,

Correspondence concerning this paper should be addressed to Doğuş Öksüz (dco24@cam.ac.uk)

Recent work by Schepens and colleagues shows that the linguistic distance~~s~~ between learners' L1s and additional languages learned later in life (i.e. L2s and L3s) can predict proficiency scores in speaking tests (Schepens, Van Hout & Jaeger, 2020). This finding, based on over 56 L1s, shows that not only individual features, but also the aggregate of L1-L2/L3 similarities or differences for a range of linguistic features (lexical, morphological or phonological) influence learning outcomes. One question arising is how linguistic distance might affect the acquisition of individual features, rather than broad outcomes like speaking proficiency scores. Does the acquisition of individual features depend solely on the presence/absence of a congruent element in the L1 (e.g. Murakami & Alexopoulou, 2016), or do broader typological differences guide the way learners approach the input and influence the acquisition of individual features? To address these questions, we focus on the acquisition indefinite articles in L2 English. Importantly, adult L2 learners often experience difficulty in the acquisition of articles with large individual variation (Murakami & Alexopoulou, 2016).

Method: a) *Measuring the linguistic distance.* We draw a distinction between *general distance*, capturing broad typological realisations and *domain distance* capturing variation in the realisation of nominals and articles. For example, we examine whether properties like count/mass distinction or classifiers influence the acquisition of articles. To measure the syntactic variation, we use the generative parameter as our unit of measurement and calculate identities (*i*) and differences (*diff)* between L1 and L2 regarding features of parametric variation (Longobardi & Guardiano, 2009). Specifically, we divide the number of identities by the sum of identities and differences (LD= (*/i*+*diff*).
b) Data. We used a subset of the EFCAMDAT, a written, learner, error-tagged corpus of English. The subset was 37 million words in size, including writings from 112,064 different learners. We targeted 11 native-language groups: Portuguese, Chinese, German, French, Italian, Japanese, Arabic, Russian, Mexican Spanish, Korean and Turkish, with proficiency levels from A1 to B2 (see also Tables 1 and 2). We selected these groups to provide a typologically diverse set for comparison, at the same time choosing languages with a sufficient amount of data. As a measure of accuracy, we employed suppliance in obligatory context scores - the ratio between correct uses and omission errors. We first obtained obligatory contexts in general and then specified existential contexts. We retrieved error-tagged texts and converted them to corrected texts.

Results. Linear mixed-effects models are used to determine the extent to which linguistic distances between learners' L1s and L2 alongside the absence or presence of articles in learners' L1s affect developmental patterns of accuracy. Accuracy scores were modelled as a function of several variables including L1-L2 distance (general and domain), proficiency, and absence/presence of articles in L1s. Preliminary results show that the presence or absence of articles in learners' L1s affects the overall accuracy of article use (see Figure 1). This lends empirical supports to previous findings that learners whose L1s have articles reached higher accuracy scores (Murakami 2013). We also observed some effects of contexts – learners reach higher accuracy scores when providing indefinite articles in existential sentences like 'there is a teacher in the entrance'. We observed that existential contexts are not vulnerable to L1 influence. As a way forward, we will focus on the potential effects of linguistic distances between learners' L1s and L2 English on the accuracy of article use.

References
Longobardi, G. & Guardiano, C. (2009). Evidence for syntax as a signal for historical relatedness. *Lingua 119*(11), 1679 – 1706.

Murakami, A., & Alexopoulou, T. (2016). Longitudinal L2 development of the English article in individual learners. In A. Papafragou, D. Grodner, J. Trueswell, & D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1050-1055).

Schepens, J., van Hout, R., & Jaeger, F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition 194,* 104056.

Table 1. Target L1 Groups

L1 Groups	Number of learners	Number of writings	Number of words produced
Brazilian Portuguese	52259	295144	17450610
Mandarin Chinese	14044	80378	5012814
German	6364	29393	2302805
French	5490	23391	1664064
Italian	5746	26459	1929994
Japanese	2558	12360	837859
Arabic	6935	29450	1616648
Russian	8168	35408	2437060
Mexican Spanish	8810	52238	3160870
Korean	711	2759	200093
Turkish	1979	7810	488202

Table 2. Proficiency levels

Proficiency levels	Number of learners	Number of writings	Number of words produced
A1	64805	305748	12894531
A2	37067	164035	10929844
B1	21617	92141	8837698
B2	8898	32866	438846

Note. For more information about the EFCAMDAT, please visit here

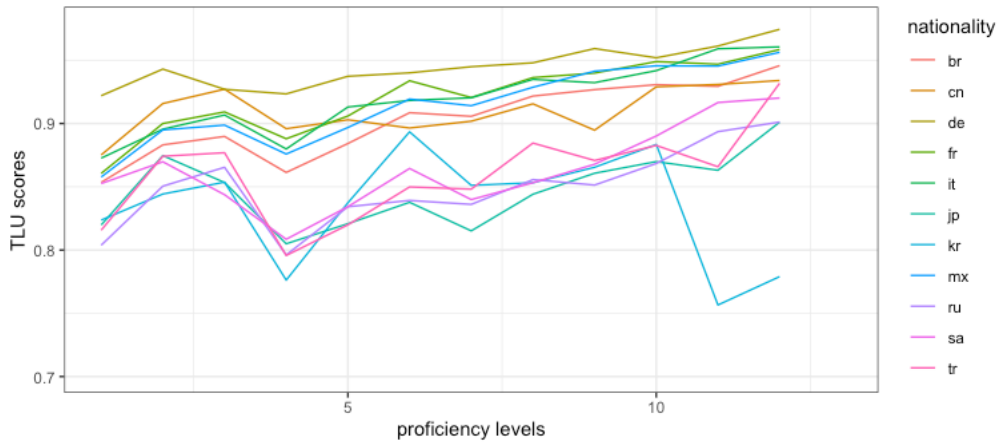

Figure 1. The accuracy of indefinite articles by L1 groups