Corpora-based lexicography in service of overcoming errors in the production of prepositions in L2

Noam Ordan1,2, Omaima Abboud2, Rani Abboud3, Yifat Ben-Moshe1, Ilan Kernerman1

1 K Dictionaries, Tel Aviv

2 The Academic Arab College of Education, Haifa

3 Technion – Israel Institute of Technology, Haifa

3   
E-mail: noam@kdictionaries.com, omaima.abboud@gmail.com , rani.3abboud@gmail.com , [yifat@kdictionaries.com](mailto:yifat@kdictionaries.com)

Abstract

We discuss the main results of a corpus-based study on the use of prepositions in L2 production and make suggestions on how to treat them in bilingual lexicography.

Keywords: parallel corpora; learner corpora; bilingual corpora

1. Introduction

A recent study conducted by Ordan et al. (work in progress) presents the difficulties and the ensuing errors in the production of prepositions in English as a foreign language for speakers of Arabic. This brief paper discusses the major issues of the study and suggests lexicographic approaches for overcoming the problems raised

1. The study and its findings

Parallel corpora provide multiple translations of words and phrases from L1 to L2 (and vice versa). Ordan et al. used an alignment algorithm (Och, 2003) that traverses Arabic-English parallel lines taken from a large corpus (Tiedemann and Nygaard, 2004). The output is a table that includes the word and phrase in L1, their translations to L2, and the probability in which these translations occur.

After extracting the probabilities of translation candidates from a parallel corpus, the authors moved on to learners’ corpora. The first corpus accumulates the TOEFL exams of students with 11 different mother tongues, including Arabic (Tetreault et al., 2013). The high probability of generating *in* given في is reflected in the common use of phrases such as “in [at] the same time” by many Arab students in the TOEFL exam, and can be seen in Figure 1. Such usage echoes the various English equivalents where في (or the less formal proclitic ب) is used. Out of 53 occurrences of the phrase “the same time”, 37% are preceded by the preposition *in* (the rest are correctly preceded by *at*).

Figure 1: Frequency rates of the English translations of the Arabic preposition في.

The authors then focused on a small-scale learner corpus compiled from short essays in English written by applicants to the Academic Arab College for Education in Haifa (ArabCC). The subjects consisted of 33 female high-school graduates from the north of Israel, aged 18-24. All of them were native-speakers of Arabic, and fluent in Hebrew and English as additional languages. Their writing task was untimed and was administered to each candidate individually. On average, the task took about 30 minutes to complete, and the average length of the essays was 176 words.

The findings emphasize that many prepositional errors reflect a wrong choice of the English preposition, tending to follow the most probable translation of the target Arabic preposition, e.g. “in [by] the mail”, “in [at] that level”, “in [during] the break”. Other errors ensued from redundantly adding *in* and *on* where the English verb actually takes a direct object, e.g. “convince them in the importance of”, “enjoy in her job”, “affect on”.

The frequency breakdown of the prepositions *in*, *at* and *on* in Arabic applicants’ corpus is 140, 21 and 21, respectively. Thus, the 33 students were six to seven times more likely to use the preposition *in* than *at* or *on*. The authors normalized the frequencies of these prepositions by the total number of prepositions in the whole corpus and compared it to a web-crawled corpus whose size is approximately 19 billion tokens collected over many registers and assumed to give a more general representation of English use (English Web, cf. Jakubíček et al., 2013). Table 1 indicates the total number of prepositions in each corpus, showing that Arab learners are 1.8 more likely to use *in* than the alleged English speaker. This high rate of overuse indicates that this word fulfils the role of other optional prepositions and, indeed, the authors noted that *on* is used only 0.62 as frequently in ArabCC compared to English Web.

|  |  |  |
| --- | --- | --- |
| **preposition** | **English Web** | **ArabCC** |
| * *in* * *on* * *at* | 13.27%  5.77%  3.46% | 23.39%  3.58%  3.58% |

Table 1: The occurrences of *in*, *on* and *at* normalized by the total number of prepositions

1. Suggestions

The findings reported above are connected to two vulnerable aspects of bilingual lexicography: (a) the same type of errors tend to be made by the majority of learners and they correlate to (or reflect) the frequency of the English preposition as an equivalent of the Arabic one; and (b) the range of possible translations is wider than the information usually provided in (learners’) dictionaries, including digital versions.

To address these problems and improve the function of bilingual learners’ dictionaries for production purposes, we suggest to enrich the information that relates to the L2 translation equivalents in the following ways:

1. Provide an indication of the degree of frequency of the translation equivalents by ordering the translations according to their frequency rate (from more to less frequent) andapply visual means, such as different font color intensity for the translations (starting with stronger color for higher frequency and reducing the intensity according to diminishing frequency; similar color intensity would indicate that the translations are of equal frequency).
2. Provide as many examples as possible for use of the L2 item in various contexts (including expressions, collocations, etc.) based on parallel corpora of the relevant language pair, thus reducing the misuse of prepositions as described above, by enabling learners to look for L2 expressions of similar semantic and syntactic environments as those they have in mind in L1.

In addition, indicating the sense distribution of the translation equivalents could facilitate the production in L2, both for the sake of capturing the exact quality and range of a specific meaning and for avoiding ambiguity. The inclusion of such information in digital bilingual dictionaries has become quite standard nowadays, and can be obtained easily by linking the translations to their corresponding L1 dictionary entries.

1. References

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The tenten corpus family. In 7th International Corpus Linguistics Conference (pp. 125-127).

Och, F. J. (2003). Giza++ software.

Ordan, N., Abboud, O., and Abboud, R. (work in progress). Looking at *translation potential* for second language acquisition.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 48-57).

Tiedemann, J., and Nygaard, L. (2004). The OPUS Corpus-Parallel and Free: http://logos. uio. no/opus. In LREC.