Longitudinal investigation of vocabulary development in learner writing

Stefania Spina (University for Foreigners Perugia, Italy) and Anna Siyanova-Chanturia (Victoria University of Wellington, New Zealand)

Recent years have seen a surge of interest in learner corpus research (e.g. Granger, Gilquin, & Meunier, 2015; Paquot & Granger, 2012). Much of this research has focused on the use and development of vocabulary, both single words and collocations. However, much of the work has been done with learner data collected at one point in time. Longitudinal studies are still rare, such that researchers have called for a greater emphasis on studies conducted over a period of time with the same group of learners (e.g. Laufer & Waldman, 2011; Paquot & Granger, 2012). In addition, many of the current studies are limited in a number of ways. First, researchers have for most part focused on upper intermediate, or advanced learners. How the many and varied aspects of vocabulary use (e.g. the use of single words, collocational knowledge development, lexical diversity, etc.) develop in less proficient learners is still poorly understood. Second, the topics in the groups being compared (e.g. less vs. more proficient second language [L2] learners, native vs. non-native writers, etc.) have often not been controlled for. This could have potentially introduced a confound affecting the results reported. Third, most of the longitudinal studies that have looked at production of L2 collocation have employed only a handful of participants (e.g. Crossley & Salsbury, 2011; Li, Eskildsen, & Cadierno, 2014; Li & Schmitt, 2009; Yuldashev, Fernandez, & Thorne, 2013; but see Siyanova-Chanturia, 2015). Thus, most of the studies looking at collocation use and development have been case studies or cross-sectional studies, potentially limiting our understanding of the learning process. In addition, it is noteworthy, that the majority of learner corpus studies with a focus on vocabulary have looked at English as a L2. Relatively few studies have investigated L2s other than English.

To address the above concerns in the current learner corpus research focusing on vocabulary, a large-scale[1] longitudinal corpus of L2 Italian (first language [L1] Chinese) was collected. In total, 175 learners contributed two essays to the corpus. One essay was written at the beginning of a six-month course of Italian, and the other was written at the end of the course. The students were enrolled in a full-time course of Italian as a second language, which took place at a university in central Italy. Students of three proficiency levels contributed to the corpus: A1 (n=39), A2 (n=86), and B1 (n=50). All students came from China and were between 17 and 33 years of age

[1] Large-scale in terms of the number of learners who participated in this longitudinal study. While the resulting corpus is relatively small, the large number of learners (n=175) by far surpasses participant pools used in earlier longitudinal learner-corpus studies.

(mean=20.5, *SD*=2.7; 105 females). On average, the students spent 1.7 months in Italy (range 0.5-5, *SD*=0.69) prior to writing the first essay. The exact same 175 students who wrote the first essay also wrote the second essay. The students who only wrote one of the two essays were not included in the corpus. Three similar essay topics were offered: 1) My first impression of Italy and Italians, 2) My hobbies: what do I usually do in my free time, 3) My last holidays. The students were instructed not to write on the same topic more than once. Hence, all students chose two of the three topics. Finally, the students were taught by the same group of teachers at the same university. Thus, it can be said that the corpus creation addressed the issues of topics, teaching style and learning environment. Importantly, a range of proficiency levels is represented in the corpus: A1, A2, and B1 (according to CEFR). The total size of the corpus is circa 97,000 words (data collection Level 1: A1=7,126, A2=22,851, B1=15,903; data collection Level 2: A1=9,487, A2=24,117, B1=17,386).

The main aim of the present investigation was to examine vocabulary use – in terms of single words and collocations – in essays collected at the beginning of the course (Level 1) versus those collected at the end of the course (Level 2), separately for the three proficiency levels[2]. Thus, we were interested in the progress (if any) for A1 learners at Level 1 vs. Level 2, A2 learners at Level 1 vs. Level 2, and B1 learners at Level 1 vs. Level 2. To this aim, a number of analyses were conducted.

First, the POS distribution (nouns, adjectives, adverbs, and verbs) was analysed. The use of the POS was found to be comparable with the exception of nouns. Level 1 essays were found to contain significantly more nouns than Level 2 essays for A1 and A2 writers (but not B1).

Second, we looked at the number of tokens per essay and the number of tokens per sentence in Level 1 essays versus Level 2 essays. Level 2 essays were found to contain significantly more tokens than Level 1 essays in A1 and B1 writers (but not A2).

Third, we used the Guiraud index (e.g. Guiraud, 1954) as an index of lexical diversity. This index was used instead of type/token ratio because it compensates the systematical decrease of the number of tokens when texts to compare have different lengths (e.g. Van Hout & Vermeer, 2007). Level 2 essays were found to have a consistently higher Guiraud index. The differences between Level 1 vs. Level 2 were found significant across all three proficiency levels, but were particularly prominent in A1 and B1 writers.

The above results suggest that Level 1 and Level 2 essays appear to differ in terms of lexical diversity and the students' ability to produce longer pieces of writing, rather than in terms of POS distribution. This trend is particularly and consistently noticeable in beginner (A1) and intermediate learners (B1).

Finally, we looked at L2 learner use of collocations of different types (e.g. noun+adjective). To this aim, all learner items were first extracted

[2] It needs to be noted that no learner advanced from one CEFR level to the next between data collection points. So, all A1 learners remained A1, A2 remained A2, and B1 remained B1.

automatically from the learner corpus. A L1 reference corpus – Paisà (Lyding et al., 2014) – was used to extract frequencies of learner items in a representative corpus of L1 Italian. These frequencies were also used to calculate measures of association strength: *t*-score and mutual information. For the data analysis, we opted for mixed-effects modelling (Baayen, Davidson, & Bates, 2008). Preliminary analysis suggested a relatively comparable collocation usage, in terms of L1 frequencies and measures of association strength, for Level 1 vs. Level 2 across the three proficiency levels. It appears that while six months were sufficient for the students to be able to produce longer pieces of writing and to exhibit greater lexical diversity, this period of time might have not been sufficient for these learners to improve substantially in their use of L2 collocation (although some improved was observed).

In sum, the present large-scale longitudinal investigation of L2 Italian writing has provided a rich picture of how vocabulary use evolves in an Italian as a second language environment. Despite the relative brevity of the course – only six months – we observed clear developmental patterns, as suggested, for example, by longer sentences and essays, as well as higher lexical diversity indices. In line with previous research, however, our analysis also showed that L2 collocation learning can be slow and uneven (e.g. Laufer & Waldman, 2011).

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Crossley, S. A., & Salsbury, T. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Teaching*, 49, 1-26.

Granger, S., Gilquin, G., & F. Meunier (2015). *Cambridge Handbook of Learner Corpus Research*. Cambridge: CUP.

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning*, 61, 647e672.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of Second Language Writing*, 18, 85-102.

Li, P., Eskildsen, S., & Cadierno, T. (2014). Tracing an L2 learner's motion constructions over time: a usage-based classroom investigation. *The Modern Language Journal*, 98(2), 612-628.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. *Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, April 2014,* 36-43.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System*, 53: 148-160.

Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton & J. Treffers-Daller (Eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: CUP, 93-115.

Yuldashev, A., Fernandez, J., & Thorne, S. L. (2013). Second language learners' contiguous and discontiguous multi-word unit use over time. *The Modern Language Journal*, 97, 31-45.