Bundles in learner corpora: what type and token analyses can reveal

This paper aims at presenting the analysis of four-word bundles extracted from native and non-native corpora of argumentative essays. Although the corpora size are very different, we claim that the running of statistical tests to compare both the bundle types and token can be revealing of the problems learners might have.

Several Corpus Linguistics studies have focused on lexical bundles recently. Some of them deal with bundles produced by native speakers: a) in the university – oral and written discourse - (Biber et al. 2004; 2006; 2009); b) in different academic areas – electric engineering, biology, administration, applied linguistics (Hyland 2008); c) in business contexts – genre based analysis of business report (Berber Sardinha 2003); and d) in academia, where Simpson-Vlach and Ellis (2010) propose a list of the most commonly used bundles in academic registers. This Academic Formulas List (AFL) contains 435 bundles divided in 18 subcategories, which forms the basis for the research presented in this paper. This list enabled us to classify bundles in types, that is, according to their functional characteristics. Previous studies have considered frequency to discuss bundle types across registers (Biber et al. 2004), but even when they present bundles produced by non-native speakers (Chen and Baker, 2010) no discussion is put forward in relation to the implications of such numbers for a better understanding of how learner corpora might differ from native-speaker corpora. The main aim of the present paper is to discuss the relevance of analyzing both types and tokens of bundles when research focuses on comparing bundles produced by native and non-native speakers, and in the specific case of this investigation, in argumentative essays. Our data consisted of 18 different learner corpora, namely the Louvain Corpus of Native English Essays (LOCNESS), the 16 ICLE sub-corpora (Granger et al. 2009) treated as one learner corpus, as well as, Br-ICLE, the Brazilian sub-corpus of ICLE, which is still being compiled and does not form part of the consolidated ICLE distribution. These corpora constitute of argumentative essays summing up more than 4 billion words, their sizes being: LOCNESS 324005, ICLE 3768527 and Br-ICLE 159182 words. The research methodology included the following steps. First, bundles of 3 and 4 words were extracted from each corpus with scripts specially developed for our research project. Second, bundles were categorized manually and automatically according to the AFL, both its broad categories (referential expressions, stance expressions and discourse organizing functions) as well as its 18 specific subcategories (e.g. Intangible and tangible framing attributes and quantity specification). Third, we identified the most frequent categories in each corpus. Fourth, we compared the most frequent bundles across the different corpora with a list of bundles for all corpora combined. Fifth, we detected the significant differences in terms of types of bundles across the broad categories. Sixth, we ran statistical tests to identify differences within each category. Seventh, token frequency analysis was done to investigate the extent to which they could reveal significant differences among the subcategories even if the type analysis had not revealed such differences. This paper adopts a conservative cut off point (Cortes 2008) and presents the bundles that occurred at least 20 times per million words. Results, based on the analysis of 4-word bundles, show that: a) there is a total of 676 four-word bundles as far as the broad categories are concerned and the differences among the corpora are statistically significant ($X^2 = 17.126$, $df = 4$, $p = 0.002$); b) there is a significant difference among the corpora when all subcategories are compared together ($X^2 = 79.624$, $df = 34$, $p = 0.000$); c) when we looked for statistically significant differences in bundle frequency types whithin the categories, in other words, among the subcategories, we detected that only the referential expression category presents relevant results ($X^2 = 35.2$, $df = 10$, $p = 0.000$); d) yet, statistical significant differences were found when the tokens were accessed to detect different use among the corpora in relation to the subcategories (e.g. chi-square test results for the referential expression subcategory "Specification of intangible framing attributes" across the three corpora was x^2 = 1821.35, df= 2, p= 0); e) due to the findings about the tokens in the subcategories in the three corpora, we did a qualitative analysis of the bundles. It was found that the bundle internal lexical variation in the same structure, for the same subcategory, may be responsible for the high or low frequency of such category. Internal lexical variation is common in LOCNESS for the "Specification of intangible framing attributes" subcategory but not in the same extent in the learner corpora. The structure DT + NN + that + DT, for example, occurs with internal variation of the noun (e.g. *the fact that the, the rest that the*) in all corpora. The pattern IN (in) + DT (the) + NN + IN (of) is present in the LOCNESS and in the ICLE corpora, but only in LOCNESS do they occur more than 20 million words. Even when they occur in ICLE, most of them are present at a level lower than 20 words per million (*in the course of*, *in the hands of*, *in the eyes of*) or do not appear at all (*in the light of*, *in the face of*). On the other hand, the pattern DT + NN + IN (of) + IN (*the issues of whether*) occurs only in Br-ICLE more than 20 times per million words, which may reflect a pedagogical emphasis in some contexts where the essays were collected. From a pedagogical perspective, analyzing the occurrence of these bundles through their frequency, even when the corpora size difference is considerable, can bring interesting insights to the English as a Foreign Language (EFL) community. Such a description would, for instance, possibly determine how and which bundles should be addressed in the classroom. We conclude that both bundle type and a bundle token analyses should be done for researchers to spot the overuse and underuse of bundles, which can be a great contribution to our understanding of the problems students face in producing academic texts. A bundle analyzer is being developed by one of the members of the research group; however, the results presented here were mainly categorized manually. More bundles need to be classified and the analyzer should be further tested for its efficiency level. Once we are confident of its classification capacity, we will make it available to teachers with guidelines about the pragmatic function and the usage of the bundles.