Academic Collocations in Computer Science Research Articles: A Corpus-based Study

Ping-Yu Huang (Ming Chi University of Technology, Taiwan)

In this article, we report results coming from two sorts of research works. The first involved examining the distribution of items on the Academic Collocation List (ACL, Ackermann & Chen, 2013) in a computer science corpus which contains more than 1,300 articles from high-quality journals. Second, for certain ACL entries which were found to appear rarely in the domain-specific corpus, we investigated whether corpus-derived collocation clusters were able to provide semantically similar alternatives (e.g. *academic community* as an alternative for the ACL item *academic circle*). The two sorts of works taken together suggest an innovative approach to collect academic collocations for a particular domain; English for specific purposes (ESP) instructors can first identify which ACL items are frequent (or non-frequent) in a selected domain, and, for non-frequent ones, utilize our collocation-cluster techniques to acquire domain-specific academic usages.

Distribution of ACL entries in computer science texts

Ackermann & Chen's (2013) Academic Collocation List arguably has been the most well-established collocation list compiled for academic purposes to date. Compared with similar works, the ACL is of higher pedagogical value because it adopted a rigid selection process which involved expert judgment. Ackermann & Chen utilized a four-stage approach to extract collocations from the Pearson International Corpus of Academic English (PICAE). The written curricular part of the PICAE, specifically, consisted of texts from four fields of study: applied sciences and professions, humanities, social sciences, and natural/formal sciences, with each covering seven disciplines. To gather candidate collocations, the authors first performed computational analyses using frequency, mutual information (MI), and t-score as main measures. Next, the collected lexical pairs were "refined" with the ones holding lower MI and t-scores or consisting of non-target part-of-speech (POS) combinations (e.g. determiner-noun) being removed. Human intervention was introduced at the third stage during which only the collocations judged as of higher pedagogical use were retained. The last step, systematization, was adopted to make the ACL more systematic and useful. Several function words were added to collocations (e.g. *'be' generally accepted*) if deemed necessary. The results of such four-stage process were a 2,468 collocation list. Being examined in the source corpus as well as a general-purpose comparison corpus, the ACL exhibited a 14-times higher frequency in the former, suggesting the high academic nature of the ACL.

Well-established as it is, however, to our knowledge whether the ACL is indeed useful across different professional domains has rarely been investigated or reported. The current study, serving as one of the first to empirically examine the distribution and usefulness of the ACL, focused on computer science, a discipline covered by the written curricular part of the PICAE. Our purposes, as specified earlier, were to explore whether the ACL entries frequently appeared in the selected discipline and which entries were particularly frequent/non-frequent. Totally, our computer science corpus (CSC) comprises over 14 million running tokens coming from texts of high-quality journals. To ensure that the CSC texts are representative of current written discourse in computer science, we used only the journal papers published from 2014 to 2016. The CSC contains articles from

Table 1. Numbers and Percentages of Non-frequent ACL Items in CSC

	ADJ-Noun	ADV-ADJ	ADV-Verb	Noun-Noun	Verb-ADJ	Verb-Noun	Verb-ADV	Overall
Numbers of Non-frequent ACL Items in CSC	526	35	23	16	4	19	3	626
Percentages of Non-frequent ACL Items in CSC	29.7%	28.2%	16.4%	25.8%	13.3%	6.1%	10.3%	25.4%

twelve major computer science sub-domains, including, for example, artificial intelligence and human-computer interactions. For each sub-domain, we consulted four high impact factor journals and extracted about 30 articles from each of them. The collected articles were further "refined" with some unwanted sections (e.g. authors' affiliations and references) being excluded. We applied Stanford Log-linear Part-Of-Speech Tagger (Toutanova, et al., 2003) on all CSC sentences which then enabled us to check whether the different POS combinations on the ACL showed up in them.

Here we focus on and report non-frequent ACL items in the CSC, as shown in Table 1. We adopted the normed frequency (0.2 times per million) used by the ACL; consequently, an ACL item was considered non-frequent if it appeared only once or never appeared in the CSC. Overall, although we did find that (1) several academic usages (e.g. *experimental result*, 1011 times) were highly frequent and (2) collocations related to information technology (e.g. *natural language*, 246 times) or statistics (e.g. *statistically significant*, 319 times) appeared very often, 25.4% of ACL items were not frequently used in computer science. In terms of different POS combinations, most verb-based pairs (e.g. *make available*, 201 times, and *address issue*, 286 times) were high- or medium-frequency ones whereas many adjective- or noun-based collocations appeared rarely in the CSC. One of the reasons for the latter finding is straightforward: many adjective- or noun-based ACL collocations seemed closely related to social sciences. Although Ackermann & Chen (2013) attempted to adopt the ones appearing across disciplines, a certain number of their entries still appeared to be highly humanities- or social sciences-relevant (e.g. *capitalist economy* and *culturally specific*). Our results generally support Hyland & Tse's (2007) claim that "coverall" lexical lists do not reflect the real needs of ESP students and what they need should be a lexical repertoire particularly collected for them.

We in our CSC data also observed some domain-specific word meanings (e.g. *value* referring to quantity as in *high value*) and word choice (e.g. using *stage* rather than *phase* as in *initial stage*). These phenomena again support Hyland & Tse's (2007) suggestion of studying disciplinary rather than general lexical conventions.

Automatic collection of domain-specific academic collocations

In the second part of this study, we investigated whether it was likely to automatically generate domain-specific academic collocations based on corpus-derived collocation clusters. Our collocation-cluster techniques, developed based on Cowie & Howarth's (1996) notions of "overlapping collocations", utilize a hybrid approach taking into consideration both word co-occurrences and semantic information to establish networks of collocations. As two words (e.g. verb-noun pairs such as *achieve purpose*) are selected, our collocation-cluster system begins to automatically search a corpus in order to identify collocates for both words (e.g. *goal/quality/objective* for *achieve* and

state/accomplish/define for *purpose*). Next, the two identified word groups are filtered with only the ones which "share" the most collocates with the target two words being left. This process results in a complete and manageable cluster. The words embedded in such clusters, finally, are further ranked in order of semantic relevance; that is, the more semantically related to the target words, the higher ranking (e.g. *attain/accomplish* for *achieve* and *goal/objective* for *purpose*). As the cluster involving *achieve purpose* suggests, an important function that collocation clusters are expected to perform is to detect/correct English learners' collocation errors. Our system basically can show what pairs should be avoided (e.g. *attain purpose*) as well as what pairs are correct usages (e.g. *achieve goal*) which are semantically similar to the searched words.

In this study we applied the collocation-cluster techniques to ESP research and explored, for the ACL entries which were non-frequent in the CSC, whether corpus-derived clusters could automatically provide alternative usages. From the 626 non-frequent collocations, we tested the techniques on 60 randomly selected verb-noun or adjective-noun combinations. Those selected pairs were further checked or replaced by others to ensure that the 60 tested collocations did not include highly humanities- or social sciences-relevant ones (which apparently should not be expected to appear frequently in computer science texts). Each collocation then was fed into our system using the CSC as the main dataset to generate collocation clusters. The results collected were rather promising and encouraging. For the 60 collocations, collocation-cluster techniques successfully provided 50 semantically-similar alternative word pairs, showing a high rate of 83%. Some examples for the alternatives include *construct argument* for the ACL entry *develop argument*, *demonstrate capability* for *demonstrate competence*, *primary concern* for *central concern*, *critical evaluation* for *critical examination*, etc. These results again confirm Hyland & Tse's (2007) viewpoint that people in different domains tend to have preferred expressions. If computer science students rely only on the ACL to learn academic collocations, it is likely that they will use several non-domain-specific academic pairs (e.g. *profound effect*, 0 times in CSC) instead of the ones that their colleagues prefer to produce in written academic texts (e.g. *significant effect*, 238 times).

Conclusion

The current research analysed the distribution of ACL entries in computer science research articles and reported an empirical study in which we successfully utilized collocation-cluster techniques to collect domain-specific academic collocations. The two types of results taken together suggest an innovative approach to gather academic collocations for a particular domain: ESP instructors can first examine which ACL entries are frequent and which are non-frequent in a domain-specific corpus and, for the latter, make use of corpus-derived collocation clusters to acquire semantically-similar alternatives. We in our presentation also discuss future improvements of collocation-cluster techniques, which we expect to make the techniques even more useful for both corpus linguists and ESP practitioners.

References

Ackermann, K. & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes,* 12(4), 235-247.

Hyland, K. & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly,* 41(2), 235-253.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.