Part-of-speech patterns in research introductions: A cross-disciplinary study

(Kathy) Ling LIN (Shanghai Jiao Tong University) & Ming LIU (Hong Kong Polytechnic University)
kathyll@sjtu.edu.cn, ming1.liu@polyu.edu.hk

This study applies the Part-of-Speech-gram (PoS-gram) procedure to the examination of language patterning and variability in a largely conventionalized part-genre (i.e., research introductions). A *PoS-gram*, as defined by Stubbs (2007, p. 91), is "a string of part-of-speech categories", "the tokens of which are strings of words that have been annotated with these PoS tags" (Pinna & Brett, 2018, p. 107). Stubbs (2007) considered it as a type of "routine phraseology", in addition to *n-grams* and *phrase-frames*. Yet, as phraseology is generally defined in corpus linguistics research as "the recurrent co-occurrence of words" (Clear, 1993, p. 277) and the compositional unit of a PoS-gram is a PoS category (grammatical category) rather than a word form, PoS-grams in our understanding may arguably not be a type of phraseology. Accordingly, we only treat it as a phraseology-related concept, since the exponents of each PoS-gram may be potential phraseology and the identification of it can be an effective way to extract recurrent phraseologies and patterns (Pinna & Brett, 2018).

Based on 400 article introductions of computer engineering (CE) and cognitive linguistics (CL) collected from AntCorgenGen 1.1.2 (Anthony, 2019), the study has identified key PoS-grams and their associated lexico-grammatical frames, using the written academic component of British National Corpus as the reference corpus. In the identification and concordance search of key PoS-grams, Sketch Engine with their modified English TreeTagger PoS tagset was adopted (Kilgarriff et al., 2014).

Findings are summarized as follows. First of all, the PoS-grams with high keyness scores have been successfully identified for introductions of both disciplines, with their representative lexicogrammatical frames and phraseologies highlighted, which has empirically validated the phraseological tendency and idiomaticity of language use in academic genres (Sinclair, 1996). Second, the analysis reveals key PoS-grams shared in CE and CL introductions, e.g., those associated with the step "purposive announcement" (viz., *IN DT JJ NN VBD TO* and *DT JJ NN VBD TO VV*), as well as the discipline-specific ones such as the PoS-gram for structure-outlining only found in CE introductions (viz., *DT NN VBZ VVN RB VVZ*). In addition to identifying sets of characteristic lexicogrammatical frames and phraseologies that could be directly transformed into EAP pedagogical input, the PoS-gram analysis has also helped revealing contrasting language styles in introductions of the two disciplines. The apparently more compact language use has been noted in CE introductions than in CL introductions, as evidenced in the total absence of the *that*-clause but the strong presence of the *to*-infinitive clause and the prepositional phrase instead in tokens of top-ranking key PoS-grams identified in CCE. Contrastingly, in CCL, the use of the *that*-clause is far more frequent, e.g., three out of the four key PoS-grams for realizing the step of topic summarization do contain it. The more compressed language style of academic introductions in CE could also be perceived from the particularly intensive use of the construction "noun +noun(+noun) …" as well as the pre-modifications and/or post-modifications of noun phrases in them.

Compared to various forms of multi-word sequences like n-grams, the PoS-gram has the unique strength of grouping phraseologies with similar or identical structure and discursive functions and

yet either recurrent or varying lexical choices under the co-selected grammatical categories. The advantage enriches analyses and helps yield pedagogically useful findings, in that patterning and variability is revealed not only in the overall function, structure and composition of PoS-grams but in such aspects of their recurrent or diversified tokens. This study illustrates the innovative application of corpus-based PoS-gram procedure to academic genres, which may inspire a promising new line of inquiry and the current genre pedagogy.

Acknowledgements

This work was supported by Shanghai Pujiang Program (Project no.: 2019PJC067) and the Philosophy and Social Science Planning Program of Shanghai (Project no.: 2017EYY007).

References

Anthony, L. (2019). AntCorGen (Version 1.1.2) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Clear, J. (1993). From Firth principles: Computational tools for the study of collocation. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 271-292). Amsterdam: John Benjamins.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7-36.

Pinna, A., & Brett, D. (2018). Constance and variability: Using PoS-grams to find phraseologies in the language of newspapers. In J. Kopaczyk, & J. Tyrkkö (Eds.), *Applications of pattern-driven methods in corpus linguistics* (pp. 107-130). Amsterdam: John Benjamins.

Sinclair, J. M. (1996). The search for units of meaning. *Textus, 9*(1), 75-106.

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89-105). Amsterdam: Brill Rodopi.