Fixedness and Variability: Using PoS-grams to Study Phraseology in Newspaper Articles

Antonio PINNA and David BRETT
Dipartimento di Scienze Umanistiche e Sociali
Università degli Studi di Sassari
Sassari, Italy
dedalo@uniss.it, dbrett@uniss.it

The pervasiveness of phraseology in human language, as demonstrated in numerous studies conducted over the last fifty years (Firth, 1957; Pawley & Syder, 1983; Sinclair, 2004; Wray, 2002), has wide-reaching implications for second and foreign language learning. Learning a language, above all a specialised version of it, requires as a necessary condition, knowledge of the phraseology particular to it (Swales 1990).

The advent of Corpus Linguistics has contributed greatly to our understanding of phraseology, and several techniques have been developed to elicit patterns from strings of word forms, such as *n*-grams, skip-grams and conc-grams. These three typologies vary along an exact-fuzzy axis, in terms of either lexis or syntax, or both, with the first being a concatenation of identical strings of word forms, while the third allowing considerable syntactic flexibility rotating around the same lexical items.

The technique used in this study can be considered an even more flexible query type: the Part-of-Speech-gram (usually abbreviated to PoS-gram) is a string of Part-of-Speech categories (Stubbs, 2007: 91), the tokens of which are strings of words that have been annotated with these PoS tags. Hence, in each slot of the PoS-gram, any word can occur as long as it belongs to the PoS category of that particular slot. By casting a considerably looser net than that of the *n*-gram and the skip-gram, PoS-grams are potentially very useful in revealing relatively long sequences that fly below the statistical radar of the former techniques. PoS-grams can be used to highlight the use of syntactic structures that are typically employed for certain functions, that latter aspect being testified to by the repeated occurrence in the same slot of items that may be identical, synonymous or from the same semantic domain.

Using the PoS-gram technique, this paper investigates recurrent syntactic and lexical patterns in two c. 500,000 token corpora compiled from different sections of the Guardian newspaper: Travel and Crime. The texts were tagged for Part of Speech and then 6 unit PoS-grams were formed starting from each token of the texts. The PoS-grams obtained from each corpus were then quantified and compared, also with a database of PoS-grams obtained from the 100M token BNC. A large number of PoS-grams were found to occur in the sample corpora with statistical significance at the highest levels of certainty. These will be illustrated and discussed in an attempt to explain the typical syntactic and lexical patterns of the genres. For example, the PoS-grams typical of the Travel section were almost exclusively composed of noun and prepositional phrases, indicating a preoccupation in the genre with (the description and evaluation of) entities, mostly places. The Crime section results, on the other hand, displayed a far higher presence of verbs, more specifically, passive constructions in the simple past, hence the genre appears to be more process oriented than that of Travel.

The pedagogical implications of this research concern Sinclair's (1991) well-known idiom principle. PoS-gram analysis elicits sequences in particular domains which are by and large fixed in some slots, while not wholly open, but rather 'ajar' as it were in others. From a pedagogical point of view this approach has the potential to raise awareness of the fact that learning to be effective communicators within the constraints of a given genre requires that learners be able to reproduce and reformulate its typical phraseological sequences. Indeed, learners of English for Special Purposes may well benefit from the presentation of not only

prefabricated constructions, but also those in which a certain amount of variability is permitted, mirroring the scenario envisaged by Ellis (2008) whereby "category learning is optimized by an initial low-variance sample centered upon prototypical examples" from which generalized productive schemata are then developed.

References

Ellis, N. C. (2008). 'Phraseology. The periphery and the heart of language.' In: Meunier, F. & Granger S. (eds). *Phraseology in Foreign Language Learning and Teaching.* Amsterdam: John Benjamins. 1-13.

Firth, J. R. (1957). *A Synopsis of Linguistic Theory, 1930-1955.* Oxford, Blackwell.

Pawley, A. & Syder, F. H. (1983). 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency.' In: Richardson, J.C. & Schmidt, R. W. (eds). *Language and Communication.* London: Longman. 191-225.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: Language, corpus, and discourse*. London: Routledge.

Stubbs, M. (2007). 'An example of frequent English phraseology: distributions, structures and functions.' In: Facchinetti, R. (ed.). *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, 89-106.

Swales, J.M. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge. Cambridge University Press.