Abs-84

Federico Gaspari (University of Bologna, Italy)

# Combining the analysis of lexical bundles and POS-n-grams: a phraseological comparison of the BNC and ukWaC

Several studies have looked into the phraseology of English by focusing on lexical bundles, i.e. sequences of words that occur a certain number of times in a corpus, regardless of their structural completeness or function (e.g. Biber, 2006; Biber & Conrad, 1999; Cortes, 2004; Hyland, 2008; Partington & Morley, 2004; Stubbs, 2007). A complementary strand of research has considered the syntactic-grammatical configuration of English phraseology looking at POS-n-grams, i.e. complexes formed by strings of specific grammatical categories (e.g. William Fletcher's "Phrases in English" database derived from the BNC supports the search for patterns of between 1 and 8 POS tags - see Fletcher, 2007). Surprisingly, however, only very little research has investigated the combination of these two interrelated levels of phraseological analysis as a means of describing and comparing corpora in the same language, one exception being Bernardini et al. (2010:36ff) who examine the differences between native and non-native institutional academic English.

This paper presents an analysis of lexical bundles and POS-n-grams in the BNC (Aston & Burnard, 1998) and ukWaC (Ferraresi et al., 2008) as a basis to contrast the phraseological make-up of the two corpora. ukWaC is a web-derived corpus of English containing more than 2 billion tokens, which has already been compared to the BNC by Baroni et al. (2009) in terms of lexical coverage, text types and subject matters, and by Ferraresi et al. (2008) with a qualitative study of salient nouns, verbs and adjectives. In addition, Sharoff (2006) built web-derived corpora of a similar size to the BNC for a number of languages, using the BNC as a benchmark to evaluate the composition of the English corpus based on text typology and word lists. Finally, Ferraresi et al. (2010) showed that ukWaC and the BNC were similarly helpful in performing lexicographic tasks aimed at dictionary compilation.

Following a discussion of the appropriate length and frequency cut-off point for lexical bundles and POS-n-grams, the paper contrasts these two dimensions of phraseology in the BNC vs. ukWaC. The investigation looks at the most common lexical bundles vis-a-vis the high-frequency POS-n-grams within each of the two corpora, comparing the extent to which the retrieved constructs overlap. The conclusion points out the main phraseological similarities and differences between the BNC and ukWaC, providing new insights into the likeness of these two corpora, which is of interest to researchers using them for a range of theoretical and applied purposes. Finally, we discuss from a methodological perspective the advantages and the potential of combining the analysis of lexical bundles and POS-n-grams both to describe the phraseology of individual corpora and to compare phraseological features across corpora in the same language.

References

Aston, G. & L. Burnard (1998) 'The BNC Handbook: Exploring the British National Corpus with SARA'. Edinburgh: Edinburgh University Press.

Baroni, M, S. Bernardini, A. Ferraresi & E. Zanchetta (2009) "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". 'Language Resources and Evaluation' 43(3):209-226.

Bernardini, S., A. Ferraresi & F. Gaspari (2010) "Institutional academic English in the European context: a web-as-corpus approach to comparing native and non-native language". L. López, Á. & R. Crespo Jiménez (eds) 'Professional English in the European context: The EHEA challenge'. Bern: Peter Lang. 27-53.

Biber, D. (2006) 'University language: a corpus-based study of spoken and written registers'. Amsterdam: John Benjamins.

Biber, D. & S. Conrad (1999) "Lexical bundles in conversation and academic prose". H. Hasselgård and S. Oksefjell (eds) 'Out of corpora: Studies in honor of Stig Johansson'. Amsterdam: Rodopi. 181-189.
Cortes, V. (2004) "Lexical bundles in published and student disciplinary writing: Examples from history and biology". 'English for Specific Purposes' 23:397-423.
Ferraresi, A., S. Bernardini, G. Picci & M. Baroni (2010) "Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation". R. Xiao (ed) 'Using Corpora in Contrastive and Translation Studies'. Newcastle: Cambridge Scholars Publishing. 337-359.
Ferraresi, A., E. Zanchetta, M. Baroni & S. Bernardini (2008) "Introducing and evaluating ukWaC, a very large Web-derived corpus of English". S. Evert, A. Kilgarriff and S. Sharoff (eds) 'Proceedings of the 4th Web as Corpus Workshop - Can we beat Google? (WAC-4) LREC 2008'. Marrakech, Morocco, 1 May 2008. 47-54.
Fletcher, William H. (2007) "Implementing a BNC-Compare-able Web Corpus". 'Proceedings of the 3rd web as corpus workshop'. Louvain-la-Neuve, Belgium, 15-16 September 2007. 43-56.
Hyland, K. (2008) "As can be seen: Lexical bundles and disciplinary variation". 'English for Specific Purposes' 27: 4-21.
Partington, A. & J. Morley (2004) "From frequency to ideology: Investigating word and cluster/bundle frequency in political debate". B. Lewandowska-Tomaszczyk (ed) 'Practical Applications in Language and Computers – PALC 2003'. Frankfurt am Main: Peter Lang. 170-192.
Sharoff, S. (2006) "Creating general-purpose corpora using automated search engine queries". M. Baroni & S. Bernardini (eds) 'Wacky! Working papers on the Web as Corpus'. Bologna: Gedit. 63–98.
Stubbs, M. (2007) "An Example of Frequent English Phraseology: Distribution, Structures and Functions". Facchinetti, R. (ed) 'Corpus Linguistics Twenty-five Years On'. Amsterdam: Rodopi. 89-105.