A lexical bundle approach to comparing languages: Organizational and stance markers in English and French

Sylviane Granger
Université catholique de Louvain, Centre for English Corpus Linguistics

The field of phraseology has expanded rapidly in recent years. Originally centred on the most ‘colourful’ types of units (idioms like a pain in the neck or proverbs like a bird in hand is worth two in the bush), it now encompasses a wide range of much more mundane units which were not previously considered. The expansion of the field is largely due to the use of powerful corpus linguistic techniques which make it possible to extract typical patterns of word combinations automatically. One of these techniques is the extraction of word n-grams, i.e. sequences of contiguous n words (2 words, 3 words, etc.) from a given corpus. Using this method it is possible to identify what Biber et al (1999: ch. 13) call ‘lexical bundles’, i.e. the most frequent recurring sequences of words in a register. A large number of studies, mostly focused on academic settings, have highlighted the major role that these prefabricated units play in discourse. Lexical bundles may take different structural forms (phrasal: the total number of, as a result of, and things like that; clausal: is likely to be, as shown in, I would like to) and fill a range of functions which Biber et al (2004) group into three main categories: (1) referential bundles which make direct reference to physical or abstract entities, or to the textual context itself (a lot of people, in the United States); (2) discourse organizers which reflect relationships between prior and coming discourse (with this in mind, this is why); and (3) stance bundles which express attitude or assessment of certainty (I don’t want to, it is possible to). The last two types of markers are part of the more general notion of ‘metadiscourse’, which Hyland (2005: ix) defines as the use of language to “organise texts, engage readers and signal attitudes to the material and the audience”. The lack of salience that characterizes many lexical bundles constitutes a challenge for both foreign language learners and translators or interpreters who may be led to produce awkward-sounding phrases, directly transferred from their mother tongue or the source language (Chen & Baker 2010, Bal 2010, Lee forthcoming). However, transfer remains a largely hypothetical factor, as systematic contrastive analyses of lexical bundles in different languages are very rare. This is a pity as languages differ markedly in their use of metadiscourse (cf. Sultan 2011, Zarei & Mansoori 2011) and lexical bundles are an efficient way to access those differences, as shown by Cortes’s (2008) comparison of history writing in English and Spanish.

The purpose of my presentation is to demonstrate the value of the lexical bundle approach for crosslinguistic research. My focus will be on the use of organizational and stance markers in comparable corpora of English and French, i.e. corpora consisting of original texts in the two languages matched by criteria such as genre, time of publication, etc. (Johansson 2007). As French is described as more explicitly conjunctive and emphatic than English (Vinay & Darbelnet 1995 [1958]: 234ff and 220ff; Delisle 1993: 432) and generally more verbose, my hypothesis is that metadiscursive bundles will tend to be more frequent in French than in English. I will also investigate the balance between impersonal it-constructions expressing modality (it is true that/il est vrai que) and personal I/we-references marking authorial presence (I believe that/je crois que) in each language. It is difficult to formulate a hypothesis in this area as the English-French contrastive literature contains seemingly contradictory claims: Vinay & Darbelnet (ibid: 216) contend that French favours subjective representation
of reality whilst English tends towards objective representation (On était au commencement de février vs. It was the beginning of February), while Chuquet & Paillard (1987: 141) highlight the preference for structures with animate subjects in English and impersonal structures in French (Of course everyone is free to go window-shopping in France vs. Il est certes loisible à chacun, en France, de faire du lèche-vitrine).

My investigation differs from previous cross-linguistic studies of metadiscourse in two major ways. First, it focuses on longer metadiscursive markers while most studies focus on single words (admittedly, however) or compound-like units (in fact, on the other hand). Secondly, it does not rely on a pre-established list of markers but uses a fully corpus-driven method to identify them. The analysis is based on two subcorpora of English and French extracted from a version of the Europarl corpus which clearly identifies the source vs. target status of the two languages (Cartoni et al 2011, Cartoni & Meyer 2012). Using the WordSmith Tools text retrieval program (Scott 1996), I extracted the lexical bundles of 3 or more words automatically from the two subcorpora and subjected them to structural and functional analysis based on taxonomies inspired by Chesterman (1998), Biber et al (1999 & 2004), Biber & Barbieri (2007), Hyland (2005) and Cortes (2008).

As amply demonstrated in the literature (Biber et al 1999, Ninisha 2007), the quantity and quality of lexical bundles - and of metadiscursive markers in particular - are highly sensitive to genre. The second stage of the investigation aims to assess to what extent the similarities and differences uncovered by the analysis of the Europarl subcorpora reflect general characteristics of the two languages or whether they are limited to the genre of parliamentary debates. To achieve this aim, the analysis was replicated on corpora representing a different genre, viz. editorials from quality papers in English and French extracted from the Mult-Ed corpus and the results compared with those based on Europarl.

For English, additional insights on the impact of genre can be gained from a comparison of the Europarl results with studies of lexical bundles in written EU documents (Trebits 2009 and Jablonkai 2010).

The concluding part of the presentation sums up the main results of the study, suggests avenues for future research and considers ways of applying the methodology to a number of fields, in particular bilingual lexicography.

References


---

1 I would like to express my gratitude to Bruno Cartoni for giving me access to this version of the corpus.

2 For a description of Mult-Ed, see http://www.uclouvain.be/en-cecl-multed.html


