The contribution of corpus-based phraseology to translation studies: from experiments to theory

Jean-Pierre Colson

The notion of phraseology is now used across a wide range of linguistic disciplines: Phraseology (proper), Corpus Linguistics, Discourse Analysis, Pragmatics, Cognitive Linguistics, Computational Linguistics. It is, however, conspicuously absent from most studies in the area of Translation Studies (e.g. Delisle 2003, Baker & Saldanha 2011). The paradox is that many practical difficulties encountered by translators and interpreters are directly related to phraseology in the broad sense (Colson 2008, 2013), and this can most clearly be seen in the failure of SMT-models (*statistical machine translation*) to deal efficiently with the translation of *set phrases* (used here as a generic term for all categories of phraseological constructions, from collocations to proverbs).

Although corpus-based and computational phraseology still need to be clearly delineated from other concurrent disciplines, a possible way of narrowing the gap between phraseology and translation studies is proposed here: the recourse to experiments involving on the one hand set phrases and, on the other, evidence from parallel translation corpora or SMT-machines such as Google Translate. We will argue that both phraseology and translation studies have much to gain from this cross fertilisation, because both disciplines are regularly criticised for their lack of coherent terminological description and for the insufficient number of reproducible experiments they involve. The aim of this paper is not to draw up an exhaustive list of the possible experiments showing the interweaving of phraseology and translation studies, but to propose directions for future research involving a number of key issues that are posed by phraseology and are illustrated by translation practice.

A first series of experiments relating to this subject matter concerns the problems posed by phraseology to human translation. Decoding phraseology in the source text is far from easy for translators and interpreters, all the more so as they are usually not native speakers of the source language. Also, finding a natural formulation in the target language and avoiding *translationese* requires an excellent mastery of the phraseology of the target language. I will argue that experiments with translation corpora may precisely shed some light on some crucial notions of phraseology and of translation studies. Experiments have shown that translation errors due to phraseology are legion in many translation corpora, even in the official translations of the European Union. A contribution of corpus-based phraseology would therefore consist in making human translators aware of the pitfalls of phraseology in the source text. Even experienced professionals sometimes fail to detect the fixed or semi-fixed character of a source text construction. Experiments along these lines should therefore also include the creation of large, multilingual phraseological databases, which brings us back to two serious shortcomings of computational phraseology:

1. There is no universally accepted algorithm for the automatic extraction of phraseology, especially not for ngrams larger than bigrams.
2. There is no consensus as to the proportion of set phrases in relation with the rest of the vocabulary: according to Jackendoff (1995), there are about as many fixed expressions as there are single words in the dictionary, but others (such as Mel'čuk 1995) hold the view that fixed expressions far outnumber single words.

I will argue in that respect that algorithms derived from text mining and information retrieval techniques (Baeza-Yates, R. & B. Ribeiro-Neto 1999) can be efficient and (computationally) cost-effective in order to build up unfiltered collections of recurrent fixed or semi-fixed phrases, from which translators could gain information about the number of set phrases in the source text. Such an algorithm has been proposed in Colson (2014), and a provisional database of about 700,000 English set phrases (tokens) has been assembled, which seems to confirm that Jackendoff's view about the total number of fixed expressions was not correct.

A second series of experiments that would turn out to be profitable to a better theoretical understanding of both phraseology and translation studies, has to do with the specific problems posed by phraseology to automatic translation. Phraseology has only recently been identified as one of the main sources of errors in automatic translation systems, including the most recent SMT-systems (Monti, Mitkov, Corpas Pastor & Seretan 2013). I will however point out that the theoretical underpinnings of phraseology are at stake in order to provide a coherent explanation for the serious shortcomings in the automatic translation of sentences containing phraseology. The crux of the matter seems to be the complex interplay between association and frequency in fixed expressions. Recent evidence shows that, contrary to what is assumed by most statistical scores, there should be no relationship between the statistical association of the grams constituting a set phrase, and its frequency in a huge corpus. The countless examples of wrong translations of phraseologically rich sentences by Google Translate, for instance, all point to the fundamentally wrong way in which ngrams were traced down, namely by giving the highest priority to frequency.

Further experimentation should also shed some light on the overall statistical distribution of set phrases in large corpora. The well-know zipfian distribution of words in a corpus poses theoretical problems as far as phraseology is concerned. Corpus-based studies (Baroni 2008) indicate that the distribution of ngrams themselves may display a Zipf-Mandelbrot curve. This is an important theoretical challenge to the theory of phraseology and also to semantics, having therefore consequences on the way meaning may be expressed in different languages and be adequately translated from one language into another. I will point out that a general theory of phraseology, as outlined by Mejri (2006), may offer a new insight into the statistical underpinnings of both morpheme associations (in words) and of word association (in set phrases).

References

Baeza-Yates, R. & B. Ribeiro-Neto (1999). *Modern Information Retrieval*. New York: ACM Press, Addison Wesley.

Baker, M. & G. Saldanha (eds.) (2011). *Routledge Encyclopedia of Translation Studies.* New York: Routledge.

Baroni, M. (2008). Distributions in text. In: A. Lüdeling & M. Kytö, (eds.), *Corpus linguistics. An international handbook*. Berlin, New York: Walter de Gruyter, p. 803-821.

Baroni, M., Bernardini, S., Ferraresi, A. & E. Zanchetta. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*, 43, p. 209-226.

Colson, J.-P. (2008). Cross-linguistic phraseological studies: An overview. In: Granger, S. & F. Meunier (eds.), *Phraseology. An interdisciplinary perspective*. John Benjamins, Amsterdam / Philadelphia, p. 191-206.

Colson, J.-P. (2010a). The Contribution of Web-based Corpus Linguistics to a Global Theory of Phraseology. In: Ptashnyk, S., Hallsteindóttir, E. & N. Bubenhofer (eds.), *Corpora, Web and Databases. Computer-Based Methods in Modern Phraselogy and Lexicography*. Hohengehren, Schneider Verlag, p. 23-35.

Colson, J.-P. (2010b). Automatic extraction of collocations: a new Web-based method. In: S. Bolasco, S., Chiari, I. & L. Giuliano, *Proceedings of JADT 2010,Statistical Analysis of Textual Data*, Sapienza University of Rome, 9-11 June 2010. Milan, LED Edizioni, p. 397-408.

Colson, J.-P. (2013). Pratique traduisante et idiomaticité : l'importance des structures semi-figées. In : Mogorrón Huerta, P., Gallego Hernández, D., Masseau, P. & Tolosa Igualada, M. (eds.), *Fraseología, Opacidad y Traduccíon. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation* (Herausgegeben von Gerd Wotjak). Frankfurt am Main, Peter Lang, p. 207-218.

Colson, J.-P. (2014). Set phrases around *globalization* : an experiment in corpus-based computational phraseology. Paper presented at *CILC 2014, 6th International Conference on Corpus Linguistics*. University of Las Palmas de Gran Canaria, 22-24 May 2014.

Corpas Pastor, G. (2013). Detección, descripción y contraste de las unidades fraseológicas mediante

tecnologías lingüísticas. In Olza, I. & R. Elvira Manero (eds.) *Fraseopragmática*. Berlin: Frank & Timme, p. 335-373.
Delisle, J. (2003). La traduction raisonnée. Ottawa: Presses de l'Université d'Ottawa.
Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 133-165.
Mejri, S. (2006). Polylexicalité, monolexicalité et double articulation. *Cahiers de Lexicologie*, 2 :209-221.
Mel'čuk, I. 1995. Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schroeder (eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 167-232.
Monti, J., Mitkov, R., Corpas Pastor, G. & V. Seretan (eds) (2013). Workshop Proceedings: *Multi-word units in machine translation and translation technologies*, Nice 14th Machine Translation Summit.