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

Colson, Jean-Pierre

Abstract
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: ...

Document type: Communication à un colloque (Conference Paper)

Référence bibliographique

Colson, Jean-Pierre. The contribution of corpus-based phraseology to translation studies: from experiments to theory. Europhras 2015 (Malaga, du 29/06/2015 au 01/07/2015).
The contribution of corpus-based phraseology to translation studies: from experiments to theory

Jean-Pierre Colson
Université catholique de Louvain (Louvain-la-Neuve)
We refer here to phraseology as the study of set phrases in the broadest sense, including partly fixed phrases (routines and formulae, collocations), and also very fixed phrases (idioms and proverbs).

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 (…).
What about Translation Studies?

Most publications on Translation Studies mention the problem of *expressions / idioms / collocations* but they do not refer to phraseology as a theory or discipline.

Example 1: Delisle, J. (2003). *La traduction raisonnée: expressions* are treated as a part of the lexicon.
Example 2: phraseology is also most conspicuously absent from a major reference work in the field, the *Routledge Encyclopedia of Translation Studies* (Baker and Saldanha 2011), and the same holds true of collocations.
The interest for phraseology (at least for collocations) in translation studies came mainly from corpus linguistics (example: W. Teubert, 2002. The Role of Parallel Corpora in Translation and Multilingual Lexicography).
On the other hand, computational linguistics is now showing a growing interest for phraseology, particularly against the backdrop of automatic translation.

2. Computer, corpus, phraseology and translation studies

* As they show many contact points, phraseology and translation studies have to gain from cross fertilisation

* Both disciplines are regularly criticised for their lack of coherent terminology or of reproducible experiments (Čermák 2001, Baker & Salhanha 2011)
A number of experiments are crucial for exploring the complex interaction between phraseology and translation studies, and they may reveal new theoretical insights, while also providing practical solutions.

Three types of experiments are presented here, as well as preliminary results of a personal research project.
3. Phraseology and human translation: experiments

- 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.
- Finding a natural formulation in the target language and avoiding *translationese* (Tirkkonen-Condit 2002) requires an excellent mastery of the phraseology of the target language.
* Experiments with translation corpora may precisely shed some light on some crucial notions of phraseology and of translation studies.
* Translation errors due to phraseology are present in many translation corpora, even in the official translations of the European Union
Example:

Cost-cutting and cutting corners caused the biggest environmental disaster in history.

Réduire les coûts et arrondir les angles ont engendré le plus gros désastre écologique de l'histoire. (europarl.eu / linguee.com, 01/07/2015)

cut corners
Fig. to take shortcuts; to save money or effort by finding cheaper or easier ways to do something.

arrondir les angles: arranger les choses, réduire les difficultés

=> Correct translation: Rogner sur les coûts, lésiner sur les coûts
Example:

Above all it is important to avoid cutting corners, which easily happens with the ongoing and multiple evaluations that are nowadays often required.

Il faudrait éviter avant tout de traiter les choses de manière superficielle, ce qui est souvent le cas avec les évaluations [...] (eur-lex.europa.eu / linguee.com, 06/04/2015)
Example:
If we are going to move the goalposts, if that is the will of this General Conference, let us do that before the whistle [...]

Il Si nous voulons déplacer les poteaux de but, si telle est la volonté de la Conférence générale, alors faisons-le avant que la partie n'ait commencé.
(unesdoc.unesco.org / linguee.com, 06/04/2015)

=> changer les règles du jeu
There are also numerous examples of errors of this kind in translations produced by students (Colson 2010c): phraseology is not decoded, and this results in a literal translation, a *calque*.

Further experimentation is necessary: phraseology extraction, psychological perception of phraseology by translators and interpreters, practical tools
4. Phraseology and machine translation: experiments

* 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 machine 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)
Example: results from Barreiro et al. (2013) with OpenLogos and Google Translate (language combinations: English, French, Italian, Portuguese)
<table>
<thead>
<tr>
<th>System</th>
<th>Lang pair</th>
<th>OK</th>
<th>ERR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OL</td>
<td>EN-FR</td>
<td>40</td>
<td>48</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>EN-IT</td>
<td>36</td>
<td>83</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td>EN-PT</td>
<td>60</td>
<td>96</td>
<td>156</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>136</td>
<td>227</td>
<td>363</td>
</tr>
<tr>
<td>GT</td>
<td>EN-FR</td>
<td>70</td>
<td>38</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>EN-IT</td>
<td>59</td>
<td>47</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>EN-PT</td>
<td>67</td>
<td>47</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>196</td>
<td>132</td>
<td>328</td>
</tr>
</tbody>
</table>
According to those results, Google Translate wrongly translates phraseology in about 40 PERCENT of the cases.

Indeed, it doesn’t take more than a few seconds to try a sentence containing a set phrase and to find an obviously wrong translation:
* A lot of things just come out of the woodwork.
* GT: Beaucoup de choses vient de sortir de la boiserie (=> apparaissent comme par enchantement)
* Let us give credit where credit is due.
* GT: Donnons crédit lorsque le crédit est dû. (=> il faut rendre à César ce qui appartient à César)
* You would be up the creek without a paddle
* GT: Vous seriez jusqu'à la crique sans pagaie(=> vous seriez dans le pétrin)
Other examples of set phrases that are wrongly translated (in all languages) by Google: **come hell or high water**, **clutching at straws**, **the credits roll**, **cut the mustard**, **cutting corners**, **a dab hand**, **a dead giveaway**, **in deadly earnest**, **death by a thousand cuts**, **snatch defeat from the jaws of victory**, **do a double take**, **does what is says on the tin**, **don’t mention the war**, **don’t quote me on this**, **a cheap shot**, **the chickens are coming home to roost**, **chuffed to bits**
5. From experiments to theory

- Such experiments with human and machine translation have highlighted the importance of set phrases in language use.
- There is a broad consensus that phraseology represents a major feature of language, but the exact percentage it represents is still highly controversial.
According to John Sinclair, the idiom principle (1991) or phraseological tendency (1996) is the general rule:

“Most normal text is made up of the occurrence of frequent words, and the frequent senses of less frequent words. Hence, normal text is largely delexicalized, and appears to be found by exercise of the idiom principle, with occasional switching to the open-choice principle.” (1991: 113)

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.
Hanks (2010): “The meaning potential of a word consists of a puzzling mixture of terminology and phraseology: contextual dependence and contextual independence”
Erman and Warren (2000): prefabs represent about 55% of the texts they have analyzed.

The method they use for identification is counting the number of words that are part of prefabs.

They also point out that “The identification of prefabs is difficult. There are two main reasons why this is so. One is that what is a prefab to some members of a language community need not be a prefab to all members. Some prefabs will be known to practically all native and fluent speakers; others will be more limited in dispersion and entrenchment.” (Erman and Warren 2000: 33)
If the identification of phraseology by humans is so difficult, one possible solution would be to turn to automatic extraction.

However, according to Gries (2013), after “50-something years of work on collocations”, the results are still disappointing and “after many decades of 'more of the same', (…) it is time to explore new ways of studying collocations”
I have therefore proposed a new score (Colson 2015) for the automatic extraction of phraseology: the Corpus Proximity Ratio (CPR)
The Corpus Proximity Ratio (CPR), J.-P. Colson 2015

\[
cpr = \frac{n \left( x_{i_1} x_{i_2} x_{i_3} ... x_{i_n} \right)}{n(\max (\|x_i - x_j\|) \leq W)}
\]

\[
x_{i_1}, x_{i_2}, x_{i_3}, ... x_{i_n}
\]
Examples: *take the road, hit the road, fork in the road* (PerlPr)

A database of about 700,000 English candidate collocations has been assembled that way; it lies on the border between low-frequency lexical bundles and phraseology.

Examples: a bit, art, back to, bank, barrel, bridge, bring, jet, road
CPR meets four criteria recommended by Gries (2013) for the improvement of automatic extraction of collocations:

- The measure is directional
- The methodology uses recurrence across corpora
- It is extendable (extended) to multiword expressions
- There may be a pyscholinguistic foundation in the Firthian principle of attraction, and in the comparison with association databases
Experiments along these lines should not be restricted to just a few examples, because only a general view, with huge databases of linguistic material, can corroborate the hypotheses.

It should also be pointed out that computational phraseology (linguistics) depends a lot on the testing of algorithms, and this also pertains to computer science and its recent developments.
Example: testing the CPR-score for hit the road on a corpus of 200 millions words takes the following steps:

* Read whole corpus (about 500,000 A4 pages!), check all offsets (positions in file) for ‘hit’, ‘the’, ‘road’
* For all of the offsets of ‘hit’: check if there is an instance of ‘the’ within a given distance (20 words) and, if so, if there is an instance of ‘road’ within a given distance (20 words)
* Compute the sum of relevant results, check frequency of ‘hit the road’, compare results
* Average computing time per request: 0.07 second
Experiments with the CPR-score make it possible to visualise phraseology, a crucial step in decoding the source text (translators), and a contribution to the theoretical debate on the proportion of phraseology.

Examples with texts in English, French, Spanish.
* From a theoretical point of view, experiments along these lines may also shed some light on the overall statistical distribution of set phrases in large corpora.

* There are many areas of disagreement in the linguistic literature on the statistical or probabilistic foundation of morpheme, word, ngram, chunk or phrase distribution: “Prefabs, it must be remembered, are not like phonemes and morphemes, or noun phrases and verb phrases. They are probabilistic, some more than others” (Erman and Warren 2000: 33)
This is a highly debatable matter, as experiments with the statistical scores used for collocation extraction point to the contrary.

“*If there is one Central American country whose football has come on leaps and bounds in recent years, it is unquestionably Panama.*” (fifa.com, May 2012)

Extraction of set phrases (IdiomSearch): *has come on leaps and bounds, etc.*, but WHAT ABOUT unquestionably? According to several authors (e.g. Beck and Mel'čuk 2011), there are also *morphological phrasemes.*
<table>
<thead>
<tr>
<th>phrase</th>
<th>CPR</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>come on leaps and bounds</td>
<td>0.42</td>
<td>10</td>
</tr>
<tr>
<td>spill the beans</td>
<td>1.00</td>
<td>15</td>
</tr>
<tr>
<td>un – question – ably</td>
<td>0.98</td>
<td>206</td>
</tr>
<tr>
<td>un – think – able</td>
<td>0.96</td>
<td>276</td>
</tr>
<tr>
<td>un – like – ly</td>
<td>0.74</td>
<td>4529</td>
</tr>
<tr>
<td>what – ever</td>
<td>0.99</td>
<td>7454</td>
</tr>
<tr>
<td>un – this – ly</td>
<td>0.00</td>
<td>0</td>
</tr>
</tbody>
</table>
From a theoretical point of view, the automatic extraction of phraseology poses the question of the statistical nature of language, a crucial issue in statistical machine translation.

According to the Zipfian law (Zipf 1949), the general distribution of words displays a very limited number of high frequency items, a fair amount of average frequency words, and a long tail of words with extremely low frequency (an army of dwarves).
* Although the matter is still controversial, Zipf himself interpreted his law as evidence for the *principle of least effort* in language.

* This principle is closely related to the *principle of economy*: language re-uses the same elements in order to avoid linguistic inflation (Martinet’s *double articulation* at the level of morphemes and phonemes, polysemy, phraseology).
In 1953, the mathematician Mandelbrot proposed a slight correction to Zipf’s law; the law of Zipf-Mandelbrot is now the most widely used version

\[ f(w) = \frac{C}{(r(w) + b)^{a}} \]

where \( f(w) \) represents the frequency of a word, and \( r(w) \) its frequency rank. \( C \) and \( a \) are constants that are set empirically according to the data. \( C \) is normally set to the highest frequency value obtained and \( a \) has been set at 1.09 for the British National Corpus. As for \( b \), Baroni (2008) recommends an empirical adaptation by increasing it according to the results, with a typical increase of \( b=1 \) for the 20 highest frequency ranks.
Mandelbrot’s interpretation of Zipf’s law was precisely that word combinations in language follow statistical principles; to put it differently, the basis for phraseology is (largely) statistical...

If this is the case, n-gram frequencies should also display a Zipfian distribution

This is precisely suggested by Baroni (2008)
Experiments tend to show that this is indeed the case, even at the level of small texts.

Example:

“I normally drink between 18 and 24 units per week. I consider the week worryingly quiet if I haven’t had cause to drink one bottle of wine by Wednesday. Tuesday feels far enough into the week to be a heavy night, then the next night is usually a day of relative rest. If I don’t have another big night on Thursday, I will let loose on Friday or Saturday. That’ll be a good bottle. Sunday is generally a recovery night. I pondered the dry week ahead. I realised I had not had one in four years. I felt like a voyager setting off to a featureless land. At 23, did I really need alcohol to enjoy my evenings?” (The Times Online, August 03, 2005)
x stands for the log$_{10}$ of the rank of the items (shown in decreasing order of frequency) and y for the log of the frequency for each item. If the Zipf-Mandelbrot principle applies here, such a log-log table should (for mathematical reasons) display a straight line, with an abrupt fall at the right end of the table and some minor irregularities along the line (Mandelbrot’s corrections).

The next figure presents the log-log results for the bigrams (BigramsDrinking), as well as two projections according to the law of Zipf-Mandelbrot. We follow here the same method as Ha/Sicilia-Garcia/Ming/et al. (2002), who have computed bigram frequencies for the whole Brown Corpus (1,000,000 words): no b factor is used, and a is set to 0.76 (in the legend of Figure 1: ZipfMan076). By way of comparison, Figure 1 also presents a projection with $a=1$ (ZipfMan1).
The provisional database of around 700,000 phrases for English displays a tendency toward the Zipf-Mandelbrot distribution
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. 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).