***R. Salkie***

# DISCOVERING PHRASEOLOGICAL PATTERNS IN TRANSLATION CORPORA

**Abstract.** We are not making as much use of translation (parallel) corpora as we could, and our use of them is not sophisticated enough. We propose a new method, based on *semantic specificity* and *phraseological patterns*. We can identify lexical items in different languages which are similar in meaning but different in their phraseology: these items can be identified in translation corpora, and are likely to be of interest to bilingual lexicographers.

## Missed opportunities

Translation corpora – collections of texts in one language and their translations in another language – contain a vast amount of rich and interesting information about language contrasts. Serious corpus resources and tools are now finally becoming available, notably OPUS[1]. In an earlier paper[2], I surveyed the use of translation corpora across the board, in fields such as machine translation, language learning, and contrastive linguistics. One glaring gap – then as now – is bilingual lexicography, a field where people talk about using translation corpora but have made almost no use of them in reality. This is a pity for practical reasons: the potential of corpus data for enriching bilingual dictionaries is enormous, with the prospect of improved help for language learners and translators.

[1] *Tiedemann J.* 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), 2214-2218. Online: **http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf**

[2] *Salkie R.* 2008. How can lexicographers use a translation corpus? Presented at Using Corpora in Contrastive and Translation Studies (UCCTS), Hangzhou, Sept 2008. Online: **http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf**

Here is a simple example. The French word *permettre* has a wider range of meanings than its English cognate *permit*: as well as the deontic sense «give permission», *permettre* also means «make something possible». The *Oxford-Hachette French-English Dictionary*, one of the best of its kind, distinguishes these senses as «donner l'autorisation» and «donner les moyens». However, it gives the deontic sense first, despite the fact that examples of the other sense are vastly more common. More importantly, it mostly gives translations for the second sense involving *permit* and *allow*. These are in fact very rare in the INTERSECT translation corpus[3]: examples (1) and (2) are among the few. Far more common – and much more natural in English – are examples using *let* and *with ... can*, as in (3) and (4):

*(1) La morphologie [[permet]] de détecter avec précision la localisation des protéines membranaires.*
*Morphology **permits** the precise localisation of synaptic receptor proteins.*

*(2) Monsieur le président, cela nous [[permet]] de nous mieux préparer.*
*Mr. Speaker, this **allows** us to get ready and be more efficient.*

*(3) Cet outil vous [[permet]] d'accepter un mot entier.*
*This tool **lets** you accept an entire word.*

*(4) Le feu est l'un des meilleurs outils de survie. Il vous [[permet]] de vous garder au chaud.*
*Fire is one of the best survival tools. **With** it you **can** keep warm.*

It is disappointing that useful information of this kind has hardly begun to find its way into bilingual dictionaries.

[3] *Salkie R.* 2010a. The INTERSECT Translation Corpus. Online: **http://arts.brighton.ac.uk/staff/raf-salkie/portfolio-of-major-works/intersect**. *Salkie R.* 2010b. Facing the future together: French and English in contrast. In *F. Neveu et al.* (eds.), *Congrès Mondial de Linguistique Française – CMLF 2010* (Paris, Institut de Linguistique Française, 2010). P. 1787–1798. Online: **http://www.linguistiquefrancaise.org/articles/cmlf/pdf/2010/01/cmlf2010_000157.pdf**

## Phraseological patterns

The limited use of translation corpora in bilingual lexicography is also unfortunate from a theoretical point of view, because corpus-driven techniques in the bilingual domain have the capacity to provide new insights into language just that they did in monolingual lexicography: think of John Sinclair's *idiom principle* and *open-choice principle*[4]. Because they examine large quantities of real data in detail, lexicographers who reflect on their investigations deserve the attention of everyone concerned with language.

One lexicographer not averse to theorising is Patrick Hanks, who has developed his framework in a number of recent works[5] (2012, 2013, and for a brief summary, 2009). The key notion is Corpus Pattern Analysis (CPA), which associates each different sense of a word with a set of valency and collocational preferences called a *phraseological pattern* (Hanks 2012: 76). Using the verb *throw* as an example, we can distinguish among its many senses:

| | |
|---|---|
| Sense 1 | to use your hand to send an object through the air. |
| Pattern 1 | **People** throw **hard physical objects** like stones, bricks and bottles **at** **other people and things**, typically but not necessarily with the intention of causing damage. |
| Example | *The police threw tear gas canisters at the demonstrators.* |
| Sense 2 | to force someone to go to prison as punishment |
| Pattern 2 | **The legal system** or **its agents** throw **people** **into prison**. |
| Example | *Many protestors were thrown into jail without trial.* |

[4] *Barnbrook G.* 2009. Sinclair on collocation. In R. Moon (ed.), W*ords, grammar, text: revisiting the work of John Sinclair* (Amsterdam, John Benjamins), 23–38.

[5] *Hanks P.* 2009. Norms and Exploitations: a 'double-helix' theory of language. Abstract of talk at the University of Erlangen. Online: **http://www.lexi.uni-erlangen.de/abstract_hanks.pdf**. *Hanks P.* 2012. Corpus evidence and electronic lexicography. In S. Granger & M. Paquot (eds.), *Electronic Lexicography* (Oxford, OUP), 57–82. *Hanks P.* 2013. *Lexical analysis: norms and exploitations*. Cambridge, MA: MIT Press.

Hanks argues that native speakers store these patterns as «phraseological prototypes» in their heads, rather than as distinct senses: the context narrows down the meaning in particular cases. In a bilingual context, the aim is to present equivalents for these phraseological patterns – «to offer realizations in another language for phraseology that cannot be translated literally, word for word»[6].To do this in a data-driven way, we need to extract equivalent phraseological patterns from a translation corpus. How can we do that?

The most commonly used methods will not help us. Most contrastive research based on translation corpora employs a limited range of strategies. Typically, analysts extract the equivalents of a word or phrase, classify them, calculate the frequency of each one, and analyse the factors which determine the choice of each equivalent. An example of high quality research using this methodology is Celle (2011) on modal adverbs in English and French.[7]

These studies are encouraging and valuable, but we can do more. As Marzo et al. remark, contrastive corpus-based studies require «a multi-methodological approach which is objective and verifiable»[8]. We need a more sophisticated range of strategies for extracting information from translation corpora. In particular, we need strategies for comparing phraseological patterns across languages.

## Semantic specificity

In Salkie (2008) I proposed a method for exploiting translation corpora based on word frequency. Building on that work, the present

[6] *Hanks* 2012: 60.

[7] *Celle A.* 2011. The intersubjective function of modal adverbs: a contrastive English-French study of adverbs in journalistic discourse. In *K. Aijmer* (ed.), *Contrastive pragmatics* (Amsterdam, John Benjamins), 23–36.

[8] *Marzo S., Heylen K., de Sutter G.* 2012. Developments in corpus-based contrastive linguistics. In *S. Marzo, K. Heylen & G. de Sutter* (eds.), *Corpus studies in contrastive linguistics* (Amsterdam, John Benjamins), 1–6.

paper proposes and illustrates another method based on *semantic specificity*. Helge Dyvik had already put forward the idea that «words with wide meanings ought to have a larger number of translations than words with narrow meanings» (2002: 311): he suggests, for instance, that the word *good* will have more translation equivalents than *tasty*.[9] This claim makes intuitive sense, but it must be treated with caution. Firstly, a quick look in a monolingual thesaurus reveals many more synonyms for *good* than for *tasty*: a contrastive analysis adds little to this unsurprising observation. Secondly, other factors play a part in shaping the number of translation equivalents of a word: for example, expressions relating to emotions will contrast strongly across languages no matter how specific they are. Thirdly, Dyvik's claim ignores frequency: the more frequent a word, the more it will accumulate different translation equivalents because of the vagaries of usage in different contexts.

Suppose, however, that we integrate Hank's theory of phraseological prototypes with Dyvik's claims about specificity. If Hanks is right, each sense of a word is associated with (in principle, can be reduced to) a distinct pattern. A word with a wide meaning should have more such patterns than a word with a narrow meaning (other things being equal, as noted in the previous paragraph). Patterns involve valency (grammatically obligatory arguments) and collocates. If we can compare these across languages, any significant differences will be lexicographically interesting – they will count as «nice surprises» to use the technical term introduced in Salkie (2008).

## Reporting verbs in French and English

Here is an example from the domain of verbs used to report speech. The French verb *expliquer* («to explain») is more specific in meaning than the verb *dire* («to say»). Since these two words have

[9] *Dyvik H.* 2002. Translations as semantic mirrors: from parallel corpus to Wordnet. In *K. Aijmer & B. Altenberg* (eds.)*, Advances in Corpus Linguistics: papers from* ICAME 23, (Amsterdam, Rodopi), 311–326. Online: **http://www.hf.uib.no/i/LiLi/SLF/Dyvik/ICAMEpaper.pdf**

obvious equivalents in English, it is possible in principle to investigate the patterns of the French words in a monolingual French corpus and those of the English words separately in a comparable English corpus. This is a useful – in fact, indispensable – step, but a translation corpus gives us access to a wider spectrum of possibilities because we can compare the phraseological patterns of source and target items, including «surprising» equivalents. The aim is to locate the level of specificity of meaning where interesting «surprises» can be revealed.

We looked first at equivalents of *expliquer* in a balanced subset of the INTERSECT corpus. From 49 examples, 25 English translations used a form of *explain*. The other 24 included:

(5) *Ceci avait l'avantage de tout [[expliquer]].*

(6) *That had the advantage of* ***accounting for*** *everything straight away.*

(7) *Comme [[expliqué]] au chapitre précédant, il existe deux systèmes pour râper, broyer ou trancher.*

(8) *You can use the two slicing systems* ***described*** *in the last chapter to grate and chip as well.*

In these instances, the verbs used in English are not the «expected» one, but the patterns are otherwise indistinguishable from the French ones: in (6) we find the word *everything* (cf. *tout* in [5]), while (8) has *in the last chapte*r (cf. *au chapitre précédant* in [7]). Consider, however, these examples:

(9) *En chuchotant il m'a retrouvé et on s'est alors [[expliqué]] tous les deux.*

(10) *He came up whispering and we* ***talked****.*

(11) *Mme Hall, sans hésiter, lui [[expliqua]] les difficultés du pays, et la conversation s'engagea.*

(12) *Mrs. Hall, nothing loath,* ***answered his questions*** *and developed a conversation.*

In these examples the patterns of the English verbs *talk* and *answer* are very different from those of *expliquer*. These examples are therefore of particular interest to bilingual lexicographers, though they are rare for *expliquer*, these being the only two out of the 49 instances.

For the verb *dire* we focused on the form *dit*. Of 196 instances in the corpus, around half (*n* = 91) were translated using *said* – in this respect, no different from the situation with *expliquer* where the obvious equivalent *explain* accounted for about half the examples. And once again we find congruent patterns with other verbs:

(13) *... les membres de son parti et les néo-démocrates nous ont [[dit]] que la question était beaucoup trop brûlante pour en saisir la Chambre des communes.*

(14) *... members of his party and the New Democratic Party* ***told*** *us that it was much too sensitive a subject to be brought before the House of Commons.*

(15) *Puis, faisant allusion à l'intérêt que beaucoup de jeunes Asiatiques portent au taoïsme, il [[dit]], d'une voix plus grave: «La vieille pensée chinoise les pénètre plus qu'ils ne le croient».*

(16) *Then, alluding to the interest many young Asians have in Taoism, he* ***continued*** *in a lower tone: «Ancient Chinese thought permeates their existence more than they think».*

But we also find many instances of very different patterns in French and English, such as:

(17) *... il y a de bonnes raisons pour penser qu'elle est antérieure de quelques semaines au début du journal proprement [[dit]*

(18) *... there is good reason to believe it was written some weeks before the diary* ***itself****.*

(19) *Le genre Plasmodium se divise en deux sous-genres: Plasmodium proprement [[dit]] et Laverania.*

(20) *The genus Plasmodium can be further divided into two subgenera: Plasmodium* ***in the strict sense*** *and Laverania.*

(21) *Pour San Tapeta, on ne pouvait donc pas se tromper, il avait [[dit]] vrai, c'était tout droit devant soi.*

(22) *There was no missing the road to Santa Tapeta;* ***he was quite right****, you simply followed your nose.*

(23) *«On pensera pour vous mon ami! Tenez-vous-le pour [[dit]]».*

(24) *«We will think for you, my friend.* ***Don't forget it****».*

(25) *Un chien commença à hurler, quelque part devant une ferme, au bas de la route, un long hurlement sonore qu'on aurait [[dit]] provoqué par la peur.*

(26) Then a dog began to howl somewhere in a farmhouse far down the road, a long, agonized wailing, **as if** from fear.

Thus from the perspective of phraseological patterns, the verb *dire* with its relatively non-specific meaning is more interesting for a bilingual lexicographer than a more specific verb like *expliquer*: the less specific verb yields a wider range of phraseological contrasts.

## Projectiles in English and German

Consider now the verb *throw*, as discussed by Hanks, and the more specific verb *fire*, with their German equivalents. Concentrating on the throwing or firing of physical objects, the corpus equivalents for *fire* were exclusively *feuern* or *schiessen*, except for one example where the object in question was not a military projectile:

(27) Furthermore, a one-in-a-million figure assumes that one million Cassini space probes have been [[fired]] into space, and only one Cassini space probe malfunctioned.

(28) Außerdem wird mit der Zahl 1:1.000.000 suggeriert, daß eine Million Mal ein Cassini in den Weltraum **geschickt** wird und es dabei nur zu einem einzigen Unfall kommt.

Here perhaps the military overtone of *feuern* or *schiessen* would have been inappropriate. For the more general verb *throw*, on the other hand, there were more «surprising» equivalents, including these:

(29) *They [[threw]] stones and bottles at policemen and shot tracer bullet.*

(30) *Sie* ***schleuderten*** *Steine und Flaschen auf Polizisten und schossen mit Leuchtspurmunition.*

(31) *«To-morrow you'll go on a journey with a stone tied round your neck», went on Frederick, and [[threw]] another stone at the dog.*

(32) *«Morgen sollst du auf die Reise mit einem Stein am Halse», fuhr Friedrich fort und **stieß** nach dem Hunde.*

(33) *He [[threw]] a bombshell.*

(34) *Was er sagte, **wirkte** wie eine Bombe.*

Again we find examples like (30), where the phraseological patterns of *schleudern* and *throw* are almost identical; but also examples like (32) and (34), where the German pattern is very different.

## Implications

Semantic specificity is relevant for bilingual lexicographers. Words with a more general meaning will tend to have a wider range of equivalents and can potentially enrich the information in dictionaries. No one has ever suggested that translation corpora can provide insights into large numbers of lexical items. For a small number, however, these corpora can open up exciting perspectives. If corpus specialists can find ways to direct the attention of lexicographers to this small number, everyone will benefit. This paper has suggested one way to do this.