Comparing Phraseologisms: Building a Corpus-Based Lexicographic Resource for Translators

Laura Giacomini
University of Heidelberg
laura.giacomini@iued.uni-heidelberg.de

Abstract

Today there is still a significant need for specific lexicographic resources in digital form, as they can remarkably improve access to data and actively assist the process of text production. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms and similes in a bilingual dictionary for translators.
Keywords: phraseologisms; corpora; translation

1 Introduction

The characteristic formal stability and contentual figurativeness of phraseological expressions as a result of cultural encoding best reflect a society's deeply rooted patterns of world interpretation. Given the strong presence of phraseologisms in the lexicon of a language and the translatability issues they inevitably raise, it is necessary to support the practice of translation by means of specific and up-to-date lexicographic resources. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms, similes, in a bilingual lexicographic resource for translators. It is based on a larger study carried out in 2013 at the Australian National University (Canberra) on the language pair Australian English (AuE) – Italian and recently published (Giacomini 2014). Translatability issues arise from semantic obscurity linked to the presence of culture-specific words and concepts.

General language dictionaries usually define similes and other multiword expressions through a paraphrase, without providing synonymic idioms even if these are available. This prevents translators from finding functionally equivalent phraseological data, which would be particularly useful for reproducing semantic and pragmatic features as well as the overall familiarity of the multiword expression in the target language. In the case of language pairs with phraseology reflecting distinctive cultural marks (Wierzbicka 2010), the exploration of corpus data can efficiently support the selection of adequate phraseological equivalents through reliable quantitative measures, thus forming a useful dictionary basis.

2 Object and sources

Similes are based on an explicit comparison between entities and are semantically related to metaphors, in which resemblance becomes implicit and one thing is understood and experienced in terms of another (Lakoff/Johnson 2003: 21 ff., Wikberg 2008: 128). In the usually regular syntactic structure of similes, the resemblance relation between the two compared entities is expressed by a connective, mostly *like*/*as* in English and *(così) come* in Italian. In addition, similes can reveal a different phraseological status: they can be defined either as collocations or as semi-idioms, according to the transparency of the comparison.

Two comparison patterns in AuE have been considered, the first being similes containing the phrase *full as* (e.g. *full as a tick*) meaning a) "having no empty space", b) "having eaten to one's limits or satisfaction", or c) "drunk", and the second involving a single culture-specific lexeme as the second compared entity, mostly a native animal. The semantic pivot is the contextual reading of the word that designates the shared property (*tertium comparationis*) and that determines the referential object of the multiword expression as a whole, both on the denotative and the connotative level. However, whereas similes belonging to the first pattern are made up of elements that are semantically transparent in their literal and figurative meanings, both for the English and the Italian native speaker, the others confront the translator with the presence of *realia* (cultural keywords) involving a culture- or environment-specific referent (Peters 2007: 249-251).

AuE similes were chosen on the basis of their relevance in a large-scale digital corpus of full-text Australian general news sources[1], major general language dictionaries of AuE, and selected dictionaries of idioms or colloquialisms. Italian monolingual general language dictionaries and a comparable newspaper corpus (articles published between 2000 and 2013 in major Italian newspapers, totalling around 980 million words) were also employed for this purpose. Due to their stable structure, similes can be split into bigrams. The closest Italian equivalents of the semantic bases (e.g. *full* ≈ *pieno*, *sazio*, *ubriaco*) can be used to query the corpus for their collocators. The absolute frequency of the extracted bigrams can be compared with their log likelihood value, which provides reliable information on the association strength of a certain bigram and thus on its suitability as a phraseological equivalent (cf. Dobrovol'skij 2009[2]). The results of data analysis are displayed in Table 1 according to an onomasiological procedure, which assigns Italian phraseological units to the concepts expressed by the AuE similes. Up to five equivalents for each concept are shown and arranged according to their absolute frequency F in the corpus and the log likelihood ratio LL.

1 National and regional newspaper texts covering the period 1985 to 2012, the Australian Corpus of English (ACE), and the Trove database (National Library of Australia).

2 Log likelihood has been chosen because of its reliability with sparse data, which is the case of the chosen words in the AuE corpus (for the topic of LL and normal distribution cf. McEnery et al. 2006, 53).

to be full as a goog/ state school (hat rack)/ catholic school/ fat lady's sock	
(a) "full/ overcrowded"	essere pieno zeppo (1799/>100), essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come una botte (1/9)
to be full as a goog/ tick/ boot/ fairy's phone book/ fat lady's sock	
(b) "full up/ satiated"	essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come un otre (1/17)
(c) "full/ drunk/ intoxicated"	essere ubriaco fradicio (567/>100); avere bevuto come una spugna (5/21); essere pieno come un otre (1/17), essere pieno come una botte (1/9)
to be mad as a cut snake, to be pissed as a parrot	
(d) "angry/ nervous"	essere (incavolato/...) nero (245/>100); essere arrabbiato/ incavolato/... come una bestia (27/>100)/ una belva (7/>100)/ una iena (6/>100)/ una biscia (5/>100)
to be pissed as a parrot: cf. meaning (c)	
to be mad as a cut snake, to be mad as a gum tree full of galahs	
(e) "crazy/ eccentric"	essere fuori di testa (1548/>100), essere tutto matto (343/>100), essere pazzo/ matto da legare (96/>100), essere pazzo/ matto come un cavallo (27/>100), essere fuori come un balcone (27/>100)
to be (as) game as Ned Kelly	
(h) "game/ brave/ bold"	avere coraggio da vendere (154/>100), avere un coraggio da leone/leoni (130/>100); avere il coraggio di un leone (7/59); essere coraggioso come un leone (5/83)
to be flat out like a lizard drinking	
(j) "fully extended"	essere/ stare lungo disteso (60/39)
(k) "with the utmost effort"	col massimo impegno (251/>100); impegnarsi al massimo (139/>100); col massimo sforzo (9/42); sforzarsi al massimo (2/7)
(l) "very busy"	essere pieno/ oberato di lavoro (404/>100), essere pieno di impegni (376/>100)
(m) "at full speed"	cf. (i)
to be miserable as a bandicoot	
(n) "wretchedly unhappy"	essere un povero diavolo (243/>100), essere un povero Cristo/cristo (238/>100)
(o) "contemptible"	-
(p) "needy"	essere povero in canna (139/>100), essere povero come Giobbe (2/31)

Table 1: Corpus data in the target language with F and LL values.

3 Translatability

AuE similes turn out to have close counterparts among phraseological Europeanisms, even though this may happen to varying degrees of equivalence. *Full* usually retains in the simile both its literal and figurative meanings, thus determining the total or partial compositionality of the multiword expression and contributing to its transparency.

Compositionality can be stated on the denotative (either literal or figurative) and connotative semantic level, but not always on the pragmatic level. The observations concerning compositionality of the multiword expression are also true of the second comparison pattern, in which a variable adjective (e.g. *mad*, *full*, *miserable*...) or a verb (e.g. *to shoot through*) designating shared property is combined with a culture-specific entity, used with a predicative or an adverbial function. In the case of similes belonging to this pattern, *realia* inevitably produce a lexical gap. Compatible data in the target language and culture cannot be sought for in terms of denotatively equivalent phraseological expressions, which is possible in the case of the *full*-pattern, but, at the most, in similes sharing the semantic pivot (*matto*, *veloce*, *coraggioso*, etc.) and with an equal degree of compositionality.

4 Lexicographic representation

The extracted clusters of equivalence candidates disclose the presence of alternative comparative patterns in Italian. For instance, we have prepositional phrases headed by *da* (*avere un coraggio da leone*) or *di* (*avere il coraggio di un leone*). In the Italian language, a significant part of the extracted lexical components in similes and other idioms stereotypically refer to animal behaviour and belong to a common European cultural heritage.

Dictionary data have been a useful resource to identify initial information on some widespread phraseologisms, but they fail to cover context-dependent phraseological variation and variability. The comparable newspaper corpus, instead, has revealed that, in concrete language usage, non-lexicalized and not conventionalised collocative or semi-idiomatic variants performing more specific message functions are constantly created on the basis of already existing patterns. The evaluation of corpus data can also disclose differences in phraseological distribution among languages (e.g. a strong tendency of the Italian languages towards metaphorical comparisons for purely descriptive purposes) and point out recent lexical formations which have not yet been recorded in dictionaries (cf. *essere fuori come un balcone*) but are already perceived by native speakers as familiar.

Corpus analysis in the target language supports the creation of a lexicographic basis for a bidirectional dictionary that is suitable for both translation directions. On the one hand, it activates passive knowledge in the native speaker of Italian by providing him/her with pragmatic tags in the source language and a wide choice of equivalents in the target language. On the other hand, it supports the AuE native speaker who is performing an active translation task by 1) allowing for a statistic evalua-

tion of the word combinations, 2) tagging equivalents with pragmatic marks and, most of all, 3) categorizing phraseologisms with varying idiomatic range (*pieno zeppo*, *pieno da scoppiare*) and distinguishing them from non-phraseological material (cf. Wiegand 2002: 52-53). Among non-phraseological equivalents are often single lexical items, usually an emphasised adjective (*strapieno*, *affollatissimo*; *sbronzo*) that can be specifically sought for in syntagmatic or paradigmatic dictionaries and further tested for phraseological strength. Every time a semantically and pragmatically equivalent phraseologism is missing, dictionary users are provided with an open set of non-phraseologisms, which function as reproducible syntactic models (e.g. superlative adjectives) and are particularly helpful for non-native speakers of the target language.

In order to take full advantage of the rich corpus materials and its bifunctionality, the dictionary should be designed as a digital resource, which should allow the translator to access lexicographic data along different combinable criteria, grasp the semantic connections existing between phraseological expressions, and retrieve unabridged corpus examples for each of them, in both the source and the target language.

The first two goals, *data accessibility* and *the disclosure of semantic connections*, can be primarily achieved through a coherent onomasiological macrostructure, which should group phraseologisms together along a common denotative/connotative meaning, and a systematic mediostructure, the aim of which should be to link each meaning to the correspondent phraseologisms and vice versa. The entry examples below show how lexicographic data in the section AuE-Italian can be displayed in a functional microstructural frame, and arranged based on a specific kind of search input (Table 2 according to a specific concept, Table 3 according to a specific phraseologism)[3].

PHRASEOLOGISMS MATCHING THE CONCEPT IN THE SOURCE LANGUAGE	**EQUIVALENTS IN THE TARGET LANGUAGE**
to be full as a goog *coll.* ≈ **state school (hat rack)** *coll.* ≈ **catholic school** *coll.* ≈ **fat lady's sock** *coll.*	PHRAS: essere (pieno) zeppo, pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come una botte *coll.* ◆ essere pienissimo, affollatissimo, strapieno *coll.*, stracolmo

Table 2: Search input: the concept FULL/OVERCROWDED.

3 ◆ marks the division between phraseological and non-phraseological equivalents, ≈ indicates similes referring to the same concepts.

CONCEPTS MATCHING THE PHRASEOLOGISM IN THE SOURCE LANGUAGE	EQUIVALENT PHRASEOLOGISMS IN THE SOURCE LANGUAGE	EQUIVALENTS IN THE TARGET LANGUAGE
FULL/OVERCROWDED	≈ **state school (hat rack)** *coll.* ≈ **catholic school** *coll.* ≈ **fat lady's sock** *coll.*	PHRAS: essere (pieno) zeppo, pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come una botte *coll.* ◆ essere pienissimo, affollatissimo, strapieno *coll.*, stracolmo
FULL UP/SATIATED	≈ **tick coll.** ≈ **boot coll.** ≈ **fairy's phone book coll.** ≈ **fat lady's sock coll.**	PHRAS: essere pieno come un uovo *coll.*, pieno da scoppiare *coll.*, pieno come un otre *coll.* ◆ essere pienissimo *coll.*, strapieno *coll.*, stracolmo *coll.*
FULL/DRUNK	≈ **tick coll.** ≈ **boot coll.** ≈ **fairy's phone book coll.** ≈ **fat lady's sock coll.**	PHRAS: essere ubriaco fradicio; avere bevuto come una spugna *coll.*; essere pieno come un otre *coll.*, una botte *coll.*

Table 3: Search input: the phraseologism to be full as a goog.

The modular microstructure includes the following items: concept, phraseologism in the source language, phraseologisms in the source language referred to the same concept, and equivalents in the target language (subdivided into phraseological and non-phraseological equivalents). Pragmatic tags are added to a phraseologism or equivalent whenever required.

According to the lexicographic corpus, these Italian similes do not have a marked level of usage. However, a glance at their concordances in the newspaper corpus reveals a frequent tendency towards a colloquial register. In comparison with the source text similes, a generally more neutral level of usage has to be clearly stressed. The equivalents are selected on the grounds of their statistical relevance in the corpus and are not meant to cover the whole spectrum of equivalence in the target language. For the professional translator, they constitute the starting point from which further translation proposals can be generated.

CONCEPT	PHRASEOLOGISM	CONCORDANCES
FULL/ OVERCROWDED	**PHRAS: to be full as a goog *coll.***	The carpark in a certain flat pack emporium, starting with I and ending with A, middle letters K and E, was **as full as a goog**. We drove about 4000km with five adults and a lot of luggage and although the car was **as full as a goog** it was a good performer. She took a chance and opened up Swansea cafe. **Full as a goog**. And she must be doing something right because her business is a finalist in the Cafe category. How weird, though, that Old Trafford can hold only - even when it's **as full as a goog** - 23,000? They reckon they could have sold 70,000 tickets for the last day.
	PHRAS: essere pieno come un uovo *coll.*	Non possiamo farvi entrare – gridano gli organizzatori ai tornelli - dentro è **pieno come un uovo** e non si respire. La platea è quella di lavoratori provenienti da tutta la regione, ieri mattina al PalaDozza (**pieno come un uovo**) E ci piacerebbe che il palasport fosse **pieno come un uovo** (i biglietti numerati sono già stati esauriti ieri in prevendita acclamato come una star nella sua tappa aretina del tour in camper. **Pieno come un uovo** l'auditorium del palaffari

Table 4: Links to corpus concordances.

In order to account for context-dependent variation in meaning and register, each equivalent needs to be hyperlinked to the correspondent corpus concordances both in the source and in the target language, which provide the translator with a large-scale database of real language examples (cf. Table 4 for corpus concordances of phraseologisms related to the concept FULL/OVERCROWDED).
This concept-based macrostructure could also constitute the architecture of a multilingual resource aimed at the representation of a core of cultural scripts shared by different languages (for recent research on Europeanisms cf. Piirainen 2012, Reichmann 2001).

5 Conclusions

Today there is still a significant need for specific lexicographic resources in digital form for translators, as they can remarkably improve access to data and actively assist the process of text production. In the best-case scenario, such resources could be integrated, together with other dictionaries and language tools, in multi-layer databases, allowing for advanced and customised search options.
This study shows that syntactic and semantic patterns can be effectively extracted from corpora and serve as lexicographic data in a digital resource which is specifically designed for supporting translation of culture-specific word combinations thanks to an onomasiological/conceptual macrostructure and a systematic mediostructure. Moreover, the study demonstrates that a corpus-based procedure is able to adequately account for phraseological variation and variability.

6 References

Dobrovol'skij, D. (2009), Zur lexikografischen Repräsentation der Phraseme (mit Schwerpunkt auf zweisprachigen Wörterbüchern). In Mellado Blanco, C. (ed.), *Theorie und Praxis der idiomatischen Wörterbücher*, Lexicographica Series Maior, pp. 149-168

Giacomini, L. (2014), Languages in Comparison(s): Using Corpora to Translate Culture-Specific Similes. In: SILTA Studi Italiani di Linguistica teorica e Applcata, Pacini Editore 3/2013.

Lakoff, G./Johnson, M. (2003), *Metaphors We Live By*, Chicago/London, University of Chicago Press.

McEnery, T. et al. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Milton Park/Abingdon/Oxon, Routledge.

Peters, P. (2007), Similes and other evaluative idioms in Australian English". In Skandera P. (ed.), *Phraseology and Culture in English*, Berlin, Mouton de Gruyter, pp. 235-256.

Piirainen, E. (2012), Widespread Idioms in Europe and Beyond: Towards a Lexicon of Common Figurative Units, Frankfurt am Main, Peter Lang.

Reichmann, O. (2001), Das nationale und das europäische Modell in der Sprachgeschichtsschreibung des Deutschen, Freiburg (Schweiz), Universitätsverlag.

San Vicente, F. (ed.), *Lessicografia bilingue e traduzione*, Milano, Polimetrica

Wiegand, H.E. (2002), Äquivalenz, Äquivalentdifferenzierung und Äquivalentpräsentation in zweisprachigen Wörterbüchern: Eine neue einheitliche Konzeption. In *Symposium on Lexicography XI: Proceedings oft he Eleventh International Symposium on Lexicography*, Copenhagen, pp. 17-57.

Wierzbicka, A. (2010), Experience, Evidence, and Sense: The Hidden Cultural Legacy of English, OUP.

Wikberg, K. (2008), Phrasal similes in the BNC. In Granger S., Meunier F. (eds.), *Phraseology: An Interdisciplinary Perspective*, Amsterdam/Philadelphia, John Benjamins, pp. 127-142.