# Language technology tools and the Phrasal Lexicon

María Fernández-Parra
Swansea University, UK

https://doi.org/10.36505/ExLing-2012/05/0013/000219

## Abstract

This paper explores the possibility of resorting to language technology to automatically extract lists of phrasal items instead of terms. Such items come in multiple and varied forms and constitute a very large part of what we say and write. Two types of tools, concordancing software (*Lexicon* feature in *Déjà Vu*) and automatic terminology extraction software (*MultiTerm Extract* and *PhraseFinder* from *SDL Trados*, *PlusExtract* from *Wordfast*, *Araya*, *ExtPhr32*), were tested on a purpose-built corpus, and their performance was compared to manually extracting the items. The results of the experiments suggest that many of these tools can yield high recall and that since phrases are highly variable, compared to terms, any software designed for phrases should take such variability into account.

Keywords: language technology, phraseology, automatic extraction, computer-assisted translation (CAT).

## Introduction

Since Becker (1975: 62) established that the mental lexicon of English speakers contains more than 25,000 phrases, estimates of the number of items in the phrasal lexicon has increased since then. For example, Glucksberg (2001: 68) proposed a figure of 80,000 items, based on Jackendoff (1995), but this figure may still be under-representative of the vast numbers of phrases in our lexicon, because we constantly coin new phrases to deal with new language needs. Phrases take on special importance because a large proportion of what we say and write is made up of them.

In addition, phrases are extremely diverse; it has been calculated that there are more than 200 terms to describe them, in a bewildering array of terms (van Lancker-Sidtis & Rallon 2004: 211) for somewhat overlapping concepts. Examples of phrasal items range from social formulae (e.g. *Good morning*), collocations (e.g. *striking difference*), idioms (e.g. *hit the hay*), dual phrases (e.g. *facts and figures*), comparisons (e.g. *good as gold*), proverbs (e.g. *birds of a feather flock together*), commonplaces (e.g. *money talks*), complex prepositions (e.g. *in spite of*) and many others.

Therefore, in this paper I explore the possibility of resorting to language technology in order to automatically compile lists of phrases which can then be used, for example, by translators, terminologists and linguists in general to create bilingual vocabulary lists. Such lists can also have multiple uses in computer-assisted translation (CAT).

## Overview of the software

Two types of tools were used in the experiments, concordancing software (*Lexicon* feature in *Déjà Vu*) and automatic terminology extraction software (*MultiTerm Extract* and *PhraseFinder* from *SDL Trados*, *PlusExtract* from *Wordfast*, *Araya*, *ExtPhr32*). The novel side of this research consisted of applying such tools to phrasal items, rather than to terminological items. Below is a brief introduction to each tool; for a more detailed overview and workflow of each tool, see Fernández-Parra (2012).

### Lexicon in Déjà Vu

Déjà Vu was developed in 1993 by a Spanish telecommunications engineer, Emilio Benito (Nogueira 2004), and is currently provided through his Madrid-based company Atril (*www.atril.com*). It has a unique built-in feature to CAT tools, the Lexicon, which is an index of all words and phrases from a text, to be used alongside terms from a terminological database. Although the Lexicon is not a typical concordance tool, that is the use made of it here.

### MultiTerm Extract and PhraseFinder (from SDL Trados)

SDL Trados (*www.translationzone.com/en*) is the result of the merger of two companies, Trados and SDL International in 2005 (DePalma 2005: 4). Its main product, a CAT tool also called SDL Trados, is the current leader in the CAT tools market (e.g. Cocci 2009, García 2005). MultiTerm Extract was inherited from Trados and is a statistics-based extraction tool, whereas PhraseFinder was inherited from Trados and is mainly a linguistics-based extraction tool.

### PlusExtract (from Wordfast)

Wordfast (*www.wordfast.com*) was developed by Yves Champollion in 1999 (Wassmer 2008) and it is described as a simpler and free alternative to SDL Trados on the Wordfast web site. PlusExtract is a statistics-based term extraction component in +Tools (or PlusTools), a small but powerful set of tools in Wordfast that can be downloaded for free as a standalone suite.

### Araya

Araya is a statistics-based bilingual terminology extraction tool developed by Dr. Klemens Waldhör in 2002 according to its own web site (*www.heartsome.de*). Although Araya only performs bilingual term extraction, in this research only the monolingual part was used, in order to compare its performance to that of the other tools.

### ExtPhr32

ExtPhr32 is a freeware program for statistical monolingual term extraction developed by Prof. Tim Craven from the University of Western Ontario, Canada, in the 1990s. It can be downloaded from Prof. Craven's web site, *http://publish.uwo.ca/~craven/freeware.htm*.

## Setup of the experiments

The research followed two stages. First, a purpose-built corpus was collected of about 200,000 words and every phrasal item was extracted manually. In total, 1,985 types and 4,183 tokens were found, and these figures were then used as benchmarks for comparison.

In the second stage, the corpus was processed with each of the tools. Because there were several settings and combinations of settings possible in each tool, the corpus was processed 387 times, that is, once with every setting and combination of settings available in each tool.

## Evaluation of the results

Because of the variety of settings available in each tool, only the setting to select the maximum number of words per returned string was used when comparing the results obtained (*Max setting* for short), since it is provided by all the tools. For the evaluation of the results, the Max setting is taken into account together with measures of precision, recall and F-measure as described by Manning and Schütze (1999).

The Lexicon (concordance-like method of extraction) achieved full recall with a Max setting of 6, but precision was extremely low (0.3% at best). This means that every target item was retrieved, but the user would have to search through a long list of items in order to find the target ones.

The statistical extraction tools (MultiTerm Extract, PlusExtract, Araya and ExtPhr32) produced varying recall (from 13% to 98%) but they returned the highest precision (16.6% with MultiTerm Extract). This means that almost all the target items can be retrieved, especially with ExtPhr32 and PlusExtract (98% recall each), but post-editing tasks are considerably reduced, since the list of items to search through is shorter.

PhraseFinder (the linguistics-based tool) returned very low recall (9% at best) and low precision (1.8%). This result is hardly surprising if we consider that linguistic methods of extraction have a number of built-in rules to extract terms specifically, which tend to consist of noun phrases, rather than the general phrases I tried to extract with it, many of which consisted of verbal phrases. In order to improve this result, new rules would have to be built into the program specifically designed for every type of grammatical phrase, more specifically verbal and prepositional phrases, which in my corpus amounted to 75% of the total number of phrases.

One problem common to all tools was that of variable phrases such as verbal phrases. For example, the component words of the phrase *break the ice* do not always occur adjacent to one another and *break* may occur in a number of forms, e.g. *broke*, *was breaking*, etc. This means that the Max settings which produced the best results ranged from Max 6 to Max 20, to

allow for varying amounts of intervening material between the component words.

Although the similarities between terms and phrases in general provided a theoretical framework at the start of the research, the results obtained rather highlight the differences between them. There is considerable scope for further research on the full extent of variability in phrases but, in the meantime, I hope to have shown that the combination of language technology and phraseology is a fruitful one.

## References

Becker, J.D. 1975. The Phrasal Lexicon. Proceedings of the 1975 ACL Workshop on Theoretical Issues in Natural Language Processing. Cambridge, MA.

Cocci, L. 2009. CAT Tools for Beginners. Translation Journal 13, 4.

DePalma, D.A. 2005. Gala Report: SDL Trados Merger Survey Results. GALAxy Newsletter.

Fernández-Parra, M. 2012. Formulaic Expressions in Computer-Assisted Translation. PhD thesis, Swansea University, UK.

García, I. 2005. Long Term Memories: Trados and TM Turn 20. JoSTrans, the Journal of Specialised Translation 4, 18-31.

Glucksberg, S. 2001. Understanding Figurative Language. From Metaphors to Idioms. Oxford, Oxford University Press.

Jackendoff, R. 1997. The Architecture of the Language Faculty. Cambridge, MA, Newbury House.

van Lancker-Sidtis D. and Rallon G. 2004. Tracking the Incidence of Formulaic Expressions in Everyday Speech: Methods for Classification and Verification. Language & Communication 24, 207-240.

Manning, C. D. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA, MIT Press.

Nogueira, D. 2004. In Memoriam: Emilio Benito. Translation Journal 8, 2.

Wassmer, T. 2008. Wordfast 5.5 Classic and a First Glance at Wordfast 6.0. Retrieved from *www.localizationworks.com/DRTOM/Star/STAR.html* on 20 March 2012.