Discovering clines of variation

The location and analysis of non-canonical forms in general reference corpora

Gill Philip
University of Bologna – Italy

Keywords: categorization, fixedness, units of meaning, variation, word play

1. Introduction

As the growth of large corpora and the resulting move away from the serendipitous collection of language examples makes available more and more data from which language norms, and deviations from them, can be identified, studies into phraseology continually reaffirm the fact that “so-called ‘fixed phrases’ are not in fact fixed” (Sinclair 1996:83). Yet, in spite of this recognition, the discovery of non-canonical forms in corpora is still generally considered to be a matter of good fortune (Moon 1998:51). Further to this, the fact that the data which can be extracted is heavily restricted to that which is entered into the search query is cited all too often as one of the drawbacks of corpus-related research into the phenomenon (Moon 1996:252; Deignan 1999:197). After all, how can one search for something without knowing what that something is? This paper presents procedures for the location of non-canonical ‘fixed’ phrases, and puts forward some arguments in favour of their inclusion in phraseological analysis.

2. Searching for variant forms

Corpus query syntax varies from programme to programme, but a factor shared by the majority of applications is that search procedures are rarely exploited to the full. Although tagging allows fairly abstract searches to be carried out in the search for grammar patterns (see, for example, Hunston & Francis 1999), the retrieval of lexical variation is all too often limited to word forms and lemmas rather than phraseological structures (Moon 1996; Cia&gni & Coffey 2000:550). But this need not be the case: serendipity can be replaced by the systematic, chance by trial and error.

This paper presents a method for the retrieval of variant forms from computer corpora (in this case, the Bank of English), demonstrating how key-words, wild cards and phrasal structure all contribute to the identification of underlying schemas which can in turn be translated into search queries. It is worth noting that the same principles are relevant to the use of advanced queries in Internet search engines, with the result that the linguist is able to extend and verify his/her findings on a much larger, if less homogeneous, data set – a factor of considerable importance when dealing with multi-word strings, where the location of sufficient examples for analysis can often pose a problem.
3. The linguistic value of phraseological variation

If finding the data is in itself a barrier to the study of variation, so too is the fairly low esteem in which it is held in language description. Perhaps surprisingly, canonical forms of idioms and other figurative phrases are actually quite uncommon in language corpora and are, as a general rule, outnumbered by their corresponding non-canonical forms. Despite this observation, non-canonical forms are still considered to be exceptions to the norm. They are peripheral to the interests of lexicography and foreign language teaching because they can ultimately be reduced to a canonical form, and it is this underlying canonical form, not its variants, which requires documentation and learning.

3.1. The semantic effects of variation

At the far extreme of the scale of variability lie puns and other types of word plays. Unlike their less showy counterparts, these attract considerable interest on the part of linguists. The many layers of meaning that they involve are held in place by an underlying canonical structure which appears to generate an infinite variety of novel utterances – see Partington (1998:121-143) on the exploitation of fixed phrases in newspaper headlines; see also Moon 1998; Philip 2003. In the study of word play, novel utterances are typically contrasted with their canonical relatives, all but ignoring the relationship that they have with the common-or-garden variation created on the fly by language users. A large swathe of language is overlooked as a result of it being neither normal nor exceptional enough to merit attention. Yet, it is precisely to this little-studied middle-ground that phraseology scholars can turn in order to discover more about language and the use that its speakers make of it.

The average language user can and does manipulate conventional structures to the particular communicative situation, but contrary to what our intuition might suggest, this does not necessarily imply that double meanings, humour, irony or other deliberate textual effects are created. What the study of variant forms shows is the extent to which learned forms can be adapted to make them contextually appropriate, whether this be in order to disambiguate, specify or reinforce meanings.

3.2. Factors affecting phraseological variation

In the view taken in this paper, canonical phrases are seen as a particular sub-set of collocational frameworks (Sinclair & Renouf 1991). This means that they provide a structure which permits the productive variation of some slots while resisting change to others. The most revealing productivity occurs along the paradigmatic axis where terms are substituted not only by members of the same semantic set but also by apparently unrelated terms. This phenomenon can be accounted for by the Class Inclusion Hypothesis (Glucksberg & Keysar 1993; Glucksberg & McGlone 1999), which posits that although classes are traditionally considered to be taxonomic, in the case of metaphorical language, they are attributive. If the relationship between the substituted term and the canonical one is based on common attributes, then everyday, unmarked variation would appear to be functioning along different lines from those which are intuitively believed to be in operation. This contributes significantly to the difficulties encountered in trying to predict how variant forms will manifest themselves. It also helps to explain why computational models of language find non-canonical forms difficult to account for, whereas most language users find both their interpretation and production unproblematic.
The study of canonical forms alongside non-canonical forms in all their guises also highlights the relationships which hold between phraseological items and their co-textual environments. In fact, the analysis of corpus data demonstrates that the central element of an extended unit of meaning (Sinclair 1996), typically taken to be a single word, can just as readily take the form of an entire phrase, canonical or otherwise (Philip 2003). The data suggests that although it is possible for a canonical phrase to occur in an unusual co-text, thus triggering contextually-relevant interpretations, it is extremely rare for a non-canonical form of a phrase to occur in a context which differs markedly from the norm (Philip 2003). When it does occur, it tends to occur in transcribed spontaneous speech when two similar structures are fused during on-line processing, resulting in a 'crack' in the phraseological priming (Hoey 2005:11).

4. Conclusions

It is the unpredictability of non-canonical forms that makes them difficult to extract from a corpus, and this certainly hinders attempts to study them comprehensively. Although this paper shows that their location is not governed by mere happenstance, the manual extraction of the type described here still poses a challenge to automation: automatic extraction requires a degree of stability which non-canonical forms do not always comply with. However, the study of variant forms can contribute fruitfully to existing linguistic knowledge bases, providing a semantic dimension that has yet to be explored in corpus studies. This in turn will contribute towards future success in automatic extraction of variant forms from corpora, as well as to shaping and refining our understanding of how natural languages really work.

References

PHRASEOLOGY 2005
The many faces of Phraseology
An interdisciplinary conference

PHRASÉOLOGIE 2005
La phraséologie dans tous ses états
Colloque interdisciplinaire

Louvain-la-Neuve, 13-15 October 2005

EDITORS
C. COSME, C. GOUVERNEUR, F. MEUNIER & M. PAQUOT

2005
<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blumenthal Peter</td>
<td>Profil combinatoire des mots: analyse contrastive</td>
<td>1</td>
</tr>
<tr>
<td>Gross Gaston</td>
<td>Réflexions sur la notion de figement</td>
<td>5</td>
</tr>
<tr>
<td>Heid Ulrich</td>
<td>Computational Phraseology: Approaches to the computational analysis and representation of phraseological units and to their extraction from text corpora</td>
<td>13</td>
</tr>
<tr>
<td>Kennedy Graeme</td>
<td>Focus on Form and Meaning: Phraseology in Foreign Language Learning and Teaching</td>
<td>17</td>
</tr>
<tr>
<td>Sinclair John</td>
<td>The phrase, the whole phrase, and nothing but the phrase</td>
<td>19</td>
</tr>
<tr>
<td>Wray Alison</td>
<td>Looking at the WHY in Phraseology: a psycholinguistic perspective on patterns in text</td>
<td>23</td>
</tr>
<tr>
<td>Adolphs Svenja</td>
<td>Multi-word units and second language speech fluency</td>
<td>27</td>
</tr>
<tr>
<td>Alba-Salas Josep</td>
<td>'Hacer miedo' and 'dar miedo': a Corpus-based Diachronic Study of Spanish Collocations with 'hacer/dar' + State Noun</td>
<td>31</td>
</tr>
<tr>
<td>Arnaud Pierre, Ferragne Emmanuel, Lewis Diana &amp; Maniez François</td>
<td>[Adj+N] sequences in attributive or NP-final positions: observations on lexicality</td>
<td>35</td>
</tr>
<tr>
<td>Augusto Maria Celeste</td>
<td>'Sourd comme une porte' ou 'sourd comme une caille'? Une approche contrastive des expressions phraséologiques du type comparatif en portugais et néerlandais</td>
<td>39</td>
</tr>
<tr>
<td>Ballier Nicolas</td>
<td>MWUs licensing noun complement clauses in English</td>
<td>43</td>
</tr>
</tbody>
</table>
Barlow Michael
Collocate

Biorci Grazia & Cini Monica
Tirare su un bambino'... L'élever ou le soulever? Aspects phraséologiques dans les verbes syntagmatiques

Boers Frank & Stengers Hélène
Adding sound to the picture: an exercise in motivating the lexical composition of idioms in English, Dutch and Spanish

Bolly Catherine
Constructions récurrentes, collocations et séquences figées avec le verbe à haute fréquence 'prendre': Pour une méthode «mixte» d'analyse de corpus (FL1 et FL2)

Čermák Frantisek & Blatná Renata
Grammar idioms

Coffey Stephen
Exploring the phraseological environments of the particle 'a' in modern English

Colson Jean-Pierre
Improving translation practice by the semi-automatic extraction of set phrases

Cosme Christelle & Gilquin Gaëtanelle
Free and bound prepositions in a contrastive perspective: the case of 'with' and 'avec'

Coxhead Averil
"I like to learn verbs because you only have to learn one word"

Davis Boyd & Lunsford Ronald
Metonymy in Alzheimer's speech: when the whole is more than the sum of its parts

DeCesaris Janet & Battaner Paz
Idioms and Fixed Expressions in Learner's Dictionaries: Proposals for Spanish

De Cock Sylvie
The routine aspects of speech and writing, a comparative corpus-driven study of native and nonnative performance data

Dister Anne, Fairon Cédrick & Watrin Patrick
Recherche d’expressions figées dans la presse. Méthode et observations

Diwersy Sascha
La combinatoire lexico-syntaxique des noms de qualité agentive en français et en allemand
Dubroca Galin Danielle
La phraseologie au service de la mercatique 101

Dziadkiewicz Aleksandra
Analyse de phraséologismes pragmatiques français et polonais en vue d'une traduction automatique 105

Evert Stefan
Empirical Research on Association Measures – the UCS Toolkit 111

Evert Stefan & Krenn Brigitte
Exploratory Collocation Extraction 113

Eyckmans June, Kappel Jenny, Stengers Hélène, Demecheleer Murielle & Boers Frank
Formulaic Sequences and Perceived Oral Proficiency: Putting a Lexical Approach to the test 117

Farø Ken
On Motivation/Iconicity and Arbitrariness of Idioms 121

Fedulenkova Tatiana
Isomorphism and Allomorphism of English, German and Swedish Phraseological Units Based on Metaphor 125

Fedulenkova Tatiana & Anufriyeva Maria
Some Tendencies in Lexicalisation of English Phraseological Units 129

Fiedler Sabine
Phraseology as a Learning Aim and Aid: Language Play in the Classroom 133

Fitzpatrick Tess
An exploration of factors affecting the acquisition and use of formulaic sequences in the L2 137

Forsberg Fanny
Prêt-à-parler: le rôle des séquences préfabriquées en français parlé L2 et L1 141

Frath Pierre & Gledhill Christopher
Qu'est-ce qu'une unité phraséologique? 145

Fujimura Itsuko & Nakao Hiroshi
Noms composés et choix de l'article 'de' et 'des' en français 149

Gilquin Gaëtanelle
To take or not to take phraseology into account: the place of multi-word sequences in corpus data and experimental data 155
The phraseological patterns of high frequency verbs in advanced English for General Purposes: new perspectives

Granger Sylviane
Pushing back the limits of phraseology: how far can we go?

Gustawsson Elisabeth
'Straw-clutchers', 'kite-flying' and 'trail blazing': on derivations of verbal idioms in the British National Corpus

Hammer Francoise
Le diable et le bon Dieu - rencontres phraséologiques

Handl Susanne
Collocation - a question of direction and weight

Hanks Patrick
Resonance and the Phraseology of Metaphors

Heyvaert Liesbeth
A phraseological approach to nominalization: issues of collocation and colligation

Hümmer Christiane
Meaning and Use: a Corpus-based Case Study on German Idiomatic Multi-Word Expressions

Ishida Priscilla
A method for the contrastive analysis of idioms: Japanese and English idioms of anger

Jaglińska Anna
Idiomaticity in learner language: a study of the incidence and variability of prefabs in advanced EFL learners' writing

Jautz Sabine
The Use of Expressions of Gratitude in Radio Phone-ins and Radio Interviews in British and New Zealand English: Instances of Politeness or of Impoliteness?

Jones Steven, Murphy Lynne, Paradis Carita & Willners Caroline
Googling for Opposites: a phraseological approach to assessing antonym canonicity

Kerz Elma
A construction grammar approach to the analysis of lexico-grammatical patterns in academic texts
Klein Jean & Lamiroy Béatrice
Relations systématiques entre paires d'expressions verbales figées à travers quatre variétés de français

Kozak Katarzyna
Idioms of paralanguage: a comparative study

Kübler Natalie
Phraséologie verbale transdisciplinaire et spécifique: essai de classification en sciences de la terre

Lecler Aude
J'ai la mémoire qui flanche, j'me souviens plus très bien... Le défigement: réinvestissement et réinitialisation dans le cycle phraséologique

Legallois Dominique
Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons: la grammaire de construction et la grammaire contextualiste

Levison Michael & Lessard Greg
Generating Complex Verb Phrases: an Extension of Burger's French Verbal Morphology

Lindquist Hans & Levin Magnus
FOOT and MOUTH: the phrasal patterns of frequent nouns

Maclagan Margaret, Davis Boyd & Tillard Gina
Fixed phrases in the speech of patients with dementia

Manerko Larissa & Tuarmenskaya Angela
Spatial Conceptualization of Human Body Parts in English Phraseology

Martí Solano Ramón
Variation and Exploitation of Phraseological Units of Body Parts in the British Press: Types of Modification

Martin Willy
A Unified Approach to Semantic Frames and Collocational Patterns

Mason Oliver
Automatic Identification of English Multi-Word Units

Meunier Fanny
Fine-tuning the lexical approach

Moon Rosamund
'White as snow, red as blood, black as ebony': conventionalized English similes and corpus evidence
Mortchev-Bouveret Myriam
Étude phraséologique du verbe CASSER : fonctions lexicales et classes sémantiques pour un encodage lexicographique

Mudraya Olga, Piao Scott S.L., Löfberg Laura, Rayson Paul & Archer Dawn
English-Russian-Finnish Cross-Language Comparison of Phrasal Verb Translation Equivalents

Murano Michela
La phraséologie dans les préfaces des dictionnaires bilingues français-italien, italien-français

Naciscione Anita
Cognitive Aspects of Visual Representation of Phraseological Image

Neff van Aertselaer JoAnne
Contrasting English-Spanish discourse phrases: A corpus study

Nenonen Marja
Prototypical idioms: evidence from Finnish

Nuccorini Stefania
"Absolutely no? Absolutely not!" On some contrastive aspects in the phraseology of English and Italian

Oakey David
Absolute and pragmatic frequencies of word combinations in academic journal articles in English

Omazić Marija
Conceptual Motivation and Processing of Idioms and Idiom Modifications

Oncins-Martínez José
Another plea for 'phraseology'. Or how literary texts can contribute to the development of phraseology: the case of Shakespeare

Orliac Brigitte
Extraction automatique de collocations en langue de spécialité: une approche combinée

Osborne John
Phraseology effects as a trigger for errors in L2 English: the case of more advanced learners

Paquot Magali
EAP vocabulary in learner corpora: a cross-linguistic perspective

Parizoska Jelena
False pairs in phraseology
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pazos Bretaña José Manuel &amp; Pamies Bertrán Antonio</td>
<td>Convergence de critères statistiques et grammaticaux pour la récupération d'unités phraséologiques en corpus électronique</td>
<td>331</td>
</tr>
<tr>
<td>Pecman Mojca</td>
<td>Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions</td>
<td>335</td>
</tr>
<tr>
<td>Philip Gill</td>
<td>Discovering clines of variation: the location and analysis of non-canonical forms in general reference corpora</td>
<td>339</td>
</tr>
<tr>
<td>Piébard Sophie &amp; Bestgen Yves</td>
<td>Identification automatique des marqueurs globaux du discours par l'analyse des expressions récurrentes</td>
<td>343</td>
</tr>
<tr>
<td>Piirainen Elisabeth</td>
<td>Widespread Idioms in Europe and Beyond: Toward a Cross-linguistic and Cross-cultural Research Project</td>
<td>347</td>
</tr>
<tr>
<td>Pontonx Sophie de</td>
<td>Expressions idiomatiques françaises et allemandes: vrais et faux amis. Une aide à l'apprentissage des phrasèmes</td>
<td>351</td>
</tr>
<tr>
<td>Prat Zagrebelsky Maria Teresa</td>
<td>Difficulties in dealing with phraseological patterns on the part of Italian advanced EFL learners: a corpus-based approach</td>
<td>357</td>
</tr>
<tr>
<td>Reboul-Touré Sandrine &amp; Pagnier Thierry</td>
<td>Rouge comme une tomate' ou 'bleu comme une orange': collocations et créations dans le cadre de la didactique du français langue maternelle</td>
<td>359</td>
</tr>
<tr>
<td>Renouf Antoinette</td>
<td>Issues of automatic phrase retrieval in web text</td>
<td>363</td>
</tr>
<tr>
<td>Römer Ute</td>
<td>A phraseology-driven approach to identifying evaluation in a book review corpus</td>
<td>369</td>
</tr>
<tr>
<td>Romero Fresco Pablo</td>
<td>The analysis of phraseological translation in a parallel audiovisual corpus: a proposal for a dictionary of idioms in use (English-Spanish)</td>
<td>373</td>
</tr>
<tr>
<td>Sabban Annette</td>
<td>Some considerations on the culture-boundness of phraseology</td>
<td>377</td>
</tr>
</tbody>
</table>
Schmitt Norbert
Processing advantages of formulaic sequences vs. non-formulaic sequences in text passages

Segermann Krista
Les unités phraséologiques comme unités d'apprentissage dans l'enseignement du français langue étrangère

Siepmann Dirk
Phraseology and Learners' Dictionaries: Coverage, Generativity, Semasiological vs. Onomasiological Approaches

Simpson Rita & Ellis Nick
An Academic Formulas List (AFL)

Stengers Hélène, Eyckmans June, Horemans Arnout & Boers Frank
Optimising mnemonic strategies through CALL: a tool called Idiom Teacher

Stubbs Michael
The most natural thing in the world: quantitative data on multi-word sequences in English

Svensson Maria Helena
Un critère de figément très nuancé: la non-compositionalité

Tschichold Cornelia
A computational lexicography approach to phraseologisms

Tutin Agnès
Collocations du lexique transdisciplinaire des écrits scientifiques: annotation et extraction des propriétés syntaxiques

Valetopoulos Freiderikos
Pour exprimer son état émotionnel en grec: le cas des locutions verbales

Van der Wouden Ton
On the phraseology of stop words

Vangehuchten Lieve
L'identification de l'unité phraséologique en terminologie: énoncé du problème à partir de l'analyse d'un corpus de discours économique d'entreprise

Vásquez González Juan Gabriel
"The company a word keeps": collocates as a means for categorizing Old English
Vergely Pascale
*Mise au jour de structures régulières dans un corpus oral de communications de travail: l'exemple de la navigation aérienne*

Verlinde Serge, Selva Thierry & Binon Jean
*Les unités polylexicales: de leur détection à leur apprentissage*

Wanzeck Christiane
*The Etymology of Phraseological Units*

Wikberg Kay
*Similes in the BNC*

Williams Geoffrey
*The Good Lord and his works: A corpus-based study of collocational resonance*