Phrasal creativity viewed from an IT perspective

ANTOINETTE RENOUF

Introduction

If there is one particularly active area of linguistic study and debate at the moment, it is in the study of phrases. Rarely has such a gamut of terminology referring to essentially the same phenomenon (albeit from different standpoints) been generated and so long sustained: from 'multi-word units' to 'compounds: 'chunks: 'bundles: 'phraseologismes' and so on.

This paper is the result of yet another investigation of phrases. I recently conducted an informal study into our *WebCorp* tool (http://www.webcorp.org.uk/wcadvanced. html) (Renouf, 2002, Kehoe & Renouf, 2002; Renouf *et aI*, 2005 & forthcoming) and its potential for retrieving abundant and enlightening instances of up-to-date phrasal productivity, variability and creativity from web-based text. The purpose was to test the pattern-matching functions of the software with a view to improving them in the light of better understanding of the structure of phrases. I decided to focus on established fixed expressions, idiomatic, clausal and tending towards aphorism and proverb. Established fixed expressions were assumed to be appropriate vehicles for studying both convention and creativity, since in actual language use, most if not all so-called frozen expressions can be and regularly are modified.

That initial investigation showed that a wealth of information about conventional, variable and creative phrasal use can be garnered from web text with a flexible pattern

search mechanism. It also indicated that both conventional and creative uses of language operate according to clear rules, suggesting some ways in which automated pattern search could be improved. However, it also showed that, both in the core and at the boundaries of the phrase, the rules involved are not all amenable to surface text search routines. Lexical, collocational and grammatical requirements can often be predicted by the user and a matching procedure devised, but the observed phrases also revealed semantic and discoursal features which are not so readily identifiable by automated means. Another problem, both for the human to specify, and for the computer to recognise, is the fact that phrases exist at different degrees of fullness of articulation, from minimal and allusive, to fully expressed and contextualised.

This paper will seek to identify some of the less tractable aspects of phrasal behaviour which are faced in automated pattern - matching, particularly in the process of enhancing a first-stage surface search through the application of linguistic knowledge.

1. Data and Method

This was an initial investigation, and the sample of conventional phrases was selected arbitrarily, although they tended towards the more idiomatic, proverbial and lexically rich. They were edited down to form templates, patterns with wildcards, according to the existing search options offered by the *WebCorp* tool devised by my research unit. The editing made assumptions about the phrase with regard to potentially variable or additional elements, substituting these with a wildcard or a set of inflectional variants.

It should be said that the purpose was not to carry out a quantified, corpus-linguistic study, since the WebCorp tool is currently on-line, and thus unable to supply any sensible statistics of occurrence. The relatively casual approach was felt adequate for the intended qualitative study. The phrasal output was primarily extracted from the 'UK broadsheet newspaper' domain, using an optional 'domain' filter, with a wider search where a richer crop was desired. The results presented in this paper have been edited in order to highlight an indicative sample of the phrasal features that must be dealt with in improved pattern-matching.

2. Classification of phrasal rules

The data have been tentatively placed into three categories. These are intended to reflect and correspond to
2.1 phrase-internal constraints (lexis, colligation)
2.2 boundaries and completions (discourse)
2.3 beyond-phrase completions (grammar)
The findings in the first two categories will probably be familiar. The third section, however, yields facts that may be new to the reader.

2.1 phrase-internal constraints

2.1.1 Lexico-semantic constraints

This section illustrates some types of lexico-semantic constraints operating within a phrase, information which must be taken into account in developing automated phrasal search.

The first idiomatic phrase was selected out of passing curiosity, in response to a statement once made by Cruse (1986) to the effect that there was very restricted variability available for certain idiomatic phrases, of which *kick the bucket* was one, and *chip on the shoulder* was another. My expectation was that a wide range of variability would actually be found in real text. On the assumption that the conventional pattern was something like *he~ got a (big) chip on his shoulder*, and that the creative variations would be concentrated in the modification of the object noun phrase, the pattern submitted for search was *[chip on * * shoulder]*. The results are shown in Figure 1.

1. their newspapers good name dragged down by the weight of the enormous chip on their reporters shoulder
2. Drinks make you overlook the huge chip on the bartender~ shoulder. '
3. Heartbroken lover, with a great big chip on his lonely shoulder
4. The north-eastern support with the justified chip on its collective shoulder
5. I *have a big chip on my genealogical shoulder about such omissions*
6. Houston would bring a 43-year chip on the Astros' shoulder
7. Matthews' treatment of Michelle Malkin was the hidden chip on Millers shoulder

Figure 1: search results for pattern *[chip on * * shoulder]*

Figure 1 shows that the truth lies somewhere between Cruse's and my own intuitions: there is more variability than Cruse allows but less that I anticipated, as well as unexpected restrictions on the type of variation. There is indeed frequent modification of the object NPs, but in the case of objNP 1, *chip*, there is a semantic restriction to expressions of size: e.g. *large, massive, big, huge;* while in the case of objNP2, *shoulder*, the modification generally requires the semantics of sociology, with reference to social group characteristics, such as *collective, teams', feminist.*

In the phrase (X *is) a storm in a teacup*, the two NPs allow limited substitution by other nouns alluding to the particular context to which the idiom is applied, but the substitution within the search pattern *[a * in a * cup]* is typically *tempest* for *storm* and *coffee* for *tea*. Less predictable is the use, shown in Figure 2, of downplaying modifiers such as *just, was all, no more than;* and the modalisation of the verb BE.

1. This issue is just a storm in a tea cup
2. It could be termed a storm in a tea cup,
3. You may drink Horlicks but in Burnley this would be a storm in a tea cup 4. Is EIB a storm in a coffee cup?
5. the cybercafe domain was all a storm in a coffee cup
6. this situation is just a tempest in a coffee cup
7. A possible espresso tax here is creating a tempest in a coffee cup

Figure 2: search results for the pattern *[a * in a * cup]*

The idiomatic phrase *[drive* + NPhum + *round the bend]*, with wildcards inserted for the variable elements, as in [X *[dr{ilo]ve[slnl] him [al]round the] bend]*, is realised in web-based text as shown in Figure 3.

1. she really drives me round the bend with her nagging
2. Conor asks Stacey for help as his mum drives him round the bend
3. Maybe losing his wife was enough to drive him round the bend
4. He affirms that US driving rules literally drive him round the bend
5. over used, that stupid drumbeat drove him round the bend
6. The noise drove her around the bend
7. fury over lorry drives residents round the bend
8. we quit - you've driven us round the bend
9. we'll drive you round the island or take you shopping
10. Her Majesty was driven twice round the Mews yard
11. 'sick' Diana pic drives critics round the Benz

Figure 3: Search results for the pattern X *[di[i\o]ve[s\n\] him [a\]round the] bend*

The output indicates that the agent or subject of the clause, is not realised as an arbitrary or entirely free choice, but is conventionally a noun phrase which has the semantics of 'potential to be a cause: typically realised in web text by one or other of the semantic components of: +female; +bureaucracy; +behaviour; +insecurity; +noise; +computing technology. This pattern also raises the problem of ambiguity, in that there is similar lexical realisation for literal (9, 10 above) and metaphorical uses. However, a deeper search of source text would find semantic clues or signals (Renouf, 2002) such as the addition of an adverbial, e.g. twice in 10, in the literal use, or the absence of a causative NP by which to differentiate the two.

The idiomatic phrase pattern X *gets short shrift* is assumed to be relatively fixed, and thus the minimal and assumed core of the pattern *[shrift]*, was submitted. The results in Figure 4 show this assumption to be correct, indicating that the variation in the verb tends to be between the *get* and *give*, and the recipient subject NP of *give* is typically human and the indirect object usually a pronoun and usually denoting an abstract noun *(growth, proposals, EU, song, issue, chemistry).* As with *storm in a cup,* the phrase seems to function as an assertion, and thus to licence modality *(tends to, may, suspect, think)* in use.

1. growth tends to get short shrift, but some deserve short shrift
2. proposals may get shorter shift from schools who top league tables
3. the European Union gets even shorter shrift.
4. to give someone short shrift is to give very little attention to them
5. *I gave the issue short shrift by trying to short hand it.*
6. We think it was paid short shrift.
7. Inorganic chemistry receives short shrift

Figure 4: search results for the pattern *[shrift]*

2.1.2 Fronting

Another feature of apparently more lexically fixed expressions in web text is internal syntactic variation. Figure 5 illustrates this with reference to the previously mentioned pattern *[shrift]*, in cases where the verb, whether either *give* or *get*, is ellipted, and the core lexis, the word *shrift*, fronted.

1. Ishmael's shrift would have been short
2. their shrift was an equally short one when caught.
3. after being driven out, his shrift was frequently a short one.

Figure 5: search results for the pattern *[shrift]*, showing fronting

2.1.3 Semantic and colligational constraints of ordering The next example concerns the idiomatic phrase pattern *[make somebody look like]*, as in *'John makes Paul look like an amateur'*. The search pattern which was applied was *[makes "look* like], and the results are illustrated and classified in Figure 6.

Text: NP (+very good) makes NP (+good) look like NP (+bad)

1. with this stunning CD, bunch of tuneless novices
2. XMLTV makes VideoPlus look like hard work

text: NP (+ very good) *makes* NP (+ bad) *look like* NP(+good)

1. Pastry shortcut makes beginners look like pros

Figure 6: search results for the pattern *[makes "look like]*

The results in Figure 6 reveal that there are clear semantic relations required of, and more importantly imposed on, the NP 1 and the NP2 in the pattern. Specifically, semantic value of relative importance or virtue is assigned to them by the phrasal framework, used to promote the impressiveness of NPl.

2.1.4 Inflection specific constraints

Figures 7 and 8 illustrate the idiomatic phrase [X *NOT KNOW Y from* Z], and reveal a fact that all corpus linguists are aware of: namely that each inflection of a lemma has its own existence and behaviour in language use. In Figure 7, we see that the pattern containing the present tense verbal inflection *doesn't* generates NPs referring to relatively more or less famous people and things.

1. *she doesn't know Tim Couch from Bill Clinton*
2. *she doesn't know Santa Cia us from Donald Rumsfeld*

Figure 7: search results for *[*DOESN'T know * from]*

In contrast, Figure 8 shows that the variant pattern with the past tense inflection *didn't* is not associated with fame, but is heavily associated with the use of adverbials which are sentence-initial and which contextualise the main clause as being in contrast with a previous state of ignorance alluded to metaphorically but not specified.

1. *Eight months ago, I didn't know one note from another*
2. *When I put that hat on, I didn't know a rose from a juniper*
3. *I hadn't read /. R. Tolkein's works, so I didn't know an elf from an ent* or *a hobbit from a troll*

Figure 8: search results for [* *DIDN'T know * *from]*.

2.2. Grammatical boundaries and completions

The last section indicated some internal lexico-semantic features of so-called fixed expressions which have to be taken into account in automated search. This section deals with conventions at the phrase boundary which we are not necessarily conscious of, possibly since our intuitions are geared to speech, and which thus cause both users and systems problems to predict or specify.

2.2.1 Clause completion

The first convention concerns clause completion. It seems that the idiomatic retort in speech is typically couched in more complete syntax when occurring I nwritten text.

2.2.1.1 with S, *V and Adv (instrument)*

The phrase which typically takes the form *[better than a poke in the eye (with a rusty nail/with a sharp stick, etc)],* when submitted to web text search as the search pattern *[better than a * in the],* invokes a clausal completion which opens with an existential verb construction and closes with an adverb of instrument, as shown in Figure 9.

1. *It was better than a poke in the eye with a rusty nail*
2. *that, dear friends, is a whole lot better than a poke in the eye with a sharp stick*
3. *it's definitely better than a poke in the eye with a sharp stick.*
4. *This is certainly better than a poke in the eye with a sharp bit of metal*
5. *just about everything is better than a poke in the eye with a uranium-depleted shell casing*

Figure 9. search results for pattern *[better than a * in the]*

2.2.1.2 Cl completion with S, *V (NP O/A)*

Similarly, with the pattern *[a * like the * of a]*, which can be minimally expressed in speech as a comment or complaint in the form *'mouth like the bottom of a parrot's cage!'*, we find in the written examples from web text that the phrase is usually expressed within a complete syntactic unit, as shown in Figure 10.

1. *I woke up with a bad head and a mouth like the bottom of a zoo cage*
2. *I've got a mouth like the bottom of a budgie's cage*
3. *I woke up in the gutter with a mouth like the bottom of a cockie's cage*
4. *I've got a mouth like the bottom of a pterodactyl's cage*
5. *Beer, man. I've got a mouth like the bottom of a bat's cave*
6. *Hung over. Has a mouth like the bottom of a baby's pram*
7. *I crawled out of my tent with a mouth like the inside of a lime-burner's clog*
8. *He has a mouth like the inside of a Turkish wrestler's jock strap*

Figure 10: search results for pattern *[a * like the * of a]*

It is not surprising that half the examples in Figure 10 are syntactically complete, since they occur in narrative discourse, which typifies journalism, but the other half involve the present tense, and three the first person, as though from conversation of the chat room variety, and might have been more elliptical. This mixture of written and semi-spoken chat-room and other text is a characteristic of the web which adds a layer of complexity or unpredictabiIity to automated search.

2.3. Discourse Boundaries

This section deals with requirements beyond the boundary of the phrase itself. These are not lexical or grammatical, but discoursal, completions. Though predictable and describable at discourse level, they are problematic to identify automatically, since the lexis is not controlled. They tend to occur within a clausal or coordinated clausal structure, however, as well as being marked by certain characteristic discoursal and semantic signals.

Draft

2.3.1 Discourse Prefacers

Typically, the discoursal completions take the form of prefacers. Two instances were actually apparent in Figure 10 above, in lines 5 and 6, where the phrasal idiom is prefaced by an explanation for the condition described - *'Beer, man:* and *'Hung over:* If we take the phrase *its better than a kick in the pants/ a poke in the eye with a rusty nail* and submit it as the search phrase *[better than a * in the]*, we find that it retrieves instances of the pattern which function as a clause of concession stating that things could be worse, and which are coordinated contrastively by *but* with a prefacing clause which specifies the ideal situation which has not been achieved, or which ought really to have been achieved.

1. *Given our desperate state at the beginning,* I *think most of my companions would agree that it was better than a kick in the pants.*
2. *By no means will it guarantee you anything, but its better than a poke in the eye.*
3. *Well, I'm about all guessed out and none of the above strike me as particularly satisfying, but as they say it might be better than a kick in the whatever*

Figure 11: search results for pattern *[better than a * in the]*

The phrase typified as *he doesn't know his arse from his* elbow can be rendered in the simplified search pattern [* *doesn't know * * from* *]. When this is applied to web text search it produces another illustration of clausal completion by prefacing. In particular, it reveals a clausal prefacer which indicates the attitude of the writer towards the state of ignorance, including his/her own (6,8) expressed by the phrase. This attitude may be that the state is lamentable, probable, general, salient, significant, obvious and so on, as illustrated in Figure 12.

1. *Who, as far as we can tell, doesn't know his ass from his elbow*
2. *If you ask me, this guy doesn't know sh*t from shinola!*
3. *The trouble is, he doesn't know a flush from a fold*
4. *The point is, Dean doesn't know his ass from his elbow*
5. *Generally, a new student doesn't know his brass from his oboe*
6. *Heck,* I *didn't know a molecule from an element*
7. *The point was, these guys didn't know Peter Leo from Galileo (no relation).*
8. *Of course,* I *was a complete novice.* I *didn't know one clay from another.*

Figure 12: search results for pattern *[doesn't know * *from]*

Figure 13 demonstrates that the same pattern *[doesn't know * * from]* can alternatively retrieve instances prefaced by a clause of negative evaluation which, since it refers to a human entity, is attached to the phrase itself by the relative pronoun *who.*

1. *You're just an ignorant hick who doesn't know his chad from a hole in the ground*
2. *they see you as a* 15 *year old who doesn't know his ass from a hole in the ground*
3. *You're a stupid high-school kid who doesn't know his ass from a hole in the ground*
4. *I'm a total idiotic standards-wonk bonehead who doesn't know my A from my elbow*
5. *She's some wimp of a woman who doesn't know her head from a hole in the ground*
6. *You're one of those tunnel minded individuals who doesn't know your 'ars' from a hole in the ground*

Figure 13: search results for pattern *[doesn't know * *from]*

The phrase generally expressed as *better than a poke in the eye* can be retrieved in written web text by the pattern *[better than a * in the]*. The resultant output can be seen in Figure 14. This reveals that whereas in speech the phrase is a brief and humorous retort, in writing it is articulated as a complete clause, and typically prefaced by a clause with the discourse function of marker of introduction, signalling metalinguistically that the phrase has authority but is perhaps not linguistically original.

1. *My motto is A Fly in the Sky is better than a Walk in the Stalks*
2. I *guess I'm thinking the old* "1 *in the hand is better than a bunch in the bush"*
3. *As my Dad used to say, "Its better than a kick in the head."*
4. *In case you were in doubt, Shinola is better than a hole in the head*

Figure 14: search results for pattern *[better than a * in the]*

The phrase *not for all the tea in China* can be retrieved by the search pattern *[not for all the * in]*. The results, as shown in Figure 15, reveal that in written text this phrase is commonly prefaced by a clause which states that something will not be done, and that the phrase itself serves adverbially to emphasise the fact.

1. I *could never make you stay, not for all the tea in China*
2. *he would never pick cotton again, not for all the money in the world*
3. *He didn't have a chance with Neil, not for all the Ale in the world*
4. *We wouldn't cut a boy off: no, not for all the plate in the country, sir*
5. *A dog can not distinguish shades of color like we can, not for all the weenies inthe world*
6. I *will not help you, not for all the stars in your body*
7. I *would never go back to the way things were, not for all the ambrosia in Olympia*

Figure 15: search results for pattern *[not for all the * in]*

Conclusion

The investigation revealed that phrasal variability has clear rules, lexical, grammatical and discoursal. These are conventions operating within the phrase, as well as at its boundaries and beyond. The specific words in a phrase, their meanings and sounds and appearances, can be played with, by substitution, in creating phrasal variants. The data show that a phrase conforms to conventions in patterning beyond its specifically worded boundary, adopting various discourse strategies for merging into surrounding text. These conventions of variability and creativity are known and intuitively conformed to by the native (or proficient non-native) speaker. Unfortunately, neither these rules nor a reliable definition of what constitutes a complete phrase are so easily deducible by automated systems. But at least the access to large numbers of these phrasal phenomena which is facilitated by the *WebCorp* tool and others allows a detailed inspection of the phenomenon, and thus presents an important step in the study of phrasal variability in real text.

References

CRUSE, D. A. (1986): *Lexical semantics.* Cambridge University Press, Cambridge. . KEHOE, A. & RENOUF, A. (2002): *WebCorp: Applying the Web to Linguistics and Linguistics to the Web.* World Wide Web 2002 Conference, Honolulu, Hawaii, 7-11 May 2002.

· RENO UF, A. (2001) : 'Lexical signals of word relations'. In, Scott, M. and Thompson, G. (eds.) *Patterns of text: in honour of Michael Hoey,* John Benjamin Publishing Co, Amsterdam/ Philadelphia, p. 35-54.

. RENOUF, A. (2002): *'WebCorp:* providing a renewable data source for corpus linguists: in Granger, S. & Petch- Tyson S. (eds.). *Extending the scope of corpus-based research: new applications, new challenges.* Amsterdam & Atlanta: Rodopi. 39-58.

· RENOUF, A., KEHOE, A and BANERJEE, J. (2005) : 'The WebCorp Search Engine: A holistic approach to web text search: *in Electronic Proceedings of CL200S,* University of Birmingham.

· RENOUF, A., KEHOE, A. & BANERJEE, J., (forthcoming): *'WebCorp:* an integrated system for web text search' in Nesselhauf, C (ed.), Corpus Linguistics and the Web. Amsterdam & Atlanta: Rodopi.

WebCorp: http://www.webcorp.org. uk/wcadvanced.html