Cognitive entrenchment, idiomaticity and corpora : from theoretical issues to practical examples

Jean-Pierre Colson

Université catholique de Louvain

Get caught red-handed

面红耳热 **Miàn hóng ěrrè** (face red ear hot: flush with shame)

1. Introduction

- There are many fascinating interactions between cognitive linguistics and figurative meaning, for instance (in the case of phraseology) by studying metaphors across cultures and languages
- This contribution, however, will try to investigate this topic from another point of view: linguistic corpora and statistical association
- No matter which theoretical underpinnings are used for studying cases such as ***get caught red-handed*** or面红耳热, there is no hiding the fact that **they behave just the same** on large comparable linguistic corpora:
- ***BE caught red-(?)handed*: 0.93 (association) 14 (freq)**
- 面红耳热：**0.92 (association) 22 (freq)**

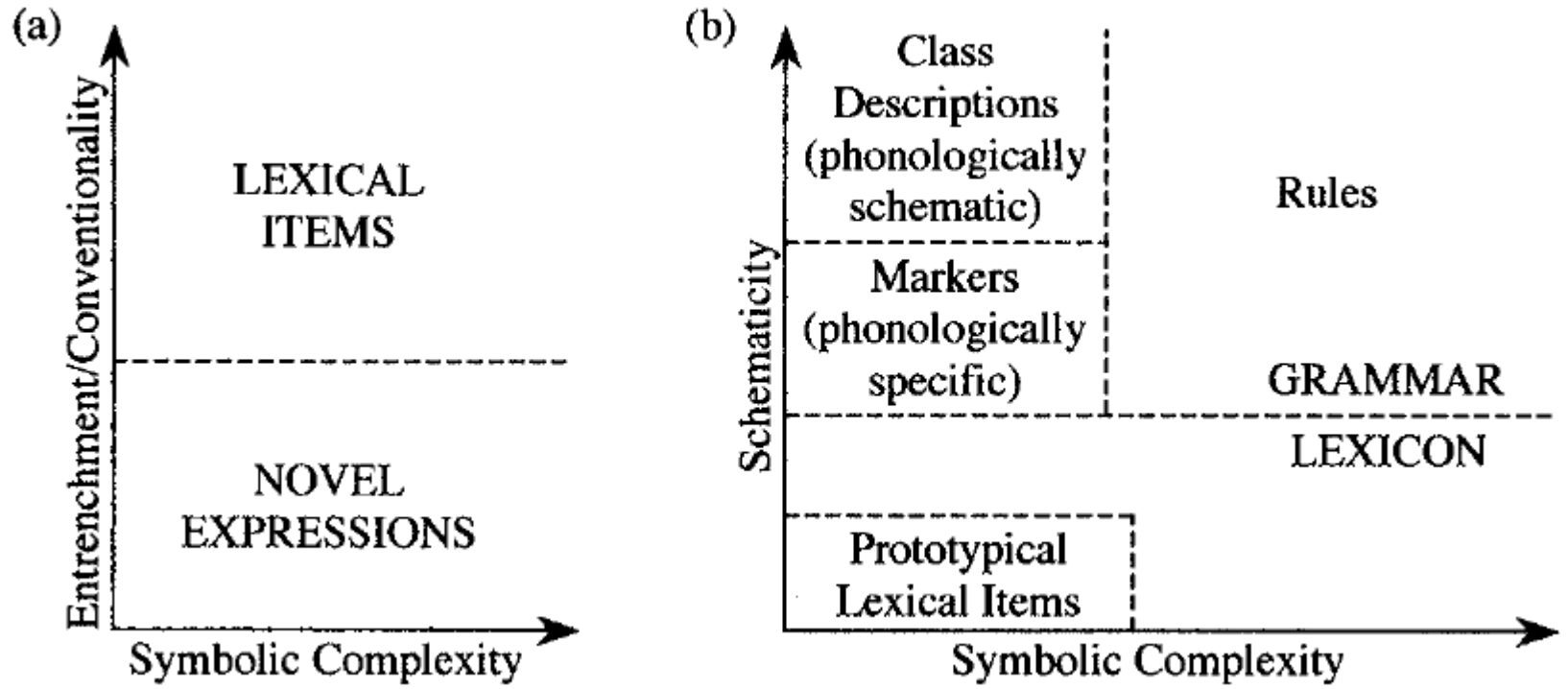
- From a theoretical point of view, however, this matter is particularly complex
- Let us start from a central concept of cognitive linguistic: *entrenchment*

1. ENTRENCHMENT: 它是什么? (What is it?)

- It should first be pointed out that the English notion of (cognitive) entrenchment is unknown to general dictionaries of English, even to the Oxford Dictionary (see attachment).
- It therefore comes as little surprise that it is particularly difficult to translate into other languages…
- In cognitive linguistics, entrenchment is related to one of the four general cognitive processes that (also) play a role in language: automatization (the other processes are: association, schematization and categorization, see Langacker 1987; 2008)

- In cognitive grammar, Langacker (2008:27) defines entrenchment as follows: *"Automatization is the process observed in learning to tie a shoe or recite the alphabet: through repetition or rehearsal, a complex structure is thoroughly mastered, to the point that using it is virtually automatic and requires little conscious monitoring. In CG parlance, a structure undergoes progressive **entrenchment** and eventually becomes established as a unit. Lexical items are expressions that have achieved the status of units for representative members of a speech community".*
- Langacker (2008:32) further distinguishes between entrenchment and conventionality*: "For ease of discussion, I am conflating two parameters that eventually have to be distinguished: entrenchment or unit status (pertaining to a particular speaker) and conventionality (pertaining to a speech community).*

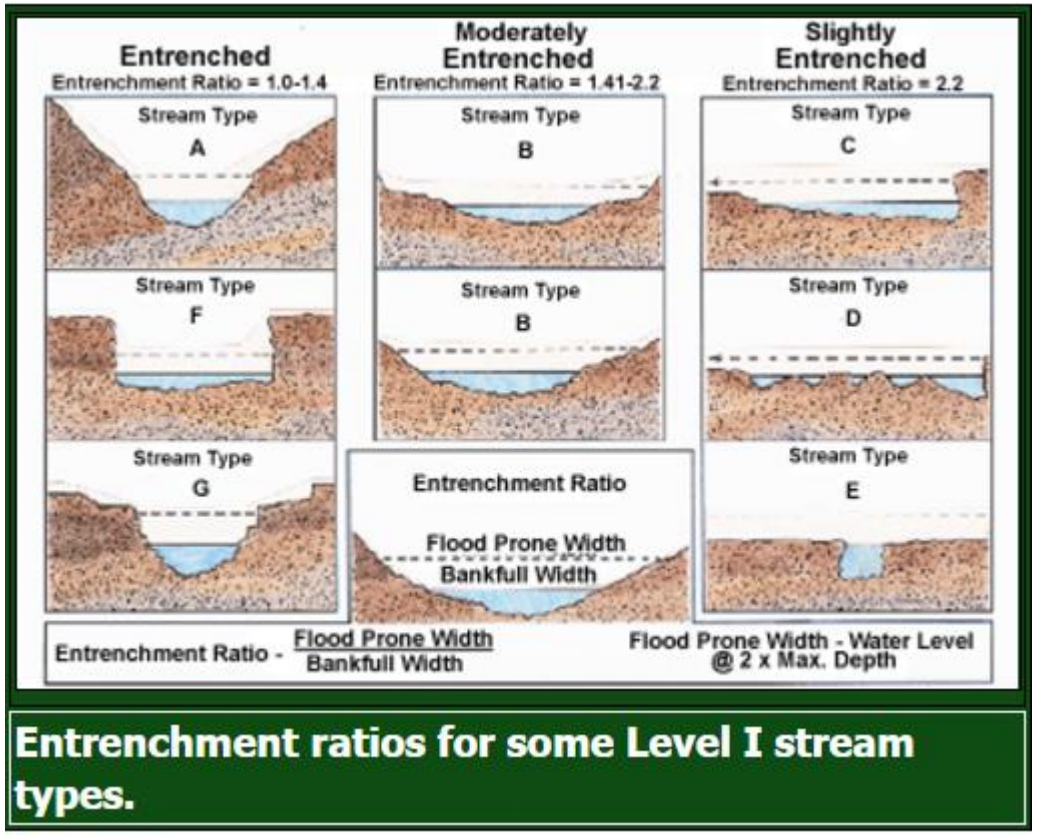
- For him (2008:32), all grammar consists of symbolic assemblies that can be situated along 3 main parameters (symbolic complexity, schematicity /



- A further elaboration of cognitive grammar was construction grammar (CxG)
- Constructionist approaches are varied (BCG, CG, CCG, RCG…) but they share the basic notion of CONSTRUCTION:
- *Constructions are Saussurean signs, i.e. "conventional, learned form-function pairings at varying levels of complexity and abstraction" (Goldberg 1995, 2006, 2013 p. 17): Word, partially filled word (pre-N, V-ing) (=morpheme for some researchers), idiom (=PU): filled, partially filled, minimally filled, abstract constructions such has Ditransitive, Passive...*

- *What did Mina buy Mel? (Goldberg 2013:28)*
- 11 constructions (Ditransitive construction, Nonsubject Question construction, Subject-Auxiliary Inversion construction, VP construction, NP construction, Indefinite Determiner construction, Lexical constructions: Mina, buy, Mel, what, do)
- *"It's constructions all the way down"* (Goldberg 2006:18)
- Continuum between lexicon and syntax: *constructicon* (Fillmore 1988, Goldberg 2003)

- In a sense, ALL CONSTRUCTIONS ARE IDIOMS:
- *"What may license referring to some constructions as idioms and not others is merely a reflection of the fact that effects of idiomatic variation are best observable in partially schematic complex constructions – however, this does not make them fundamentally different in nature from other constructions (Wulff 2013:285)*
- Constructions that contain slots are schematic (OH:2): a comparative construction has one substantive element (than), and the rest is schematic (many slots); an idiom such as "X take Y for granted" is more schematic than "X spill the beans".
- However, entrenchment receives different slightly different or more specific definitions in CxG



Entrenchment ratios for some Level I stream types.

- For Goldberg (2013:247), referring also to Croft and Cruse (2004) and Bybee (2013), token frequency determines the degree of entrenchment of "individual substantive word forms"
- But for others (Booij 2013, Baayen 2003, Hay and Baayen 2005, Barðdal 2008; 2011), it is type frequency that correlates with the degree of entrenchment
- For Wulff (2013:279), *schematic idioms* (i.e. idioms or phrasemes with at least one schematic slot: *break DET ground, take DET course, cross DET mind…*) are of particular interest, because they show a 'multi-dimensional continuum' of formally and semantically irregular and cognitively entrenched expressions (note: if entrenched, what about frequency?)

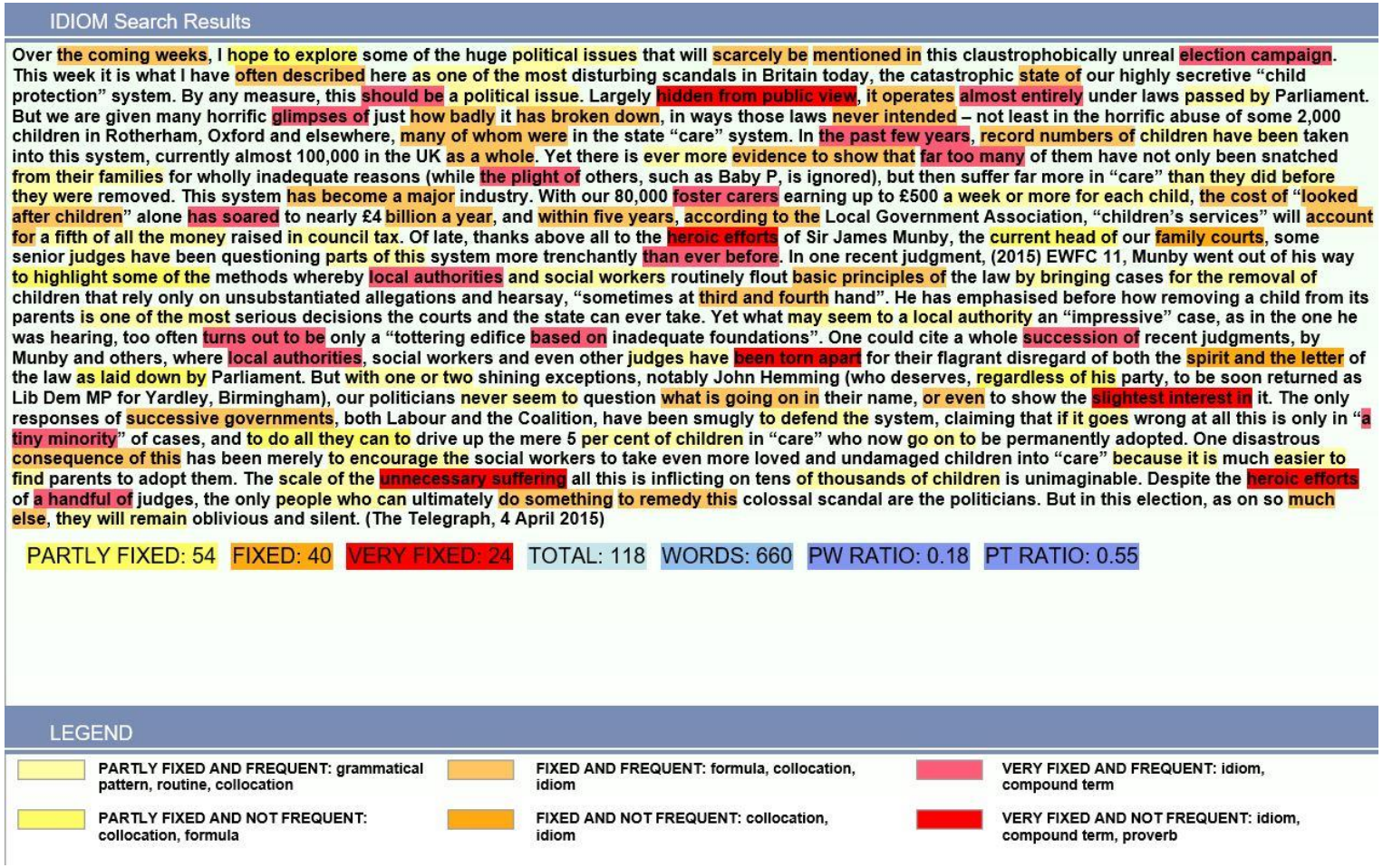
- For several researchers in CxG, the constructicon is a probabilistic network of nodes and inheritance links:
- ***Croft (2013:213): "constructions are a pairing of syntactic structures with semantic structures"… "This mapping is complex, partly arbitrary, and probabilistic (as a crosslinguistic generalization).***
- ***Stefanovitsch (2013:306) "(…) the combinatory capacity of concrete lexical constructions (words) and more abstract grammatical constructions (phrasal and clausal structural schemas) is not (or not exclusively) a categorical issue, but one that must be framed in probabilistic terms"***

- This probabilistic and statistical aspect of CxG has only been investigated by collostructional analysis (Gries 2007, 2013; Gries & Stefanowitsch 2003, 2004a, 2004b, 2010; Stefanowitsch 2006, 2013), a methodology that makes it possible to quantify association strength in constructions, and is derived from collocational approaches used in corpus linguistics
- A few findings:
- Stefanowitsch (2013:292): ***"(…) is shows that statistical association between verbs and Argument Structure constructions (and words and constructions in general) exist and that verbs differ quite drastically with respect to the degree of this association".*** –
- Stefanowitsch (2013:306): ***"the combinatory capacity of concrete lexical constructions (words) and more abstract grammatical constructions (phrasal and clausal structural schemas) is not (or not exclusively) a categorical issue, but one that must be framed in probabilistic terms"***

- *“A separate issue is whether p-values from inferential statistical tests can be used directly as measures of association strength. In collostructional analysis, it is assumed that this is indeed possible, since p-values assess the joint influence of corpus size and distribution and are therefore a natural measure of cognitive entrenchment”.* (Stefanowitsch 2013:304)

2. The *IdiomSearch Experiment* (JP Colson, 2016)

http://idiomsearch.LSTI.ucl.ac.be



- Corpus-based computational phraseology has recourse to **huge corpora**, as it has been demonstrated that, while set phrases as a whole represent a high percentage of word combinations in any text (Sinclair 1991), the frequency of a given set phrase in a corpus is usually low (Colson 2007, Moon 1998).
- Computational phraseology largely relies on the statistics of word co-occurrence (Evert 2004), but also explores other techniques derived from **information retrieval**, such as clustering (Baeza-Yates & Ribeiro-Neto 1999).

$$cpr = \frac{n\,(x_{i_1} x_{i_2} x_{i_3} \ldots x_{i_n})}{n(\max\limits_{x_{i_1}, x_{i_2}, x_{i_3}, \ldots x_{i_n}} (||x_i - x_j||) \leq W}$$

The Corpus Proximity Ratio (CPR), J.-P. Colson 2016

- For the extraction of phraseologisms in the broad sense, CPR already yields promising results, showing the discrepancy between fixedness and frequency
- The algorithm should still be improved, but it can already extract an impressive array of compound terms, collocations, idioms, formulas, clichés

- Examples: *a bit, air, back to, bank, barrel, bourgeois, bring, jet, perfect*
- An (originally unwanted) side effect of this research on extraction of phraseology was precisely the sociolinguistic picture implied by many phraseologisms: *art, beer, bridge*

3. Is the *cpr* score a measure of entrenchment / conventionality? What is the interaction with figurative language?

- If language is *"constructions all the way down"* (Goldberg 2006:18), measures of association strength should be able to work at all levels of complex constructions in much the same way
- In *construction morphology* (Booij 2010, 2013), ***"A constructional idiom is a (syntactic or morphological) schema in which at least one position is lexically fixed, and at least one position is variable" (Booij 2013:258)***
- Examples: *whatever, unlikely, unthinkable, unquestionably, come on leaps and bounds, spill the beans*
- Results (*cpr-score* and freq on 200 MW web corpus):

	CPR	Freq
come on leaps and bounds	0.42	10
spill the beans	1.00	15
un – question – ably	0.98	206
un – think – able	0.96	276
un – like – ly	0.74	4529
what – ever	0.99	7454
un – this – ly	0.00	0

Nf200mw	cpr	freq
dans sa totalité	0,85	344
en tout et pour tout	0,4	159
totale-ment	1	14759
entière-ment	1	9251
complète-ment	1	12720

- The results indicate that there is indeed a continuum between those various constructions
- They are COMPATIBLE with construction grammar, but also with formulaic language (Wray 2002, 2008), and with the hypothesis of ***a third articulation of language*** (Mejri 2006) consisting of "words":
  - Simple morphemes: *dry, house, in*
  - Sequences of free and bound morphemes: *unquestionably, incapable, aforementioned*
  - Sequences of two or more free morphemes (phraseology): *civil servant, easy rider, run of the mill, walks of live, take it easy, spill the beans*

- Preliminary results also suggest an important point about the relationship between the *cpr-score* and **entrenchment**: **in the case of Chinese, the score makes it possible to perform acceptable segmentation**
- **Example:**
- 中國在由計劃經濟向市場經濟轉軌過程中
- **(Zhōngguó zài yóu jìhuà jīngjì xiàng shìchǎng jīngjì zhuǎnguǐ guòchéng zhōng**
- *China in the transition from a planned economy to a market economy* (Google Translate)
- This confirms the **probabilistic network** of constructions posited by CxG, as well as the notion of "**statistical semantics**"

- If statistical scores are a good indication of entrenchment (Stefanowitsch 2013), we may wish to investigate the link between figurative meaning and entrenchment / statistical association
- Example: a selection of n-grams with high statistical scores and their frequency on a 200 million word corpus
- All those n-grams are recurrent, highly associated, and are therefore probably entrenched and conventionalized, but few of them display highly figurative language
- If, on the other hand, we start from very figurative idioms, we get very high scores as well

Idiom	frequency	cpr-score
add insult to injury	47	0.96
at the drop of a hat	40	1.00
back to the drawing board	67	0.97
barking up the wrong tree	7	0.88
beat about the bush	13	1.00
beat around the bush	8	1.00
keep up the good work	82	0.90

- These results indicate that figurative idioms / phrasemes / PUs are very entrenched, but this works only in one direction
- From a theoretical point of view, it is also a confirmation of Wulff (2013:279): they show a ‘multi-dimensional continuum’ of formally and semantically irregular and cognitively entrenched expressions

3. Conclusions

- These experiments confirm one of the tenets of *information retrieval: meaning can be accessed by means of statistics, probably because it is partly statistical in nature*
- Figurative meaning cannot (yet) be addressed directly by statistical methods, but it normally implies high cognitive entrenchment / statistical assocation scores
- The striking similarities in association strength across languages (see intro) suggest that similar cognitive principles underly entrenchment / fixedness