Linguistic reproducibility and cognitive entrenchment: corpus-based evidence and applications for language teaching and translation

Jean-Pierre Colson

Université catholique de Louvain

1. Introduction

Reproducibility, stability

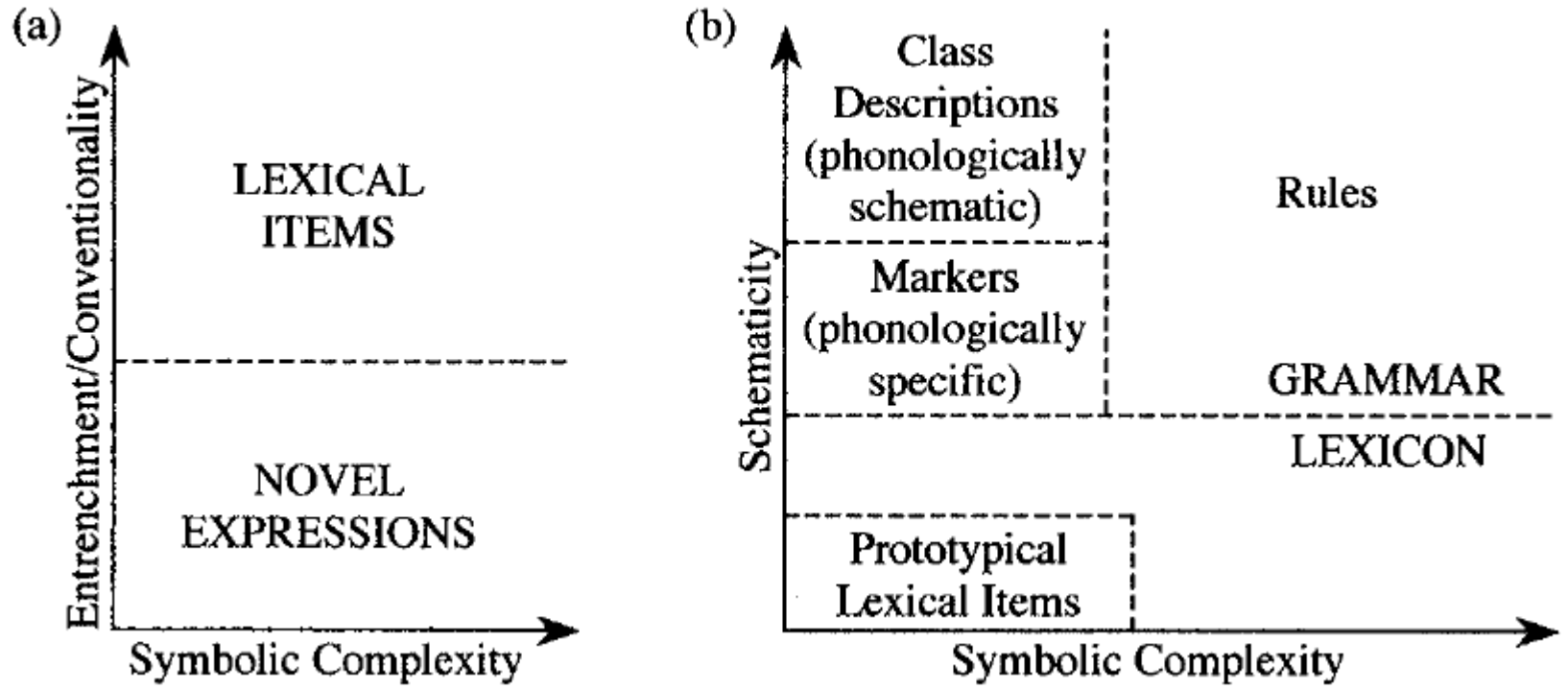
- The notions of ***reproducibility*** and ***stability*** have been used in the Russian phraseological tradition, at least since Vinogradov (1947): *"the very fact of stability and semantic limitation of phraseological units shows that in reality they are used as ready phraseological units, which are reproducible, not constructed anew, in the speech process"* (Vinogradov 1947: 160).
- Just like ***words***, phraseological units are functionally repeatable in different situations, and are retrieved from memory as a whole
- Reproducibility has also been studied from a ***cognitive*** point of view (Artomonova 2016): *"A reproducible unit is a unit tending to possess some invariant character, i.e. "a stable image, a stereotype…, a continual verbal symbol, able to unfold into a whole segment of the 'picture of the world', which is expressed by a word, a morpheme, a root, a phrase"* (Karaulov 1987: 181)

ENTRENCHMENT

- 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 / specificity and entrenchment / conventionalilty):



- 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

- 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"***

Reproducibility and entrenchment

- To sum up, ***reproducibility*** and ***entrenchment*** show many similarities:
  - STABLE units in the individual's cognitive system and in the language community (*conventionality*)
  - HOLISTIC units, retrieved as a whole from memory
  - VARIED units, such as a morpheme, a word, a lexical or syntactic construction can be characterized by reproducibility / entrenchment

- Reproducibility, just like entrenchment, might therefore be measured by probabilistic methodologies such as *collostructional analysis*

2. The *IdiomSearch Experiment* (JP Colson, 2016) http://idiomsearch.LSTI.ucl.ac.be

- I have introduced the ***cpr-score*** for the automatic extraction of MWE and formulas (Colson, 2017)
- http://idiomsearch.LSTI.ucl.ac.be

$$cpr = \frac{n(w_1, w_2, \dots, w_n)}{n\left(x_{t_1} = w_1, x_{t_2} = w_2, \dots, x_{t_n} = w_n \,\middle|\, \max(t_{i+1} - t_i) \leq W;\ i = 1, \dots, n-1\right)}$$

Figure 1. The *cpr-score*

- Exactly the same methodology (*cpr-score)* actually makes it possible to ***segment Chinese***
- Examples: the English phrase *have a look at* is identified in a 200 MW English corpus, as well as the Chinese word 个人计算机 (*gèrén jìsuànjī*, personal computer) in a 200 MW Chinese corpus; the association threshold determines segmentation, which is (also) a cultural phenomenon (e.g. *PhraseoSegmenter*)

- Statistical scores such as the *cpr-score* make it possible to extract recurrent statistical combinations in the broad sense, covering all formulas and fixed expressions, and even CWS (associations of morphemes)
- If CxG is right, the association can be traced back at various levels of abstraction and schematicity
- Let us check this point

3. Clues as to automatic extraction of constructions

- According to CxG, the **probabilistic network of constructions** is valid at various levels of abstraction and schematicity
- Part of that complex interplay between morpho-syntactic features can easily be captured by applying clustering methods to the tagged corpus. The same algorithm that we used for **CWS** (the cpr-score) can check the association between parts of constructions and specific tags, as shown in table 3.

	Cpr-score	**Frequency**	**Window (*w*)**
it is *w* ADJ *w* what	0.12	428	4
it is *w* amazing *w* what	0.52	11	4

Table 3: *cpr-score* and frequency of a schematic construction and a derived MWE.

- This MWE, a specific lexical (and partly idiomatic) construction actually inherits (in CxG parlance) from the **more schematic construction it is ADJ what**. As shown in table 3, we can measure a weaker association at this more schematic level as well. Other examples of schematic constructions that were extensively studied in the literature on CxG (Hoffmann and Trousdale, 2013) include the **Ditransitive construction** (e.g. give a book to someone) and the **All-cleft /Wh-cleft construction** (as in all he had to do was to arrive on time). As illustrated by table 4, our POS-tagged corpus also yields association scores for these constructions.

	Cpr-score	**Frequency**	**Window (*w*)**
NOUN VERB *w* NOUN *w* NOUN[10]	0.27	163979	5
all PRONOUN *w* VERBPast *w* VERBPast[11]	0.29	460	7

Table 4: *cpr-score* and frequency for the Ditransitive and All-cleft construction

- Up till now, the approaches to grammar and lexicon have been rather conservative and in contradiction with recent findings from the cognitive and constructionist approaches
- A global approach to constructions (in the decoding of the source text and in language learning) may imply some **awareness raising** of the complex network of constructions that is at stake in language
- As a tentative step, manipulating schematic, partly schematic and fixed constructions in corpora may be useful to learners and curriculum designers

- **Examples:**
- **take RP IN: 0.06 589**

  look RP IN: 0.31 1402

  come RB IN: 0.09 1433

  is JJ TO VB: 0.05 14721

  do DT NN: 0.08 3503

  make DT NN: 0.28 9890

  have DT NN IN NN: 0.05 2403

  play DT NN: 0.12 1543

  the NN TO VB: 0.05 38506

Conclusions

- Analyzing reproducibility / entrenchment from the perspective of computational phraseology / construction grammar reveals a complex network of **probabilistic constructions**, with some **patterns** being more phraseological than others
- For translators and language learners, it is important to access the relative entrenchment of constructions, and to gain some insight into the network
- The fact that these **phraseological patterns** are not simply frequency-based, but interact with a **probabilistic network**, is evidence for **construction grammar** (especially CCG based on entrenchment), in which the theoretical implications of phraseology must still be further explored (syntax as coded and idiomatic combinations)