Jean-Pierre Colson
Université catholique de Louvain
Belgium
https://orcid.org/0000–0003–1267–3491

Phraseology and Cognitive Entrenchment: Corpus-based Evidence and Applications for Language Teaching and Translation

Abstract. Cognitive entrenchment, originating from cognitive grammar, actually comes very close to other theoretical notions such as reproducibility, fixedness or idiomaticity. By the means of experiments carried out on huge linguistic corpora, computational phraseology makes it possible to find partial evidence for the theoretical notions, and to offer at the same time practical tools to language users in general. This paper provides evidence for the probabilistic nature of the network of constructions. Indeed, the same statistical score, the *cpr-score*, developed in the first place for the extraction of phraseology, turns out to yield significant results for other types of constructions: lexical ones (in the case of Chinese word segmentation), cultural references and named entities, and even more schematic or abstract patterns underlying syntactic constructions.

Key words: *phraseology, entrenchment, construction grammar, corpora, translation*

1. Introduction

Perhaps one of the most striking features of phraseology is that researchers coming from a wide range of theoretical backgrounds have reached a similar conclusion: much of what we say or write consists of (at least) partly idiomatic constructions. Spontaneously, native speakers will put together elements of meaning which, according to their mastery of the linguistic system, are used together as a set of ready-made structures.

An overview of the multiple paths and tracks of phraseological research leading to similar conclusions falls beyond the scope of the present contribution. We would just like to take the example of a key issue underlying idiomatic constructions of any type, namely the nature of the attraction between the elements of a phraseme (or phraseological unit).

In the Russian phraseological tradition, the notions of *reproducibility* and *stability* have been used in that respect, at least since Vinogradov (1947: 160): "the very fact of stability and semantic limitation of PUs (PUs) shows that in reality they are used as ready PUs, which are reproducible, not constructed anew, in the speech process".

Just like words, PUs are seen as functionally repeatable in different situations, and are retrieved from memory as a whole. In the Russian tradition, reproducibility has also been studied from a cognitive point of view:

> 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).

These statements bear a striking resemblance to the notion of *entrenchment* as it has been used within the theoretical framework of cognitive linguistics and construction grammar. 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). Much in the same way as an activity tends to become a habit, a linguistic structure may undergo progressive entrenchment and eventually become established as a unit. This is valid at the lexical level (for the traditional notion of words) but also at higher levels of complexity. 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)". Crucially, all grammar consists of symbolic assemblies that can be situated along three main parameters: symbolic complexity, schematicity/specificity and entrenchment/conventionality (Langacker 2008: 32).

The first parameter, symbolic complexity, may be roughly explained by the length of the structure (containing more or fewer symbolic elements; for instance *merry* is less complex than *merry-go-round*). Schematicity, as opposed to specificity, refers to the possibility of using other elements paradigmatically at a given position or *slot*, or of modifying the existing element by inflection. Thus, *long time no see* is fully specific, while *take X into account* contains two specific slots (*into* and *account*), one schematic slot (*X*, the direct object) and one partly schematic slot (*take*, as the verb may be conjugated). *Entrenchment/conventionality* refer, as mentioned above, to the unit status of the assembly (or construction), as in the case of *book* but also of *long time no see*.

A further elaboration of cognitive grammar was provided by construction grammar (CxG), in which a number of different versions may be differentiated (e.g. Berkeley Construction Grammar, Cognitive Construction Grammar, Cognitive Grammar, Radical Construction Grammar, Sign-Based Construction Grammar, Fluid Construction Grammar; for an overview, see Hoffmann & Trousdale 2013). These approaches are varied, but they share the basic notion of constructions, defined as follows. They are Saussurean signs, i.e. "conventional, learned form-function pairings at varying levels of complexity and abstraction" (Goldberg 2013: 17). A construction may therefore be a word, a partially filled word (*pre-N*, *V-ing*) or morpheme, an idiom (in the general sense of a phraseological unit), but also a more abstract structure such as the transitive or passive construction.

A crucial point with respect to phraseology is that, as pointed out by Wulff (2013), all constructions are, in a sense, 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).

In other words, the idiosyncrasies associated with almost any construction make them in a sense (at least partly) idiomatic. Think, for instance, of the various ways of asking what the time is, even in European languages: *What time is it?* may sound like a purely grammatical construction, but the point is that this specific pairing of form and meaning (the very definition of a construction) is purely conventional in English, and a look at German and Dutch (resp. *Wie spat ist es? / Hoe laat is het?*, literally 'How late is it?') suffices to see that other languages use other conventional pairings of form and meaning for this everyday phrase.

It should also be pointed out that *entrenchment* has received slightly different definitions in CxG. For Goldberg (2013: 247), token frequency determines the degree of entrenchment of "individual substantive word forms". In other words, entrenchment can simply be measured by the number of occurrences of the tokens in a corpus. But for others (e.g. Booij 2013), 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.

To sum up, reproducibility and entrenchment show many similarities, as the notions are applied to:

- 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.

It is also clear that both notions come very close to *fixedness*, which has been widely used in the phraseological tradition (Burger et al. 2007), because fixed words in a phraseological unit are stable and are supposed to constitute one unit. However, most versions of CxG go one step further, because they also view words and even morphemes as stable and holistic constructions.

Entrenchment remains largely a theoretical hypothesis, which is very hard to prove from a purely scientific point of view. However, *collostructional analysis* has already provided some clues in that direction.

This probabilistic and statistical methodology (for an overview, see Gries 2013; Stefanowitsch 2013), makes it possible to quantify association strength in constructions, and is derived from collocational approaches used in corpus linguistics. The results tend to show that there is some statistical association between verbs and Argument Structure constructions (and words and constructions in general) and that verbs display very different pictures of association. Even the combination of lexical constructions and more abstract grammatical constructions may be of a probabilistic nature (Stefanowitsch 2013).

2. Statistical experiments around entrenchment

2.1. Extraction of phraseology

I have proposed the *cpr-score* for measuring the association strength between words in a phraseological unit (Colson 2017; 2018). As indicated in figure 1, for any ngram of length 2 to n, it basically measures the average distance between the component grams in a huge corpus. The exact number of occurrences of the grams (without a window) is divided by the number of occurrences within a given window *W*, that is experimentally set according to the average word length in a language (for English, it is typically set at 20 words). Thus, the score ranges from 0 to 1, with a significance threshold that can experimentally be set at 0.065.

Figure 1. The *cpr-score*

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

This metric seems to be complex when expressed in mathematical terms as in figure 1, but it actually tries to simulate, by using very large corpora, the general human principle that elements displaying strong semantic links will tend to occur very close to each other. This simple idea was already expressed by the famous British linguist John R. Firth, who stated that "You shall know a word by the company it keeps" (Firth 1957: 11).

The *cpr-score* has been implemented in a freely accessible web application, *IdiomSearch* (http://idiomsearch.lsti.ucl.ac.be), allowing the user to enter a source text and to receive an approximation of the most common PUs in the text (including formulaic language). In much the same way as collostructional analysis, the *cpr-score* can be seen as a measure of the degree of association prevailing between words within PUs, i.e. the degree of entrenchment/conventionality of those constructions.

The crucial point is that, if the predictions of construction grammar are correct, the scores yielding significant results for one type of constructions (in this case PUs) should also work for other constructions, including partly schematic, schematic and even abstract constructions.

2.2. Chinese word segmentation

In Colson (2018), the *cpr-score* has been tested against Chinese word segmentation. It should be reminded that (Mandarin) Chinese is an unsegmented language, which means that there is no blank space between *words* as we understand them in Western languages. For instance, a *personal computer* (two words in English) is written as one sequence of characters in Mandarin Chinese (simplified): 个人计算机 [*gèrénjìsuànjī*]. As a matter of fact, we should be very cautious not to be Eurocentric when having recourse to traditional linguistic notions. Words, for instance, are in themselves very controversial when applied to very different languages such as Chinese (Dixon & Aikhenvald 2002). In the traditional vision of their own language, Chinese native speakers often consider that any Chinese character or *han* is a word, which used to be the case in classical Chinese. In modern Mandarin Chinese, however, it is generally agreed that most words (or at least what corresponds to the Western notion of words) consist of two characters, and some of three or more. As there is in the language itself a fuzzy border between lexical constructions (words) and grammatical or phraseological ones, Chinese is a particularly interesting object of study for construction grammar and for phraseology, all the more so if we take into account the fact that it is the most spoken language in the world, and that it relies upon a very rich and ancient culture. Most Chinese words and phrases display complex cultural features. For instance, a *university teacher* is in Mandarin Chinese (simplified): 大学老师 [*dàxuélǎoshī*]. The literal meaning of those characters is: big – learn(ing) – old – master: an old master (i.e. a teacher) of the big learning (i.e. of higher education).

When applying corpus or computational linguistics to Chinese, the first basic task is *segmentation*: a sequence of Chinese characters must be separated into words, in order to be processed and understood by users or algorithms. How should this segmentation be carried out? There is no general agreement on this point.

The state-of-the art tools for segmenting Chinese are circular: they are based on existing lists such as those found in dictionaries, or on models derived from hand-annotated data. In many cases, however, the lists contradict each other, and so do native speakers. Large-scale experiments have shown that the average degree of agreement between native speakers is just 75 percent (Sproat et al. 1996; Ying Xu et al. 2010). In addition, a native speaker who is asked to segment the same text again after a few weeks, will often segment it in a different way. In the case of the *personal computer*, some Chinese segmentation systems or native speakers will consider 个人计算机 [*gèrénjìsuànjī*] as one word, while others will separate 个人 [gèrén] (personal) from 计算机 [*jìsuànjī*], computer.

In Colson (2018), the *cpr-score*, previously used only for the extraction of PUs, was applied to the segmentation of Chinese texts. The efficiency of the methodology was checked by the state-of-the-art methodology: the results are measured against a gold standard provided by native speakers, and they are automatically evaluated by a computer program. In this case, the gold standard and the evaluation program were the freely available datasets from the second International Chinese Word Segmentation Bakeoff (Emerson 2005). When a gold test is available, as in this case, the results of the automatic extraction are checked, as is the case for the extraction of phraseology, against precision and recall. Recall checks whether all the structures that had to be recognized were indeed identified, whereas precision checks if every identification is indeed a correct one. For instance, if there are 2 dogs and 2 cats in a room, and your algorithm checking the number of cats claims that there are 4 cats, the recall is 100 percent, because every cat has been recognized as such, but the precision is just 50 percent, because the 2 dogs were wrongly identified as cats. Finally, the F-measure (or F1-measure) computes an average between precision and recall.

Measured against the MSR-dataset of the Bakeoff (Emerson 2005), the segmentation of the Chinese texts on the basis of the *cpr-score* (Colson 2018) reached a recall of 0.749, a precision of 0.658 and an F-measure of 0.70. Of course, those figures are less good than those obtained by state-of-the-art Chinese segmenters, but it should be emphasized that these rely on existing lists or dictionaries, and are not corpus-based. On the contrary, our experiment with *cpr* was purely corpus-driven: a web corpus of about 200 million words was assembled for the purpose of the experiment, and the algorithm just relied on that corpus for recognizing words. To our best knowledge, those precision and recall results for the automatic segmentation of Chinese are the best ones that were ever obtained by means of a purely unsupervised and corpus-driven method. Besides, a recall of 0.749 and an F-measure of 0.70 come pretty close to the average degree of mutual agreement for segmentation reached by Chinese native speakers (0.75).

What can we learn from this about entrenchment, constructions and phraseology? It will be recalled that exactly the same methodology (extraction from a corpus by means of the *cpr-score*) was applied to the detection of PUs (Colson 2017) and to the segmentation of Chinese (Colson 2018). Applying the same metric yields quite comparable results: most PUs can be extracted from a text, and most Chinese words as well. This confirms the very fuzzy border between phraseological and lexical constructions. In European languages, we often take it for granted that words are combined with each other by means of grammatical rules, but very different languages such as Chinese illustrate how our Eurocentric view should relativized. Thus, even common Chinese words such as *boy*, 男孩 [*nánhái*] or *woman*, 女人 [*nǚrén*] might equally be considered as collocations, as they resp. mean 'male child' and 'female people'.

Indeed, the statistical method shows that, in many respects, Chinese words behave just like PUs, which they are at the end of the day, if we take the constructionist view that the very associations of morphemes into words are entrenched and idiomatic. In construction morphology (Booij 2013), a constructional idiom is defined as "a (syntactic or morphological) schema in which at least one position is lexically fixed, and at least one position is variable" (Booij 2013: 258).

2.3. Extraction of cultural PUs

As the whole set of constructions of a language or *constructicon* is seen by most researchers in CxG as a complex and probabilistic network interacting with all aspects of the language community, many constructions are also entrenched and idiomatic because of specific references to culture (in particular history). Extracting very entrenched constructions on the basis of idiomaticity (as in the IdiomSearch experiment) or of lexical associations (as for the segmentation of Chinese) should therefore also work for compound terms displaying a reference to tradition, history, culture or society in general.

In Colson (2016), the same methodology was used for the extraction of PUs around globalization in 6 languages. The study revealed the emergence of candidate PUs around globalization, a major notion in our society, as in *unfettered globalization* or *in the era of globalization*.

In addition to such recent PUs or compound terms referring to society, a whole host of cultural, and in particular historical or geographical references can be extracted with the *cpr-score* by having recourse to large linguistic corpora (of at least 200 million tokens). This includes most compound named entities (proper nouns) denoting famous people or cities, but also historical notions such as *the partition of Poland*.

To illustrate this point, table 1 below displays the *cpr-score* and the frequency (number of occurrences) of a number of PUs, including communicative formulas, collocations, idioms, but also named entities, and cultural PUs. All those results were extracted from the same corpus: a web corpus of 1.4 billion tokens (the freely available ukWaC corpus, Baroni et al. 2009).

The figures displayed under Table 1 illustrate how various types of phraseological units in the broad sense display significant statistical scores in the same corpus, despite their number of occurrences. While *long time no see, run of the mill, it takes two to tango, the chickens have come home to roost* clearly belong to phraseology, the first elements in the table are cultural PUs. The *partition of Poland* refers to history, *New Mexico* is an American state, and it is also called *Land of Enchantment* (on American number plates). *The Black Country* is the region around Birmingham (UK) and part of the *West Midlands*. The *Industrial Revolution* and *Sturm und Drang* are two periods in European history.

Considering all those examples from a cognitive point of view, it is clear that they can all be seen as very entrenched, specific, complex constructions, because their association score is very high. Such evidence gained from corpora confirm that a very complex network of constructions, including cultural and social ones, is at stake in language.

Table 1. Association and frequency of varied PUs in a 1.4 billion word corpus (ukWaC)

	Cpr-score	Frequency
partition of Poland	0.73	22
New Mexico	0.70	1796
Land of Enchantment	0.67	18
the Black Country	0.49	1099
the West Midlands	0.60	1071
the Industrial Revolution	0.83	2769
Sturm und Drang	1.00	53
long time no see	0.64	98
run of the mill	0.92	1005
it takes two to tango	0.92	107
the chickens have come home to roost	0.73	8

Source: own research.

2.4. Extraction of schematic constructions

According to CxG, the probabilistic network of constructions is valid, as we have seen, at various levels of abstraction and schematicity. If we wish to find evidence for this claim in large linguistic corpora, we should therefore check whether association scores such as those found for Chinese word segmentation and for other categories of specific constructions (Table 1) also hold for more schematic or abstract constructions.

Let us start from the example of the very common idiomatic construction *as white as snow*. Obviously, this stereotyped comparison is very entrenched in the linguistic competence of any native speaker of English. He/she will certainly also be aware of other similar cases like *as clear as crystal, as good as gold, as stupid as a donkey*, etc.

If linguistic corpora are a reflection of the native speaker's mastery of the complex network of constructions, we should be able to find a trace of these associations by means of our statistical score. The missing link, in this case, is just the use of *POS-tagged* corpora. Following the claim of CxG about the existence of abstract constructions, we will assume that a tag (such as *Noun*, *Adjective*, *Verb* etc.) will also be open, in a measurable way, to statistical associations that will reflect the complex construction network.

In the following examples, a randomly selected portion of 120 million words (tokens) from the ukWaC corpus (Baroni et al. 2009) was tagged by means of the Stanford POS Tagger[1]. Table 2 shows the association and frequency results for the idiomatic construction *as white as snow* and for the more abstract construction *as ADJ as NOUN*. The window (w) corresponds to the maximum number of words that is allowed between each token.

Table 2. Association and frequency of a PU and its abstract construction

	Cpr-score	Frequency	Window (*w*)
as white as snow	1.00	11	0
as ADJ as NOUN	0.53	429	2

Source: own research.

Thus, the association score for the abstract construction *as ADJ as NOUN* turns out to be already significant (0.53, with a significance threshold at 0.065). This means that anyone using a sufficiently large linguistic corpus could predict, by means of the algorithm, that this structure is very entrenched in English. Besides, *as white as snow* clearly inherits, in CxG parlance, from a more abstract construction, because it is a particular case of a pattern that belongs to the natural constructions of English.

3. Possible applications to language teaching and translation

As already advocated by Michael Lewis (1993, 1997), awareness raising of phraseology, by means of confrontation with corpora and varied linguistic data, offers new perspectives for learning foreign languages or for translating them.

As we have seen in section 1, construction grammar confirms many of the findings of phraseology, while giving it a solid theoretical grounding. The implications for language teaching and translation are numerous, because the very structure of language turns out to be very different from the vision given by more traditional approaches. In particular, the notion of grammar as a separate entity largely disappears, as there is a cline from lexicon to syntax. Although the experiments presented in section 2 are not, strictly speaking, evidence for construction grammar, they are quite compatible with it. There is presently no better theory of language that can explain

[1] We used version 3.9.1 of the Stanford POS tagger (https://nlp.stanford.edu/software/tagger.shtml).

the similarities in the behavior of very different constructions such as words, idioms, named entities, idiomatic constructions, etc.

If these findings are confirmed by other studies, it also means that we should start from a very different perspective for learning foreign languages and for translating them. The IdiomSearch experiment, briefly discussed in section 2.1., already offers several new possibilities to (advanced) language learners and translators, thanks to the mere detection of a great many PUs in any source text. It is generally admitted that advanced learners will learn a lot by reading in the foreign language, but they are often misled by sentences in which they fail to detect the figurative and idiomatic meaning.

Consider, for instance, the following excerpt from a British newspaper (The Guardian, 23 December 2018)[2]:

> It is notable that this latest iteration of fantasy Brexit is most often promulgated by ministers, such as Andrea Leadsom, who have no responsibility for delivering essential services. Even these Brexiters don't deny that a no-deal outcome would present a big challenge to government on multiple fronts. In the light of their recent performance, how confident are you that our masters of disaster could cope?

The IdiomSearch tool makes it possible to extract from this passage the following PUs and communicative formulas: *It is notable that / iteration of / is most often / promulgated by / a big challenge / In the light of / how confident are you that / masters of*. The communicative formula *How confident are you that* is an interesting example, because it is unlikely that even advanced learners reading this text will recognize it as a recurrent formula, unless their attention is focused on it by a teacher or by a tool.

Verbal constructions will also serve to illustrate the benefit that can be drawn from a manipulation of corpora by means of the *cpr-score*. If we take the traditional view that grammar is a major part of language structure, with for instance transitive constructions like *He takes the money*, we should expect a very similar behavior for most high frequency verbs, as in the basic pattern: a verb, followed by a determiner, followed by a noun (VERB, DET, NOUN). However, using the same methodology and the same corpus as in Table 2 yields the following results.

As can be seen in Table 3, *do* and *make* are very often followed by a direct object in the form of a determiner and a noun, as in *do the work*. However, a close look at the *cpr-score* reveals that the situation is quite different between

[2] https://www.theguardian.com/commentisfree/2018/dec/23/bluff-blackmail-brinkwomanship-why-a-no-deal-brexit-is-still-on-the-cards

Table 3. Association and frequency for a number of transitive verbal constructions

	Cpr-score	Frequency
do DET NOUN	0.08	3503
make DET NOUN	0.28	9890
play DET NOUN	0.12	1543
seize DET NOUN	0.67	335
take DET NOUN	0.26	8118

Source: own research.

these two verbs: in terms of CxG, this construction is much more entrenched with *make* (as the *cpr-score* is 0.28) than with *do* (*cpr-score*: 0.08). This also means that, taking the variety of examples of this construction into consideration, there is a much higher proportion of phraseology with *make* than with *do*. A brief look at the most frequent examples with *make* thus yields the following examples.

Table 4. Frequency of verbal constructions with *make* (120 MW web corpus)

Frequency	Verbal construction
779	make a difference
486	make any changes
296	make a decision
283	make a donation
268	make an appointment
192	make every effort
168	make a claim
162	make a note
133	make a complaint
111	make a contribution
101	make a profit
91	make a booking
90	make a difference
89	make a living
88	make any difference
86	make a start
86	make an impact
85	make a sudoku

Source: own research.

As shown by Table 4, many of the most frequent transitive constructions with *make* are at least partly idiomatic (e.g. *make a decision, make a claim, make a living, make any difference*), which explains why the overall association score for the abstract construction is so high (Table 3). The picture is different with *do* in the same construction:

Table 5. Frequency of verbal constructions with *do* (120 MW web corpus)

Frequency	Verbal construction
268	do the job
173	do a lot
163	do the work
98	do the rest
94	do the things
92	do the trick
70	do a bit
60	do the things
53	do the work
49	do some work
47	do this thing
46	do a job
46	do these things
44	do the following
40	do all things
39	do any harm
39	do the initials
39	do the rest
37	do some research
32	do this work
32	do the job

Source: own research.

Among the most frequent transitive constructions with *do*, we note an opposite tendency: there are many weakly or non-idiomatic examples, such as *do a lot, do the rest, do the things, do a bit, do this thing, do the rest, do this work.*

The kind of information provided by Table 3 (association scores for abstract verbal constructions), exemplified by a look at the relative frequencies of specific examples, provides a picture of grammar that is compatible with CxG and with phraseology. Not only are specific verbal constructions very entrenched (e.g. *make a claim*, *make a start*), but the underlying pattern, i.e. the abstract construction itself is more or less entrenched, depending on the verb. The point made here is just valid for one type of transitive construction, but it might be extended to other aspects of the cline ranging from grammar to lexicon.

4. Conclusions

Recent developments in computational phraseology and in construction grammar converge on the existence of a complex network of probabilistic constructions, which is at the same time the reflection of the relative cognitive entrenchment of those constructions. Although the notion of entrenchment, inherited from cognitive grammar, might be further specified, it displays many theoretical and practical similarities with the notions of reproducibility, fixedness and even idiomaticity. Indeed, the only observable feature of all those theoretical notions in huge linguistic corpora is the high degree of statistical association of the constructions.

In this contribution, we have shown that very similar types of association can be found at the level of phraseological units, of lexical constructions (as illustrated by Chinese word segmentation), at the level of cultural constructions, and even at the more schematic or abstract level of underlying syntactic patterns. The only general theory of language that offers an explanation for these similarities is construction grammar, but the contribution of phraseology to the theoretical debate is also of paramount importance. Even if we just take traditional phraseology into account, there is no denying that recurrent associations can also be traced back, which confirms the overall importance of a statistical approach.

From a theoretical point of view, this is not to say that statistics are intrinsically present in constructions, in phraseology or in semantics, because they might just be an indirect way of describing the arbitrary pairings of form and meaning. Recent developments in artificial intelligence might however point in the other direction: meaning in itself may turn out to be far more statistical in nature than was previously thought.

On the practical side, learning and teaching a foreign language, or translating languages, may profit from tools allowing for complex statistical manipulation on the basis of huge corpora. More than ever, the big data approach turns out to be of the essence in applied linguistics. It is often fascinating to see that corpus-based data contradict traditional views on many aspects of syntax or lexicon. However, there is a need for more practical tools adapted to language professionals and not just to computer scientists and engineers. The *IdiomSearch* project mentioned in this paper was meant as a tentative step towards that goal, but new user-friendly interfaces are necessary between the big data and actual language use.

Bibliography

Alonso Almeida, Francisco; Ortega Barrera, Ivalla; Quintana Toledo, Elena; Sánchez Cuervo, Margarita E. (eds) 2016. *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*. Newcastle: Cambridge Scholars Publishing.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Journal of Language Resources and Evaluation*. 43: 209–226.

Booij, Geert. 2013. Morphology in Construction Grammar. In: Hoffmann, Thomas; Trousdale, Graeme (eds.). 255–273.

Burger, Harald; Dobrovol'skij, Dmitrij; Kühn, Peter; Norrick, Neal (eds.) 2007. *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*. Berlin / New York: De Gruyter.

Colson, Jean.-Pierre. 2016. Set phrases around globalization: an experiment in corpus-based computational phraseology. In: Alonso Almeida, Francisco; Ortega Barrera, Ivalla; Quintana Toledo, Elena; Sánchez Cuervo, Margarita E. (eds). 141–152.

Colson, Jean-Pierre. 2017. The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In: Mitkov, Ruslan (ed.). 16–28.

Colson, Jean-Pierre. 2018. From Chinese word segmentation to extraction of constructions: two sides of the same algorithmic coin. In: Savary, Agata; Ramisch, Carlos; Hwang, Jena D.; Schneider, Nathan; Andresen, Melanie; Pradhan, Sameer; Petruck, Miriam R. L. (eds). 41–50.

Dixon, Robert M.W.; and Aikhenvald, Aleksandra Y. (eds.) 2002. *Word: A Cross-Linguistic Typology*. Cambridge UK: Cambridge University Press.

Emerson, Thomas. 2005. The second international Chinese word segmentation bake-off. In: *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*. 123–133.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*. 1–32.

Hoffmann, Thomas; Trousdale, Graeme (eds.) 2013. *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Goldberg, Adele. 2013. Constructionist Approaches. In: Hoffmann, Thomas; Trousdale, Graeme (eds.). 15–31.

Hoffmann, Thomas; Trousdale, Graeme (eds.) 2013. *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Gries, Stefan. 2013. Data in Construction Grammar. In: Hoffmann, Thomas; Trousdale, Graeme (eds.). 93–108.

Karaulov, Yu N. 1987. *Russkii yazyk i yazykovaya lichnost* [Russian language and linguistic personality]. Moscow: Nauka.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar*, Volume I, Theoretical Prerequisites. Stanford, California: Stanford University Press.

Langacker, Ronald W. 2008. *Cognitive Grammar – A Basic Introduction*. Oxford: Oxford University Press.

Mitkov, Ruslan (ed.) 2017. *Computational and Corpus-based phraseology*, Lecture Notes in Artificial Intelligence 10596. Cham: Springer International Publishing.

Lewis, Michael. 1993. *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, Michael (ed.) 1997. *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

Savary, Agata; Ramisch, Carolos; Hwang, Jena D.; Schneider, Nathan; Andresen, Melanie; Pradhan, Sameer; Petruck, Miriam R. L. (eds) 2018. *Proceedings of the Joint Workshop on Linguistic Aotation, -Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 25–26 August, Santa Fe (New Mexico, USA), Association for Computational Linguistics.

Stefanowitsch, Anatol. 2013. Collostructional Analysis. In: Hoffmann, Thomas; Trousdale, Graeme (eds.). 290–306.

Vinogradov, Viktor V. 1947. *Russkij jazyk: Grammaticheskoe uchenie o slove* [The grammatical studies of the word]. Moscow: Gosudarstveoye uchebno-pedagogicheskoye izdatel'stvo Ministerstva prosveshcheniya RSFSR.

Wulff, Stephanie. 2013. Words and idioms. In: Hoffmann, Thomas; Trousdale, Graeme (eds.). 15–31. 274–289.

Frazeologia i kognitywne ucieleśnienie: korpusowe dowody i ich zastosowanie w dydaktyce języków obcych i tłumaczeniu

Streszczenie

Kognitywne ucieleśnienie, wywodzące się z gramatyki kognitywnej, jest właściwie bardzo bliskie innym teoretycznym pojęciom takim, jak odtwarzalność, stałość czy idiomatyczność. Za pomocą eksperymentu przeprowadzonego na dużym korpusie językowym, komputerowa frazeologia umożliwia zarówno znalezienie częściowych dowodów dla pojęć teoretycznych, jak i zaproponowanie praktycznych narzędzi dla użytkowników języka. Niniejszy artykuł przedstawia dowody na probabilistyczną naturę sieci konstrukcji. Okazuje się, że statystyczny wynik *cpr-score*, opracowany przede wszystkim do ekstrakcji frazeologizmów, daje istotne wyniki dla innych typów konstrukcji: leksykalnych (w przypadku segmentacji chińskich słów), odniesień kulturowych i nazwanych jednostek, a nawet bardziej schematycznych czy abstrakcyjnych wzorów będących podstawą konstrukcji składniowych.