

The *IdiomSearch* Experiment: Extracting Phraseology from a Probabilistic Network of Constructions

Jean-Pierre Colson^(✉)

Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium
jean-pierre.colson@uclouvain.be

Abstract. This paper reports the preliminary results of an experiment carried out on a large scale for the extraction of PUs (phraseological units, also called idioms) from large web corpora in four languages (English, Spanish, French, Chinese). The use of a new algorithm based on metric clustering techniques, of optimized database storage and of interaction with users and researchers by means of a web application, made it possible to reach high precision scores for most common PUs in the four languages, while further experimentation is still necessary for establishing recall levels with long n-grams. In the meantime, the freely accessible web application makes it possible to visualize the high proportion of phraseology in the broad sense (or of formulaic language): about 30 to 60% of the newspaper articles tested in the experiments consisted of PUs. The most surprising results, however, came from Chinese: as the algorithm had to be changed for taking into account the associations between morphemes, the methodology used made it possible to partly confirm, from a statistical point of view, one of the major claims of construction grammar: the existence of a probabilistic network of constructions, from morphemes to idiomatic phrases.

Keywords: Collocation extraction · Construction grammar · Statistical Semantics · Information retrieval · Collostructions

1 Introduction: Theoretical Background

One of the paradoxes of recent research in corpus and computational linguistics is that the theoretical underpinnings are rarely addressed in detail, as if applying algorithms to an object (language) whose very structure is so controversial were a matter of evidence. In such a complex matter as recurrent linguistic sequences, it is besides necessary to have recourse to huge linguistic corpora and to take linguistic diversity into account.

Foreign language learners and translators are often impressed by the overwhelming importance of prefabricated elements and figurative senses in language. For more than 50 years, corpus and computational linguists have therefore been trying to extract those elements automatically from corpora. However, as pointed out by [13], the results of the automatic extraction of *collocations* in the broad sense are still disappointing, and new avenues of research ought to be explored. Indeed, another paradox of the research on automatic extraction of recurrent sequences in language is that *you don't know exactly what you (and the algorithm) should actually extract*.

Phraseology as defined by [4] encompasses all *phraseological units* or *phraseologisms*, from *collocations* (weakly idiomatic phrases) to proverbs, but *collocation* is widely used in corpus or computational linguistics [13] as a generic term (covering all *phraseological units*). In the same way, *idiom* has sometimes been used for very idiomatic phrases [4] but is also a generic term for all set phrases of a language [17].

Yet another way of describing all language blocks that are accessed as one element of meaning is the notion of *formulaic language* [22]: a *heteromorphic* lexicon makes it possible to store linguistic material of different sizes (morphemes, words, multiword strings); besides, [22] claims that, by default, native speakers will use formulas and will only break down linguistic material into smaller components when they have specific needs (the principle of NOA, Needs Only Analysis). Adults will, on the whole, tend to analyze input and store smaller lexical units. Using formulaic language is considered as a way of promoting the speaker's own interests, because larger and holistic units, with pragmatic and cultural associations, make it possible to exercise a better control over the interpretation of the message by the hearer. In other words, formulaic language is a way of manipulating the hearer. The secondary and somehow artificial character of written language, as opposed to original dialogues, is stressed.

In addition to the great terminological diversity around the notion of phraseology, its extraction raises another thorny issue: where does a phraseological unit (PU) start and where does it end? On the one hand, many PUs are discontinuous (as in *the more you ... the more you* or *X take Y into account*); on the other hand, the exact beginning or the end of the PU are not always straightforward (as in *to VERB (and to VERB) (...)* was with *N/Pro the work of an instant*).

Finally, developments in cognitive linguistics over the past 30 years and among constructionist approaches in particular [17] have proposed a radical new approach to the study of constructions in general. For construction grammar [10, 12], all *constructions* are Saussurean signs, i.e. conventional and learned pairings of form and function at varying levels of complexity and abstraction. Thus, constructions include partially filled words or morphemes (e.g. *pre-N* or *V-ing*), words, but also idioms in the generic sense of PUs: filled (*spill the beans*), partially filled (*take X for granted*), minimally filled (*the more... the more*), and even abstract constructions such as the Ditransitive Verb Construction (*give X to Y*), or the Passive Construction.

The constructionist approaches to language have fundamentally changed the vision of PUs, because of the proposed continuum between lexicon and syntax, the *constructicon* [9, 11]: all language sequences are basically of the same type, as they are all *constructions*, ranging from very schematic and abstract constructions such as the passive, to *substantive* or specific constructions (morphemes and words), some of which are complex (idioms, i.e. PUs). According to this approach, PUs are in no fundamental way different from ordinary words or from syntactic constructions; on the opposite, as pointed out by [23], all constructions are in a sense *idioms*. Thus, partially schematic complex constructions (e.g. *X SPILL the beans (on/about Y)*) are traditionally labelled idioms (or PUs), but this is for [23] simply due to the fact that such constructions make the effects of idiomatic variation very clear, the main point being that they are not fundamentally different from other constructions.

An even more radical view is taken by [7]: all constructions are language-specific, and categories are besides construction-specific. On the basis of a rigorous analysis of

various languages of the world, the author comes to the conclusion that there is no such thing as, for instance, a universal passive construction (similarities can only occur between cognate languages) or a universal *verb* category.

Those fascinating and challenging insights were not only gained through introspection and comparison of relevant examples, but also by means of rigorous psycholinguistic experiments, and were besides confirmed by the *collostructional* approach.

This methodology [14, 15, 21], makes it possible to quantify association strength in and between constructions, and is derived from collocational approaches used in corpus linguistics. For instance, [21] shows that there is a statistical association between verbs and Argument Structure constructions and that verbs display very different association strengths within those constructions.

In constructionist approaches, grammar is conceived as a network of constructions, of which the nature is largely probabilistic [8, 21].

These theoretical and practical issues were the starting point of the *IdiomSearch* experiment. In this paper, a summary is given of its methodology, main results, and possible further developments.

2 Methodology

In order to gain fresh insights into both the practical aspects and the theoretical underpinnings of extracting PUs from corpora, we chose to use a *big data* approach. In the first place, it has been demonstrated by several studies such as [5, 19] that large linguistic corpora (of 100 million tokens or more) are necessary in order to be confronted with various examples of the most common PUs. Besides, as pointed out by [13], the dispersion across corpora has not sufficiently been taken into account, and many studies are of a very limited scope, as opposed to the huge number of PUs in a single language.

[13] also lays stress on additional limitations of current methods for extracting PUs from a corpus: most statistical measures are not directional (they consider *that you* and *you that* as the same PU), they are not easy to reconcile with psycholinguistic or cognitive principles, and they are not easily extendable to longer sequences such as trigrams, fourgrams, let alone 6-grams.

Within the framework of the *IdiomSearch* experiment, it was therefore decided to use the *cpr-score* (*Corpus Proximity Ratio*) described in [6]. This experimental score is non-parametric and directional, as it is derived from information retrieval [1], and more specifically from metric clustering techniques (Fig. 1):

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

Fig. 1. The *cpr-score*

It basically corresponds to the average distance between the component grams of an n-gram, given a window W , set between 20 and 50 tokens according to the language. In order to compute it, it suffices to keep a trace of all the offsets (positions in the text file) of the n-gram and its component grams in a corpus. For instance, the PU *spill the beans* occurs 14 times on a 200 million token web corpus of English; this exact frequency is then divided by the frequency of *spill the beans* within a window of a maximum of 20 tokens between the grams, which yields 15 occurrences. Dividing 14 by 15 gives a *cpr-score* of 0.93. The significance threshold for PUs has been set experimentally at 0.40, while scores as low as 0.065 still yield partly fixed phrases or elements of phrases, as explained in [6].

Table 1. *cpr-score* for a few English idioms

Idiom	Frequency	<i>cpr-score</i>
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

Bigrams such as *easy rider*, *New York*, *sharp criticism* etc. are easy to explain to informants having to evaluate the results of automatic extraction. Once the level of trigram is reached, and in particular for longer n-grams, making clear to non-linguists what we mean exactly by PUs (or collocations, or formulas and so on) is not an easy matter, especially if the diversity of languages is taken into account.

This is the reason why *recall* was difficult to measure, in the absence of reliable gold standards in several languages, as is the case for most automated tasks in NLP (natural language processing). *Precision* on the basis of native speaker judgment or dictionaries, on the other hand, is very high for very idiomatic phrases (*cpr* > 0.40). The novel feature of *cpr* is that it is very stable, *no matter what the length of the n-gram is* (at least between bigrams and 7-grams). Thus, in Table 1, the very fixed PUs or *idioms* chosen randomly in the dictionary all yielded a very high *cpr-score* on a 200 million token web corpus¹.

At the lower end of the spectrum, however, linguistic structures yielding partly significant association scores (with a *cpr* between 0.065 and 0.40) were problematic. As a rule, separating the wheat from the chaff at the left hand of the phraseological spectrum is nigh on impossible, as already pointed out by [4].

In order to shed new light on the interplay between lexis and grammar, and particularly with respect to phraseology [4], formulaic language [22] and construction grammar [17], the *cpr-score* was computed on large web corpora of 200 million tokens.

¹ The web corpora used in the *IdiomSearch* experiment were assembled using the WebBootCat tool provided by the Sketch Engine (<http://sketchengine.co.uk>), on the basis of seed words and following the methodology described in [2].

As mentioned in note 1, assembling the corpora happened in an automated way, with a balanced list of seed words. Web corpora of 200 million tokens were created for English, French, Spanish and Chinese² (Mandarin, simplified spelling).

For each language, all n-grams ranging from bigrams to 7-grams were extracted from the corpus, with a frequency threshold of 3 occurrences (for 200 million tokens). The *cpr-score* was computed for each selected n-gram. Thanks to an optimization of the database requests by means of a query likelihood model [18], this took no more than one week per language, with an average time for each request of 0.07 second on a Linux machine. Several regex-based algorithms were passed through the results, in order to deal with the problem of phraseological encapsulation: smaller n-grams (bigrams, trigrams) may be included in the results of longer n-grams (e.g. *take the rough* as a partial result for *take the rough with the smooth*); in this case, n-grams included in larger n-grams were discarded, except if their association score was high, which is sometimes the case if competing PUs are at stake (e.g. *take a walk* and *take a walk down memory lane*).

In order to allow easy access to researchers and students, all selected n-grams were put in a database, freely accessible from a web application³. The user is presented with results highlighted in colors ranging from pale yellow (partly fixed and frequent) to deep red (very fixed and not frequent). The number of words in the results is also indicated, as well as the number of phrases per word (PW ratio) and the number of phrases in the text (PT ratio). The PT ratio is computed by checking how many words (tokens) of the text are included in PUs. Thus, a PT ratio of 0.45 means that just 55% of the words of the text are not included in phrases.

3 Results and Discussion

3.1 The Proportion of Phraseology in Texts

Before providing an overview of the results produced by the IdiomSearch experiment, it may be useful to start from a small text fragment, in order to show what types of PUs the algorithm is able to extract: a few lines from an opinion piece on Brexit, published in a British newspaper⁴.

“Don’t get me wrong, I – and I’m sure many other Labour voters – consider Brexit to be the biggest act of political self-sabotage in my lifetime, with stark consequences for my generation and those following it. Were it to be miraculously cancelled I would be over the moon. But there’s a sense that some would like everything to be about Brexit and only Brexit. When it comes to what people care about in 2017 – and vote on the basis of – that simply is not the case.”

The algorithm used by the IdiomSearch application makes it possible to extract the following PUs on the fly.

² For Chinese, the corpus does not consist of 200 million Chinese characters (*hans*), but of 200 million Chinese words (as tokens).

³ IdiomSearch is accessible on the web at: <http://idiomsearch.LSTI.ucl.ac.be>.

⁴ The Guardian, <http://www.theguardian.com>, 7 August 2017.

Table 2. Examples of extracted PUs

PUs	Frequency	<i>cpr-score</i>
Don't get me wrong	124	0.99
Many other	3507	0.55
Be the biggest	179	0.20
In my lifetime	56	0.66
Consequences for	407	0.52
Would be over the moon	3	0.75
Sense that	2365	0.26
Would like	12456	0.80
When it comes to	2026	0.57
Care about	1714	0.46
On the basis of	2752	0.73
Is not the case	449	0.23

As one of the aims of the experiment is to receive feedback from users, and to improve the algorithm and the selection of corpora, no systematic survey of different registers or text types has been carried out so far, but more than 200 newspaper articles have been tested in the different languages, confirming that such a method yields a high percentage of phraseology (as expressed by the PT ratio) in newspaper articles. For opinion pieces published by British or American newspapers, the overall PT ratio, reflecting in other words the total percentage of phraseology that could be extracted by the algorithm, lies between 0.30 and 0.55. For the whole of the example text above (803 tokens), it was precisely 0.50 (i.e. roughly half of the text consisted of PUs). Table 3 presents the results (number of tokens and PT ratio) for 5 newspaper articles (comments on the news).

A comparison was also made with about 200 fragments (of comparable length) from available corpora of spoken English. An example with 5 texts is given in Table 3, comprising 5 randomly selected passages from the *Corpus of Spoken, Professional American-English (Athelstan)*⁵.

A word of caution is necessary in the interpretation of these results. It should be stressed again that the extraction of PUs by means of the *cpr-score* crucially depends on the reference corpus used. As mentioned in the preceding section, a balanced English corpus of 200 million tokens was used for the experiment, but there is no guarantee that all phrases from a specific text are also present on the reference corpus; if that is not the case, they can of course not be extracted by the program. The PT ratio is therefore an indication of the minimal percentage of phraseology (in the broad sense).

It may also come as a surprise that fragments from spoken corpora did not show major differences with written corpora as far as the total percentage of phraseology (PT ratio) is concerned. This is partly due to the topics that were discussed during the interviews: the Athelstan fragments (Table 4), for instance, contain interviews about university topics. Many differences appear in the types of PUs used, with many typical

⁵ Athelstan Homepage, <http://www.athel.com/cspatg.html>, last accessed 2017/08/09.

Table 3. *PT ratio* (percentage of phraseology) for 5 newspaper articles in English

	Tokens per text	<i>PT ratio</i>
Article one ^a	743	0.41
Article two ^b	789	0.41
Article three ^c	730	0.48
Article four ^d	660	0.56
Article five ^e	769	0.55

^aThe Guardian Homepage, <http://www.theguardian.com>, 2015/10/05.

^bThe Guardian Homepage, <http://www.theguardian.com>, 2015/03/31.

^cThe New York Times Homepage, <http://www.nytimes.com>, 2016/01/13.

^dThe Daily Telegraph Homepage, <http://www.telegraph.co.uk>, 2015/04/04.

^eThe Daily Telegraph Homepage, <http://www.telegraph.co.uk>, 2015/05/16.

phrases of spoken English (*Thank you very much, I would strongly suggest, I'm not sure, I wondered if*), but the average percentage is not very different from the results obtained for the written texts.

From a theoretical point of view, these results partly confirm the hypothesis that about 50% of all texts consist of phraseology in the broad sense, the *idiom principle* [20], as the *PT ratio* (see examples in Tables 3 and 4) reaches figures between 0.35 and 0.60.

Table 4. *PT ratio* (percentage of phraseology) for 5 texts from the *Athelstan* corpus

	Tokens per text	<i>PT ratio</i>
Text one	830	0.50
Text two	795	0.56
Text three	793	0.39
Text four	809	0.49
Text five	759	0.48

The examples of extracted PUs, as illustrated in Table 2, also provide convincing evidence for the existence of a network of statistical association between the elements of complex structures such as idioms (*spill the beans*), grammatical collocations (*care about, consequences for*), communicative formulas (*don't get me wrong, when it comes to*), which is compatible with construction grammar [17].

3.2 Experiments with Spanish, French and Chinese

Within the framework of the *IdiomSearch* experiment, similar tests were conducted for Spanish, French and Chinese (Mandarin, simplified spelling).

For Spanish and French, the results were comparable to those obtained for English, with roughly the same percentages of phraseology in the broad sense. As often in computational linguistics, the main difficulties were technical ones: compiling web corpora by means of the robot required special attention for possible errors of encoding⁶. As Spanish and French are also Indo-European, segmented and inflectional languages, it comes as no surprise that the algorithm was able to work in much the same way as for English. Chinese, on the other hand, represented in many regards a daunting challenge for the *IdiomSearch* experiment.

Chinese is, in the first place, an *unsegmented* language: there are no blanks between words. Modern Mandarin Chinese remains largely an isolating language (there is a very low morpheme per word ratio, and no inflectional morphology). In classical Chinese, one character (*han*) corresponded to one word, but most words in modern Chinese consist of two characters, and sometimes more. The situation is therefore rather complex, which makes it also particularly interesting for testing linguistic hypotheses.

According to [7], constructions but also categories are language-specific, and Chinese is often cited as an example of a language apparently functioning in a totally different way for grammatical categories such as Noun and Verb. As pointed out by [22], several studies have besides confirmed that Chinese native speakers make different decisions when they have to segment a text into words, and that even the same persons do not always confirm their first choices. Therefore, words –if they exist at all, function in a very different way in Chinese.

For these reasons, it was not an easy matter to adapt the *IdiomSearch* algorithm to Mandarin Chinese, as the *cpr-score* is based on the average distance between words in a corpus. As a temporary solution, it was decided to consider the distance between Chinese characters, which made it possible to reach very high precision scores on the basis of established Chinese phrases. Table 5 shows the frequency and the *cpr-score* for a few common *chengyu*, the 4-syllable idioms [16], from which study the examples are borrowed.

The examples of *chengyu* in Table 5 clearly show the achievements of the *cpr-score* for very fixed Chinese phrases: contrary to what might have been expected, the score works particularly well. Being non-inflectional, mostly isolating and having under-gone few influences from other languages, Mandarin Chinese is actually well suited for testing linguistic extraction algorithms, in spite of some technical adaptations.

An important finding of the *IdiomSearch* experiment, thanks to extensive testing with Chinese, is that statistical association of morphemes/words is partly discontinuous, **even within established phrases**. This contradicts the intuition that adding one element at a time to a *n*-gram allows to narrow down the probabilities. Not only is frequency of minor importance in the statistical association of words or morphemes,

⁶ The computational issue is well known: many web pages contain Unicode errors; the robot assumes that the downloaded web page is in Unicode, but the errors remain and appear in the web corpus.

Table 5. *cpr-score* for a few Chinese *chengyu*

<i>Chengyu</i>	Frequency	<i>cpr-score</i>
小心翼翼 <i>xiǎo-xīn-yì-yì</i> ^a	1442	1.00
自怨自艾 <i>zì-yuàn-zì-yì</i> ^b	100	0.99
趁热打铁 <i>chèn-rè-dǎ-tiě</i> ^c	87	1.00
孤掌难鸣 <i>g -zhǎng-nán-míng</i> ^d	33	1.00
杯弓蛇影 <i>bēi-gōng-shé-yǐng</i> ^e	33	0.97
破釜沉舟 <i>pò-fǔ-chén-zhōu</i> ^f	153	0.99

^aLiterally: small-heart-respectful-respectful; meaning: being careful, with respect, taking precautions [16].

^bLiterally: self-hate-self-refrain; meaning: to repent, to be sorry for one's deeds [16].

^cLiterally: profit-heat-beat-iron; meaning: strike while the iron is hot [16].

^dLiterally: solitary-palm-difficult-resonate; meaning: who is without support is doomed to fail [16].

^eLiterally: cup-bow-snake-reflection; meaning: be alarmed for nothing [16].

^fLiterally: break-cauldron-sink-ship; meaning: be decided to win or lose, cross the Rubicon [16].

but there is no strict continuity between the elements. Suppose, for instance, that ABCD is a common PU or *constructional idiom* in English. Thanks to the *cpr-score*, it is possible to measure the statistical association between A + B + C + D. It is also possible to compute the score for A + B + C. Intuitively, one may be tempted to think that A + B + C is incomplete (as D is missing), and that the statistical score for ABC will therefore be lower than for ABCD. This is indeed often the case, but there are many counterexamples showing that the statistical association is much more complex, as internal PUs may interact with the overall score.

Table 6 illustrates this point for the communicative English phrases *long time no see* and *the next thing I knew*, and for the Chinese proverb 书中自有黄金屋⁷.

For both English phrases and for the Chinese proverb in Table 6, one can clearly see that adding a gram to the sequence, although it brings the sequence closer to the complete phrase, does not necessarily yields a higher statistical score at each level. These examples also suggest that the best method for extracting longer PUs is not bottom-up but top-down: starting at the level of a gram, adding one gram at a time, and checking the statistical association on the corpus at each level, will not yield good results, because the association is sometimes discontinuous, as between *long time* and *long time no*. On the contrary, the method used in the IdiomSearch experiment was top-down, in the sense that all n-grams (ranging from bigrams to 7-grams) were extracted (with a frequency threshold of 3 occurrences for 200 million tokens); the association was measured for each n-gram at all levels, which made it possible to

⁷ shū zhōng zì yǒu huángjīn wū, *A book holds a house of gold*.

Table 6. *cpr-score* for successive n-grams within the phrases *long time no see* and *the next thing I knew*

n-gram	Frequency	<i>cpr-score</i>
long time	3604	0.68
long time no	15	0.12
long time no see	4	0.80
the next	16344	0.56
the next thing	95	0.20
the next thing I	24	0.20
the next thing I knew	13	0.46
书中	6731	0.47
书中自	57	0.17
书中自有	51	0.66
书中自有黄	28	0.72
书中自有黄金	28	0.72
书中自有黄金屋	28	0.72

extract even idiomatic 7-grams. The difficulty, however, remains for cases such as the Chinese proverb in Table 6, because the maximum frequency and association scores are already reached at the level of the 5-gram, whereas the full proverb is a 7-gram. This is one of the reasons why the algorithm has to be slightly adapted for each language, applying the general principle of the *baboushka* (Russian nesting dolls), or *encapsulation*: for long PUs such as 7-grams, a fine-tuned analysis of the associations between the internal grams is crucial.

3.3 A Probabilistic Network of Constructions

As mentioned in the introduction, constructionist approaches view grammar as a complex network of constructions, and several researchers hold the view that this network is based on probabilistic principles [8, 21].

Thanks to the specific problems posed by the Chinese language (as it is non-inflectional and unsegmented), some additional experiments were carried out within the framework of *IdiomSearch*.

In the first place, applying the statistical association score to an unsegmented language poses the question of the artificial segmentation in European languages. *Tea cup*, for instance, can be written in one word or two words, *à l'aéroport* is considered as a sequence of 3 words in French but the Spanish equivalent *al aeropuerto* as a two word sequence.

A widely accepted view in construction grammar [3] is precisely that *constructional idioms* exist both at syntactic and morphological level. A constructional idiom is then defined as a syntactic or morphological schema in which at least one position is fixed [3]. This is, for instance, the case in *This is the life*, but also in adjectives such as *un-believable* (in which *-able* cannot be replaced by *-ible* or other suffixes). Constructional morphology views complex words just as complex syntactic constructions.

Table 7. *cpr-score* for the association of morphemes in a few English words

Morphemes	Frequency	<i>cpr-score</i>
Accept-able	3855	0.97
Afford-able	2030	0.99
Approach-able	365	0.73
Circum-scrib-ed	107	1.00
Friend-ship	1399	0.97
Ir-respons-ible	505	0.92
System-atic	1893	1.00
Un-intellig-ible	104	0.74
Un-precedent-ed	104	1.00

All constructions are partly fixed, but there is a cline from schematic to substantive constructions.

The specific claims about constructional morphology [3] have been made on the basis of solid examples and of cognitive experimentation, but the question arises of their statistical foundation in corpora. In order to test this hypothesis on a wide scale, it would be necessary to apply morphological segmentation to a whole corpus. In the meantime, preliminary experiments with the *cpr-score* indeed confirm that many words composed of several morphemes display association scores that are quite comparable to those obtained for PUs, especially for fully idiomatic PUs or *idioms*.

In the examples presented under Table 7, a number of English words were treated as separate morphemes, and the *cpr-score* was computed for their association.

As illustrated by the examples from Table 7, there is indeed very little difference between the statistical associations prevailing within idiomatic PUs (Table 1) and within morphological constructs (Table 7).

One may actually go one step further, as [3] does, and consider that constructions such as the English past tense construction ‘*have + past participle*’ are also constructional idioms, in which the auxiliary is fixed and the participle slot schematic. Again, this hypothesis can be supported by the *cpr-score*. We may, for instance compute the score for a given form, say *has*, followed by a maximum of two tokens, followed by the suffix *-ed*, indicating many regular past participles. If the above mentioned structure is indeed a constructional idiom, we will expect a very significant score ($cpr > 0.40$). This is indeed the case: our 200 million token corpus yields a score of 0.58 for 52350 occurrences. In other words, the *construction itself* can somehow be captured by the *cpr-score*, as it can be extended to 7-grams or even higher.

4 Conclusion

As the extraction of phraseological units/idioms that consist of more than two words is fraught with a wide range of difficulties, both practical and theoretical, the IdiomSearch Experiment sought to determine if a corpus-driven experimental score, the *cpr-score* [6], based on techniques derived from information retrieval, could yield acceptable

results for different languages. The experiment was carried out on English, Spanish, French and (Mandarin) Chinese.

The recourse to large web corpora of 200 million tokens each, to optimized data-base storage and to a user-friendly web application has already made it possible so far to receive extensive feedback from students and other researchers, while also shedding fresh light on the theoretical underpinnings of any attempt to derive associative meaning from n-grams.

Although the statistical score can still be improved, as well as the qualitative and quantitative aspects of the web corpora, the preliminary results indicate that most common PUs and even a high number of relatively rare and very fixed PUs can be extracted by the *cpr-score* for European languages such as English, Spanish and French. The fact that the results are at present slightly better for English than for Spanish and French may be due to two main reasons. First, there are many more pages in English on the Web⁸, which might explain that robots assembling pages on the basis of seed words in specific combinations yield more representative results for English; the second reason has to do with technical issues around the encoding of special characters in languages other than English.

The results obtained for Chinese are however particularly interesting, because they confirm that we should relativize our Eurocentric vision of language as being assembled from words and syntax. The whole cline of statistical associations, measured by the *cpr-score*, starts at the level of morphemes and ends up in schematic constructions, which is quite compatible with and may even serve as evidence for the claims made by the constructionist approaches to language. Thus, extracting phraseology from corpora looks like an achievable target, but the whole enterprise may only make sense against the backdrop of a probabilistic network of constructions.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press/Addison Wesley, New York (1999)
2. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *J. Lang. Res. Eval.* **43**, 209–226 (2009)
3. Booij, G.: Morphology in construction grammar. In: Hoffmann, T., Trousdale, G. (eds.) *The Oxford Handbook of Construction Grammar*, pp. 255–273. Oxford University Press, Oxford/New York (2013)
4. Burger, H., Dobrovolskij, D., Kühn, P., Norrick, N. (eds.): *Phraseologie/Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung/An International Hand-book of Contemporary Research*. De Gruyter, Berlin/New York (2007)

⁸ According to Wikipedia, English is good for 51.6% of all web pages, Spanish for 5.1%, French for 4.1%, and Chinese for 2.0%. Wikipedia homepage, https://en.wikipedia.org/wiki/Languages_used_on_the_Internet, last accessed 2017/08/17.

5. Colson, J.-P.: The World Wide Web as a corpus for set phrases. In: Burger, H., Dobrovol'skij, D., Kühn, P., Norrick, N. (eds.) *Phraseologie/Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung/An International Handbook of Contemporary Research*, pp. 1071–1077. De Gruyter, Berlin/ New York (2007)
6. Colson, J.-P.: Set phrases around globalization: an experiment in corpus-based computational phraseology. In: Alonso Almeida, F., Ortega Barrera, I., Quintana Toledo, E., Sanchez Cuervo, M.E. (eds.) *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*. Cambridge Scholars Publishing, Newcastle, pp. 141–152 (2016)
7. Croft, W.: *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford (2001)
8. Croft, W.: Radical construction grammar. In: Hoffmann, T.H., Trousdale, G. (eds.) *The Oxford Handbook of Construction Grammar*, pp. 211–232. Oxford University Press, Oxford/New York (2013)
9. Fillmore, C.H.: The mechanisms of construction grammar. *Berkeley Linguistic Soc.* **14**, 35–55 (1988)
10. Goldberg, A.: *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago (1995)
11. Goldberg, A.: Constructions: a new theoretical approach to language. *Trends Cogn. Sci.* **7** (5), 219–224 (2003)
12. Goldberg, A.: *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford (2006)
13. Gries, S.: 50-something years of work on collocations. What is or should be next *Int. J. Corpus Linguist.* **18**, 137–165 (2013)
14. Gries, S.: Data in construction grammar. In: Hoffmann, T.H., Trousdale, G. (eds.) *The Oxford Handbook of Construction Grammar*, pp. 93–108. Oxford University Press, Oxford/New York (2013)
15. Gries, S., Stefanowitsch, A.: Extending collocation analysis: a corpus-based perspective on 'Alternations'. *Int. J. Corpus Linguist.* **9**(1), 97–129 (2004)
16. Henry, K.: Les chengyu du chinois: caractérisation de phrasèmes hors norme. *Yearb. Phraseology* **7**, 99–126 (2016)
17. Hoffmann, T.H., Trousdale, G. (eds.): *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/New York (2013)
18. Manning, C.H., Raghavan, P., Schütze, H.: *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2009)
19. Moon, R.: *Fixed Expressions and Idioms in English*. Clarendon Press, Oxford (1998)
20. Sinclair, J.: *Corpus, Concordance, Collocation*. Oxford University Press, Oxford (1991)
21. Stefanowitsch, A.: Collocation analysis. In: Hoffmann, T.H., Trousdale, G. (eds.) *The Oxford Handbook of Construction Grammar*, pp. 290–306. Oxford University Press, Oxford/New York (2013)
22. Wray, A.: *Formulaic Language: Pushing the Boundaries*. Oxford University Press, Oxford (2008)
23. Wulff, S.: Words and idioms. In: Hoffmann, T.H., Trousdale, G. (eds.) *The Oxford Handbook of Construction Grammar*, pp. 274–289. Oxford University Press, Oxford/New York (2013)

Computational and Corpus-Based Phraseology
Second International Conference, Europhras 2017,
London, UK, November 13-14, 2017, Proceedings
Mitkov, R. (Ed.)
2017, XV, 463 p., Softcover
ISBN: 978-3-319-69804-5