From Chinese Word Segmentation to Extraction of Constructions: Two Sides of the Same Algorithmic Coin

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

Abstract

This paper presents the results of two experiments carried out within the framework of computational construction grammar. Starting from the constructionist point of view that there are just constructions in language, including lexical ones, we tested the validity of a clustering algorithm that was primarily designed for MWE extraction, the *cpr-score* (Colson, 2017), on Chinese word segmentation. Our results indicate a striking recall rate of 75 percent without any special adaptation to Chinese or to the lexicon, which confirms that there is some similarity between extracting MWEs and CWS. Our second experiment also suggests that the same methodology might be used for extracting more schematic or abstract constructions, thereby providing evidence for the statistical foundation of construction grammar.

1 Introduction

In many respects, constructionist approaches have led to a new paradigm in the description of language structure. Building on Langacker's cognitive grammar (Langacker, 2008), the different versions of construction grammar (CxG) converge on the notion of constructions, defined as Saussurean signs, i.e. "conventional, learned form-function pairings at varying levels of complexity and abstraction" (Goldberg, 2013: 17). A construction may be a word in the traditional sense (e.g. *book*), a bound morpheme (*pre-*, *-ing*), an idiom (*spill the beans, take the rough with the smooth*), a partially filled idiom (*take X into account*), but also an abstract construction such as the ditransitive construction or the passive. As the famous quotation goes (Goldberg, 2006: 18), "It's constructions all the way down", i.e. language structure is made of nothing else than constructions, at various degrees of abstraction and schematicity. Schematic slots are the positions in the constructions allowing for several choices (e.g. X in *take X into account*), whereas specific (or substantive) slots are fixed (e.g. *into* and *account* in the same construction).

The continuum between lexicon and syntax plays a key role in CxG: there is no strict borderline between grammar on the one hand and the lexicon on the other, and this cline has been called the *constructicon* (Fillmore, 1988; Goldberg, 2003). Thus, the constructicon includes all types of constructions, be they of a more syntactic, morphological, phonological, phraseological, pragmatic or lexical nature.

As a general theory of language, CxG has far-reaching consequences for corpus and computational linguistics. In particular, it sheds a new light on multiword expressions (MWEs), in the general sense of all word combinations displaying lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies. Indeed, all constructions are per definition partly *idiosyncratic* and in that sense partly *idiomatic*: "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)

It is worth noting that constructions are seen as a complex network, ranging from abstract to specific, from simple to complex and from schematic to idiomatic constructions. Crucially, this network of constructions is thought to be of a probabilistic nature (Croft, 2013; Stefanowitsch, 2013).

Proceedings of the Joint Workshop on
Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pages 41–50
Santa Fe, New Mexico, USA, August 25-26, 2018.

Although there may be different ways of adhering to the constructionist approach, we think that (at least) two major theoretical claims of CxG have practical consequences for any computational analysis of MWEs:

1. As morphemes are also constructions, MWEs should rather be studied as MMEs (multimorphemic expressions). This makes it possible to apply the notion of constructions to the diversity of languages (Croft, 2001), and to be coherent with the constructicon. It indeed follows from the notion of constructicon that morphology and syntax are two sides of the same coin. For Booij (2013), exactly the same principles apply to schematic and idiomatic constructions at morphological and syntactic level. Thus, "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).
2. The network of constructions is of a probabilistic nature, so that statistical associations should not (only) be measured between the component parts of MWEs, but also at various levels of schematicity and abstraction.

From a practical point of view, point 1 makes it necessary to question the validity of traditional corpus analysis based on tokens (the traditional strings of letters separated by a blank). Constructionist morphology (Booij, 2013) suggests that associations between morphemes should be studied in the same way as syntactic associations between words, which means that using algorithms based on simple tokens will yield incomplete results. Besides, point 2 implies that statistical associations may exist between schematic and specific slots of constructions, which makes it necessary to adapt the corpus data by means of POS-tagging or more sophisticated representations. Some promising results have already been achieved by *collostructional analysis* (Gries and Stefanowitsch, 2004; Stefanowitsch, 2013), a methodology that makes it possible to quantify association strength in constructions, and is derived from collocational approaches used in corpus linguistics. Key findings are the statistical association between verbs and argument structure constructions, and the probabilistic relation between abstract grammatical constructions and concrete lexical constructions, the traditional words of the language. By having recourse to statistical measures and to linguistic corpora, collostructional analysis confirms that there is a cline from grammar to lexicon, and that the global network of all constructions may (largely) be governed by probabilistic principles.

For all these reasons, the implications of CxG for corpus and computational linguistics should not be underestimated, and a number of common practices of both disciplines should be adapted if we apply the principles of CxG:

- *Tokenization*: the traditional tokens should be questioned and cross-token measures should also be tested if we take morphological constructions into account
- *Lemmatization*: if morphemes are considered as constructions (Booij, 2013), lemmas should be considered as conventional units of meaning, and their concrete realizations (with different affixes) may be as important as the lemmatized form
- *Phonological features*: most corpora do not take phonological features into account (intonation, word stress), but these are an integral part of the construction. To give just one example, *not* will in most cases be treated as a simple substantive construction *inheriting* from the abstract NEG-construction (negation), but it can also be a complex idiomatic construction in cases such as *This book is excellent. Not!* (an ironical way of expressing the opposite of the preceding clause, with rising intonation). In this example, the idiomatic construction will have a rising intonation as one of its defining features, which should ideally be rendered by corpus annotation.

According to Gries (2013), the automatic extraction of collocations (in the general sense of MWEs) from corpora has been going on for over 50 years, but has produced very mixed results. If we add one level of complexity, that of constructions as defined by CxG, the situation may get even worse. Most studies dedicated to the extraction of MWEs face the problem of the validity of the gold standard. The very notion of MWEs may receive slightly different definitions, and it co-exists with other terms such as *collocations, set phrases, idioms, phraseological units*, to name just a few. Finally, *formulaic language* (Wray, 2008) has shed new light on the importance of all kinds of MWEs in the development and psychological background of language.

For the reasons set out above, improving the automatic extraction of MWEs from linguistic corpora, while integrating the fresh insights gained from CxG, appears at best as a daunting challenge. However, we will argue that CxG provides us with the following clues as to the extraction of meaningful structures in language:

1. According to CxG, the whole network of constructions is of a probabilistic nature.
2. As a consequence, the same statistical method should yield comparable results at the various levels of abstraction and specificity.
3. Improving the extraction algorithm at one point of the probabilistic network (for instance, for the extraction of MWEs) should therefore be useful as well for other types of constructions.

It is noteworthy that the continuum from syntax to lexicon, one of the tenets of CxG, poses another major problem to the extraction of MWEs. As in information retrieval, precision and recall play a central role in automatic extraction of collocations and MWEs, but these two notions can only be tested with reference to a gold standard: native speakers provide the researcher with manual results, for instance a list of the MWEs or a subcategorization of them. However, the cline from syntax to lexicon and from abstract to specific constructions implies that such a task is nigh on impossible, because of the very high number of borderline cases between all categories of constructions.

In this paper, we will argue that working from Chinese word segmentation may offer fresh insights into the organization of the probabilistic network of constructions mentioned by CxG. Mandarin Chinese is an unsegmented language (words are not separated by a blank). This offers the advantage of a linguistic material that can be analyzed along the whole continuum of constructions, from simple and complex words to idioms and proverbs.

2 Related work

The very notion of *word* remains controversial in Mandarin Chinese (Dixon and Aikhenvald, 2002). Experiments show that native speakers of Chinese not only disagree among themselves as to the exact segmentation of all sentences, but are often unable to replicate their own previous decisions (Bassetti, 2005). It is generally accepted that there is **an agreement of about 75 % among native speakers** as to the correct segmentation of a Chinese text into words (Sproat et al., 1996; Ying Xu et al., 2010). From the point of view of construction grammar, Chinese is therefore an excellent example of the continuum between syntax and lexicon, as even native speakers are sometimes confronted with the fuzzy borderline between constructions, phrases and words, which results in unclear segmentation.

In computational linguistics and information retrieval (IR), the state-of-the art method for Chinese word segmentation (CWS) is to tokenize an input text by using a monolingual supervised model trained on hand-annotated data, e.g. the Chinese treebank (Xue et al., 2005). It should be emphasized that such a method is not quite compatible with construction grammar, as it relies, for segmenting constructions, on decisions made by native speakers and dictionaries: this means that the cline from syntax to lexicon, and the description of constructions as a whole will depend on elements of linguistic representation rather than on evidence gained from corpora.

A full data-driven and statistical approach to the segmentation of Chinese has been taken by Xu et al. (2009), who propose the *Tightness Continuum Measure.* Their approach is based on document frequencies for segmentation patterns in corpora, and has been tested for 4-grams (in this case 4 Chinese characters or *hans*). Their results confirm the continuum of 'tightness and looseness' (Xu et al., 2009: 9) for Chinese strings, but the authors do not mention the fact that this actually corroborates one of the basic assumption of construction grammar, viz. the cline from syntax to lexicon. The *Tightness Continuum Measure* has been applied to Chinese information retrieval (CIR) by Xu et al. (2010). Their results show, again with the example of Chinese 4-grams, that a segmentation based on the *Tightness Continuum* performs better for CIR. It should be noted, however, that the better scores obtained with the *Tightness Continuum* were measured with scores used in IR and not against manually segmented texts.

More recent attempts to achieve or improve CWS on the basis of algorithms involve bilingual constraints in statistical machine translation (Zeng et al., 2014) or neural networks (Cai and Zhao, 2016).

It has also been pointed out that there is a high degree of similarity between CWS and MWE extraction (Xu et al., 2010). This should come as no surprise, if we take the constructionist view that language is made up of a complex and probabilistic network of constructions, in which there is no clear border

between (free) syntax and MWEs. Any progress made in data-driven CWS may therefore have a positive impact on MWE extraction, and vice versa.

However, the problem with existing studies is that they are almost always of limited scope, and do not deal with both phenomena on a large scale, with recourse to huge linguistic corpora in several languages. An attempt to fill this gap has been proposed by the *IdiomSearch* project (Colson, 2017). A provisional web application has been designed[1] in order to test the automatic extraction of MWEs in the broadest sense from large linguistic corpora in English, Spanish, French and (simplified) Mandarin Chinese. The statistical score used is the *cpr-score* (Colson, 2017), an adaptation of a well-known technique used in IR, *metric clusters* (Baeza-Yates and Ribeiro-Neto, 1999). From a mathematical point of view, the *cpr-score* may be described as follows.

Let a given *n*-gram of length *n* be represented as ($w_1, w_2, \ldots, w_n$), with each w_i belonging to the lexicon of a given language (for example the 3-gram "*spill the beans"*). We denote the gram appearing at position *t* in the corpus by the variable x_t. Thus, $x_t = w_i$ means that the gram w_i (e.g. *beans*), from the *n*-gram ($w_1, w_2, \ldots, w_n$), is present at position *t* (represented by a long integer) in the corpus file.

We further denote as

$$n(w_1, w_2, \ldots, w_n) \triangleq n(x_t = w_1, x_{t+1} = w_2, \ldots, x_{t+n-1} = w_n) \quad (1)$$

the number of occurrences (frequency) of the **exact** *n*-gram ($w_1, w_2, \ldots, w_n$), for instance the frequency of *spill the beans*, in the whole corpus, with no other token between the component grams (excluding e.g. *spill the proverbial beans*). As indicated in the right-hand side of Equation (1), it aims to count the number of occurrences of the event $(x_t = w_1, x_{t+1} = w_2, \ldots, x_{t+n-1} = w_n)$ in the corpus.
Moreover, the expression

$$n\left(x_{t_1} = w_1, x_{t_2} = w_2, \ldots, x_{t_n} = w_n \,\middle|\, \max(t_{i+1} - t_i) \leq W;\ i = 1, \ldots, n-1\right) \quad (2)$$

counts the total number of occurrences of the component grams (e.g. *spill*, *the* and *beans*), appearing sequentially at some positions $t_1 < t_2 \ldots < t_n$, in the corpus but with the constraint that they should be separated by a window of less than $W + 1$ positions (the maximum gap window is less or equal to W). The constant variable W (maximum window length) has been experimentally set at an integer value corresponding to a distance of 20 to 50 tokens, according to the corpus and the language. For English, it is typically set at a value representing a distance of about 20 words (as tokens).

Thus, (1) will give the exact frequency of the *n*-gram with no window between the component grams, while (2) allows for a maximal window (corresponding to up to 50 tokens) between each gram. The final expression used to measure the *cpr-score* is simply the ratio between (1) and (2):

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

Figure 1. The *cpr-score*

As in the case of metric clusters, the geometric distance between relevant elements of meaning is crucial in this approach. The implementation of the *cpr-score* can be achieved by several computational techniques, for instance by measuring the position of the strings in the whole file, by having recourse to regular expressions, or by complex indexation systems. The fastest results have been obtained by implementing a *query likelihood model* (Manning et al., 2009) such as the *Lemur Project*.[2] The *Indri Retrieval Model* included in this project can thus be parameterized to compute *cpr-scores* on pre-indexed corpora. Preliminary results obtained with the *cpr-score* indicate a level of precision higher than 95 percent if measured on a list of MWEs from dictionaries. Measuring precision and recall for MWEs of length 2 to 12 in real texts, however, poses the thorny theoretical issue of what can be objectively called a MWE or an idiomatic construction. It is also worthy of note that the *cpr-score*, deriving from metric clusters, is not fundamentally different from the above mentioned *Tightness Continuum* (Xu et al., 2010),

[1] http://idiomsearch.lsti.ucl.ac.be
[2] https://www.lemurproject.org/

based on document frequencies. We would argue that the use of document frequencies is precisely another way of introducing a windowing technique, which can therefore be seen as another variant of metric clusters.

3 An experiment in Chinese word segmentation (CWS) based on MWE recognition

As we have seen in section 1, construction grammar claims that there is a cline from syntax to lexicon, and that the structure of language therefore consists of a complex network of interrelated constructions. If this theoretical claim is correct, algorithms that are designed to extract MWEs should also be able to extract lexical constructions, provided that the corpus is adapted to that purpose. For European languages, it will for instance be necessary to start from morphemes instead of (conventional) words. In the case of Chinese, construction grammar predicts that looking for larger elements of meaning against the backdrop of a network of constructions will bring segmentation (CWS) and MWE recognition very close to each other.

In this paper we report the first results of an innovative experiment designed to test this general hypothesis.

3.1 Methodology

As this experiment is an extension of the *IdiomSearch Project*, we used as a reference corpus the same Mandarin Chinese corpus: a web-based general corpus, compiled by the *WebBootCat* tool provided by the *Sketch Engine*.[3] The methodology for compiling a general web corpus on the basis of *seed words* is fully described in Baroni et al. (2009). The likewise assembled corpus of (simplified) Mandarin Chinese comprises about 1 billion Chinese characters; as most Chinese *words* found in dictionaries consist of 2 characters, and some of 3 characters or more, we can estimate about 300 million words in the reference corpus. The corpus was indexed using the *Lemur toolkit* mentioned in section 2.

As we wanted to test the validity of a general purpose statistical score designed for MWE extraction at various levels (from bigrams to 12-grams), we implemented the *cpr-score* on the indexed corpus, by means of a Perl script. As there are some limitations inherent to very frequent records on a query likelihood model, all request yielding the maximum frequency of 50,000 were treated by another section of the script, in which a regular expression implemented the *cpr-score* on a non-indexed version of the same corpus. The average processing time for every request was 0.07 second on the indexed corpus, and 1.5 second on the non-indexed corpus (running on a pc with Linux).

In order to measure the performance of the *cpr-score* for CWS, we used the well-known MSR dataset, from the second *International Chinese Word Segmentation Bakeoff* (Emerson, 2005). For computing recall, precision and F-score of the segmented text, we used the standard scoring program (Perl script) provided by the Bakeoff.[4]

As in the case of the *Tightness Continuum* (Xu et al., 2010), the methodology for segmenting the input text was purely statistical, and used no list of training words or dictionaries of any kind. It should be stressed that such a methodology is purely data-driven, and rests upon the theoretical assumption that language structure itself includes recurrent patterns of meaning that can be captured by an algorithm, with no human intervention or any decision based on linguistic norm or culture.

Our computer program implementing the *cpr-score* proceeds as follows. Each Chinese character (han) is added one at a time, and the score is computed on the reference corpus. Let us take a simple example: the Chinese word[5] 高等教育 (*gāoděng jiàoyù*, higher education). Our algorithm first considers the bigram 高等 and checks its *cpr-score* on the reference corpus: 0.64. The *cpr-score* ranges from 0 to 1, and the high significance threshold has been experimentally set at 0.40 (Colson, 2017). Then, the third gram is added, 教, and the score for the trigram 高等教 is measured: 0.84. As the score is going up, the trigram is left unsegmented. Finally, the last gram 育 is added, and the score for the fourgram 高等教育 is computed, which yields 0.90. Again, the score is going up, so that the whole fourgram is left unsegmented.

[3] https://www.sketchengine.eu

[4] http://sighan.cs.uchicago.edu/bakeoff2005/

[5] This 4-gram is considered as two words by Google Translate (https://translate.google.com) but as one word by the MSR gold standard (http://sighan.cs.uchicago.edu/bakeoff2005/)

3.2 Results and discussion

Table 1 presents the results obtained by our experimental segmenter based on the *cpr-score* (Seg-cpr) and by a state-of-the-art segmenter, the Stanford segmenter[6], for the MSR dataset.

MSR dataset	Recall	Precision	F measure
Seg-cpr	0.749	0.658	0.700
Stanford-segmenter	0.882	0.843	0.862

Table 1: Results of CWS by means of Seg-cpr and Stanford-segmenter.

As shown in Table 1, the results obtained by our experimental segmenter based on the *cpr-score* are obviously less good than those of the Stanford segmenter, but this hardly comes as a surprise, as the *cpr-score* **was not designed for CWS** in the first place. We have also stressed in section 3.1. that our methodology, contrary to state-of-the-art segmenters such as the Stanford segmenter, does not rely on segmented corpora or on dictionaries, but only on statistical attraction as measured by the *cpr-score*. Contrary to most segmenters, it is not a mirror of how language users tend to segment the language, but of how **the language itself** contains statistically significant elements of meaning.

It is besides quite striking that the recall rate obtained by Seg-cpr (0.749) comes very close to the average rate of segmentation agreement among native speakers of Chinese (0.75, as mentioned in section 1). Contrary to manually segmented corpora, or to segmenters based on dictionary learning or segmentation pattern learning, our results are objectively measured by the algorithm on an *unsegmented* reference corpus. For this reason alone, a recall of 0.749 computed from the gold standard established by Chinese native speakers is quite high.

A fine-tuned analysis makes these results even more intriguing. In 5 to 10 percent of the cases, wrong segmentation by Seg-cpr was simply due to the fact that the n-gram was not used a single time on the reference corpus of about 250-300 million words. As the *cpr-score*, on which the tool is based, requires a frequency of at least 3 occurrences, the absence of an n-gram / word from the reference corpus inevitably leads to wrong segmentation, but this is to be blamed on the corpus size, not on the algorithm.

Taking a closer look at cases of obviously wrong segmentation by Seg-cpr raises other intriguing questions. One has to do with **discontinuous sequences**, a central issue in construction grammar as well. As a matter of fact, another 5 to 10 percent of the instances of wrong segmentation by Seg-cpr is due to discontiuous statistical association. Let us take the example of a Chinese fivegram from the MSR dataset, considered as one word by the gold standard, 个人计算机 (*gèrén jìsuànjī*, personal computer). Table 2 shows the *cpr-score* and the frequency of the different levels of grams in our reference corpus.

	Cpr-score	**Frequency**
个人	0.63	97,167
个人计	0.18	171
个人计算	0.55	140
个人计算机	0.73	122

Table 2: *cpr-score* and frequency of the component grams of 个人计算机（*gèrén jìsuànjī*).

As shown in table 2, the fivegram 个人计算机 (*gèrén jìsuànjī*, personal computer), is identified as a whole as a very significant statistical association (*cpr-score* > 0.40), but working with one gram (in this case, a Chinese han) at a time reveals that the score goes down at the level of the trigram, and then up again. This is a clear example of a **discontinuous association** between successive Chinese characters, and is by no means an exception. The same situation occurs within several Chinese idioms in the source text, e.g. 付之东流 (*fùzhīdōngliú*, to lose sth irrevocably), and this also holds true of many foreign words that are transliterated into Chinese, e.g. 马克思主义 (*mǎkèsīzhǔyì*, Marxism) or 卡斯帕罗夫 (*kǎsīpàluōfū*, Kasparov). In all those cases, our experimental methodology worked gram per gram, and the *cpr-score* was therefore unable to segment correctly. A further elaboration of the methodology should

[6] The version used here is stanford-segmenter 3.8.0 (https://nlp.stanford.edu/software/segmenter.html)

address this complex issue. Results such as these are actually a confirmation of the global statistical association, as measured by the *cpr-score*, between elements of meaning in Chinese (words, collocations, idioms). They also mean that the results of our experimental Seg-cpr tool (with already a recall of 0.749 for the MSR dataset) could be further improved by introducing a more complex algorithm taking discontinuous cases into consideration. It should further be pointed out that similar cases of discontinuous association measured by the *cpr-score* have been noted for idiomatic constructions in English (Colson, 2017), e.g. *long time no see* or *the next thing I knew*.

All in all, the results of this experiment confirm our hypothesis that MWE extraction and CWS are closely related. The *cpr-score* was designed in the first place for MWE extraction, and yields convincing results for English, Spanish and French. In this experiment, we have used it for Chinese segmentation in a simplistic way, by adding one gram at a time. Even then, the overall recall rate is pretty high (0.749) and reaches the average rate of agreement between Chinese native speakers. Besides, a closer analysis reveals that taking discontinuous association into account would further increase recall and precision. From a theoretical point of view, such a complex network of probabilistic associations is quite compatible with construction grammar. The interesting cases of discontinuous associations may even provide us with some clues about the possible extraction of more complex constructions, as we will see in the following section.

3.3 Clues as to automatic extraction of constructions

As stated above, a statistical extraction method that is fully compatible with construction grammar should be able to deal with all constructions: lexical constructions (as in the case of Chinese word segmentation), idiomatic ones (e.g. MWEs), but also constructions with more schematic slots (e.g. *X take Y into account*), and maybe even abstract constructions such as the ditransitive construction.

Our clustering method (the *cpr-score*) already yields promising results for CWS and MWE extraction, but we may wish to test it further on more schematic or abstract constructions. This may indeed be beneficial to the improvement of grammatical material selection in language teaching, and may contribute to providing more evidence for the statistical grounding of construction grammar.

As a simple clustering algorithm, the *cpr-score* is non-parametric and can therefore be easily extended to longer sequences. Besides, it allows for complex implementations in databases using a query likelihood model, but also very simple ones in the form of regular expressions (*regexes*). The recourse to complex regexes makes it possible to check the *cpr-score* for schematic or abstract constructions, provided that the corpus annotation contains information on the construction under investigation.

The crux of the matter is indeed to extract constructions from corpora containing sufficient annotation techniques. As stated in section 1, a corpus used for the extraction of complex constructions should ideally include information related to intonation. In the meantime, using large POS-tagged corpora already provides us with a lot of testable material with respect to schematic constructions. Let us start from the fairly simple construction *the more… the more*. If construction grammar is right, corpora should be able to reveal that there is statistical association between them, even though the length of the window may vary. We may easily test it with the *cpr-score* by choosing a window of 8 words between the two parts of the construction, and by using our experimental program *Construction Extractor*, based on regexes.[7] In this case, the *cpr-score* obtained with a randomly selected portion (200 million tokens) of the ukWac corpus (Baroni et al., 2009) reaches 0.64 for a frequency of 1332: as might have been expected, there is indeed a measurable association between *the more* and *the more*, even with as many as 8 words between them.

Our aim was to test more complex constructions on a tagged corpus of about 100 million tokens. For this purpose, we used another randomly selected portion of the ukWac corpus (Baroni et al., 2009), and we had recourse to the Stanford POS tagger[8] for tagging it. According to CxG, the probabilistic network of constructions is valid at various levels of abstraction and schematicity. As a matter of fact, part of that complex interplay between morpho-syntactic features can easily be captured by considering the tagged

[7] The simple implementation of the *cpr-score* by means of regexes involves a division between resp. the frequency with the smaller window (in this case, 8 words) and the larger window (which we set on the basis of previous experiments at 10 times the smaller window). In Perl syntax, the regex for the frequency with the larger window may simply look like this: /the\smore\s(\S+\s){0,80}the\smore/i

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

corpus as a geometrical space in which metric clustering can be measured. In other words, the statistical clustering algorithm (in this case the *cpr-score*) will just be looking for 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.[9]

Table 3 displays the *cpr-score* and the frequency for the MWE *it is amazing what* on the 100 million word part of the ukWac corpus, tagged by the Stanford POS tagger, given a maximal window of 4 words before and after the adjective *amazing*. 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. The lowest significance threshold of the *cpr-score* has been (experimentally) set at 0.065, so that a score of 0.12 is sufficient to detect such a degree of association prevailing within more schematic constructions.

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

Table 4 displays in the first line an approximation of the ditransitive construction in the tagged corpus, just taking into account nouns followed by a verb, followed by two nouns, with windows of 5 words in both cases. Even at this level of abstraction, it is noteworthy that the *cpr-score* implemented by a simple regex is able to measure some statistical attraction. The same holds true of the All-cleft construction in the second line of table 4. In this case, we restricted the search to the presence of two verbs in the past within a window of seven words. The regex in footnote 11 was of course checked for its validity, and it yields sentences (with the Stanford POS tags) such as *All DT he PRP did VBD the DT whole JJ time NN was VBD tell VB me PRP about IN*, or *All DT he PRP had VBD really RB expected VBN was VBD now RB propped VBN up RP on IN his PRP$ bedside NN table NN.* As predicted by CxG, there is indeed a **measurable statistical attraction in the very structure of the All-cleft itself**, as the *cpr-score* reaches a significant level of 0.29.

Our preliminary research yields similar levels of attraction for other cases of schematic constructions. For instance, the Verb Object Prep construction (as in *take a lot of effort*) yields a *cpr-score* of 0.31 for a frequency of 212 001, with a window of 3 words; similarly, the As-Noun comparison construction (e.g. *as bright as stars*) displays a score of 0.53 for a frequency of 429, with a window of 2 words.

[9] The output of the Stanford POS tagger was adapted by a Perl script, replacing the underscore signs by blanks, so that a simple regex could be used for measuring the *cpr-score*: /it\sPRP\sis\sVBZ\s(\S+\s){0,40}\S+\sJJ\s(\S+\s){0,40}what\sWP/i

[10] Regex used: /NN\s\S+\sV\S*\s(\S+\s){0,50}NN\s(\S+\s){0,50}NN/

[11] Regex used: /all\sDT\S+\sPRP\s(\S+\s){0,70}VBD\s(\S+\s){0,70}VBD/i

4 Conclusions

Starting from CxG's claim that there is a cline from syntax to lexicon and a complex network of constructions in language, at various levels of abstraction and schematicity, we have performed a first experiment on Chinese Word Segmentation. Algorithms used in CWS are usually trained on hand-annotated data, and are therefore a reflection of culture and tradition. However, we wanted to test to what extent an algorithm (the *cpr*-score) used for MWE extraction would yield results for CWS. For the reference text used, our algorithm reached a recall of 0.749 measured automatically from a gold standard established by native speakers. This may hardly be due to chance, as our segmentation method implied a binary choice at every single Chinese character. Besides, our recall score reaches the average degree of agreement between native speakers of Chinese. An analysis of the wrong cases of segmentation reveals that a discontinuous methodology may still improve the overall score on the basis of the same algorithm.

Our aim was not to provide a better segmenter for Chinese, because state-of-the-art tools trained on annotated data will inevitably reach higher scores measured on the same type of annotated data. We just wanted to test the hypothesis that CWS displays many similarities with MWE. The fact that a simple implementation of the *cpr-score*, designed in the first place for MWE extraction in European languages, reaches acceptable rates for CWS is a striking conclusion, that seems only compatible with one of the tenets of CxG: words are expressions and vice versa, as all language structure is just a network of constructions.

Building on these findings, we carried out a second experiment devoted to the extraction of more schematic or abstract constructions. Our preliminary results suggest that what is valid at the level of words and expressions will also be applicable to more schematic levels, so that the *cpr-score* or other clustering algorithms may be used for identifying constructions. The next application of this methodology may be the automatic extraction of the most fixed and recurrent schematic / partly schematic / idiomatic / abstract contexts of frequent verbs or nouns, based on the same algorithm.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press /Addison Wesley, New York.

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

Benedetta Bassetti. 2005. Effects of writing systems on second language awareness: Word awareness in English learners of Chinese as a foreign language. In Vivian Cook and Benedetta Bassetti (eds.), *Second Language Writing Systems*. Multilingual Matters, Clevedon: 335–356.

Geert Booij. 2013. Morphology in Construction Grammar. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork: 255–273.

Deng Cai and Hai Zhao. 2016. Neural Word Segmentation Learning for Chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin: 409–420.

Jean-Pierre Colson. 2017. The IdiomSearch Experiment: Extracting Phraseology from a Probabilistic Network of Constructions. In Ruslan Mitkov (ed.), *Computational and Corpus-based phraseology, Lecture Notes in Artificial Intelligence 10596*. Springer International Publishing, Cham: 16–28.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

William Croft. 2013. Radical Construction Grammar. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork: 211–232.

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

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

Charles Fillmore. 1988. The Mechanisms of Construction Grammar. *Berkeley Linguistic Society*, 14: 35–55.

Adele Goldberg. 2003. Constructions. A New Theoretical Approach to Language. *Trends in Cognitive Sciences*, 7(3): 219–224.

Adele Goldberg. 2006. *Constructions at Work*. Oxford University Press, Oxford.

Adele Goldberg. 2013. Constructionist Approaches. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork: 15–31.

Stefan Th. Gries. 2013. 50-something years of work on collocations. What is or should be next … *International Journal of Corpus Linguistics* 18: 137–165.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129.

Thomas Hoffmann and Graeme Trousdale (eds.). 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Diana McCarthy, Bill Keller and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEDX on Multiword Expressions*: 73-80.

Joaquim Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Noncontiguous Multiword Lexical Units. In *Proceedings of 9th Portuguese Conference in Artificial Intelligence* (EPIA 1999): 849.

Richard Sproat, Chilin Shih, William Gale and Nancy Chang. 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics* 22(3): 377–404.

Anatol Stefanowitsch. 2013. Collostructional analysis. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork: 290–306.

Alison Wray. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press, Oxford.

Stefanie Wulff. 2013. Words and idioms. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford/NewYork: 274–28.

Ying Xu, Christoph Ringlstetter and Randy Goebel 2009. A Continuum-based Approach for Tightness Analysis of Chinese Semantic Units. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*: 569–578.

Ying Xu, Randy Goebel, Christoph Ringlstetter and Grzegorz Kondrak. 2010. Application of the Tightness Continuum Measure to Chinese Information Retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Coling 2010, Beijing: 54–62.

Xiaodong Zeng, Lidia S. Chao, Derek F. Wong, Isabel Trancoso and Liang Tian. 2014. Toward Better ChineseWord Segmentation for SMT via Bilingual Constraints. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore:1360–1369.