Contextual fingerprints of Czech and English verbs

Lucie Chlumska
Institute of the Czech National Corpus,
Faculty of Arts,
Charles University in Prague

1. Introduction

One of the opportunities the corpus linguistics has given us is to explore words and phrases within their contexts. Thanks to large corpora, linguists have, for the first time, a chance to look at the language from the syntagmatic point of view as opposed to the traditional paradigmatic focus of linguistics. Contextual research may be used in different ways. Context has always played a crucial role in translation and translation-oriented research or contrastive linguistic studies, but it can also shed light on grammatical, lexical and semantic properties of certain word groups or parts of speech.

The focus of this paper is to apply this corpus-driven contextual approach in the study of frequent Czech and English verbs and to compare and interpret the possible contextual patterns while bearing in mind the typological difference of these two languages. While English is an isolating language with a fixed word order, Czech is largely influenced by its rich inflection, which causes certain flexibility of the word order. The research is based on Cvrcek's concept of measuring lexical variability[1] and it will comprise several thousand Czech and English verbs found in the examined corpora.

2. Contextual variability

When examining the context of a lexical unit, first we have to determine what type of context is to be analyzed. In this particular case, the immediate context, i.e. three positions on both sides of the KWIC, will be researched, the assumption being that it is usually the most influential as well as the most appropriate for the purposes of this research. For example in phraseology, such context would most probably be insufficient.

It is also necessary to define the term "contextual variability". In the broadest sense, it may refer to any variability of the context and its constituents, be it semantic, syntactic or morphological. However, in this paper this term is used to describe only the lexical variability manifested in the number of alternate word forms in the immediate surroundings of the key word in context (KWIC).

[1] Cvrcek, V.: Contextual Approach to Parts of Speech. In *InterCorp: Exploring a Multilingual Corpus*. NLN Praha, 2010

The hypothesis is that the fewer different word types occur in the given position, the greater is the influence of the KWIC on the position, and thus the more information about the KWIC can be found by examining it. In other words, the difference in variability of certain positions shows us which positions are more important for the KWIC and its meaning, function or valency than others.

2.1 Definition of a contextual fingerprint

The basic assumption is that every word has its contextual setting or sum of contexts, which can be referred to as "the contextual fingerprint" of a word (Cvrcek 2010: 191). Consequently, all semantic, syntactic and formal properties of a word are reflected in this contextual fingerprint. Contextual fingerprints are unique for each and every word. The term itself is rather general; in this case, only the lexical part of a contextual fingerprint will be taken into account.

2.2 How to measure lexical variability

As mentioned above, lexical variability of the immediate context is considered a part of the contextual fingerprint of a given word. There is a difference between the traditional linguistics and the corpus approach to contextual variability. The traditional point of view has been influenced by the so-called "free combination of lexical units" and it presents the context of a word as an infinitely recursively branching tree (Cvrcek 2010: 192). It means that the more distant the position from the KWIC is, the higher contextual variability (i.e., theoretically, infinitely many alternate word forms) we have to expect there. In a corpus, however, the reality is rather different. The maximum variability of a certain position, no matter how distant, is limited by the number of instances of the KWIC. For example, a word with a frequency of 1 000 in a corpus can have a maximum variability of 1 000 alternate word forms, in any position. However, Cvrcek found out that the maximum variability of distant context (e.g. position 30) does *not* tend to be even roughly equal to the frequency of KWIC. Let us take the example of the word *large,* which has a frequency in the BNC of 33 038 instances, while the tenth position on either side of the KWIC has a variability of only 8,000 types.

The question is how to compare the contextual variability of words with different frequencies. For this reason, Cvrcek introduces a value – probable distant context variability (PDCV) – to which the variability in distant context converges. This value is, according to his research, equal to the number of types in a hypothetical subcorpus of length equal to the frequency of KWIC created randomly from the tokens of a source corpus. For example, we know that the word *large* has a frequency of 33,038 instances in the BNC. A subcorpus of the length of 33,038 tokens created by random picking of tokens from the BNC would yield about 8,111 types. This would be the PDCV value for the English word *large*. The ratio between observed variability in a certain position and PDCV is relative context variability (RCV) and this value can be used for comparison of words with different frequencies.

2.3 Contextual patterns

When we look at the examined verbs, we can calculate a unique PDCV, and then RCV (as a percentage) for each one of them so that they may be compared. For each position (from -3 to +3) we will get a percentage (RCV) indicating lexical variability, i.e. how many alternate word forms actually occurred in the corpus in the certain position compared to the PDCV. We can then sort the examined positions according to the RCV and order them from 1 to 3, from the least variable (1) to the most variable (3). The average contextual pattern is 321 KWIC 123, i.e. the first positions around the KWIC are the least variable ones etc.

LEMMA	-3	-2	-1	KWIC	1	2	3	LEFT	RIGHT
heal	83,0	73,2	40,2	0	37,8	75,1	83,4	321	123
check	83,7	82,0	31,0	0	25,3	81,6	82,4	221	122
grasp	83,5	73,7	36,4	0	26,6	88,5	67,5	321	132

Table 1: Example of the contextual patterns (LEFT and RIGHT)

Table 1 provides an example of three English verbs. The number from -3 to +3 indicates the particular position and the figure below is the RCV (in percentage). In the first line, the verb *heal* has an average contextual patterns on both sides of the KWIC (321 – 123), whereas the third verb *grasp* has a different right pattern – 132, i.e. the second position is more variable than the third one. The difference between the percentages has to be bigger than 5%, otherwise the position is assigned the same number (see the verb *check* with a pattern 221 – 122). This setting can of course be changed; it is possible to modify the borderline to find out what difference is significant.

3. Corpora used in the research

In order to be able to compare Czech and English, two corpora of exactly the same size have been used, SYN2010 for Czech and BNC for English. Although the BNC was published almost 20 years earlier, for the purposes of this research they will be considered comparable.

3.1 SYN2010

SYN2010 is a synchronic representative corpus of written Czech developed at the Institute of the Czech National Corpus in Prague and published in 2010. It comprises 100 million tokens (without punctuation). It is a balanced corpus – it contains a proportionate number of newspaper and magazine texts, fiction texts and professional literature. All newspaper and magazine texts included into SYN2010 were published in 2005–2009, each year being equally represented. SYN2010 corpus includes professional texts published after 1989. Some of the fiction texts may have been published earlier, but there is a general rule that the corpus consists mainly of newer texts, whereas the proportion of older texts is decreasing. SYN2010 is lemmatized and morphologically tagged using the latest available tools.

3.2 BNC

BNC is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. In this research, the BNC-XML edition has been used, with the university license (for Charles University in Prague).

4. Research data

The PDCV and RCV values have been calculated for *lemmas* of the given verbs, whereas the values indicate the number of alternate *word forms* in the given position.

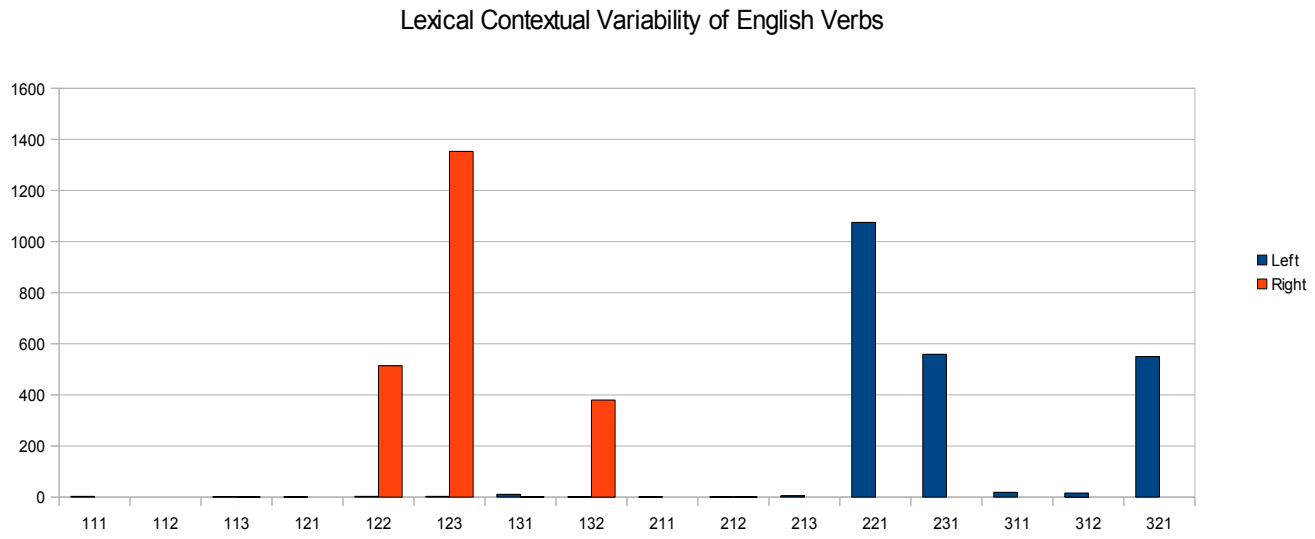
4.1 Verbs

All the verbs in the examined corpora have been used in the research with the exception of those containing non-alphabetic characters (e.g. *re-edit* etc.). The total number of 4 464 Czech verbs and 2 250 English verbs were analyzed. The significant difference in the number of verbs may be attributed to a different typology of Czech. Czech is a flective language using many affixes, which is reflected in the higher number of verbs, whereas English as an isolating language uses different means to convey shades of meaning (e.g. prepositions or adverbs).

Different typology of those two languages can also be seen in the word order. While English has a fixed word order, Czech is said to have a relatively free word order, i.e. there are no strict guidelines as to where certain parts of speech have to stand in a sentence. Naturally, there are rules for an unmarked word order, but these may often be broken (especially in spoken language) for the sake of emphasis etc. We may assume that the different word order would be somehow reflected in the lexical variability.

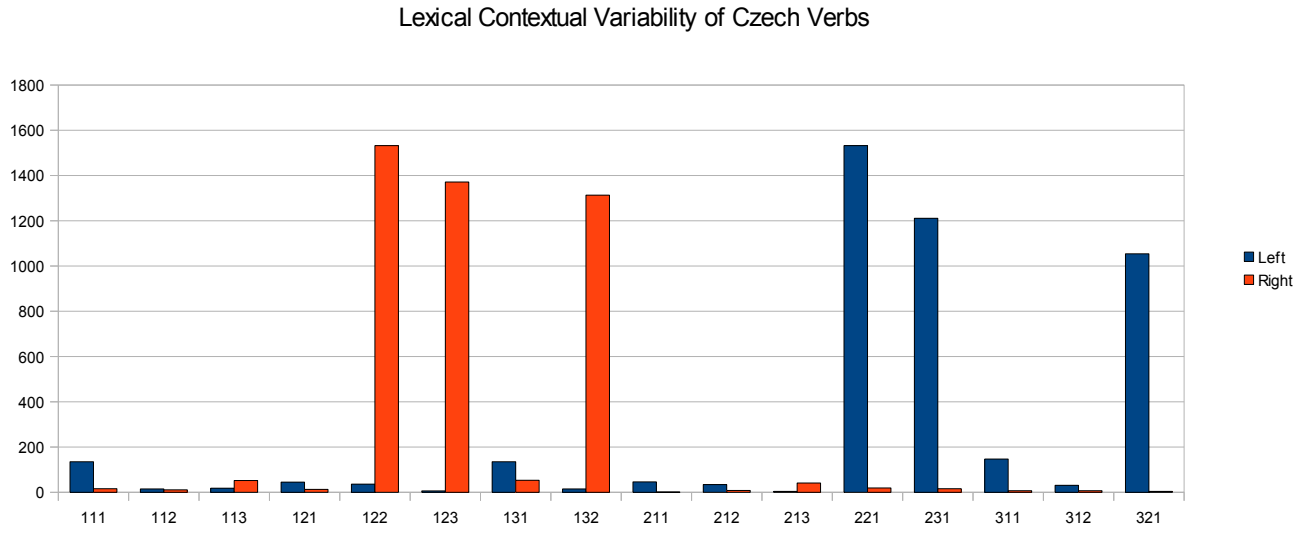
4.2 Similarities and differences

First, let us look at the most common contextual patterns in both languages. Graph 1 shows the most frequent patterns in English; graph 2 is for Czech.



Graph 1: The most frequent contextual patterns of English verbs (based on the BNC-XML)

It is obvious that the most common patterns for English are 221, 231, 321 on the left, and 123, 122, and 132 on the right. They constitute almost 97 % of the left contextual patterns and 99,8 % of the right ones. The remaining patterns seem to be insignificant.



Graph 2: The most frequent contextual patterns of Czech verbs (based on the SYN2010)

The most frequent contextual patterns shown in graph 2 are 221, 231, 321 on the left, and 122, 123, 132 on the right. They form more than 85 % of the left context and more than 94 % of the right context.

Both graphs suggest that English and Czech share the most common contextual patterns. In both languages, the three groups on each side form a vast majority of context types – the most typical and frequent. In Czech, there are more different contextual types (probably the influence of the "free" word order), but the differences are not as significant as we might expect. Verbs of minor contextual patterns (esp. in Czech) should be analyzed and described separately (e.g. 111-131> zvýhodňovat "favour sb", převýšit "exceed"etc., 231-131> rival, 131-123 > fluctuate) as we may assume that they relate to strong valency, special use, collocations etc.

Another interesting finding is that the groups on the left do not match predominantly with a particular group on the right and vice versa. It is thus wise to analyze the left and right patterns separately. There seem to be no typical left and right patterns. It only supports the general, well-established idea that the right context is different from the left one.

Within this research, it is also possible to identify values like MAX, MIN, and AVERAGE etc., which refer to the maximum, minimal and average variability and their correlation with a particular position (out of the 6 examined positions). For example, the position with minimal variability (MIN) is the same for both languages: it is the position +1. When we look at the verbs with the smallest MIN value, we will find verbs with strong valency (*tend, ought, happen*), verbs always followed by a preposition (*rely, consist*) or a subordinate clause (*vědět* "know"), or reflexive verbs – reflexiva tantum – in Czech (*odmlčet* "stop talking", *zamračit* "frown"). As expected, the maximum variability value (MAX) correlates most with the distant positions: +3 and -3 in Czech and -3, -2 for English.

The values could be further analyzed. For more precise interpretation, it would certainly be wise to compare the correlation for left and right context alone. It may provide a more subtle distinction and point out to interesting verb groups.

5. Possibilities for further research

This research indicates that the contextual variability may be a useful source of information for collocation retrieval. The least variable position is likely to be the most interesting and influential and therefore the most suitable for examining in connection with collocations.

It also provides possibilities for another verb classification. Lexical variability can be seen as a criterion for a more subtle distinction of verb groups (taking into account the role of text types etc.).

It may also serve as a basis for further comparison of contextual patterns in other languages, e.g. the Slavonic languages.

6. Conclusion

This research on lexical variability shows that Czech and English verbs share the same frequent contextual patterns. The three most common groups on each side form a vast majority of context types – the most typical and frequent. Every deviation from the typical contextual patterns indicates that the verb is likely to be somehow special, be it from the morphological, syntactical or collocational point of view.

The left and right contexts are rather independent and it is useful to analyze them separately. Finally, contextual information represents a valuable source of corpus-driven information for further research.

7. References

Cvrček, V.: Contextual Approach to Parts of Speech. In *InterCorp: Exploring a Multilingual Corpus*. NLN Praha, 2010

Czech National Corpus - SYN2010. Institute of the Czech National Corpus, Praha 2010. Accessible at WWW: <http://www.korpus.cz>.

BNC-XML edition (university license)