Corpus-based Analysis of Czech Units Expressing Mental States and Their Polish Equivalents. Identification of Meaning and Establishing Polish Equivalents Referring to Different Theories (Valency, Pattern Grammar, Cognitive Linguistics)

Elżbieta Kaczmarska
Instytut Slawistyki Zachodniej i Południowej
Wydział Polonistyki
Uniwersytet Warszawski

Abstract: The analysis is focused on Czech polysemous verbs expressing mental states. We test if different linguistic theories can lead to the closest equivalents of these verbs in Polish. The analysis proper is preceded by automatic extraction of pairs of equivalents from the parallel corpus InterCorp. These pairs constitute a kind of bilingual dictionary. The research includes automatic excerption of chosen verbs (with aligned segments). The segments are analysed manually. We check how the key verb was translated and what kinds of collocations and arguments it has. Subsequently we apply methods specific for each linguistic theory. The analysis is complemented by collocation profiles, obtained by Word Sketches that proved to be an essential tool. The analysis shows the most effective theories but also shortcomings of the proposed algorithm to spur its optimization and development..

Keywords: equivalent; psych verbs; parallel corpora, Czech, Polish

1. Introduction

Czech and Polish come from the same West Slavic language group, a part of the Slavic language family. They both share a common ancestry and their vocabularies show many similarities. Many words sound nearly the same and native speakers of both languages are able to understand them.[1] However, there are some surprising lexical contrasts between Czech and Polish, causing difficulties in language contacts. A case in point are verbs expressing mental states[2] and nouns denoting emotions and feelings, e.g.

[1] We also face the prevalent phenomenon of false friends: *sukienka* (pl) 'dress' – *sukýnka* (cs) 'short skirt'.

[2] The class of mental verbs (also known as psych verbs) includes verbs of perception, cognition and emotion (Pustejovsky 1993). In this paper, psych verbs are only an example of a group of verbs causing particular problems in the process of translation. The paper will neither analyse nor describe the nature of the psych verbs themselves. They are a relatively well-studied phenomenon in linguistics [cf., e.g., the works of Adriana Belletti & Luigi Rizzi, Agnieszka Będkowska-Kopczyk, Stefan Engelberg

mít rád (to like, to love), *být líto* (to regret, to be sorry), *toužit* (to miss, to want, to desire). In some cases it is impossible to reproduce a text (or a phrase) in the target language; a native speaker of Polish is not able to encode the meaning of such words and confronts the problem of the impossibility to express the same content in their own language which does not provide the same concept (Kaczmarska and Rosen 2014a). Instead, one can indicate a cluster of equivalents with a slight change of meaning (Lewandowska-Tomaszczyk 1984, 2013), but none of them will cover exactly the same semantic field.

Verbs expressing mental states are, in this sense, particularly problematic because of their ambiguity and subjective character. This is why parts of their meaning are lost in the act of translation and one-to-one equivalent pairs are difficult to find. It must be admitted that some of these verbs are not considered polysemous by Czechs, while they seem to have several meanings for a speaker of Polish.[3] Regarding verbs expressing mental states, one also encounters the problem of misunderstanding. In some cases it is not possible to state what a given unit means in terms of the target language.

The problems are not solved by traditional dictionaries, which provide only a limited number of equivalents (in most cases with no examples of usage). The best-known Czech-Polish dictionary (Siatkowski and Basaj 2002) presents a list of equivalents of the analyzed units, e.g.:

- *mít rád – lubić* (to like)*, kochać* (to love)
- *být líto – być żal* (to be sorry for)*, być przykro* (to feel sorry), *szkoda* (to sorry)
- *toužit – tęsknić* (to miss)*, pragnąć* (to desire)*, marzyć* (to dream).

Without additional clues, it is not possible to translate these into Polish properly. In some cases, even the context is not conclusive; e.g. the avowal *mám Tě rád* can be translated as *lubię cię* (I like you) or *kocham cię* (I love you). For a Polish speaker the unit *mít rád* has at least two quite different meanings: *kochać* (to love) and *lubić* (to like).

Martina Hřebíčková, Barbara Lewandowska-Tomaszczyk (also with Paul A. Wilson), Karel Pala & Zdena Šachová, Bożena Rozwadowska, Alexandros Tantos, Irena Vaňková, Anna Wierzbicka].

[3] In more distant languages, this is a common situation. Languages pattern the meaning in their own way.

2. Goals

The objective of the study is to find a suitable equivalent for a given unit (psych verbs) and to present an attempt at applying methods of different linguistic approaches to build an algorithm for the selection of equivalents. Our research is based on texts from the parallel corpus InterCorp (Čermák and Rosen 2012; Kaczmarska and Rosen 2014b). The experimental algorithm will involve several steps corresponding to different grammatical theories. Depending on a given verb, we can find the best equivalent at each stage.

3. Algorithm assisting in the selection of equivalents

Determining Polish equivalents for Czech verbs is a part of a larger project; within its frameworks we elaborate an algorithm assisting in the selection of equivalents of verbs, exploiting data from the parallel corpus InterCorp (Kaczmarska 2015b). By applying methods based on various linguistic theories (see. Fig. 1) we test if the meaning of a given verb depends on its syntactic characteristic (Levin 1993) and then we establish the closest equivalent based on the syntactic patterns (referring also to the syntactic structure of the potential equivalent).

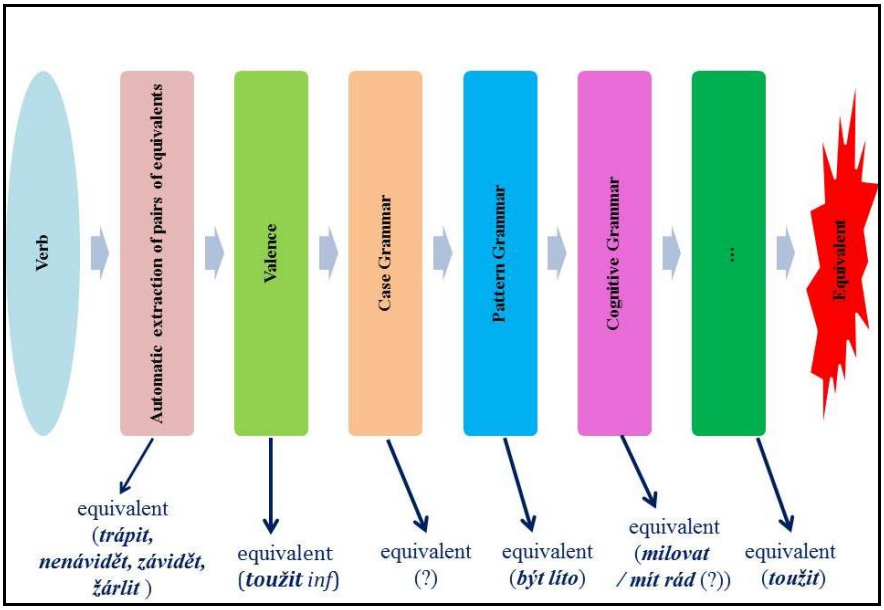

Figure 1. Steps of the algorithm for identifying equivalents

As shown, the algorithm consists of several steps. The verb, subjected to such a test, does not have to go through all the stages; the optimal equivalent can be found at each step of the analysis.

3.1 Step one – automatic extraction of pairs of equivalents

In the pilot study (Kaczmarska and Rosen 2013) we extracted pairs of equivalents from the parallel corpus InterCorp, obtaining a bilingual glossary (Och and Ney 2003; Skoumalová 2008; Jirásek 2011). At this stage we compiled clusters of Polish equivalents for each individual Czech verb. The table below (1) presents a cluster of the most frequent equivalents for the study case – *toužit*.

Equivalent of *toužit*	
pragnąć	304
chcieć	107
tęsknić	82
marzyć	70
pożądać	40
ochota	24
zapragnąć	9
pragnienie	8
tęsknota	8
zależeć	8
spragniony	7
życzyć	6
Overall	673

Table 1. Polish equivalents of *toužit*

Pilot studies have shown that for many verbs suitable equivalents can be found already at this stage; interestingly, this generally concerns verbs expressing negative feelings (Kaczmarska 2015d; see also 4.1.).

3.2 Step two – valence analysis

The aim of the analysis at this stage is to answer whether valence requirements[4] can help to identify Polish equivalents of the verbs. We

[4] In this paper, the valence is understood as a linguistic phenomenon referring to the number of arguments controlled by a verbal predicate. We define syntactic and semantic

assumed that in cases concerning some meanings the equivalent could be established on the basis of the convergence of valence requirements (Levin 1993). For this purpose, we conducted a manual analysis of the selected verb (with aligned segments) from InterCorp (Kaczmarska and Rosen 2013). We established (in each segment) how the verb was translated, how many arguments the given verb binds, what kinds of arguments they are (whether it is, e.g., a noun – if so, what kind of entity it denotes: real, abstract, human being, etc.) and how they are bound (by morphological case, preposition, infinitive, relative clause). We also checked which arguments are bound to the Polish equivalents and conducted syntactic and semantic analyses of the arguments bound by the Czech verb and by its equivalents. We expected that in most cases the results of the syntactic and semantic analyses should allow for establishing the semantically and syntactically closest Polish equivalents of the Czech verbal unit.

The verb *toužit* binds arguments using structures such as:

- *toužit po* Oabstr (abstract object)
- *toužit po* Ohum (human object)
- *toužit po / do* OR (real object)
- *toužit* + *inf*
- *toužit* + S (*aby… / po tom, aby…*)[5]

Thanks to the manual analysis of the aligned segments we could see how the verb is translated in each particular group. The study revealed that valency can influence the choice of equivalent in Polish. In the case of the analyzed verb, we concluded that the most appropriate equivalent can be identified only for *toužit* bound with an infinitive.

toužit + inf → *pragnąć*[6] + inf

Equivalent of *toužit* + infinitive	
pragnąć inf	44
chcieć inf	20
marzyć o Oabstr	4

features of analysed verbs at later stages of our research (Dębski 1982; Daneš and Hlavsa 1987; Rytel 1989; Greń and Rytel-Kuc 1991; Čermáková 2009; Urbańczyk-Adach 2001).

[5] *toužit* + sentence (*to…*)

[6] The verb *chcieć* (to want) is classified as a synonym of *pragnąć* (to desire). The difference between them lies in the intensity of the feeling.

pragnąć Oabstr	3
być pragnieniem inf	1
chętnie + S	1
mieć marzenie inf	1
mieć ochotę inf	1
pragnąć + S	1
tęsknić za (+ S)	1
zachciewać się Oabstr	1
Other	2
Overall	80

Table 2. Polish equivalents of *toužit* + infinitive

In the remaining groups the results were not decisive.[7] All the units which did not find the proper equivalent at step one would be automatically moved to the next stage of the algorithm.

3.3 Step three – Case Grammar

At this stage we identify cases – the roles played by elements bound to the verb (Fillmore 1968; Halliday 1985; Korytkowska 1984, 1992, 1993; Kaczmarska 2001). However, this step was ineffective for this particular group of verbs. The analyzed units (expressing different emotions and

[7] Complete figures were published in Kaczmarska and Rosen (2013). Translating structures with Oabstr was particularly problematic and pointed to the need for a deeper analysis of the objects. As a test, we examined two abstract objects – "big love" (*velká láska / wielka miłość*) and "exotic journey" (*exotická cesta / egzotyczna podróż*). We discovered that both are easily combined with the Czech analyzed verb: *toužit po velké lásce / exotické cestě*. We also tried to combine them with three most frequent Polish equivalents:

marzyć	*o*	*wielkiej*	*miłości*	*/*	*egzotycznej*	*podróży*
to dream	of	big	love	/	exotic	journey

tęsknić.za	*wielką*	*miłością*	*/*	*egzotyczną*	*podróżą (???)*
to miss	big	love	/	exotic	journey

pragnąć	*wielkiej*	*miłości*	*/*	*egzotycznej*	*podróży (?)*
to desire	big	love	/	exotic	journey

It is possible to combine the objects with the verb *marzyć* (to dream), but it is not correct to use them with the verb *tęsknić* (to miss). It would be acceptable only if the "big love" represented a person. Also, the verb *pragnąć* (to desire) hardly allows combination with the object of "exotic journey". The ambiguities of the results from the first step made us pursue a more detailed analysis at the next stage. The test was used also for other research (Kaczmarska 2013; Kaczmarska, Rosen, Hana & Hladká 2015).

feelings) are uniform in terms of collocability. They combine with certain arguments; we identify Experiencer and a kind of Source (or Stimulus), but on this basis we are not able to distinguish the meaning. However, this step will not be removed from the final version of the algorithm, which can be used to study other groups of verbs, where the semantic roles of the arguments bound to the equivalent and the original verbs may be significant for the differentiation of the meaning. In the case of other verbs we can identify roles such as: Agent, Beneficiary, Location, Time, Instrument, Substance, and Object (itself), and the guideline for choosing the equivalent may be its collocability with the argument and its specific role.

3.4 Step four – Pattern Grammar

"If a word has several senses, and is used in several patterns, each pattern will occur more frequently with one of the senses than the others, such that the patterning of an individual example will indicate the most likely sense of the word in that example." (Hunston and Francis 2000: 20)

The verbs we analyze are mostly polysemous. In tracking their patterns, we hope to be able to link the concrete meaning with a pattern type (understood as a repeatable combination of words).

"A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it." (Hunston and Francis 2000: 37).

We established whether there was indeed such repeatability in the corpus occurrences (Ebeling and Ebeling 2013). The manual analysis based on InterCorp indicated, i.e., two patterns of the Czech unit *být líto* (to be sorry, to regret) associated with two meanings.

If the unit *být líto* is combined with two nominal phrases (Dative and Genitive), it corresponds to the Polish equivalent *żal* (to be sorry for, to regret). If combined only with the Dative nominal phrase, and possibly with the element *to*, it corresponds to the Polish equivalent (*być*) *przykro* (to be sorry).

żal	**(*być*) *przykro***
Jak mi ho bylo líto!	*Pak mi je líto.*
Jakże mi go było żal!	*Wobec tego, przykro mi!*
Je mi ho samozřejmě líto.	*Potom nám to bylo oběma líto.*

Jest mi go oczywiście żal…	*Potem nam obu było przykro.*
Přišlo mi jí prostě líto.	*…nabídne mi sisinku a já si vezmu, protože by mu bylo líto, kdybych si nevzala…*
Po prostu zrobiło mi się jej żal.	*...zaprasza mnie na cuksa i ja biorę, bo byłoby mu przykro, gdybym nie wzięła…*
být líto* + NP_{DAT} + NP_{GEN} = *żal	***být líto* + NP_{DAT} + *to* / Ø = (*być*) *przykro***

Table 3. Patterns of *být líto* (*żal, być przykro*)

In this case, the manual analysis allows us to establish the proper equivalent. For verbs represented by a large number of occurrences, the manual analysis will not be helpful. For these units we will be able to use Word Sketches mentioned in Step six.

3.5 Step five – Cognitive Grammar

3.5.1. Dictionaries and corpora

At this stage, we try to encode the meaning of a word in terms of conceptualization (Langacker 1987, 1991, 2008; Geeraerts 2010). We analyze the unit *mít rád.*[8] In a reputable dictionary of Czech (Havránek 1989), *mít rád* is defined as *pociťovat k někomu náklonnost, lásku, milovat, mít v oblibě* (to feel affection for someone, love, to love, to like). According to these definitions, the Czech-Polish dictionary (Siatkowski and Basaj 2002) gives the following Polish equivalents: *kochać, lubić, przepadać* (to love, to like, to be found). These Polish verbs, supposedly equivalents of the analyzed Czech unit, refer to completely different feelings (emotions). For a Polish speaker, a combination of meanings "to love" and "to like" within a single expression is a strange and unfamiliar concept. In the parallel corpus InterCorp we can find more equivalents of the unit *mít rád* (*lubić*, *kochać*, *podobać się, uwielbiać, polubić, pokochać, w naszym guście*), however, they all belong to two distinct semantic fields denoting "love" and "liking".[9] The strangeness of the concept makes both its understanding and translating into Polish very difficult, and – as mentioned in the Introduction – the problem sometimes cannot be solved even in a wider context, e.g.:

[8] For *mít rád* (not fully translatable into Polish) see also Kaczmarska and Rosen (2014a).
[9] In the parallel corpus InterCorp we found 2799 occurrences of the unit *mít rád*. 66% of the occurrences were translated into Polish as a unit referring to „liking" (*lubić*) and 18% – „love" (*kochać*). The remaining 16% were translated with other units.

(cs) *Mám tě strašně rád, řekl.* (Kundera-Valcik_na_rozl)
(pl) *Strasznie cię kocham – rzekł.* (Kundera-Valcik_na_rozl)

(cs) *Kdybys mě měla ráda, nemohla by ses opičit s tím pitomým jménem*. (Grusa-Dotaznik)
(pl) *Gdybyś mnie naprawdę lubiła, nie wygłupiała byś się z tym kretyńskim imieniem*. (Grusa-Dotaznik)

(cs) *Máš-li mne jen trošku rád, shoď mne z třetího patra, dej mně tu poslední outěchu*. (Hasek-OsudyDobrehoVvSV)
(pl) *Jeśli masz dla mnie choć troszkę przyjaźni, zrzuć mnie z trzeciego piętra, udziel mi tej ostatniej pociechy*. (Hasek-OsudyDobrehoVvSV)

In these contexts we could use all equivalents offered by the Czech-Polish dictionary. Furthermore, there is a verb *milovat* in Czech, which is translated into Polish as *kochać* (to love).[10]

However, there is a certain complication regarding the translation of the analyzed unit *mít rád*. Native speakers of Polish tend to reflect the pattern from their own language in a foreign language. In Polish, "lubić" i "kochać" signify completely different feelings. The difference lies mainly in the quality, not the intensity of the feeling. In this particular situation, a native speaker of Polish can automatically attribute the meaning "kochać" (to love)[11] to the unit *milovat* and the meaning "lubić" (to like)[12] to the unit *mít rád*. In Polish there is no verb expressing feelings on the borderline of love and liking, because such a concept does not exist in this language. Consequently, the Polish language cannot offer an appropriate equivalent. While confronting Czech and Polish, we can experience, in this case, a misunderstanding.

With such a large amount of data from InterCorp, we can make an attempt at building a network of meanings (Kaczmarska 2015a).[13]

[10] In InterCorp we found also 586 occurences of the unit *milovat*: 84% translated with the unit referring to "to love" (*kochać*) and 4% – "to like" (*lubić*). The remaining 12% were translated with other units.

[11] An analysis based on InterCorp shows that Polish verb *kochać* is translated into Czech predominantly as *milovat* (over 72% occurrences out of 3497); 20% occurrences were translated as a unit reffering to "liking" (*mít rád*) and the remaining 8% were translated with other unit.

[12] In the parallel corpus InterCorp we also found 3063 occurrences of the verb *lubić*: 95% of the occurrences were translated into Czech as a unit referring to *mít rád* ("liking") and 3% – "love".

[13] The network is based on definitions from monolingual dictionaries available online: http://ssjc.ujc.cas.cz/search.php?hledej=Hledat&heslo=r%C3%A1d&sti=EMPTY&wher

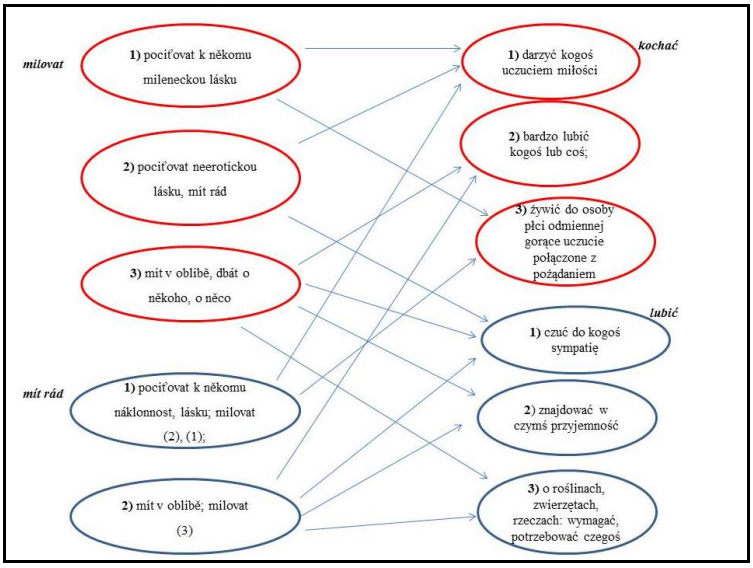

Figure 2. The network of meanings of the Czech units *mít rád* and *milovat* and of the Polish units *kochać* and *lubić*

3.5.2. A survey and the web

According to the current results of the analysis we could conclude that there are more reasons for translating the Czech unit into Polish as *lubić* than as *kochać*. Since the unit is frequent and still very problematic, we conducted a survey among Czech native speakers. The survey took place in Liberec (Czech Republic), in December 2013 (30 respondents, 19 – 24 year old). The aim of the survey was to discover the meaning of *mít rád* on the basis of the opposition with *milovat*; we also asked the direct

---

e=hesla&hsubstr=no;
http://ssjc.ujc.cas.cz/search.php?hledej=Hledat&heslo=milovati&sti=EMPTY&where=hesla&hsubstr=no;
http://sjp.pwn.pl/szukaj/lubi%C4%87.html;
http://sjp.pwn.pl/szukaj/kocha%C4%87.html. The network reflects only the way of understanding and translation of the Czech units into Polish (this is why there are only arrows pointing in one direction). We took into consideration only four described units; we realize, however, that a comprehensive map of meanings should also include other potential equivalents, such as *uwielbiać* (to adore). We treat the two units *kochać* and *lubić* as representatives for the groups of equivalents that could have the feature 'love' or 'liking'. The equivalence of meaning was based on manual analysis of occurrences (Czech originals translated into Polish), a total of 3385 examples in Czech and as much in Polish.

question if there are any differences between the two verbs. Furthermore, the respondents were requested to write what objects may be combined with the two verbs.

100% of the respondents identified semantic differences between the two verbs; *milovat* is something more than *mít rád*. However, they had a problem with the diversification of objects bound by the units; they assigned the same objects to both units. Such results could not be used in our analysis directly. We could state that the choice of an equivalent depends on a wider context. However, on the web we can find a large number of opinions concerning this problem. The Czechs discuss what the unit *mít rád* means as a declaration.[14] Consequently, we cannot say that the problem of translating *mít rád* can always be solved by a wider context. If the concept does not exist in a target language, reproducing it in this language may be connected with the loss of a part of its meaning.

We intentionally presented only vestigial elements of the Cognitive Grammar methods. The theory assumes that the analyses will be conducted manually, and for this reason it is very hard to adopt in our algorithm. In the final version of the algorithm it will be moved to the last step and only applied in exceptional cases.[15]

3.6 Step six – Word Sketches and other tools of the future

As mentioned above (Step four and Step five), the manual analysis will not be efficient in the case of verbs represented by a large number of occurrences (i.a. *toužit*). For these units, we could use Word Sketches (Kilgarriff & Tugwell 2002; Kilgarriff et al 2014).[16] The collocates of an analyzed unit are grouped according to the grammatical relations in which they occur. Word Sketches seem to be a universal tool for analyzing collocations and word combinations in terms of pattern grammar and valency.

We made an attempt to apply the tool for further analysis of the Czech verb *toužit*. The research was based on the Czech-Polish part of

---

[14] There is no doubt as to the meaning *milovat* in the same position.
(http://diskuse.doktorka.cz/mit-rad-zamilovat-se-milovat/,
http://www.poradte.cz/spolecnost/21684-milovat-nebo-mit-rad.html,
http://janajerabkova.blog.idnes.cz/c/194377/Milovat-nebo-mit-rad.html)

[15] In problematic cases, we can also refer to the explications and natural semantic metalanguage (Wierzbicka 1980, 2001) or construct an intensity scale of properties expressed by a given verb (Mikołajczuk 1997, 1999; Bratman 1987).

[16] "A word sketch is a one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour." Word Sketches are available online at http://www.sketchengine.co.uk/documentation/wiki/Website/Features#Wordsketches

InterCorp. At this point, we confront yet another problem. The Czech examples are excerpted from the vast Czech National Corpus and the Polish examples from InterCorp. The data are incomparable in terms of size. Unfortunately, at present, using Word Sketches for the Czech language is impossible in InterCorp. It is also impossible to use the National Corpus of Polish (Przepiórkowski et al 2012): its size is comparable with the size of the Czech corpus, but the Polish corpus does not offer the Word Sketches as a tool. The analysis must be conducted on the basis of one corpus and homonymous texts. The appropriate tool, which is in preparation, will cooperate with both the Czech and the Polish part of InterCorp.

Word Sketches is a promising tool for our study. We expect to be able to analyze not only objects expressed by nominal phrases, but also adverbs combined with key verbs.

4. A case study – *závidět* 'to envy' and *žárlit* 'to be jealous'

4.1. Automatic extraction of pairs of equivalents

We generated the Czech-Polish dictionary thanks to the tool Treq (available at the Czech National Corpus website).[17]

Polish equivalents of ***závidět***	
zazdrościć	**188**
zazdrość	**26**
pozazdrościć	**16**
zazdrosny	**7**
zawiść	**4**
darzyć	**1**
straszliwie	**1**
współzawodniczyć	**1**
zwyknąć	**1**

[17] Treq (available online at: http://treq.korpus.cz) generates lists of the most often equivalents of selected words. However, one should realize that the product is not an ideal dictionary, including only proper equivalents. Treq uses a fully automatic method and among proposed equivalents we also occasionally find accidental words and even punctuation marks. In previous research, we generated the dictionary ourselves (Kaczmarska & Rosen 2013).

Table 4. The most common Polish equivalents of the verb ***závidět*** based on Treq[18]

Some of the equivalents from the table do not correspond to the meaning of *závidět*: They could appear by coincidence as a result of an error of the alignment (*zwyknąć*, *straszliwie*) or they are a part of a synonymous phrase (*darzyć – darzyć uczuciem / zazdrością / uczuciem zazdrości*). Among the suggested equivalents there is also a verb *współzawodniczyć* that can be interpreted as a distant synonym of the unit *závidět*.

The most frequent equivalent is the verb *zazdrościć* (and its derivative *pozazdrościć*). The other suggestions are represented by very few occurrences. Also the noun *zazdrość* is a part of a phrase *czuć zazdrość* that means *zazdrościć*. The results are conclusive and we are able to establish the Polish equivalent at the level of our analysis.

We can also check the results with an excerpt from the Czech-Polish part of the parallel corpus InterCorp[19]. There are only 50 occurrences of the verb (from originally Czech texts translated into Polish). Most of them are translated as *zazdrościć*, which confirms the assumption.[20]

Polish equivalents of ***závidět***	**50**
zazdrościć	45
pozazdrościć	3
być zazdrosny	1
inne	1

Table 5. Polish equivalents of Czech verb *závidět* based on the Czech-Polish part of the parallel corpus InterCorp

We also generated this type of dictionary for the verb *žárlit.*

Equivalents of ***žárlit***	
zazdrosny	**141**

[18] Treq excerpted occurrences from all Czech texts and their equivalent segments.
[19] This time we excerpted occurrences only from originally Czech texts and their translations into Polish.
[20] The only occurrence with the equivalent *zazdrosny* presents an example with an ellipsis:
(cz) […] spokojen, že **mu** nemá co *závidět*…
(pl) […] zadowolony, że nie potrzebuje *być zazdrosny*… [Paral-VeletrhSplnenych]

zazdrość	**25**
zazdrościć	**14**
być	**2**
osiłek	**1**
darzyć	**1**
owszem	**1**
rywalka	**1**
zawiść	**1**

Table 6. The most common equivalents of the verb ***žárlit*** based on **Treq**

As in the case of the previous verb, there are also some wrongly aligned equivalents (*być*, *osiłek, darzyć, owszem*). The words *rywalka* and *zawiść* can be treated as elements of the structures synonymous to the concept of 'jealousy' (*zazdrość*). The most frequent equivalent is the word *zazdrosny* – a component of the phrase *być zazdrosny* (to be jealous).

We also checked equivalents of the verb *žárlit* in the Czech-Polish part of InterCorp.

Polish equivalents ***žárlit***	**59**
być zazdrosny	46
zazdrościć	7
zazdrość	4
zawiść	1
błąd	1

Table 7. Polish equivalents of Czech verb ***žárlit*** based on the Czech-Polish part of the parallel corpus InterCorp.

About 78% of occurrences include the equivalent *być zazdrosny*[21] and we can consider it as the proper equivalent.

[21] In the aligned segments, there is also the verb *zazdrościć* as an equivalent. The occurrences do not include, however, the typical object of the jealousy (expressed by the nominal phrase in Genitive), but a kind of the reason of being jealous expressed by a sentential phrase:
(cz) Povídám, jako vždycky, von na mě *žárlí*, že jsem mladší než von.

The verbs *závidět* and *žárlit* were also subjected to another investigation (Kaczmarska 2015c). A thorough analysis (both syntactic and semantic) was conducted and it confirmed that their equivalents can be found at the first step of the presented algorithm. The analysis allowed us to build a network of meanings for the analysed units:

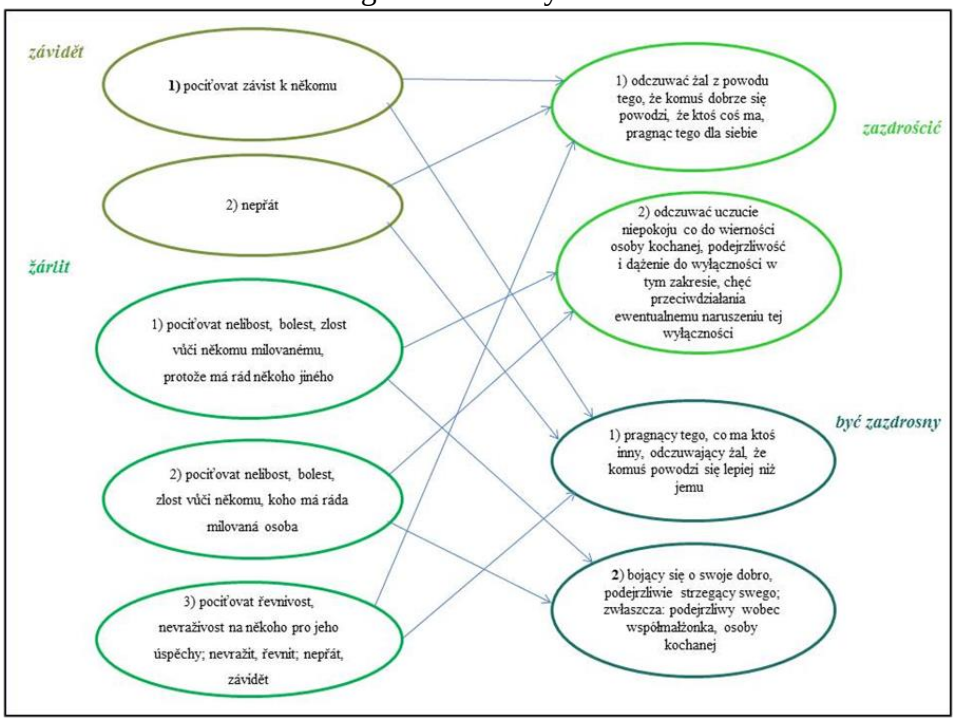

Figure 3. The network of meanings of the Czech units *závidět and žárlit* and of the Polish units *zazdrościć* and *być zazdrosnym*[22]

(pl) Powiadam, jak zawsze, on **mi** *zazdrości*, że jestem młodszy niż on [Hrabal-Prilis_hl_samot]

The verb *zazdrościć* (as equivalent of *žárlit*) appears also in constructions with ellipsis:

(cz) Právě proto, že už nechce *žárlit*, bere vážně a bez podezření jeho tvrzení!

(pl) Właśnie dlatego, że nie chce już *zazdrościć*, przyjmuje jego słowa poważnie i bez podejrzeń! [Kundera-Valcik_na_rozl]

[22] The network is based on definitions from monolingual dictionaries available online: – http://ssjc.ujc.cas.cz/search.php?hledej=Hledat&heslo=z%C3%A1vid%C4%9Bti&sti=EMPTY&where=hesla&hsubstr=no
http://ssjc.ujc.cas.cz/search.php?hledej=Hledat&heslo=%C5%BE%C3%A1rliti&sti=EMPTY&where=hesla&hsubstr=no (for the verbs *závidět* and *žárlit*)
http://sjp.pwn.pl/sjp/zazdrosc;2544740.html
http://sjp.pwn.pl/sjp/zazdro%C5%9Bci%C4%87;2544739 (for units *zazdrościć* and *być zazdrosnym*).

As mentioned in paragraph 3.1 (above), the meaning of Czech verbs expressing negative emotions and feelings (e.g. *trápit*, *mrzet*) are easier to recognize than the verbs expressing positive emotions and feelings (Kaczmarska 2015d). As a consequence, the process of translating the verbs themselves into Polish is simpler (Kaczmarska 2013). The answer as to why it happens, would require a thorough semantic analysis of two (numerous) groups of verbs – expressing both positive and negative emotions and feelings. However, conducting such an analysis in the context of this study is not possible.

5. Conclusions and perspectives

Using corpora in the process of establishing equivalents seems to be obvious and necessary. Especially a parallel corpus makes it possible to define clusters of equivalents, which are essential and fundamental for any further steps. Although a parallel corpus can assist in the development of comparative analyzes, the research is often confronted with difficulties due to incompatible tools.[23] As we found Word Sketches promising for our research, we prepared the tool for the Polish part of InterCorp, but it is not available for external users of InterCorp. Word Sketches for the Czech part of InterCorp is in preparation. We hope that Word Sketches applied to both the Czech and the Polish parts of InterCorp will be the turning point for building our algorithm.

At the later stage, we obtained rather disappointing results based on Case Grammar. The method using Case Grammar will be tested on a larger number of verbs to see if it deserves to be developed further or discarded. We also realize that the project needs a deeper cognitive analysis of the most difficult units.[24] However, the cognitive method is very difficult to implement into the algorithm and must be elaborated manually. The analysis also clearly showed the problem of "nonexistence" of a concept in the other language. Translation of such words always leads to an arbitrary decision by the translator.

[23] We face similar problems while working with monolingual corpora. Word Sketches are available for SYN (Czech National Corpus). For the Polish language, a comparable corpus is NKJP (National Corpus of Polish), but we cannot use Word Sketches for NKJP. Furthermore, the Czech and Polish corpora have different statistical functions what make corpus-based comparative analyses even more complicated.

[24] In problematical cases we can refer to explications and natural semantic metalanguage (Wierzbicka 1980, 2001) or construct a scale of the intensity of a feature expressed by a given verb (Mikołajczuk 1997, 1999; Bratman 1987).

We hope that our algorithm will be able to cooperate with machine translation tools.[25] This is why, in addition to a manual analysis of the valency requirements, we also conduct experimental trials of stochastic modeling of the choice of an equivalent on the basis of the context. We use two methods: the first – based on the context of a few lexemes before and after the keyword, and the second – on the basis of lexemes dependent directly on the keyword. The methods and results of the research are presented in a separate paper (Kaczmarska, Rosen, Hana & Hladká 2015).

References

Bratman, M. E. 1987. *Intentions, Plans, and Practical Reason*. Massachusetts: Harvard University Press.

Čermák, F. & and A. Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 13(3): 411–427.

Čermáková, A. 2009. *Valence českých substantive*. Praha: Nakladatelství Lidové noviny.

Daneš, F. & Z. Hlavsa. 1987. *Větné vzorce v češtině*. Praha: Academia.

Dębski, A. 1982. Semantyczna walencja czasownika w aspekcie konfrontatywnym. *Biuletyn Polskiego Towarzystwa Językoznawczego* 39: 79-90.

Ebeling, J. & S. O. Ebeling. 2013. *Patterns in contrast*. John Benjamins Publishing Company.

Fillmore, Ch. J. 1968. The Case for Case. In E. Bach & R. T. Harms (eds.), *Universals in Linguistic Theory*, 1-88. New York: Holt, Rinehart, and Winston.

Geeraerts, D. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

[25] Work on algorithms improving machine translation and differentiating the meanings of ambiguous units (e.g. WSD – Word Sense Disambiguation) are already carried out for a long time and known also for parallel corpora. They are mostly based on data obtained from very large corpora using various mathematical methods (mainly statistical), cf. e.g. Liang Tian et al 2014; Młodzki et al 2012; Liang Tian et al 2010; Han et al 2013; Kędzia et al 2014. Developed algorithms also use various linguistic approaches; more on this subject – Han et al 2013.

Greń, Z. & D. Rytel-Kuc. 1991. Wykorzystanie przekładów literackich w pracy nad dwujęzycznym słownikiem walencyjnym. In H. Běličová, G. Nieszczimienko & Z. Rudnik-Karwatowa (eds.), *Problemy teoretyczno-metodologiczne badań konfrontatywnych języków słowiańskich*, 69-78. Warszawa: Instytut Słowianoznawstwa Polskiej Akademii Nauk.

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Arnold.

Han, A.L-F., Y. Lu, D.F. Wong, L.S. Chao, L. He & J. Xing. 2013. Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 365-372. Association for Computational Linguistics

Havránek, B. (ed.). 1989. *Slovník spisovného jazyka českého*. Praha: Academia.

Hunston, S. & G. Francis. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing Company.

Jirásek, K. 2011. Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. In F. Čermák (ed.). *Korpusová lingvistika Praha 2011: 1 – InterCorp*, 45–55. Praha: Nakladatelství Lidové noviny, 2011.

Kaczmarska, E. 2001. Badanie struktury walencyjnej czeskich i polskich predykatów posiadających pozycję Experiencera. *Studia z Filologii Polskiej i Słowiańskiej* 37: 177-187. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Kaczmarska, E. 2002/2003. Nominalizacje odczasownikowe w języku polskim i czeskim (wybrane problemy). *Studia z Filologii Polskiej i Słowiańskiej* 38: 87-99.

Kaczmarska, E. 2010. Analiza zdolności konotacyjnych polskich i czeskich predykatów odnoszących się do strachu, złości i wstydu. In J. Goszczyńska & Z. Greń (eds.), *Res slavisticae*, 135-153. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaczmarska, E. 2012. Czeski czasownik „zdát se" w przekładzie na język polski (na podstawie badań z wykorzystaniem czesko-polskiego korpusu równoległego InterCorp). *Studia z Filologii Polskiej i Słowiańskiej*, 2012: 247-261.

Kaczmarska, E. 2014. Czy na pewno się (nie)rozumiemy? O problemach, uproszczeniach i stratach w przekładzie (na podstawie czesko-polskiej części korpusu równoległego InterCorp). In M. Benešová, R. Rusin Dybalska & L. Zakopalová (eds.), *Proměny polonistiky.*

Tradice a výzvy polonistických studií, 192-199. Praha: KAROLINUM.

Kaczmarska, E. 2015a. Mít rád czy milovat? O czeskiej miłości po polsku (in print).

Kaczmarska, E. 2015b. W poszukiwaniu znaczenia czasowników wyrażających stany psychiczne. Analiza czeskich czasowników i ich polskich ekwiwalentów – próba implementacji wybranych teorii lingwistycznych (walencja, gramatyka przypadków głębokich, Pattern Grammar, lingwistyka kognitywna). *Prace Filologiczne* (in print).

Kaczmarska, E. 2015c. O dwóch czeskich jednostkach wyrażających negatywne stany emocjonalne i ich polskich ekwiwalentach. Analiza na materiale z korpusu paralelnego InterCorp. (in print).

Kaczmarska E. 2015d. Czeskie czasowniki oznaczające stany psychiczne – sposoby ustalania polskich ekwiwalentów na podstawie korpusu równoległego InterCorp (in print).

Kaczmarska, E. & A. Rosen. 2013. Między znaczeniem leksykalnym a walencją – próba opracowania metody ekstrakcji ekwiwalentów na podstawie korpusu równoległego. *Studia z Filologii Polskiej i Słowiańskiej* 48: 103–121. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Kaczmarska, E. & A. Rosen. 2014a. *Czego nie można wyrazić w języku polskim, czyli o leksykalnych w nim brakach*, Polonica 34: 53–66.

Kaczmarska, E. & A. Rosen. 2014b. Praktyczny przewodnik po korpusie równoległym InterCorp. In M. Hebal-Jezierska (ed.), *Praktyczny przewodnik po korpusach języków słowiańskich*, 207-231. Warszawa: Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaczmarska, E., A. Rosen, J. Hana & B. Hladká. 2015. Syntactico-semantic analysis of arguments as a method for establishing equivalents of Czech and Polish verbs expressing mental states. *Prace Filologiczne* (in print).

Kędzia, P., M. Piasecki, J. Kocoń, A. Indyka-Piasecka. 2014. Distributionally Extended Network-Based Word Sense Disambiguation in Semantic Clustering of Polish Texts. In *IERI Procedia (International Conference on Future Information Engineering)* 10, 38-44. DOI: 10.1016/j.jeri.2014.09.073

Kilgarriff, A. & D. Tugwell. 2002. Sketching words. In M-H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* EURALEX: 125-137.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý & V. Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography: Journal of ASIALEX* Volume 1, Issue 1, 7–36.

Korytkowska, M. 1984. Kategoria przypadka semantycznego (na materiale języka polskiego, bułgarskiego i serbsko-chorwackiego). In V. Koseska_Toszewa & M. Korytkowska (eds.), *Studia konfrontatywne polsko-południowosłowiańskie*, 11-38. Wrocław: Zakład Narodowy im. Ossolińskich.

Korytkowska, M. 1992. *Typy pozycji predykatowo–argumentowych. Gramatyka konfrontatywna bułgarsko-polska*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Korytkowska, M. 1993. O konfrontatywnym opisie predykatorów bułgarskich i polskich (na przykładzie jednostek otwierających miejsce dla argumentu o wartości Experiencer). In V. Koseska_Toszewa & M. Korytkowska (eds.), *Studia gramatyczne bułgarsko-polskie* (5-6), 121-150. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Langacker, R. 1987. *Foundations of Cognitive Grammar, vol. 1, Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, R. 1991. *Foundations of Cognitive Grammar, vol. 2, Descriptive Application*. Stanford: Stanford University Press.

Langacker, R. 2008. *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Lewandowska-Tomaszczyk, B. (ed.). 2005. *Podstawy językoznawstwa korpusowego*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Lewandowska-Tomaszczyk, B. 1984. *Conceptual Analysis, Linguistic Meaning, and Verbal Interaction*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Lewandowska-Tomaszczyk, B. 2013. Komunikacja i konstruowanie znaczeń w przekładzie. Paper presented at the conference *Zbliżenia: językoznawstwo –translatoryka – literaturoznawstwo*, Konin, Poland, November 13-14, 2013.

Mikołajczuk, A. 1997. Pole semantyczne ‘gniewu’ w polszczyźnie (Analiza leksemów: gniew, oburzenie, złość, irytacja). In R. Grzegorczykowa & Z. Zaron (eds.), *Semantyczna struktura słownictwa i wypowiedzi*, 149-171. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Mikołajczuk, A. 1999. *Gniew we współczesnym języku polskim. Analiza semantyczna*. Warszawa: Wydawnictwo Energeia.

Młodzki, R., M. Kopeć, A. Przepiórkowski, A. 2012. Word Sense Disambiguation in the National Corpus Of Polish. *Prace Filologiczne* LXIII: 155-166.

Och, F.J. & H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1): 19-51.

Przepiórkowski, A., M. Bańko, R. Górski & B. Lewandowska-Tomaszczyk (eds.). 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo PWN.

Pustejovsky, J. 1993. *Semantics and the Lexicon*. Springer.

Rosen, A. & M. Vavřín. 2014. *Korpus InterCorp – čeština, verze 7 z 19. 12. 2014*. http://www.korpus.cz.

Rytel, D. 1989. Wybrane problemy opisu walencyjnego języka. *Studia z Filologii Polskiej i Słowiańskiej* 26: 237-247. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Rytel-Kuc, D. (ed.). 1991. *Walencja czasownika a problemy leksykografii dwujęzycznej*. Wrocław: Zakład Narodowy im. Ossolińskich.

Siatkowski, J. & M. Basaj. 2002. *Słownik czesko-polski*. Warszawa: Wiedza Powszechna.

Skoumalová, H. 2008. Extracting dictionaries from parallel corpora. In *Proceedings of The Third Baltic Conference on Human Language Technologies*, 297–301. Kaunas: Vytautas Magnus University.

Tian, L., D.F. Wong & S. Chao. 2010. An Improvement of Translation Quality with Adding Key-Words in Parallel Corpus. *Machine Learning and Cybernetics* (ICMLC) Vol. 3, 1273-1278. DOI: 10.1109/ICMLC.2010.5580888

Tian, L., D.F. Wong, L.S. Chao & F. Oliveira. 2014. A Relationship: Word Alignment, Phrase Table, and Translation Quality. *The Scientific World Journal*. Hindawi Publishing Corporation. http://dx.doi.org/10.1155/2014/438106

Urbańczyk-Adach, N. 2011. *Wariantywność walencji czeskiego czasownika*. Warszawa: Slawistyczny Ośrodek Wydawniczy.

Wierzbicka, A. 1971. *Kocha – lubi – szanuje. Medytacje semantyczne.* Warszawa: Wiedza Powszechna, 1971.

Corpora

Czech National Corpus – InterCorp. Institute of the Czech National Corpus. Available online at http://www.korpus.cz.

National Corpus of Polish. Available online at: http://nkjp.pl.