Corpus-based analysis of Czech units expressing mental states and their Polish equivalents. Identification of meaning and establishing Polish equivalents referring to different theories

Elżbieta Kaczmarska (University of Warsaw)

The analysis is focused on Czech polysemous units expressing mental states. The goal of the analysis is to test which theory can lead us to the closest equivalents of these units in Polish.

The analysis proper is preceded by automatic extraction (Och & Ney 2003) of pairs of equivalents from InterCorp, a parallel corpus (Čermák & Rosen 2012). These pairs constitute a kind of bilingual dictionary (Jirásek 2011). Then we manually analyse parallel segments (sentences) including selected words, excepted automatically from InterCorp. We check (in each segment) how the key word was translated and what kinds of collocations and arguments it has.

The aim of the first part of the analysis is to decide whether valence requirements can help to identify Polish equivalents of the verbs. A pilot study concerning the ambiguous Czech verb toužit 'to miss, to want, to desire' (Kaczmarska & Rosen 2013) was supposed to reveal if valency can influence the choice of an equivalent in Polish. It was assumed that for some senses the equivalent can be established based on the convergence of the valence requirements (Levin 1993). The hypothesis proved to be true. However, the influence of valency was not observed in all the senses of the verb.

A more extensive research is needed to establish equivalents (or cluster of equivalents) for given units (Lewandowska-Tomaszczyk 1984, 2013). To identify all the meanings of the units and finally to find a proper equivalent for each of their senses, we considered the approaches of Pattern Grammar (Ebeling & Ebeling 2013; Francis & Hunston & Manning 1996; Hunston & Francis 2000) and Cognitive Linguistics (Langacker 2008; Taylor 2002; Mikołajczuk 1999).

At the next stage of the analysis we use the Pattern Grammar methods. The verbs we analyse are mostly polysemous. In tracking their patterns, we try 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 that 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.

At the last stage we try to encode the meaning of a word in terms of conceptualization (Langacker 2008). We analyse the unit mít rád '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 fond of'. These Polish verbs, supposedly equivalents of the analysed Czech unit, refer to completely different feelings (emotions). For a Polish-speaking person, a combination of meanings 'to love' and 'to like' within a single expression is a strange and unfamiliar concept.

The results of the triple analysis let us establish the semantically and syntactically closest Polish equivalents of our study cases – Czech verbs expressing dissatisfaction.

As an outcome of this research, we aim to design an equivalent-searching algorithm, based on a syntactico-semantic analysis. The algorithm will be applied to the analysis of different words expressing mental states.

References

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

Ebeling, Jarle, and Signe O. Ebeling. 2013. Patterns in contrast. John Benjamins Publishing Company.

Francis, G. & Hunston, S. & Manning, E. (1996). Collins COBUILD Grammar Patterns 1: Verbs, HarperCollins. London.

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

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

Kaczmarska, E. & Rosen, A. (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.

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. (1984). Conceptual Analysis, Linguistic Meaning, and Verbal Interaction. Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Lewandowska-Tomaszczyk, B. (2013). Komunikacja i konstruowanie znaczeń w przekładzie. not published conference lecture, Konin 13-14.11.2013.

Mikołajczuk, A. (1999). Gniew we współczesnym języku polskim. Analiza semantyczna, Warszawa.

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

Siatkowski, Janusz, and Mieczysław Basaj. 2002. Słownik czesko-polski. Warszawa: Wiedza Powszechna.