Applying MIPVU Metaphor Identification Procedure on Czech

Dalibor Pavlas

Ondřej Vrabeľ

Jiří Kozmér

Palacký University Olomouc

annDH, Sofia, 8. 8. 2018

Czech metaphor corpus - motivation

- **Cognitive linguistics**
  - Conceptual metaphor studies are often conducted on data acquired by introspection instead of real language data

- **Corpus linguistics**
  - Detailed metaphor usage statistics for Czech language

- **Computational linguistics**
  - Resource for training and evaluation of automatic metaphor processing systems (gold standard)

Annotation Method

- MIPVU (VU Amsterdam Metaphor Corpus; Steen et al. 2010)
  - The largest metaphor resource (approx. 200K tokens; 4 genres)
  - Identifies metaphor on the word level
  - The only established method for manual metaphor identification in text
  - In smaller scale projects applied to other languages:
    - Russian (Badryzlova et al. 2013)
    - Lithuanian (Urbonaitė 2015)

MIPVU Overview

- Annotators examine each lexical unit separately and establish the basic meaning using a dictionary
- If lexical unit's (word's) contextual meaning != basic meaning, it is considered Metaphor-Related Word (MRW)
- If basic meaning of a word is:
  - a) more concrete; what it evokes is easier to imagine, see, hear, feel, smell and taste;
  - b) related to bodily action;
  - c) more precise (as opposed to vague)
- The word is marked as MRW.

Course of action for Czech project

- Train a team of annotators
- Annotate short text excerpts
- Test reliability of annotation
- Analyse errors (cases of disagreement between annotators)
- Propose modifications of the protocol – make it more suitable for Czech
- Perform second round of annotation
- Test reliability again

Annotation process

- 2 text excerpts:

    1) “Zasraný vánoce” by Michal Viewegh (fiction genre; 598 tokens)

    2) transcription of proceedings of the European Parliament (611 tokens)

- Texts were processed by white space tokenization to preliminarily determine the lexical units

- 3 annotators (2 Ph.D. students, 1 Master's student)

- Used dictionaries: *Dictionary of Standard Czech Language* (Vácha et al., 1971; *SSJČ*)
  Dictionary of Standard Czech (Kroupová et al., 2005; *SSČ*)

Reliability test I

- Reliability is measured using Fleiss' kappa
  - a statistical measure of inter-annotator agreement which corrects for chance agreement between analysts (Artstein and Poesio, 2008).

Text	Tokens	Percentage unanimous			Fleiss'κ
		Not MRW	MRW	Total	
Viewegh	598	87.46	4.85	92.31	0.65
Europarl	611	76.76	10.97	87.73	0.72
Total Fleiss' κ					0.70

- The reliability test in the Czech MIPVU project

Reliability tests comparison

- Comparison of inter-annotator agreement in other MIPVU projects

Applying MIPVU on Czech; 3 annotators, 1209 tokens Reliability test 1	Russian corpus of conceptual metaphor; 3 annotators, approx. 2000 tokens (Badryzlova et al. 2013) Reliability test 1	Russian corpus of conceptual metaphor; 3 annotators, approx. 2000 tokens (Badryzlova et al. 2013) Reliability test 2	VU Amsterdam Metaphor Corpus; 4 annotators, 1921 tokens (Steen et al. 2010) Reliability test 6
0.70	0.68	0.90	0.85

- The goal is to report performance similar to EN and RU projects

Error analysis

- POS disagreement

POS	Viewegh	Europarl	Sum of disagreement
Nouns	6	18	24
Verbs	18	30	48
Adjectives	6	6	12
Adverbs	5	4	9
Prepositions	11	16	27
Conjunctions	0	1	1
All POS	46	75	121

Text	Tokens	Percentage unanimous			Fleiss’κ
		Not MRW	MRW	Total	
Viewegh	598	87.46	4.85	92.31	0.65
Europarl	611	76.76	10.97	87.73	0.72
Total Fleiss’ κ					0.70

- The annotated excerpt of European Parliament proceedings shows more disagreements and higher inter-annotator agreement at the same time. This is caused by the fact that more than twice as many MRWs are present in the text.

Reflexive pronouns “se/si”

- used when the subject and object of the sentence are identical
  - *umyji se; I will wash myself*
- or as an integral part of a reflexive verb whose meaning they often determine
  - *prát / prát se; to wash (clothes) / to get into a fight*
  - *rozvést / **rozvést se**; to develop (an idea) / **to divorce***

Annotated sentence	*Když*	*se*	*před*	*třemi*	*lety*	*rozvedl […]*
Original MIPVU	0	0	1	0	0	1
Modified MIPVU	0	0	1	0	0	0

- the expression “se” + “rozvedl” needs to be counted as one lexical unit which is distinct from “rozvedl” (same policy as used for phrasal verbs in MIPVU)

Prepositions

- Most metaphor-rich part of speech
- Reported to account for 38.5-46.9% of MRWs (Steen et al., 2010)
- Even more homonymous in Czech
    - Hard to determine only one basic meaning

Solution

- Dividing prepositions by grammatical case e.g.:

1) *Petr stojí **za** mnou; Petr stands **behind** me* *(instrumental)*
2) *Chytil jsem ho **za** nohu; I caught him **by** the leg* *(accusative)*
3) ***Za** 2 roky to bude hotové; It will be done **in** 2 years* *(accusative)*
4) *Vyměnil jsem kolo **za** auto; I traded the bike **for** the car* *(accusative)*

- If we distinguish between "za" in instrumental (expression 1)) and in accusative 2), we can have basic meaning for each one, moreover "accusative za" standing for basic meaning of this preposition in sentences 3) and 4) which both are MRWs.

Conclusions and further steps

- Direct transferability of the MIPVU procedure to Czech language turned out to be problematic
  - similar complications were reported by researchers applying the procedure on Russian and Lithuanian
- Based on the error analysis, we have proposed several minor modifications of the guidelines in order to make them more suitable for Czech
- We plan to conduct second reliability test as soon as possible

Literature

- Arstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4): 554–596.
- Badryzlova, Y., Shekhtman, N., Isaeva, Y., Kerimov, R. (2013). Annotating a Russian corpus of conceptual metaphor: a bottom-up approach. In *Proceedings of the First Workshop on Metaphor in NLP*. Atlanta, GA: Association for Computational Linguistics, pp. 77–86.
- Kroupová, L. et al. (2005). *Slovník spisovné češtiny pro školu a veřejnost: s Dodatkem Ministerstva školství, mládeže a tělovýchovy České republiky*. Praha: Academia.
- Steen, G., Aletta, G., Dorst, J., Herrmann, B., Kaal, A. A., Krennmayr, T., Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU.* Amsterdam, John Benjamins.
- Urbonaitė, J. (2015). Metaphor identification procedure MIPVU: an attempt to apply it to Lithuanian. *Taikomoji kalbotyra*, (7):1–26.
- Vácha, J., editor, et al. (1971). *Slovník spisovného jazyka českého*. Praha: Academia.

Language Resource References

- Rosen, A., Vavřín, M., Zasina, A. J. (2017): InterCorp, 10.0, Institute of the Czech National Corpus, Charles University, Prague. Available from: http://www.korpus.cz