Mapping Czech and English Valency Lexicons: an attempt

V. Baisa, K. Pala, Z. Sitová, J. Vonšovský

Fakulta informatiky, Masarykova universita, Botanická 68a, 602 00 Brno

{baisa, pala, sitova, vonsovsky}@fi.muni.cz

This work was partially supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047

Outline

- An attempt to connect two valency lexicons
- Motivation
- Pattern Dictionary of English Verbs (PDEV), Hanks
- VerbaLex
- Related work
- Differences and similarities, example
- Conclusions

PDEV

- Pattern Dictionary of English Verbs
- Project: Disambiguation of Verbs by Collocation – University of Wolverhampton (R. Mitkov)
- adresa: {http://www.pdev.org.uk}}~\cite{hanks2013lexical}
- The technique Corpus Pattern Analysis (CPA) is strictly corpus based: approx. 1100 verbs have been annotated so far

PDEV II

- Verb patterns in PDEV consist of the basic argument structure
- Typically with semantic values stated for each of the arguments – shallow ontology is used
- Also surface valencies are given (subject, object, determiners, possesives, etc.)
- Patterns are obtained from concordancies (the source is BNC50, word sketches are used as well

VerbaLex

- Frames contain argument structure with semantic values of the arguments (semantic roles)
- Surface or morphosyntactic valencies (cases, adverbs, etc. – the source Brief was used)
- Approx. 10 500 Czech verbs with the various information – aspect, reflexivity, primary and secondary meanings, idioms, passive, active
- Complex roles, about first level 40 roles, about 900 subcategorization features taken from PWN

Vallex

- Valency lexicon from ÚFAL – approx. 6500 Czech verbs
- Related closely to PDT – exploited in its manual annotation,
- Two lexicons: PDTVallex, EngVallex
- transformation of the PropBank into the Vallex and FGD shape (functors), about 48 items
- Most of this work has been done manually (Cinková) – adaptation to the FGD notation
- Also Framenet has to be mentioned

Differences and similarities

- Direction E-Cz, some grammatical categories are different – aspect, case
- 313 single-pattern verbs in PDEV, diff. coverage
- Example with aspect: *zrychlit*, *zrychlovat* (*accelerate*)
- Some verbs are missing in PDEV (*poposednout*), some in VerbaLex (*disregard*)
- Different ontological structures are used – Top Ontology (EWN, Vossen1999), Shallow Ontology (Pustejovsky, 2006)

Differences and similarities II

- For example, group class in PDEV contains subclasses of human group, vehicle group,animal group, physical object group
- They are in their own respective categories in VerbaLex
- From 21 analyses, 9 were translated without any problems from one lexicon to another
- Another 10 were translated with some imperfections (missing frames in PDEV)

Analysis example

- {rozmazlovat (cosset)}\\{Verbalex}\\{}[1] \verb#AG <person:1> V PAT<person:1>#\\{} [2] \verb#AG <person:1> V PAT<person:1> (ACT<act:2>)#\\

- {PDEV}\\{} [1] 100\% \verb#[[Human 1]] cosset [[Human 2]] #\\HS {\it [[Human 1]] cares for [[Human 2]] in an excessively protective and fussy way}\\
  Comment: exact match \\

Discussion

- Promising: positive examples have been found
- The ontologies used in PDEV and VerbaLex are structured differently: inventories are different
- In PDEV the basic items are the individual verb lemmata, in VerbaLex – synsets
- This can be handled by obtaining appropriate (frequency) lists from VerbaLex
- The comparison of the two ontologies is a separate (challenging) task for a separate paper