Hana Žižková

Improving Compound Adverbs Tagging

RASLAN 2018

Part of Speech

- Description of the language cannot dispense with part of speech classification
- Difficulty in classifying: different thought categories, different forms, different syntactic uses
- Finding a suitable criterion
- Assignment to part of speech = assignment to the prototype of the most general meanings

Compound Adverb

- Adverbialization: Process of forming an adverb from another part of speech
- In Czech: preposition + noun (například)
  + adjective (natajno)
  + numeral (pošesté)
  + pronoun (potom)
  + adverb (dozajisté)
- One word vs. Multiword expression (do červena vs. dočervena)

How?

- Always same examples in grammars
- New compound adverbs
- Corpus probe

Approach

- Identified one-word compound adverbs tagged pos=X (dooranžova, nablint, odpředu, …)
- Manualy sorted according the prefix and ending (poanglicku, pořadě, pokrk, pošesté, poprvnı́)
- Checked if listed in Czech dictionaries
- Split in multiword expressions (do oranžova, na blint, od předu, …)
- Checked how tagged
- Checked if listed in Czech dictionaries
- Collocations

Finding

- Identified 470 forms that we thought could be compound adverbs
- Many of the one-word compound adverbs (kpředu, odposledka, zmísta, zšeda,…) recorded in existing dictionaries → not only occasionalism
- Multiword compound adverbs tagged as a preposition and initial part of speech, such as:
  - nouns (na mokro, k dobru, ob den, …)
  - adjectives (na jisto, do pevna, …)
  - adverbs (na knap, na krátce, na tajno, k stáru, …)
  - numerals (ob dva, na vícekrát, po mnohokrát, …)
  - pronouns (po svých, …)
  - prepositions (na podél, na prostřed, …)
  - verbs (do leskla, k předu, na zrz, z nenadála)

Finding

- Most of the obtained expressions = compound of preposition nouns/adjectives/pronouns/numerals in the singular (naskok, dočervena, nadálku, …)
- Four units = compound of preposition and noun in plural (nahony, sdíky, počertech, odvěků)
- Many multiword compound adverbs listed in dicionaries (na světlo, na slovo, nablint, po krk, po čertech, …)
- Some of compound adverbs have shown strong collocations (zbarvit do bíla/dobíla, zaostřovat do blízka/doblízka, holení na mokro/namokro, …)

Finding

- Tagging of multi-word compound adverbs as a preposition and seven different part of speech = inconsistent

na tvrdo (POS=R, POS=A)
na žluto (POS=R, POS=N)
na tajno (POS=R, POS=D)

Solution Suggestion

- Change of morphological tag
- Addition to the morphological dictionary
- Addition of strong collocations into the Multi-Word Expressions Lexical Database

Change of Morphological Tag

- In accordance with NOVAMORF project, new part of speech type: POS=O: an oscillating part of speech: nouns/adjectives/adverbs (sucho, mokro, modro, …)
- New subset of the SUB=s meaning compound to adverbs and numerals
- namodro: POS=D, SUB=s
- na modro: na POS=R modro POS=O, SUB=s

Addition to the Morphological Dictionary

- 177 units proposed for addition in the morphological dictionary
- 103 POS=D, SUB=s, compound adverb (domodra)
- 43 POS=O, SUB=s, oscillating, compound (do modra)
- 20 POS=C, numeral (našestkrát)
- 4 POS=D, adverb (tuty)
- 2 POS=R, preposition (naprostřed)
- 1 POS=I, interjection (doboha)
- 1 POS=J, conjunction (mezitím)
- 1 POS=T, particle (naviděnou)
- 1 POS=N, noun (podmíru, lemma podmíra)
- 1 POS=V, verb (zamražena, lemma zamrazit)

Addition of Strong Collocations into the Multi-Word Expressions Lexical Database

- Larger the MWELD is, better results in disambiguation can be reached


adverb
pronoun
noun
preposition
adjective
corpora
tagging
compound