Czech Morphological Tagset Revisited

Miloš Jakubíček, Vojtěch Kovář, Pavel Šmerk

Centrum zpracování přirozeného jazyka
Fakulta Informatiky, Masarykova Univerzita
Botanická 68a, 602 00 Brno

{jak, xkovar3, xsmerk}@fi.muni.cz

RASLAN 2011
December 2, 2011

Outline

Attributive Tagset for Czech

- Attributive
  - sequence of `xY` pairs
  - `Y` is the value of attribute `x`
  - (theoretically) the position does not matter
  - different from Prague positional tagset (`NNMP4-----A----`)
- Example
  - zaměstnance (employees, accusative) `k1gMnPc4`

Why Change?

- Usage
  - used in many subsequent tools
  - some practical problems noted
- No proper standard
  - the manual from 2006 is not very up-to-date
  - different versions of the tagset
- Practical problems
  - the current classification causes problems
  - sometimes not clear which attributes should be determined
  - problem e.g. in morphological disambiguation by syntactic analysis

Changes + Motives

- General motives
  - usability
  - simplicity
  - standardization
  - decidability, disambiguality
  - consistency
- Biggest problem
  - to agree (3 people)

Poor vs. rich tagset

- Poor tagset
  - suitable for manual annotation
  - (native speakers will be able to decide which attributes should be annotated)
- Rich tagset
  - superset of the poor tagset
  - everything we can say about the word
  - e.g. for purposes of parsing
  - including dictionary information

The changes

- remove family gender and number (gR, nR)
  - not used consistently
  - does not work in syntactic agreements
- dual number
  - the same
- substantive-adjective collision
  - e.g. červená, zavražděný
  - manually go through the list of ambiguous words
- adjective-verb collision
  - e.g. prodán, ochoten
  - manually go through the list of ambiguous words

The changes (II)

- pronoun gender and person
  - precisely specified when they should be determined
- numeral-noun collision
  - rule introduced based on agreement
  - removed unused attribute velues at numerals (xG, xH)
- removed biaspectual verbs
  - occurences of verbs is nearly always unambiguous
  - make the grey zone smaller
- ambiguity of particles and interjections
  - go through all occurences of ambiguities in the database and revise
  - nearly all ambiguities should be removed

The changes (III)

- remove kY – conditionals
  - some of them are conjunctions, some rather particles
  - divide into these classes
- remove kA – abbreviations
  - very heterogenous class, kA is not that crucial information
  - no special syntactic or semantic information
  - divide into other classes
- add kI – punctuation
  - used in `desamb` tagger
  - previously there were no tags for punctuation
  - `x` subclassification

The changes (IV)

- gender in genitive, dative, local, instrumental of plural
  - originally, we wanted to add gender to all numerals
  - then it showed that there is no gender differences in these cases in all adjectives, pronouns and numerals
  - do not determine gender in these cases
- canonical ordering
  - not defined explicitly before
  - we made it a part of the standard
  - tags can be processed as strings

Problematic issues

- Cancel numerals and pronouns?
  - behaviour very similar to nouns and adjectives
  - slight differences
  - we were unable to agree
- Gender of numerals
  - different syntactic behaviour
  - ?

Google Tagset

- Universal tagset
  - 12 tags
  - 22 languages
- Mapping

universal tag	description	attributive tags
VERB	verbs (all tenses and modes)	k5.*
NOUN	nouns (common and proper)	k1.*
PRON	pronouns	k3.*
ADJ	adjectives	k2.*, k4.*xO, k4.*xR
ADV	adverbs	k6.*
ADP	adpositions (prepositions and postpositions)	k7.*
CONJ	conjunctions	k8.*
DET	determiners	(none)
NUM	cardinal numbers	k4.*xC
PRT	particles or other function words	k9.*
X	other: foreign words, typos, abbreviations	k0
.	punctuation	kI

Conclusions

- Practically motivated changes introduced
- A Standard created
- Thank you for your attention