

# Tagsets, Corpora Annotation

## ESLLI 2013: Computational Morphology

Jirka Hana & Anna Feldman

# Overview

- Tagsets
  - Types of tagsets
  - Tagset size
  - Harmonization of tags
- Corpora annotation

# Tags and tagsets

- (*Morphological*) tag is a symbol encoding (morphological) properties of a word.
- *Tagset* is a set of tags.

# Tagset size

The size of a tagset depends on a particular application as well as on language properties.

- 1 Penn tagset (Am.English): 36 tags; VBD – verb in past tense
- 2 The Lancaster-Oslo-Bergen Corpus (LOB) (Br.English): 132 tags
- 3 Czech positional tagset: about 4000 tags; VpNS---XR-AA--- (verb, participle, neuter, singular, any person, past tense, active, affirmative)

# Types of tagsets

There are many ways to classify morphological tagsets. For our purposes, we distinguish the following three types:

- ① atomic (*flat* in (Cloeren 1993)) – tags are atomic symbols without any formal internal structure (e.g., the Penn TreeBank tagset, (Marcus et al 1993)).
- ② structured – tags can be decomposed into subtags each tagging a particular feature.
  - ① compact: Czech Compact tagsets (Hajic 2004)
  - ② positional – e.g., Czech Positional tagset (Hajic 2004), MULTEXT-East (Erjavec 2004, 2009, 2010)

# Penn Treebank Tagset

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	and, but, or	SYM	Symbol	+,%, &
CD	Cardinal number	one, two, three	TO	'to'	to
DT	Determiner	a, the	UH	Interjection	ah, oops
EX	Existential	'there' there	VB	Verb, base form	eat
FW	Foreign word	mea culpa	VBD	Verb, past tense	ate
IN	Preposition/sub-conj	of, in, by	VBG	Verb, gerund	eating
JJ	Adjective	yellow	VBN	Verb, past participle	eaten
JJR	Adj., comparative	bigger	VBP	Verb, non-3sg pres	eat
JJS	Adj., superlative	wildest	VBZ	Verb, 3sg pres	eats
LS	List item marker	1, 2, One	WDT	Wh-determiner	which, that
MD	Modal	can, should	WP	Wh-pronoun	what, who
NN	Noun, sing. or mass	llama	WP\$	Possessive wh-	whose
NNS	Noun, plural	llamas	WRB	Wh-adverb	how, where
NNP	Proper noun, singular	IBM	\$	Dollar sign	\$
NNPS	Proper noun, plural	Carolinas	#	Pound sign	#
PDT	Predeterminer	all, both	"	Left quote	(' or ")
POS	Possessive ending	's	"	Right quote	(' or ")
PP	Personal pronoun	I, you, he	(	Left parenthesis	( [, (, {, < )
PP\$	Possessive pronoun	your, one's	)	Right parenthesis	( ), }, > )
RB	Adverb	quickly, never	,	Comma	,
RBR	Adverb, comparative	faster	.	Sentence-final punc	(. !?)
RBS	Adverb, superlative	fastest	:	Mid-sentence punc	(: ; ... '-)
RP	Particle	up, off			

# Structured tagsets

- Any tagset capturing morphological features of richly inflected languages is necessarily large.
- A natural way to make them manageable is to use a *structured system*.
- In such a system, a tag is a composition of tags each coming from a much smaller and simpler atomic tagset tagging a particular morpho-syntactic property (e.g., gender or tense).

# Structured tagset: benefits

- 1 *Learnability*
- 2 *Systematic description*
- 3 *Decomposability*
- 4 *Systematic evaluation*

It is trivial to view a structured tagset as an atomic tagset (e.g., by assigning a unique natural number to each tag), while the opposite is not true.



# Structured tagsets: MULTEXT-East

## MULTEXT-East Tagset V.4 (Erjavec 2010):

- 13 languages: English, Romanian, Russian, Czech, Slovene, Resian, Croatian, Serbian, Macedonian, Bulgarian, Persian, Finno-Ugric, Estonian, Hungarian.
- Positions' interpretations vary across different parts of speech.
  - For instance, for nouns, position 2 is Gender, whereas for verbs, position 2 is VForm, whose meaning roughly corresponds to the mood.
- For example:
  - Ncmsn noun, common, masculine, singular, nominative;
  - Ncmsa--n noun, common, masculine, singular, accusative, indefinite, no clitic, inanimate.

# Structured > Positional tagsets: Czech Positional Tagset

- Tags are sequences of values encoding individual morphological features.
- All tags have the same length, encoding all the features distinguished by the tagset.
- Features not applicable for a particular word have a N/A value.
- The – value meaning N/A or not-specified is possible for all positions except the first two (POS and SubPOS).
- SubPOS generally determines which positions are specified (with very few exceptions).

# Czech positional tagset (cont.)

Position	Name	Description	Example <i>vidělo</i> 'saw'	
1	POS	part of speech	V	verb
2	SubPOS	detailed part of speech	p	past participle
3	gender	gender	N	neuter
4	number	number	S	singular
5	case	case	--	n/a
6	possgender	possessor's gender	--	n/a
7	possnumber	possessor's number	--	n/a
8	person	person	X	any
9	tense	tense	R	past tense
10	grade	degree of comparison	--	n/a
11	negation	negation	A	affirmative
12	voice	voice	A	active voice
13	reserve1	unused	--	n/a
14	reserve2	unused	--	n/a
15	var	variant, register	--	basic variant

# Czech Positional Tagset: Wildcards

Wildcards are values that cover more than one atomic value.

# Czech Positional Tagset: Wildcards

Wildcards are values that cover more than one atomic value.

---

Atomic values:

F	feminine
I	masculine inanimate
M	masculine animate
N	neuter

Wildcard values:

X	M, I, F, N	any of the basic four genders
H	F, N	feminine or neuter
T	I, F	masculine inanimate or feminine (plural only)
Y	M, I	masculine (either animate or inanimate)
Z	M, I, N	not feminine (i.e., masculine animate/inanimate or neuter)
Q		feminine (with singular only) or neuter (with plural only)

---

# Czech Positional Tagset: SubPOS position

- SubPOS values do not always encode the same level of detail:  
E.g., personal pronouns:
  - P (regular personal pronoun),
  - H (clitical personal pronoun), and
  - 5 (personal pronoun in prepositional form).
- E.g., There are eight values corresponding to relative pronouns, four to generic numerals, etc.
- It is a trade off between complexity of the tagset and linguistic adequacy

# The Russian positional tagset

Pos	Abbr	Name	Nr. of values
1	p	Part of Speech	12
2	s	SubPOS (Detailed Part of Speech)	42
3	g	Gender	4
4	y	<b>Animacy</b>	3
5	n	Number	3
6	c	Case	7
7	f	Possessor's Gender	4
8	m	Possessor's Number	2
9	e	Person	4
10	r	<b>Reflexivity</b>	2
11	t	Tense	4
12	b	<b>Verbal aspect</b>	3
13	d	Degree of comparison	3
14	a	Negation	2
15	v	Voice	2
16	i	Variant, Abbreviation	7

# Tagset size and tagging accuracy

- Tagsets for highly inflected languages are typically far bigger than those for English.
- It might seem obvious that the size of a tagset would be negatively correlated with tagging accuracy: for a smaller tagset, there are fewer choices to be made, thus there is less opportunity for an error.



# Tagset size and tagging accuracy

- Tagsets for highly inflected languages are typically far bigger than those for English.
- It might seem obvious that the size of a tagset would be negatively correlated with tagging accuracy: for a smaller tagset, there are fewer choices to be made, thus there is less opportunity for an error.
- (Elworthy 1995) shows this is not true.

# External and Internal Criteria for Tagset Design (Elworthy 1995)

- *External criterion*: the tagset must be capable of making the linguistic distinctions required in the output corpora;
- *Internal criterion*: make the tagging as effective as possible;

# Harmonizing tagsets across languages?

- e.g., MULTEXT-East, CLiC-TALP (Torruella 2002)
- What are the advantages and disadvantages?

# Harmonization: Pros

- Harmonized tagsets make it easier to develop multilingual applications or to evaluate language technology tools across several languages.
- Interesting from a language-typological perspective as well because standardized tagsets allow for a quick and efficient comparison of language properties.
- Convenient for researchers working with corpora in multiple languages – they do not need to learn a new tagset for each language.

# Harmonization: Cons

- Various grammatical categories and their values might have different interpretations in different languages.
  - E.g., definiteness is expressed differently in various languages: determiners in English, clitics in Romanian; only pronominal adjectives in Lithuanian etc.
  - E.g., plural: in Russian, only plural; in Slovenian, dual and plural.

## Summary: Tagset design challenges

- Tagset size: computationally tractable? Linguistically adequate?
- Atomic or Structural? If Structural, compact or positional?
- What linguistic properties are relevant?
  - The PDT Czech tagset mixes the morpho-syntactic annotation with what might be called dictionary information, e.g., gender;
  - The Czech tagset sometimes combines several morphological categories into one.
  - The Penn Treebank tagset has many singleton tags (e.g., infinitive *to*, punctuation).
- Should the system be standardized and be easily adaptable for other languages?

## Some annotation problems

- Truly ambiguous text.

*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)

*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO  
NOMINATIVES)

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names).



## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)
- Diachronic corpus? Lemmas change over time.  
*kóň, kuoň, kůň.* 'horse' ALL HORSE BUT AT DIFF TIME

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)
- Diachronic corpus? Lemmas change over time.  
*kóň, kuoň, kůň.* 'horse' ALL HORSE BUT AT DIFF TIME
- Clitics  
*Jste-li pojištění ...* 'If you are insured ...'

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)
- Diachronic corpus? Lemmas change over time.  
*kón, kuoň, kůň.* 'horse' ALL HORSE BUT AT DIFF TIME
- Clitics  
*Jste-li pojištění ...* 'If you are insured ...'  
*Tys to viděl?* 'You haven't seen it?'

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)
- Diachronic corpus? Lemmas change over time.  
*kón, kuoň, kůň.* 'horse' ALL HORSE BUT AT DIFF TIME
- Clitics  
*Jste-li pojištění ...* 'If you are insured ...'  
*Tys to viděl?* 'You haven't seen it?'  
*Náms to nedal.* 'You did not give it to us'

## Some annotation problems

- Truly ambiguous text.  
*Mám rád maso na šalvěji.* – M or F? (TWO GENDERS)  
*Je to v kuchyni.* – lemma *kuchyně* or *kuchyň*? (TWO NOMINATIVES)
- Expressions from other languages (company names, logos, song names, DJ's, horse names). What if it is used differently in the corpus language? (*La Manche* – fem in French, masc in Czech)
- Text with errors, text of foreigners, text which is hard to understand.  
*Tak že mislim, že kdy by byl sebe svim dítě, ...* (MANY ERRORS)
- Diachronic corpus? Lemmas change over time.  
*kón, kuoň, kůň.* 'horse' ALL HORSE BUT AT DIFF TIME
- Clitics  
*Jste-li pojištění ...* 'If you are insured ...'  
*Tys to viděl?* 'You haven't seen it?'  
*Náms to nedal.* 'You did not give it to us'

# Cohen's kappa (Cohen 1960)

- The most popular measure of agreement between two annotators.
- Takes into account (somewhat) the possibility of chance agreement.

$$\bullet \kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

$Pr(a)$  - the relative observed agreement

$Pr(e)$  - the hypothetical probability of chance agreement

$$Pr(e) = \sum_t \frac{t_a * t_b}{N}$$

$t_a$  - number of tags  $t$  assigned by annotator  $a$

$N$  - number of all tags

- Weighted kappa – gives different weights to different errors .



- (Variant of) Kendall's tau - the minimal number of operations necessary to turn one annotation into the other.
- There are other measures.
- High agreement is important but it is not everything:
  - One can use use a tagset with a single tag.
  - The annotation manual can be purely formal (Tag all sentence initial words as topics).
  - On the other hand, if iaa is below the accuracy of a tagger ...