Lionbridge

Natural Language Solutions Group,
Tampere, Finland

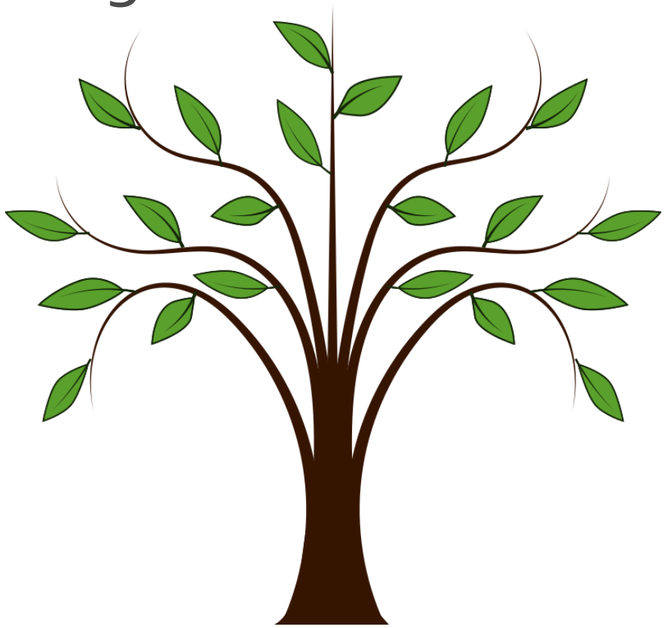
Stanford Typed Dependencies: Slavic Languages Application

Katarzyna Marszałek-Kowalewska, Anna Zaretskaya, Milan Souček

PolTal 2014, 17th-19th September 2014

Stanford Typed Dependencies: Slavic Languages Application

- Related to the treebank acquisition for the Universal Treebank Project for Google
- Based on the Stanford Dependency guidelines preparation for Polish and Russian with experience from preparing treebanks for some other languages for the Universal Treebank Project



Dependency grammar

- Sentence structure is determined by the relation between a head (word) and its dependents (words) in the sentence (unlike e.g. in constituency grammar, where phrases are the basic units)
- Dependency parsing became popular in last two decades and it is used in NLP tasks (e.g. machine translation or search methods)
- Dependency treebanks are already available for many languages, including also Slavic languages

Dependency treebanks

- Existing treebanks follow different theories and use different annotation representations

- Need for multilingual dependency treebank data in a unified format

Multilingual dependency treebanks

- Multilingual treebanks are built using different methods
  - Preparing new treebanks from scratch – expensive, time consuming
  - Parsing parallel corpora – limited text resources
  - Converting existing treebanks to consistent and unified schema

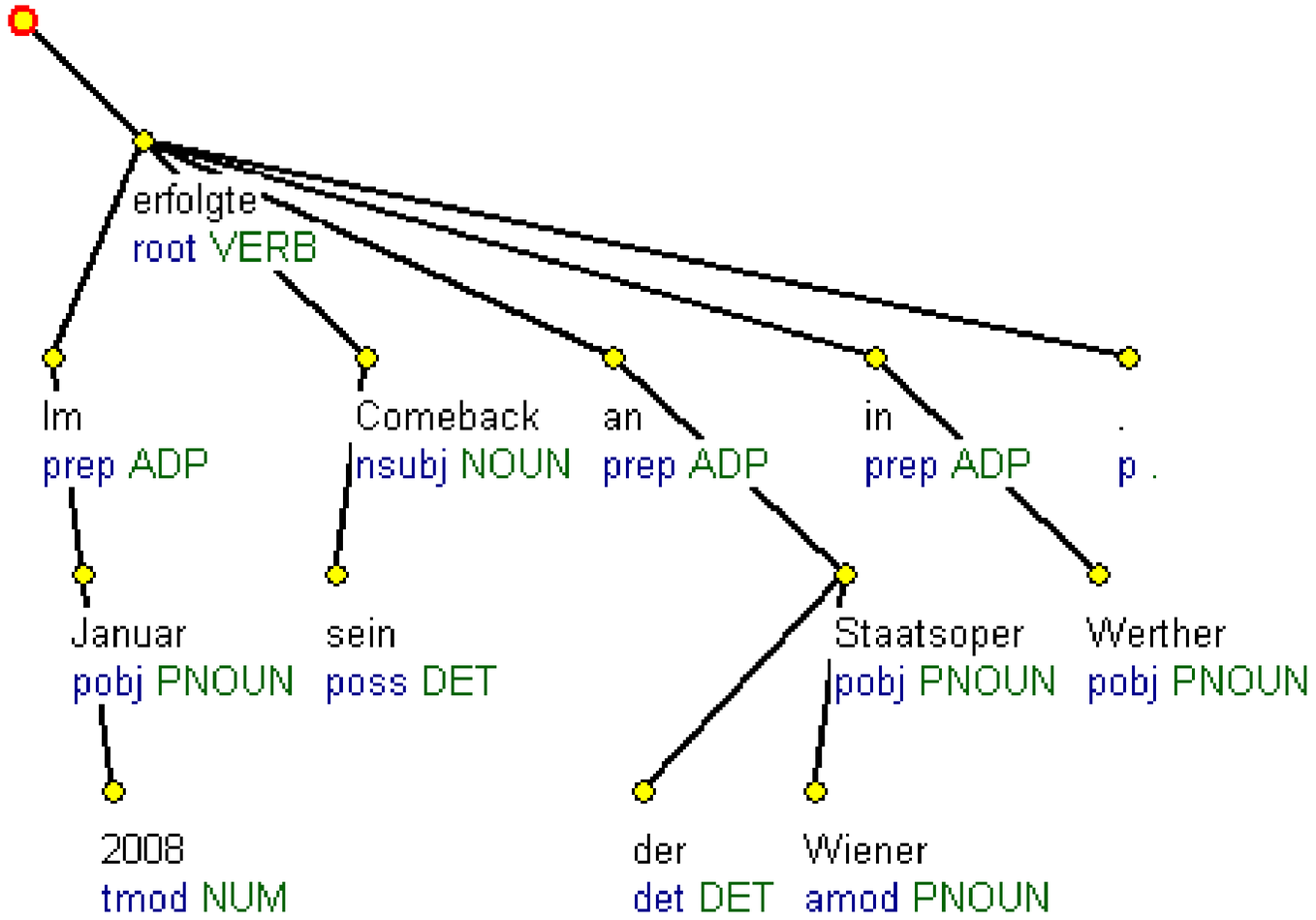
Existing sets of multilingual treebanks:

- The Universal Dependency Treebank project – 11 languages
- HArmonized Multi-LanguagE Dependency Treebank (HamleDT) – 30 treebanks

The Universal Dependency Treebank Project

- Project sponsored by Google: https://code.google.com/p/uni-dep-tb/
- A set of dependency treebanks for multiple languages annotated in basic Stanford-style dependencies
- Current Beta version has 11 languages (Brazilian-Portuguese, English, Finnish, French, German, Italian, Indonesian, Japanese, Korean, Spanish and Swedish)
- Consistent annotation schema
- POS – Petrov et al. 2012
- Dependency relations (deprel) – McDonald et al. 2013

The Universal Dependency Treebank Project



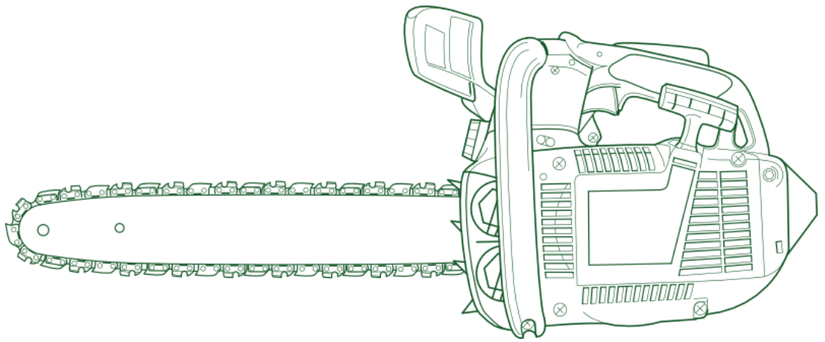
Stanford Typed Dependencies (SD)

- de Marneffe and Manning: Stanford typed dependencies manual. 2008.
- The SD representation was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations.

- Binary relation between two sentence words, semantically contentful and useful to applications.
- Where possible, relations should use notions of traditional grammar for easier comprehension by users.
- Underspecified relations should be available to deal with the complexities of real text.
- Where possible, relations should be between content words, not indirectly mediated via function words.
- The representation should be spartan rather than overwhelming with linguistic details.

SD - challenges

- The original version built for English and based on English data
- With more languages being added, new features need to be dealt with
- Most appropriate linguistic analysis vs. consistency with other languages
- Currently several different versions. With growing amount of languages involved, SD schema being reviewed and updated



SD – Slavic Languages Peculiarities

- Determiners and predeterminers
- Reflexive pronouns
- Copular verb ellipsis
- Subjects in genitive and dative
- Genitive constructions

Determiners and predeterminers

- Original SD model:

En: All the boys are here *predet*(boys, all)

- Slavic application:

Pol: Wszyscy ci ludzie “All those people” *predet*(ludzie, wszyscy)

Pol: Ci wszyscy ludzie “All those people” *predet*(ludzie, wszyscy)

In case of 2 determiners, we analyze demonstrative pronoun as *det* and interrogative pronoun as *predet*, even if the word order is switched.

Reflexive pronoun

- Object-like reflexive pronoun analyzed as an object:

Pol: Piotr umył się “Peter washed himself” *dobj*(umył, się)

- Reflexive marker of the verb analyzed as *prt*, following the English SD analysis of phrase verbs:

Pol: uśmiechnąć się “smile (oneself)” *prt*(uśmiechnąć, się)

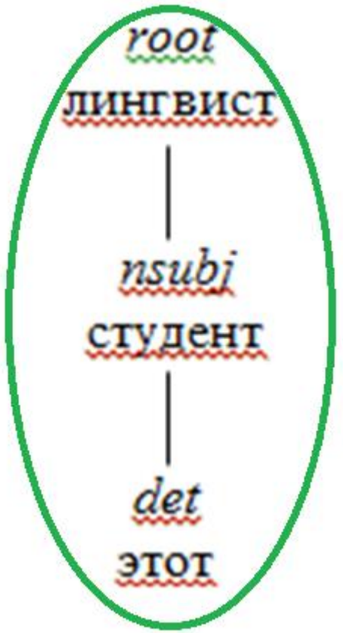
En: They broke up *prt*(broke, up)

- In Russian reflexive is attached to the verb:

Rus: Она одевается “She dresses herself”

Copula elision

- Original SD – content word is head: Bill is big *nsubj*(big, Bill); *cop*(big, is)
- Multilingual SD – function word is head: Bill is big *scomp*(is, big); *nsubj*(is, Bill)
- In Russian, copula verb is elided. The verb complement becomes the root:



Subject in genitive and dative

- A genitive NP depending on the predicative word нет, which in fact is the negated copular verb " to be " in present tense. This also works for all negated structures with verbs expressing existence:

Rus: Здесь никого нет "There is nobody" *nsubj*(нет, никого)

- A dative NPs that depend on PRED words or adverbs with predicative function:

Rus: Мне нужно уйти "I need to go" *nsubj*(нужно, Мне)

- A dative NPs with impersonal verbs in third person singular:

Rus: Ему пришлось подчиниться "He had to obey" *nsubj*(пришлось, Ему)

Genitive construction

- New deprel *gmod* added to handle cases, where a noun in genitive modifies another noun

Rus: ножка стула "leg of the chair" *gmod*(ножка, стула)

Cz: zastávka autobusu "bus stop" *gmod*(*zastávka* , *autobusu)*

- Other genitives analysed using existing relevant labels:
- Negative constructions

Rus: Я не заметил водки "I didn't notice any vodka(*gen*)" *dobj* ((не) заметил, водки)

Genitive constructions cont.

- Partitive constructions

Rus: Он дал мне денег “He gave me money(*gen*)” *dobj*(дал, денег*)*

- Genitive of quantification:

Rus: Я купил пять машин “I bought five cars(*gen*)”

Rus: Я купил много машин “I bought many cars(*gen*)”

We treat these cases in a such way that noun is always the head of the numeral/quantifier, even if the case actually is assigned by the numeral/quantifier to the noun, e.g. *num*(машин, пять), *advmod*(машин, много).

Universal Stanford Dependencies

- Several different versions of SD schema:
- de Marneffe, Manning 2008
- McDonald et al. 2013 (GSD)
- de Marneffe et al. 2013
- de Marneffe et al. 2014 (USD) – released after our article was submitted

Universal (Stanford) Dependencies: http://universaldependencies.github.io/docs/

- An online documentation and example bank for Universal Dependencies.

Universal Stanford Dependencies - Harmonization

- Although some labels outdated, they can be easier converted to the newer SD model
- In the Universal Treebank Project, harmonization post-processing applied, so treebanks are made more consistent
- With proper documentation, further adjustments are easily possible
- With more languages analyzed and added to the project, SD guidelines need updates. The online guidelines make this process easier and accessible to different teams

Universal Stanford Dependencies - Harmonization

Deprel	Gloss	GSD	USD
agent	Agent	adpmod	case
complm	Clausal complement marker	mark	mark
gmod	Genitive modifier	poss	poss
nn	Noun compound modifier	compmod	compound/name
num	Numeric modifier	num	nummod
number	Element of compound number	num	nummod
ocomp	Object complement	acomp/attr/cop	cop/xcomp
pobj	Prepositional object	adpobj	nmod
poss	Possession modifier	poss	case/poss
predet	Predeterminer	det	predet
prep	Prepositional modifier	adpmod	case
prt	Phrasal verb particle	prt	prt
purpcl	Purpose clause modifier	advcl	advcl
quantmod	Quantifier phrase modifier	advmod	advmod
scomp	Subject complement	acomp/attr/cop	cop/xcomp
tmod	Temporal modifier	advmod	advmod/tmod

Universal Stanford Dependencies - Future

- More languages and domains are added and released to public
- SD schema further improved in order to serve multilingual purposes
- Established multilingual resource for NLP tasks

Lionbridge

Natural Language Solutions Group
Tampere, Finland

Thank you!

Contacts:

milan.soucek@lionbridge.com

k.marszalek.kowalewska@gmail.com

azaretskaya@gmail.com

www.lionbridge.com

http://blog.lionbridge.com

http://twitter.com/Lionbridge

http://www.facebook.com/Lionbridge