Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian

Željko Agić and Danijela Merkler

University of Zagreb
Faculty of Humanities and Social Sciences

2013-06-13
(TSD 2013, Plzeň, Czech Republic, September 2013)

Motivation

- original HOBS tagset
  - basic
    - Atr, Adv, AdvAtr, Apos, AtrAdv, AtrObj, Atv, AtvV, AuxC, AuxG, AuxK, AuxO, AuxP, AuxR, AuxT, AuxV, AuxX, AuxY, AuxZ, Coord, ExD, Obj, ObjAtr, Pnom, Pred, Sb
  - extended
    - Adv_Ap, Adv_Co, Adv_Pa, AdvAtr_Co, Apos_Ap, Apos_Co, Apos_Pa, Atr_Ap, Atr_Co, Atr_Pa, AtrAdv_Co, AtrObj_Ap, Atv_Co, Atv_Pa, AtvV_Co, AtvV_Pa, AuxC_Ap, AuxC_Co, AuxC_Pa, AuxP_Ap, AuxP_Co, AuxP_Pa, AuxV_Co, AuxY_Pa, AuxZ_Co, AuxZ_Pa, Coord_Ap, Coord_Co, Coord_Pa, ExD_Ap, ExD_Co, ExD_Pa, Obj_Ap, Obj_Co, Obj_Pa, Pnom_Ap, Pnom_Co, Pnom_Pa, Pred_Ap, Pred_Co, Pred_Pa, Sb_Ap, Sb_Co
- improved HOBS tagset
  - adds 1 basic, 11 tags overall (*Sub** tags)
  - enables more explicit annotation
- the case of Slovene DT and JOS Corpus

Motivation

- new treebank
  - based on SETimes.HR corpus
    - manually tokenized, lemmatized and MSD-annotated
    - MTE v4, v5 MSD tagset
  - newspaper text
  - aiming at state-of-the-art dependency parsing
- new tagset
  - Pred, Sb, Obj
  - Adv, Atr, Ap, Prep
  - Atv, Aux, Pnom
  - Co, Sub
  - Elp, Oth, Punc
- compare SETimes.HR treebank with HOBS
- make it publicly available

Treebank stats

treebank	features	sent's	tokens	types	lemmas	MSDs	afuns
HOBS	full without *Sub**	4 626	117 369	25 038	12 388	914	70
	full with *Sub**	4 626	117 369	25 038	12 388	911	81
	basic without *Sub*	4 626	117 369	25 038	12 388	914	27
	basic with *Sub*	4 626	117 369	25 038	12 388	911	28
SETimes.HR	full MSD	2 488	56 334	13 409	6 901	804	15
	reduced MSD	2 488	56 334	13 374	6 943	665	15
	POS	2 488	56 334	13 374	6 943	12	15

Inter-annotator agreement

treebank	features	LAS	UAS	LA	κ(LA)
HOBS	full with *Sub**	78.89	89.16	84.07	0.839
HOBS	basic with *Sub*	82.05	89.16	88.83	0.884
SETimes.HR	full MSD	86.11	91.29	92.51	0.920

HOBS basic with *Sub*			SETimes.HR		
afun pair	frequency	pct	afun pair	frequency	pct
Obj Adv	48	17.65	*Obj Atr*	24	15.09
Obj Atr	18	6.62	*Adv Oth*	19	11.95
Sb ExD	11	4.04	*Obj Adv*	16	10.06
AuxG ExD	8	2.94	*Adv Atr*	11	6.92
Sb Atr	8	2.94	*Pnom Pred*	8	5.03
Adv Atr	7	2.57	*Pred Aux*	8	5.03
Atr Sb	7	2.57	*Pnom Sb*	4	2.52
other	165	60.67	other	69	43.40

Experiment

- follows CoNLL 2006 and 2007 guidelines
  - treebanks split into tenfold training and testing sets
  - 5.000 tokens or 200 sentences per test set
- used only MSTParser
  - free word order languages favor graph-based approaches
    - non-local dependencies, non-projectivity
  - non-projective MST parsing: `decode-type:non-proj`
  - second order features: `order:2`
- goals
  - observe difference between treebanks
    - special emphasis on main categories: *Pred*, *Sb*, *Obj*
    - tagset influence on data-driven parsing
  - not a parser investigation

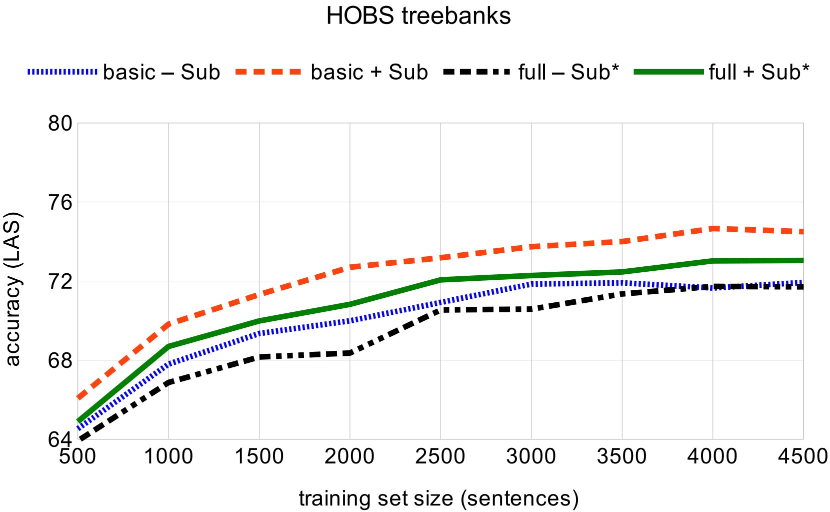
Overall accuracy

treebank	features	LAS	UAS	LA
HOBS	full without *Sub**	71.71	80.34	81.75
	full with *Sub**	73.04	81.10	82.85
	basic without *Sub*	71.93	79.98	84.65
	basic with *Sub*	74.50	81.41	86.87
SETimes.HR	full MSD	77.13	83.08	88.82
	reduced MSD	77.49	83.58	89.00
	POS	74.56	81.59	85.87

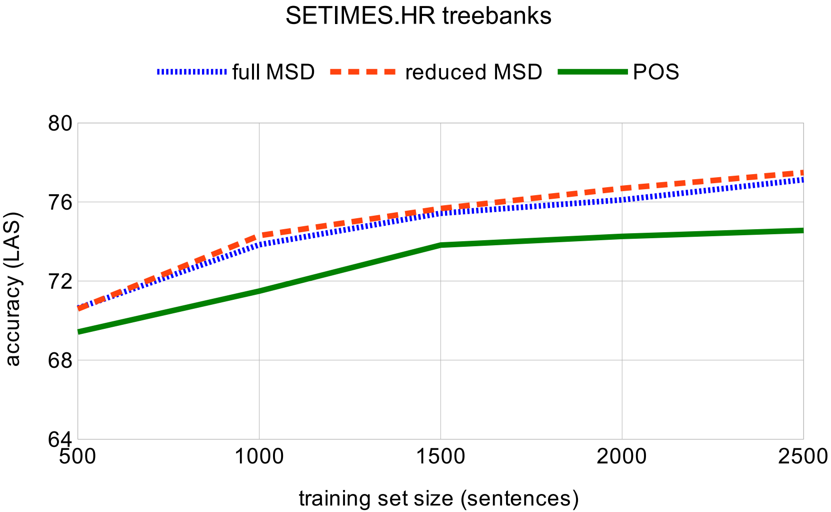
Accuracy for matching tags

	HOBS without *Sub*			HOBS with *Sub*			SETimes.HR		
afun	LAS	UAS	pct	LAS	UAS	pct	LAS	UAS	pct
Adv	65.88	84.81	9.98	68.33	88.33	8.99	61.57	84.72	4.72
Ap – Apos	38.10	47.62	0.64	36.84	42.11	0.64	89.60	92.00	3.05
Atr	81.61	88.29	28.7	83.06	89.18	25.8	80.75	88.39	26.5
Co – Coord	48.21	49.23	4.15	56.85	59.39	4.18	46.00	48.00	2.87
Obj	62.81	79.40	8.39	70.06	87.65	6.53	74.10	89.76	7.25
Pnom	58.73	80.95	1.51	60.61	77.27	1.74	65.75	73.97	2.03
Pred	65.89	72.87	4.76	80.69	82.19	9.29	86.58	88.10	9.32
Prep – AuxP	69.85	70.50	9.28	71.54	71.94	9.99	74.04	75.11	9.44
Sb	68.85	81.26	7.84	73.99	82.37	7.01	75.56	82.87	6.61
Sub	–	–	–	72.91	73.89	4.04	65.22	65.76	3.81

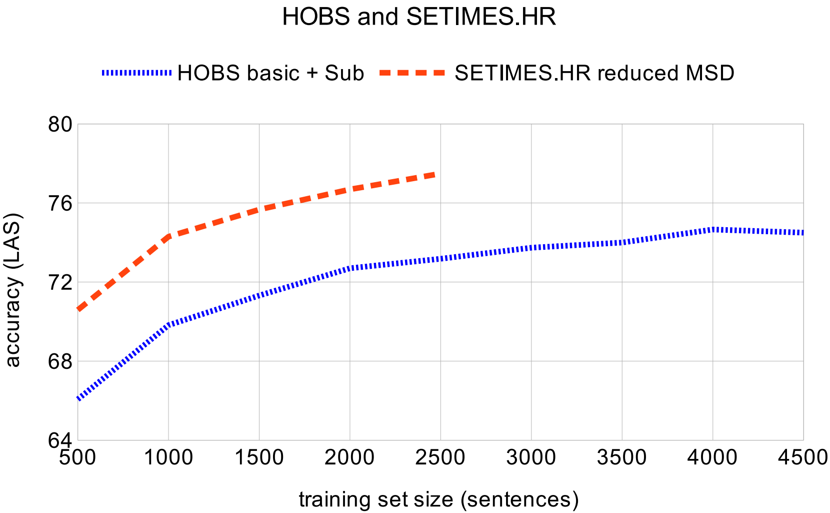
Learning curves



Learning curves



Learning curves



Conclusions and future work

- work done
  - created a new treebank of Croatian
    - new and simpler syntactic formalism
    - higher inter-annotator agreement, easier annotation
  - parsing models
    - state of the art for Croatian dependency parsing
    - publicly available
    - CC-BY-SA-3.0 licence
    - http://nlp.ffzg.hr/resources/models/
- work underway
  - enlarge the treebank
    - was 2500 sentences, now 3600 sentences
    - resolve annotation inconsistencies
  - try better parsers
    - freely available parsers combining local and non-local features
    - push LAS above 80%, overall and for *Sb*, *Obj*

Thanks! ☺