UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Research Group on Artificial Intelligence

Annotating MWEs in Treebanks: The Case of Hungarian

Veronika Vincze

University of Szeged
Department of Informatics

vinczev@inf.u-szeged.hu

PARSEME meeting, Warsaw, Poland – 18 September 2013

Introduction

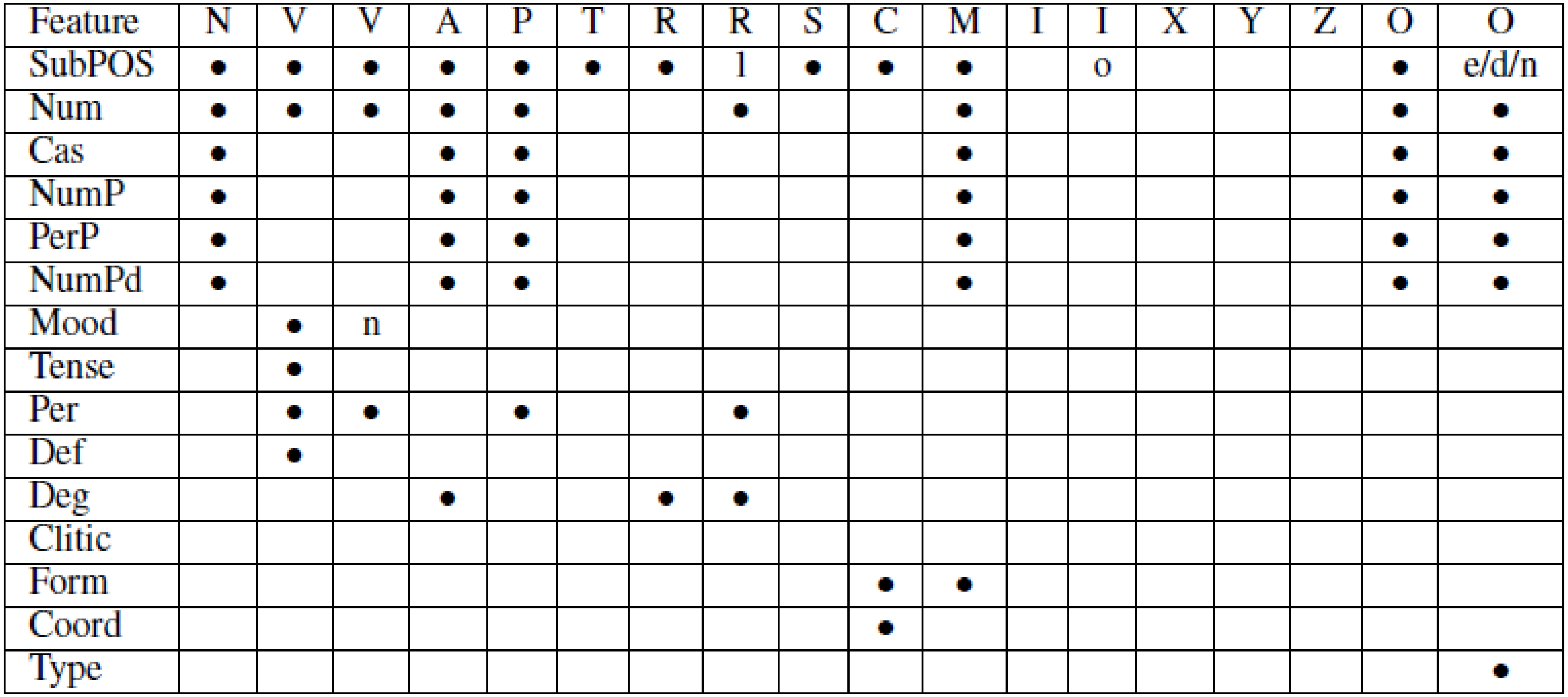
- Why to annotate MWEs in corpora?
  - Gathering real-world linguistic data
  - Training/testing of NLP tools
- how several types of multiword expressions are annotated in Hungarian constituency and dependency treebanks
  - light verb constructions
  - multiword named entities
  - multiword numbers
  - verbs with verbal prefixes

Morphosyntactic features of Hungarian

- Morphologically rich language
- Nominal declination (nouns, adjectives, numerals)
- Verbal conjugation: tense, mood, person, number, definiteness of the object
- Several hundreds of word forms for each lemma
- Grammatical relations encoded primarily by case suffixes:

lánc “chain” – *lánccal* (chain-INS)
“with (a/the) chain”

Morphological features



Feature	N	V	V	A	P	T	R	R	S	C	M	I	I	X	Y	Z	O	O
SubPOS	●	●	●	●	●	●	●	l	●	●	●		o				●	e/d/n
Num	●	●	●	●	●			●			●						●	●
Cas	●			●	●						●						●	●
NumP	●			●	●						●						●	●
PerP	●			●	●						●						●	●
NumPd	●			●	●						●						●	●
Mood		●	n															
Tense		●																
Per		●	●		●			●										
Def		●																
Deg				●			●	●										
Clitic																		
Form										●	●							
Coord										●								
Type																		●

Hungarian word order

- No fixed word order
- Information structure is reflected in word order (theme-rheme, old-new)

Péter szereti Marit. Peter love-3SgObj Mary-ACC
'Peter loves Mary.'

Péter Marit szereti. 'It is Mary who Peter loves.'

Marit szereti Péter. 'It is Mary who Peter loves.'

Marit Péter szereti. 'It is Peter who loves Mary.'

Szereti Péter Marit. 'Peter LOVES Mary (and not hates).'

Szereti Marit Péter. 'Peter LOVES Mary (and not hates).'

Word order and LVCs

- Light verb constructions: noun + verb combinations
- May not be adjacent:

Figyelembe vette a döntést.
(consideration-ILL take-PAST3SG the decision-ACC) ‘he took the decision into consideration’

Figyelembe kellett neki **venni** a döntést.
(consideration-ILL must-PAST3SG he-DAT to.take the decision-ACC) ‘he had to take the decision into consideration’

LVCs in the Szeged Treebank

- Szeged Constituency Treebank: the largest Hungarian database which has been manually POS-tagged and annotated for constituency structures (Csendes et al. 2005): 82K sentences, 1.2M tokens
- also manually annotated for light verb constructions (Vincze and Csirik, 2010) based on linguistic criteria, e.g.
  - Variativity: a verbal counterpart can substitute the LVC (*döntést hoz – dönt / make a decision – decide*)
  - Nominal component derived from a verb
- not only object + verb pairs but other cases (subject, oblique) + verb pairs as well
- nominal and participial forms of LVCs also annotated:
  - *figyelembe vétel* (consideration-ILL taking) ‘taking into consideration’
  - *döntéshozatal* (decision.making) ‘making of a decision’ – one word MWE!!!
  - *döntéshozó* (decision.maker/making) ‘(someone) who makes a/the decision’

Syntactic boundaries

- the annotation was carried out on the syntactically annotated treebank
- phrase boundaries were also taken into consideration when marking LVCs
- adjectives and other modifiers of the nominal head may be also included

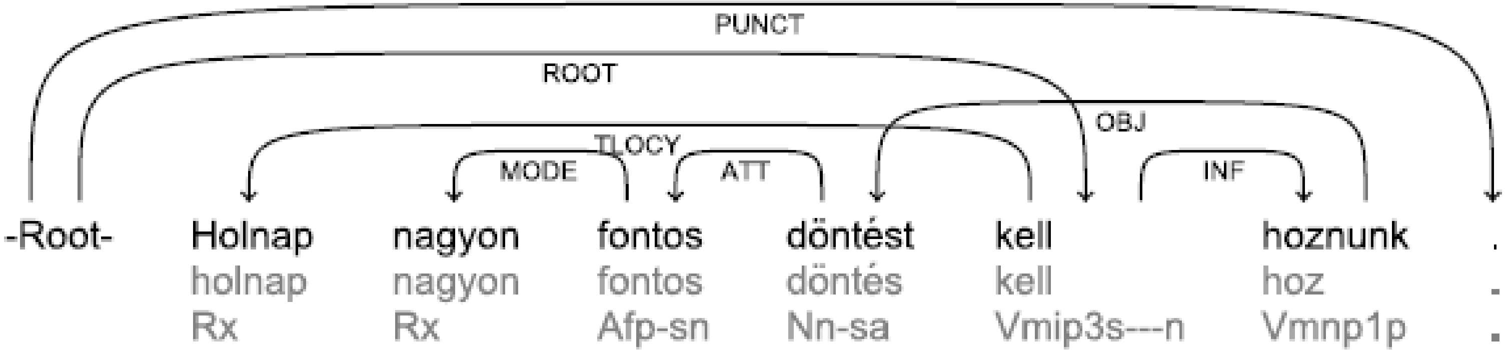
[LVC [NP nehéz döntést] hozott]

difficult decision-ACC make-PAST3SG

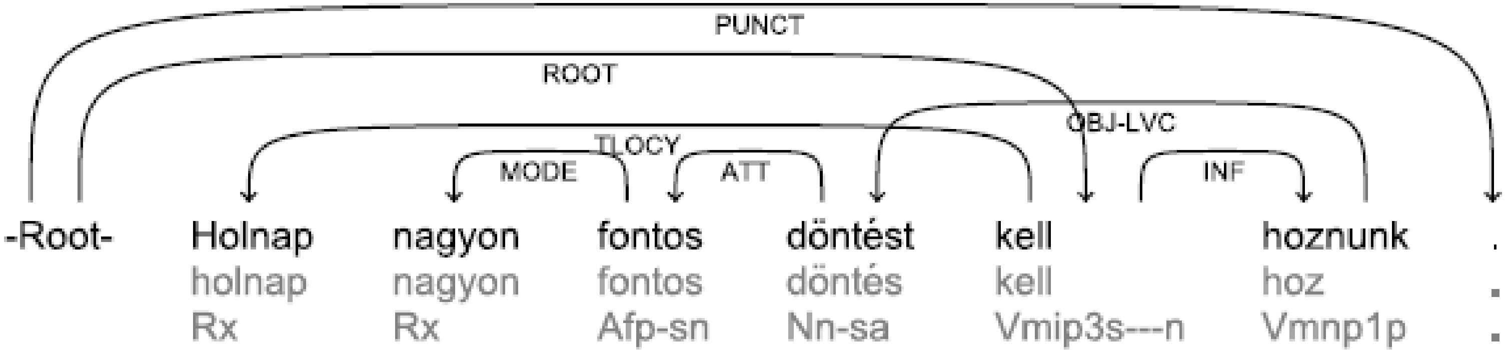
‘he made a difficult decision’

Conversion to dependency format

- manually annotated dependency version of the same treebank (Vincze et al., 2010)
- the two manual annotations (dependency trees and LVCs) were mapped
- dependency relations were enhanced with LVC-specific relations that can be found between the two members of the constructions (no adjectives etc. included)
- e.g. the relation **OBJ-LVC** can be found between the words *döntést* (decision-ACC) and *hoz* “bring” within the LVC *döntést hoz* “to make a decision”



-Root-	Holnap	nagyon	fontos	döntést	kell	hoznunk	.
	holnap	nagyon	fontos	döntés	kell	hoz	.
	Rx	Rx	Afp-sn	Nn-sa	Vmip3s---n	Vmnp1p	.



-Root-	Holnap	nagyon	fontos	döntést	kell	hoznunk	.
	holnap	nagyon	fontos	döntés	kell	hoz	.
	Rx	Rx	Afp-sn	Nn-sa	Vmip3s---n	Vmnp1p	.

Extension of dependency labels

- special attention was paid to the participle forms of LVCs
- for verbal LVCs, it is the dependency label of the noun that is extended with the LVC notation (e.g. OBJ-LVC)
- this is not straightforward in the case of past participles

figyelembe *nem* ***vett*** *szempontok*

"aspects that were not taken into consideration"

figyelembe OBL-LVC

a támogatásról ***hozott döntés***

"the decision made on the support"

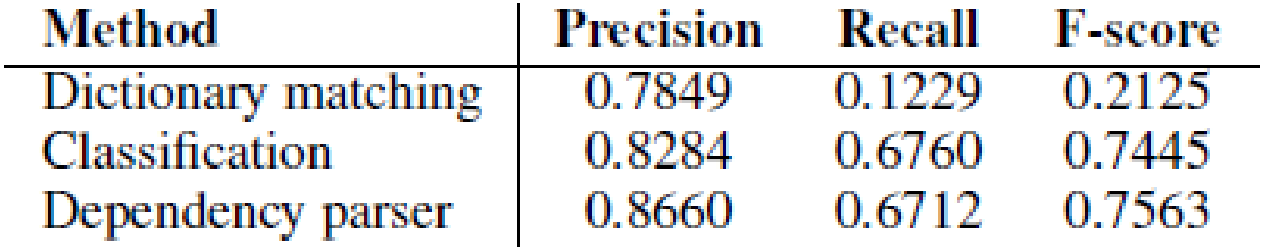
hozott ATT-LVC

- such cases were manually relabeled

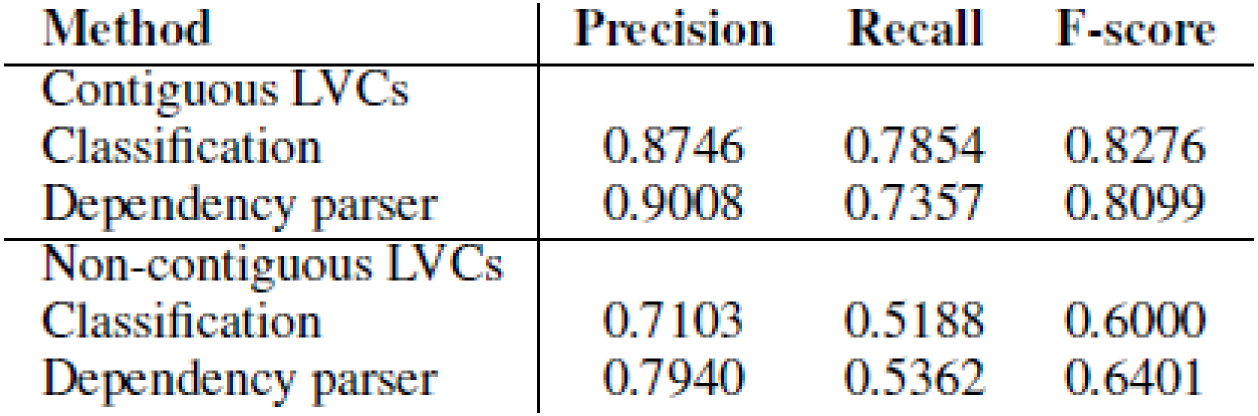
LVC detection by dependency parsing

- excursion to WG3 ☺
- dependency treebank extended with LVC-specific labels
- the Bohnet dependency parser was applied to identify LVCs in the legal subdomain of the corpus
- 10-fold cross validation
- Baseline: state-of-the-art LVC detector for Hungarian

Results



Method	Precision	Recall	F-score
Dictionary matching	0.7849	0.1229	0.2125
Classification	0.8284	0.6760	0.7445
Dependency parser	0.8660	0.6712	0.7563



Method	Precision	Recall	F-score
Contiguous LVCs			
Classification	0.8746	0.7854	0.8276
Dependency parser	0.9008	0.7357	0.8099
Non-contiguous LVCs			
Classification	0.7103	0.5188	0.6000
Dependency parser	0.7940	0.5362	0.6401

- high precision
- adequate treatment of non-contiguous LVCs
- Vincze, Zsibrita and Nagy T. (2013) paper at IJCNLP
- …end of excursion

Multiword named entities & numbers

- multiword named entities (*Coca Cola Ltd.*)
- multiword numbers (*42 million*)
- treated as one token in the Szeged Constituency Treebank
- split into tokens in the dependency treebank
- the last token inherited the morphological analysis of the original multiword unit
- the other tokens were automatically analyzed as a proper noun/cardinal number with default morphological features
- all the previous elements are attached to the succeeding word with an NE relation for named entities and a NUM relation for numbers

Examples

Kovács és társa kft.-nek “for Smith and Co. Ltd.”

Nn-sn Nn-sn Nn-sn Nn-sd

NE NE NE DAT

342 351 000-re “onto 342 351 000”

Mc-snd Mc-snd Mc-ssd

NUM NUM OBL

Dependency vs. constituency

- Different representations
- Constituency:
  - One token
  - Theoretically more desirable (WG1?)
- Dependency:
  - More tokens
  - More realistic from NLP aspects?
  - Tokenizers / MWE detectors should identify them?

Verbs with verbal prefixes

- a morphologically reannotated version of the Szeged Treebank is under construction…
- it will follow the harmonized morphological coding system developed for Hungarian (Farkas et al. 2010)
- the morpheme boundary in the lemmas of verbs with verbal prefixes will be distinctively marked

bejön in.come “to enter” -- *be+jön*

Some statistics

- Empirical data on MWE frequency from corpora
- In the Szeged Treebank:
  - LVC: 6734 occurrences, 1215 lemmas
  - MWNE: 8595 lemmas (49% of NEs)
  - MWNUM: 1870

Multilingual outlook

- Annotating English MWEs and NEs: Wiki50 (Vincze et al. 2011)
- Annotating LVCs in paralell corpora using basically the same guidelines:
  - SzegedParalellFX (Vincze 2012): English-Hungarian
  - JRC-Acquis (under construction): English-Hungarian-Spanish-German

Conclusions

- treatment of several types of Hungarian MWEs in constituency and dependency treebanks:
  - LVCs
  - Multiword NEs
  - Multiword numbers
  - Verbs with verbal prefixes
- it may be beneficial for developing MWE-annotated treebanks for other languages

References

Csendes, Dóra; Csirik, János; Gyimóthy, Tibor; Kocsor, András 2005: The Szeged Treebank. In: Matoušek, Václav et al. (eds.): *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, Karlovy Vary, Czech Republic, September 12-16, 2005, Springer LNAI 3658, pp. 123-131.

Farkas, Richárd; Szeredi, Dániel; Varga, Dániel; Vincze, Veronika 2010: MSD-KR harmonizáció a Szeged Treebank 2.5-ben [Harmonizing MSD and KR codes in the Szeged Treebank 2.5]. In: Tanács, Attila; Vincze, Veronika (eds.): *VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, pp. 349-353.

Vincze, Veronika 2012: Light Verb Constructions in the SzegedParalellFX English-Hungarian Parallel Corpus. In: *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, pp. 2381-2388.

Vincze, Veronika; Csirik, János 2010: Hungarian Corpus of Light Verb Constructions. In: *Proceedings of COLING 2010*, Beijing, China, pp. 1110-1118.

Vincze, Veronika; Nagy T., István; Berend, Gábor 2011: Multiword expressions and Named Entities in the Wiki50 corpus. In: *Proceedings of RANLP 2011*. Hissar, Bulgaria, pp. 289-295.

Vincze, Veronika; Szauter, Dóra; Almási, Attila; Móra, György; Alexin, Zoltán; Csirik, János 2010: Hungarian Dependency Treebank. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Vincze, Veronika; Zsibrita, János; Nagy T., István 2013: Dependency Parsing for Identifying Hungarian Light Verb Constructions. Accepted to: *IJCNLP 2013*.