POS Comes with Parsing: a Refined Word Categorisation Method

Balázs Indig[1,2] & Noémi Vadász[1,3]

[1]MTA–PPKE Hungarian Language Technology Research Group, Budapest, Hungary
[2]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
[3]Pázmány Péter Catholic University, Faculty of Humanities and Social Sciences

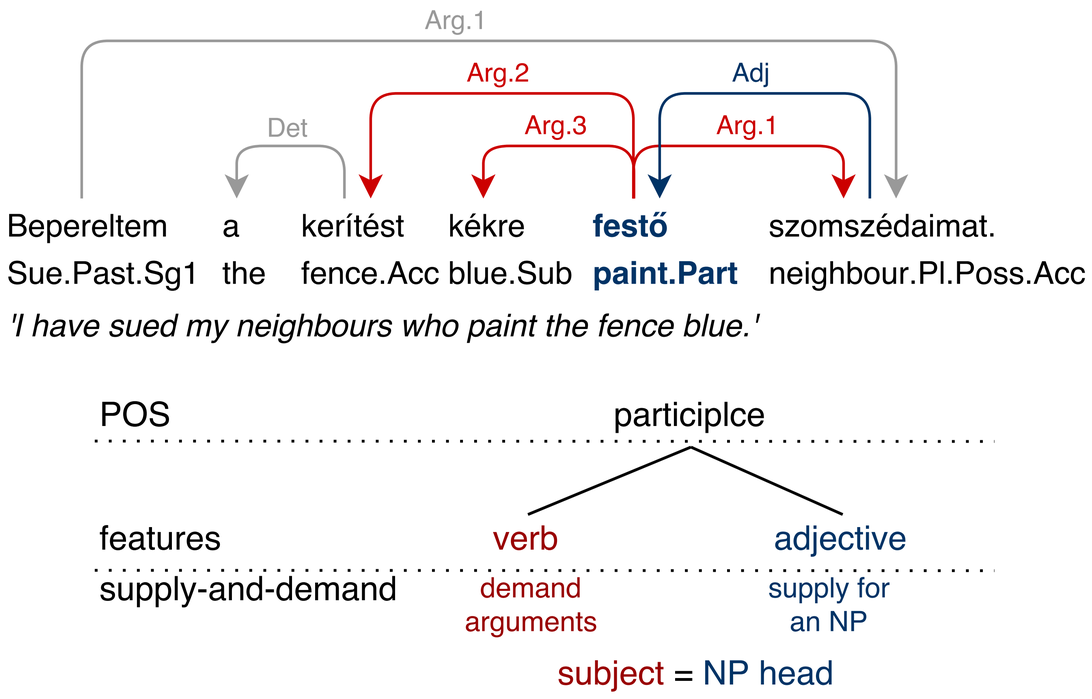
{indig.balazs,vadasz.noemi}@itk.ppke.hu

Introduction

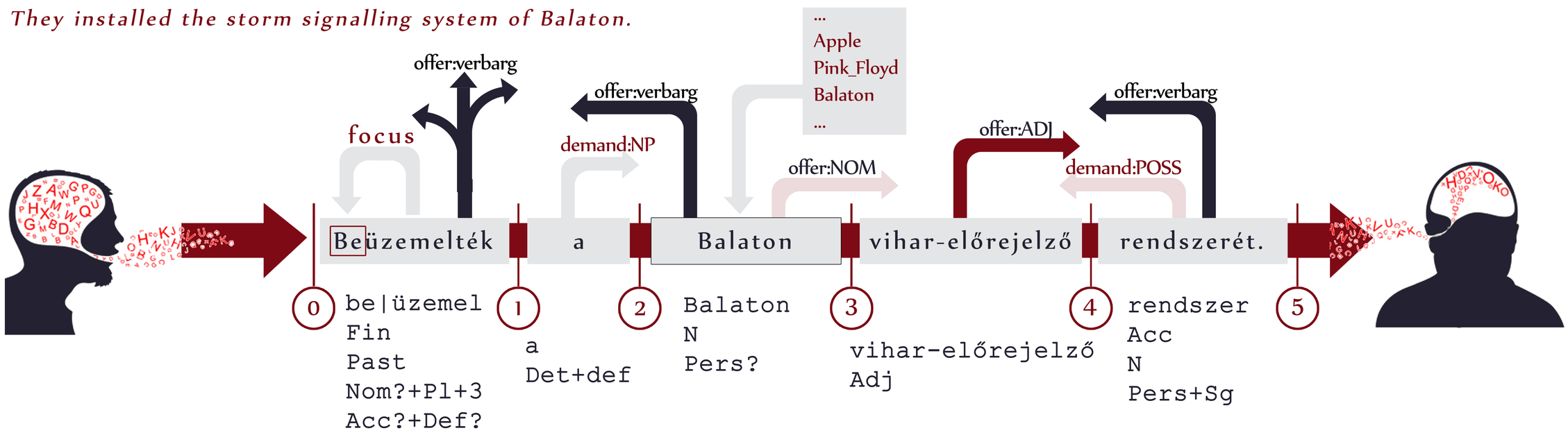
AnaGramma [1, 3] is a **psycholinguistically motivated** parser:

- **left-to-right, word-by-word** approach: as humans do
- **performance-based**: only existing constructions are preferred
- **supply-and-demand framework** [2] based on features:
  - based on lexical representation and morpho-syntactic information
  - like Word-Expert-Parsing [4]: features instead of POS tags
  - *main and subcategories*
    - based on syntactic and semantic features (statistically supported)
    - categories can start different processes during the parsing
    - the main category inherits its processes to the subcategories
    - we can mix features of classical categories within one word

Participle = Verb & Adjective



They installed the storm signalling system of Balaton.



Detached Preverbs or Adverbs

We gathered statistics from two different corpora to model Hungarian.

- Hungarian Gigaword Corpus (HGC), 709 million tokens: both edited and unedited texts from different domains
- InfoRadio Corpus (IRC), 1.953 419 tokens: edited texts of uniform domains, short political news

Positions that can be considered real preverbs:

Frequent preverbs	-2	0	+1	+2	+3
meg, ki, be, le, fel, föl, el, át, rá	0.5%	58.5%	40%	1%	0.01%

98.5% of the preverbs appear on the verb (0) or immediately after (+1) it. The others (+4–) are not real preverbs.

V.FIN	+1	+2	+3	+4	+5	+6	+7
HGC	7.527.308	163.993	5.126	1.193	267	101	27
IRC	23.552	220	-	-	-	-	-
HGC%	97.78%	2.13%	0.0666%	0.015%	0.003%	0.001%	3.5e-4%
IRC%	99.999%	0.001%	-	-	-	-	-

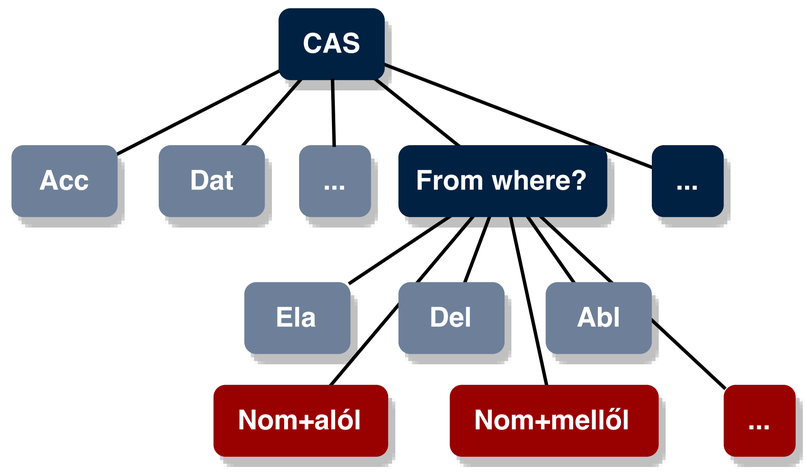
Csábítson téged a retyezáti nagy barna medve **oda** ahova akarod.
Lure you the from_Retyezát big brown bear **there** where you_want_to

'Let the big brown bear from Retyezát **lure** you **wherever** you want to.'

Postpositions as special case markers

Honnan jöttél? **'Where do you come from?'**

case or PP	abbrev.	Hungarian example	translation
elativus	Ela	*A börtönből.*	'From the prison.'
delativus	Del	*A tárgyalásról.*	'From the trial.'
ablativus	Abl	*Az ügyvédtől.*	'From the lawyer.'
postposition	Nom + **alól**	*A hegy **alól**.*	'From **under** the hill.'
postposition	Nom + **mellől**	*A Tisza **mellől**.*	'From **near** the lake Tisza.'
postposition	Nom + **mögül**	*A rács **mögül**.*	'From **behind** the bars.'



The hierarchical system of Hungarian cases and postpositions

Conclusion

We illustrate how the dissected features that build the classical POS categories start supplies and demands, still our method conforms to the existing theories.

[1] Balázs Indig and Gábor Prószéky. Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel. *Alkalmazott nyelvtudomány*, 15(1-2):29–44, 2015.

[2] Balázs Indig and Noémi Vadász. Windows in human parsing – how far can a preverb go? In Tadić Marko and Božo Bekavac, editors, *Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016)*, Dubrovnik, Croatia, sept 2016. (Accepted, In press).

[3] Gábor Prószéky, Balázs Indig, and Noémi Vadász. Performanciaalapú elemző magyar szövegek számítógépes megértéséhez. In Kas Bence, editor, *"Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére*, pages 223–232. MTA Nyelvtudományi Intézet, Budapest, 2016.

[4] Steven L. Small. Word expert parsing. In *Proceedings of the 17th Annual Meeting on Association for Computational Linguistics*, ACL '79, pages 9–13, Stroudsburg, PA, USA, 1979. Association for Computational Linguistics.