POS Comes with Parsing: a Redefined Word Categorisation Method

Balázs Indig, Noémi Vadász

October 11, 2016

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
& Faculty of Humanities and Social Sciences
& MTA–PPKE Hungarian Language Technology Research Group

Motivation: AnaGramma

AnaGramma (Prószéky and Indig, 2015; Prószéky et al., 2016) is a **psycholinguistically motivated** parsing system:

- **left-to-right, word-by-word** approach: as humans do
- **performance-based**: only real-world constructions are relevant
- **supply-and-demand framework** (Indig and Vadász, 2016) based on features:
  - based on lexical representation and morpho-syntactic information
  - like Word-Expert-Parsing (Small, 1979): 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

How does ANAGRAMMA work?

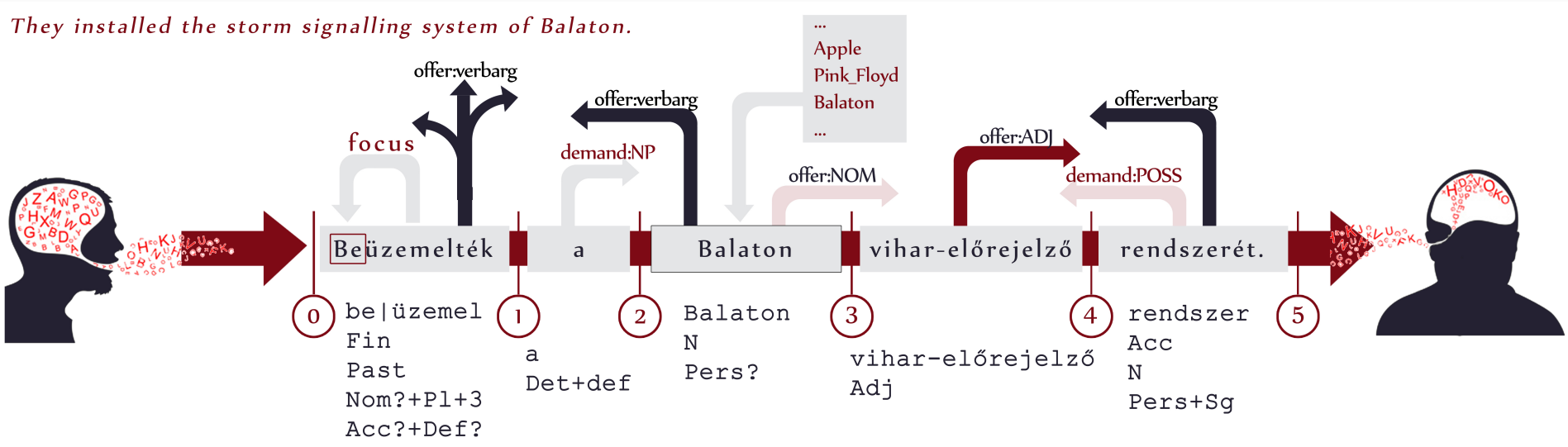

Figure 1: Supplies and demands arise subsequently

Data source: The Corpora

We used two different corpora to model the Hungarian language:

	tokens	contains
Hungarian Gigaword Corpus (HGC)	709 million	non-edited texts
InfoRádió Corpus (IRC)	2 million	edited texts only

The latter corpus contains 2-3 sentence long political and economical news, which is our target domain.

Detached Preverbs or Adverbs I.

Positions that can be considered real preverbs:

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

Table 1: Positions of some frequent preverbs

98.5% of the preverbs appear as prefixes to the verb (0) or immediately after (+1) it. The others (>+4) do not function as real preverbs.

Detached Preverbs or Adverbs II.

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%	-	-	-	-	-

Table 2: Positions of post-verbal detached preverbs

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

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

Participle = Verb & Adjective I.

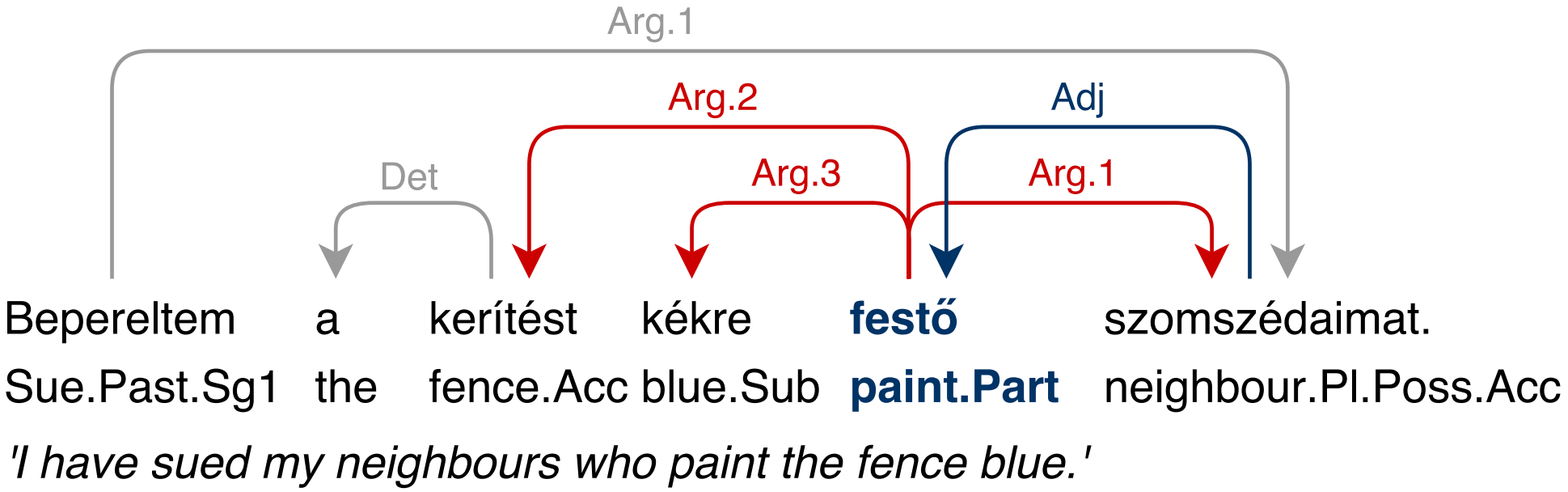

Figure 2: Supplies and demands creating dependency edges

Participle = Verb & Adjective II.

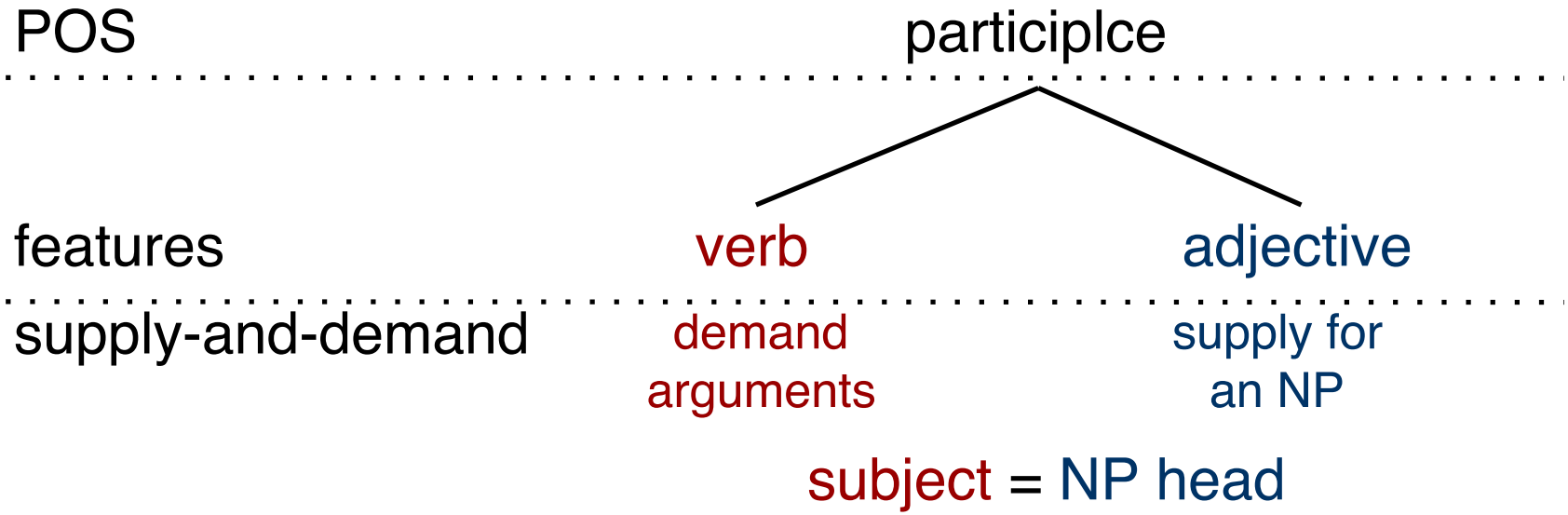

Figure 3: Verbal and adjectival components of participles

Postpositions as case markers I.

Honnan jöttél? ‘Where do you come from?’

case or PP		
elativus	*A börtönből.*	‘From the prison.’
delativus	*A tárgyalásról.*	‘From the trial.’
ablativus	*Az ügyvédtől.*	‘From the lawyer.’
Nom + **alól**	*A hegy **alól**.*	‘From **under** the hill.’
Nom + **mellől**	*A Tisza **mellől**.*	‘From **near** the lake Tisza.’
Nom + **mögül**	*A rács **mögül**.*	‘From **behind** the bars.’

Table 3: Cases and postpositions answering the question ‘From where?’

Postpositions as case markers II.

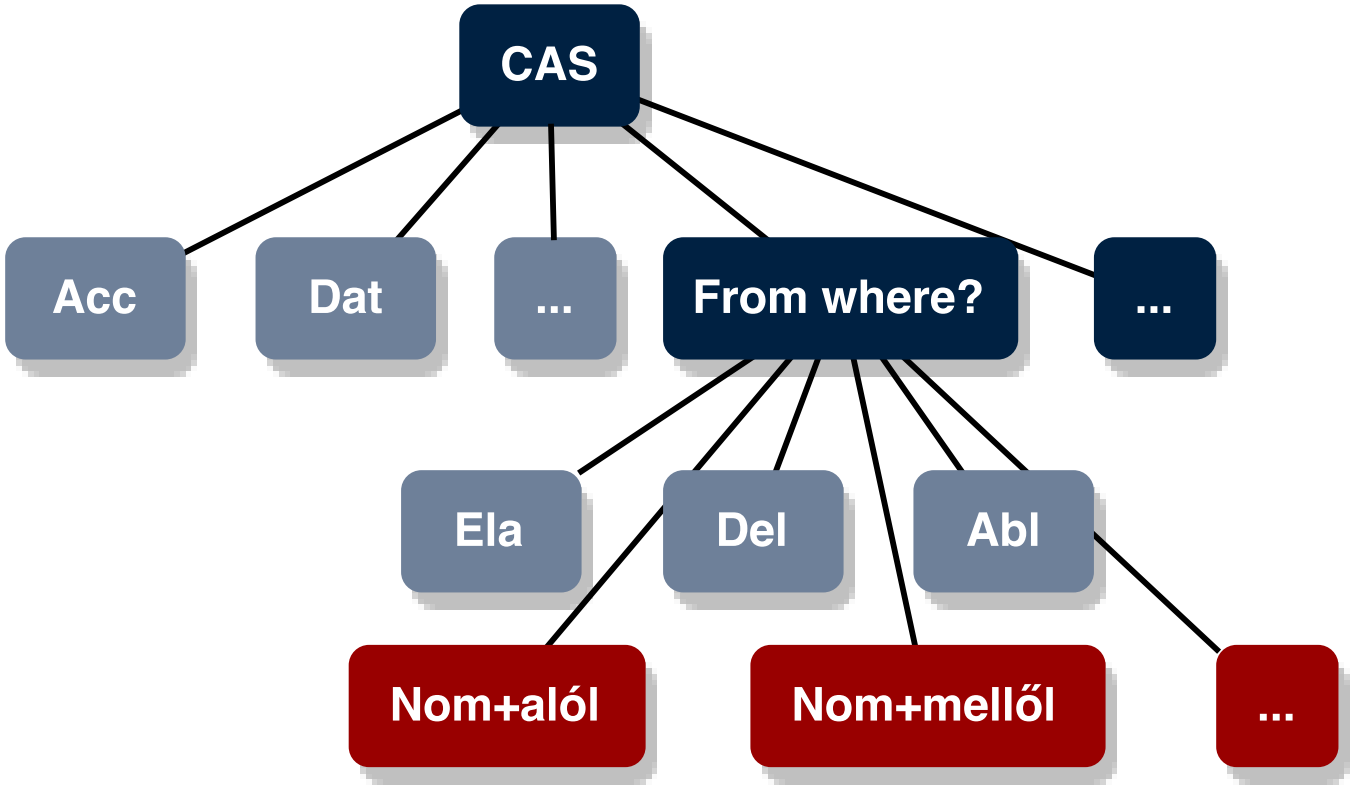

Figure 4: The hierarchical system of Hungarian cases and postpositions

Conclusion

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

- We introduced AnaGramma and the supply-and-demand framework
- We introduced our feature system conforming to existing categories
- We presented specific examples of
  - detached preverbs and adverbs
  - participles
  - postpositions as case markers

Thank you for your attention!

Questions?

References

Indig, B. and Vadász, N. (2016).

Windows in human parsing – how far can a preverb go?

In Marko, T. and Bekavac, B., editors, *Proceedings of the Tenth International Conference on Natural Language Processing (HrTAL2016)*, Dubrovnik, Croatia.

(Accepted, In press).

Prószéky, G. and Indig, B. (2015).

Psycholingusitically motivated parsing of Hungarian texts (in Hungarian) [Magyar szövegek pszicholingvisztikai indíttatású elemzése számítógéppel].

Alkalmazott nyelvtudomány, 15(1-2):29–44.

References II

Prószéky, G., Indig, B., and Vadász, N. (2016).

Performance-based parser for computational understanding of Hungarian texts (in Hungarian)[Performanciaalapú elemző magyar szövegek számítógépes megértéséhez].

In Bence, K., editor, *"Szavad ne feledd!": Tanulmányok Bánréti Zoltán tiszteletére*, pages 223–232. MTA Nyelvtudományi Intézet, Budapest.

Small, S. L. (1979).

Word expert parsing.

In *Proceedings of the 17th Annual Meeting on Association for Computational Linguistics*, ACL '79, pages 9–13, Stroudsburg, PA, USA. Association for Computational Linguistics.