Introducing the Hungarian National Corpus

Tamás Váradi, Csaba Oravecz, Bálint Sass

Research Institute for Linguistics
Hungarian Academy of Sciences
{varadi.tamas, oravecz.csaba,sass.balin}@nytud.mta.hu

Overview

- Introduction
- Background
- Design issues
- Structure of HNC
- Workflow
- Analysis and annotation
- Implementation
- Availability
- References

Introduction

- The concept of reference corpus
- Rise of national corpora
- Rise of the Internet
- Rise of social media
- Use of corpora today
  - for humans
  - for machines
- Is there still a point?

Origins: From HNC to HGC

- First version of the Hungarian National Corpus (HNC) (Váradi, 2002) developed between 1998 and 2001: "representative" (balanced) sample of the language use of the second half of the 90s, providing valuable empirical evidence for the status of the Hungarian language for theoretical analysis and language technology

- 187 million words, covering language variants inside the country and from neighbouring countries as well
  (→ the Hungarian Minority Language Corpus as a subset of the HNC)

Design Issues

Recent years: dramatic change in expectations against LRs.

- size: the dominance of data oriented methods and applications
  → more data = better performance
- quality: improved quality of language processing tools used for corpus processing
  → higher precision and finer levels of analysis and annotation in corpora
- coverage: preservation of representativity
  → subsequent samplings from language use including registers not yet covered by the HNC

Design Issues

Main objective for the HGC → increase in:

- size: extend the corpus to minimum 1 billion words,
- quality: use new technology for development and analysis,
- coverage and representativity: take new samples of language use and include further variants (transcribed spoken language data and social media).

Design Issues

- Representativity?
  Illusionary goal on this scale → balancedness (Váradi, 2002)

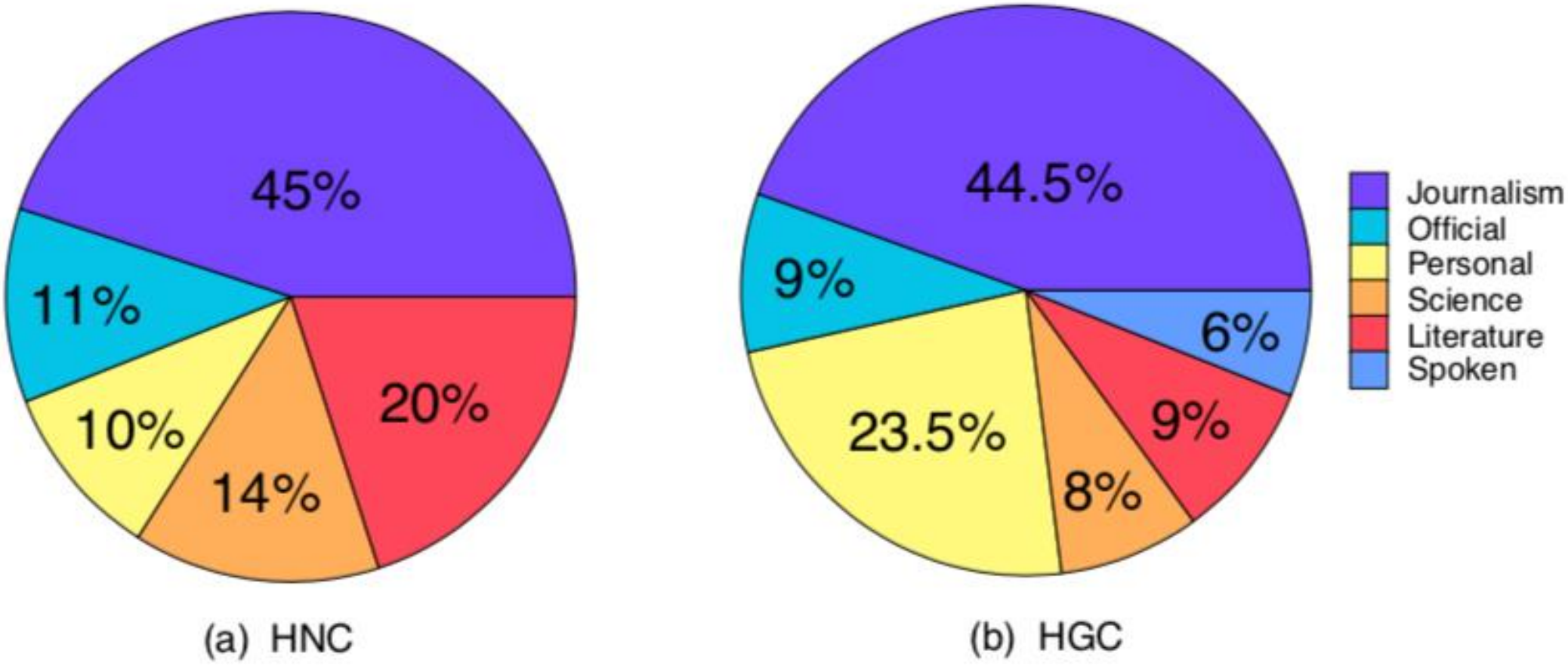
- Collection of data by web crawling or acquiring large amounts of newswire text?
  Difficult to produce a solid, balanced resource with sufficient meta-data (Baroni and Ueyama, 2006)
  → controlled, targeted resource collection, appropriate for each type of source. Easily processable source format is preferred
  → no pdf, no OCR.

- Clean IPR issues?
  Difficult, sometimes even impossible, to collect appropriate licenses
  → different availability options are offered for various sections of the HGC.

Structure of the HNC

Register	HNC	HGC	Source
Journalism	84.5	364.8	Daily/weekly newspapers
Literature	38,.2	80.6	Digital Literary Academy
(Popular) science	25.5	117.9	Hungarian Electronic Library
Personal	18.6	301.1	Social media, personal comments in forums
Official	20.9	99.0	Documents from public admin.
(Transcribed) spoken	–	76.2	Radio programs
	187.7	**1,039.7**	

Structure of the HNC



Workflow

- Preprocessing:
  - textual content and basic document structure identified in raw data
  - (near-)duplicates (Kupietz, 2005) deleted
  - non-Hungarian sections filtered out (Lui and Baldwin, 2012)
- Normalization: various renderings (eg. unicode symbols of ligatures, calligraphic letters, fancy punctuation) of characters are mapped to a (near-)equivalent base set

Output: clean XML, level 1 encoded according to a slightly modified DTD based on the Corpus Encoding Standard (Ide, 1998)

Analysis and Annotation

- Processing pipeline:
  - → tokenizer/segmenter (Huntoken)
  - → morphological analyzer (Humor (Prószéky and Tihanyi, 1996))
  - → POS tagger (modified TnT (Oravecz and Dienes, 2002))
  - → NP chunker and named entity recognizer (Varga and Simon, 2007)

Analysis and Annotation

Challenges:

- Structural ambiguity at the morphological level → simple heuristic: select the analysis with the highest number of morphemes in possibilities:

```
lehetőség[N]+ek[PL]+ben[INE]
lehető[A]+ség[_PROP]+ek[PL]+ben[INE]
*lesz[V]=le+hető[_HATO]+ség[_PROP]+ek[PL]+ben[INE]
=>lehetőség[N][PL][INE]/NP3N2
```

→ select the analysis with the highest number of morphemes

Analysis and Annotation

Challenges:

- Structural ambiguity at the morphological level → simple heuristic: select the analysis with the highest number of morphemes in possibilities:

```
lehetőség[N]+ek[PL]+ben[INE]
lehető[A]+ség[_PROP]+ek[PL]+ben[INE]
*lesz[V]=le+hető[_HATO]+ség[_PROP]+ek[PL]+ben[INE]
=>lehetőség[N][PL][INE]/NP3N2
```

  → select the analysis with the highest number of morphemes

- Productive compounding, overgeneration from the MA
  lázadó ("rebel") = láz ("fever") + adó ("tax")
  → filter rules with regular expressions

Analysis and Annotation

- Annotation format: one line per token, with annotation in tab separated columns (similar to the WaCky format (Baroni et al., 2009))
  higher level XML encoding of document structure as standoff annotation → merged with linguistic annotation to produce unified output

Analysis and Annotation

az	az	DET	D__D	compound=n;;hyphenated=n;;stem=az::DET				
token	*stem*	*MSD code*	*corpus tag*	BC	BC	az	az	B
angol	angol	A.NOM	AS_A	compound=n;;hyphenated=n;;stem=angol::A;;				
		morpheme level encoding		morphemes=ZERO::NOM				
				BCCBC	BCCBC	angol	angol	I
nyelvű	nyelvű	A.NOM	AS_A	compound=n;;hyphenated=n;;stem=nyelv::N;; morphemes=ZERO::NOM ű::_UKEP				
				CNCCF	CNCCF	Nelvű	Nelvű	I
szöveg	szöveg	N.NOM	NS3NN	compound=n;;hyphenated=n;;stem=szöveg::N;;				
		IOB encoding (of an NP)		morphemes=ZERO::NOM				
				CFCNC	CFCNC	Söveg	Söveg	I
volt	van	V.Me3	VS3PI	compound=n;;hyphenated=n;;stem=van::IGE;;				
		CV skeleton of token and stem		stemvar=vol::IGE;;morphemes=t::Me3				
				CBCC	CBC	volt	van	O
irányadó	irányadó	A.NOM	AS_A	compound=y;;hyphenated=n;;stem=ad::VERB				
		phonemic transcription		irány::N;;morphemes=ZERO::NOM ó::_OKEP				
				NCBCBCB	NCBCBCB	iráNadó	iráNadó	O
.	.	SPUNCT	SPUNCT	NA	NA	NA	NA	NA

Figure 1: “the English language text was primary”

Implementation

- Corpus engine: Manatee/Bonito corpus management system (Rychlý, 2007)

- HGC search interface: extended Bonito interface to allow for complex searches on all layers of the detailed (morphophonological) annotation (CV skeleton, phoneme classes, morpheme types, compounding etc.)

Implementation

- **Corpus engine**: Manatee/Bonito corpus management system (Rychlý, 2007)

- **HGC search interface**: extended Bonito interface to allow for complex searches on all layers of the detailed (morphophonological) annotation (CV skeleton, phoneme classes, morpheme types, compounding etc.)

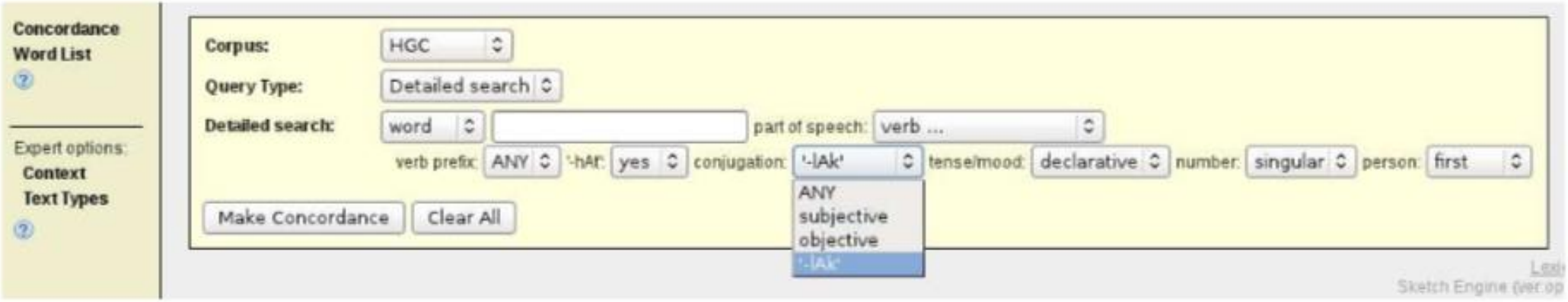


Figure 2: Searching for a specific verb form

Implementation

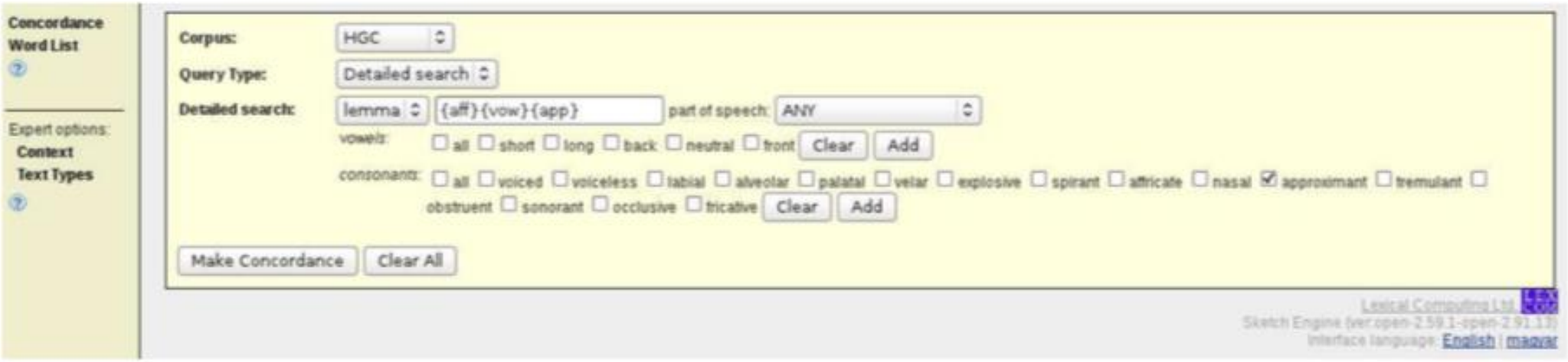


Figure 3: Searching for phonological phenomena

Implementation

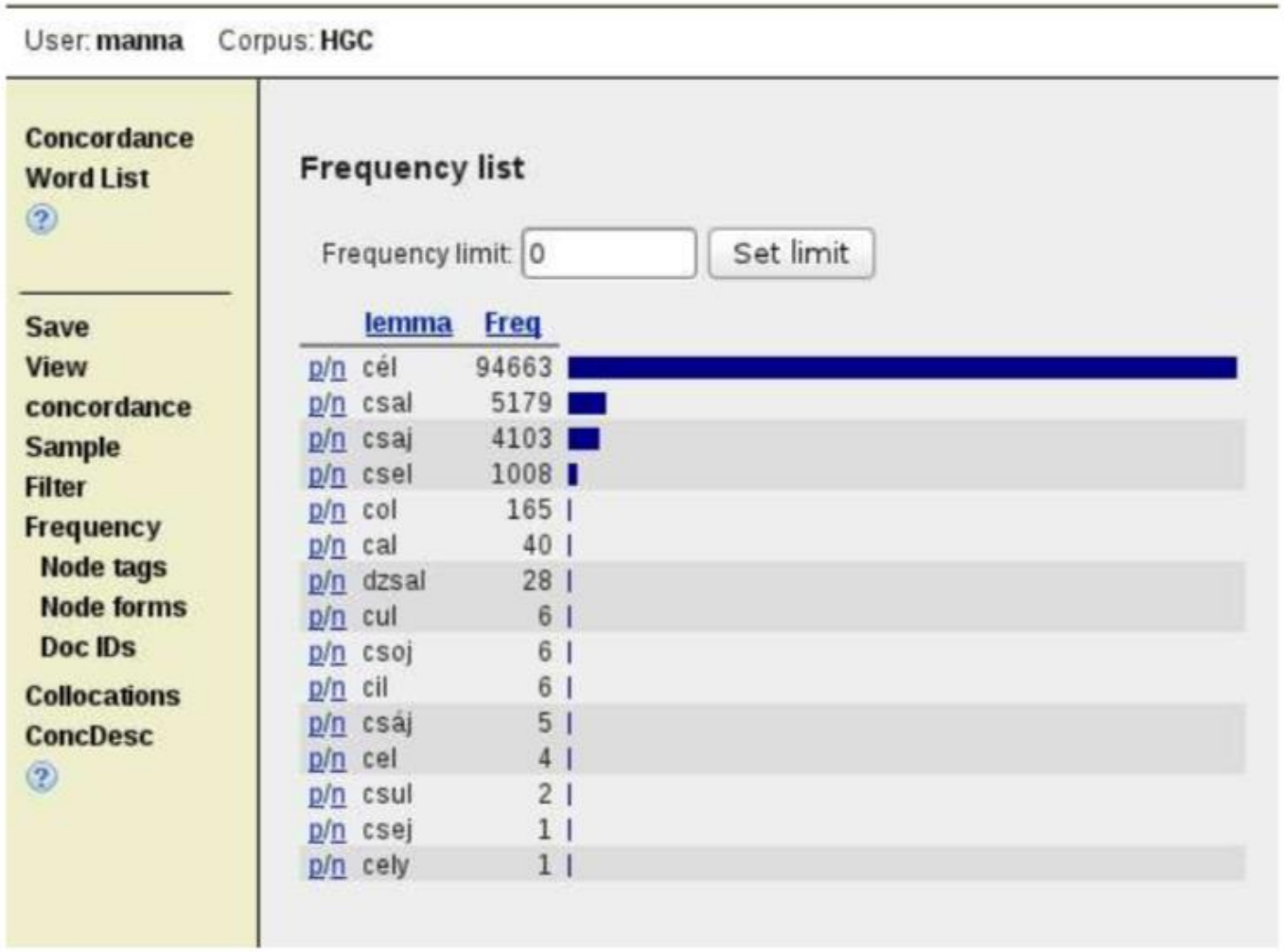


Figure 4: Frequency list generated from search result

Availability

- Accessibility: full version available only through the web search interface

http://hnc.nytud.hu

- derived resources: frequency dictionaries, collocation lists, verb subcategorization frame lexica in preparation

References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and Ueyama, M. (2006). Building general- and special-purpose corpora by web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application.*, pages 31–40, Tokyo, Japan.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Halácsy, P., Kornai, A., Németh, P., and Varga, D. (2008). Parallel creation of gigaword corpora for medium density languages – an interim report. In *In Proceedings of Language Resources and Evaluation Conference (LREC08)*.

Ide, N. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *Proceedings of the First International Language Resources and Evaluation Conference*, pages 463–470.

References

Kupietz, M. (2005). Near-duplicate detection in the IDS corpora of written German. Technical Report IDS-KT-2006-01, Institut für Deutsche Sprache.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea. Demo Session.

Oravecz, Cs. and Dienes, P. (2002). Efficient stochastic part-of-speech tagging for Hungarian. In *LREC-02*, pages 710–717, Las Palmas, Canary Islands, Spain.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English Gigaword Fifth Edition. Linguistic Data Consortium.

Prószéky, G. and Tihanyi, L. (1996). Humor – a morphological system for corpus analysis. In *Proceedings of the first TELRI seminar in Tihany*, pages 149–158, Budapest.

Rychlý, P. (2007). Manatee/Bonito – a modular corpus manager. In *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno: Masaryk University.

References

Varga, D. and Simon, E. (2007). Hungarian named entity recognition with a maximum entropy approach. *Acta Cybernetica*, 18(2):293–301.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 385–389, Las Palmas.

Wei-yun, M. and Huang, C.-R. (2006). Uniform and effective tagging of a heterogeneous gigaword corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 2182–2185, Genoa, Italy.

Thank you for your attention