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)
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 $\rightarrow$ 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 $\rightarrow$ 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)
   $\rightarrow$ controlled, targeted resource collection, appropriate for each type of source. Easily processable source format is preferred
   $\rightarrow$ no pdf, no OCR.

➢ Clean IPR issues?
   Difficult, sometimes even impossible, to collect appropriate licenses
   $\rightarrow$ different availability options are offered for various sections of the HGC.
### Structure of the HNC

<table>
<thead>
<tr>
<th>Register</th>
<th>HNC</th>
<th>HGC</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journalism</td>
<td>84.5</td>
<td>364.8</td>
<td>Daily/weekly newspapers</td>
</tr>
<tr>
<td>Literature</td>
<td>38.2</td>
<td>80.6</td>
<td>Digital Literary Academy</td>
</tr>
<tr>
<td>(Popular) science</td>
<td>25.5</td>
<td>117.9</td>
<td>Hungarian Electronic Library</td>
</tr>
<tr>
<td>Personal</td>
<td>18.6</td>
<td>301.1</td>
<td>Social media, personal comments in forums</td>
</tr>
<tr>
<td>Official</td>
<td>20.9</td>
<td>99.0</td>
<td>Documents from public admin.</td>
</tr>
<tr>
<td>(Transcribed) spoken</td>
<td>–</td>
<td>76.2</td>
<td>Radio programs</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>187.7</td>
<td>1,039.7</td>
<td></td>
</tr>
</tbody>
</table>
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 (e.g., 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)
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 \(\rightarrow\) simple heuristic: select the analysis with the highest number of morphemes in possibilities:

\[
\begin{align*}
\text{lehetőség}[N]+ek[PL]+ben[INE] \\
\Rightarrow\text{lehetőség}[N][PL][INE]/NP3N2
\end{align*}
\]

\(\rightarrow\) 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:

  \[
  \text{lehetőség[N]+ek[PL]+ben[INE]} \\
  \text{lehető[A]+ség[_PROP]+ek[PL]+ben[INE]} \\
  \text{*lesz[V]=le+hető[_HATO]+ség[_PROP]+ek[PL]+ben[INE]} \\
  \text{⇒lehetőség[N][PL][INE]/NP3N2}
  \]

  ➔ 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
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
<table>
<thead>
<tr>
<th>Token</th>
<th>Stem</th>
<th>MSD Code</th>
<th>Corpus Tag</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>az</td>
<td>az</td>
<td>DET</td>
<td>D__D</td>
<td>compound=n; hyphenated=n; stem=az::DET; morphemes=ZERO::NOM; BC BC az az</td>
</tr>
<tr>
<td>angol</td>
<td>angol</td>
<td>A.NOM</td>
<td>AS_A</td>
<td>compound=n; hyphenated=n; stem=angol::A; morphemes=ZERO::NOM</td>
</tr>
<tr>
<td>nyelvű</td>
<td>nyelvű</td>
<td>A.NOM</td>
<td>AS_A</td>
<td>compound=n; hyphenated=n; stem=nyelv::N; morphemes=ZERO::NOM ű::_UKEP</td>
</tr>
<tr>
<td>szöveg</td>
<td>szöveg</td>
<td>N.NOM</td>
<td>NS3NN</td>
<td>compound=n; hyphenated=n; stem=szöveg::N; morphemes=ZERO::NOM</td>
</tr>
<tr>
<td>volt</td>
<td>van</td>
<td>V.Me3</td>
<td>VS3Pl</td>
<td>compound=n; hyphenated=n; stem=van::IGE; stemvar=vol::IGE; morphemes=t::Me3</td>
</tr>
<tr>
<td>irányadó</td>
<td>irányadó</td>
<td>A.NOM</td>
<td>AS_A</td>
<td>compound=y; hyphenated=n; stem=ad::VERB; irány::N; morphemes=ZERO::NOM ű::_OKEP</td>
</tr>
</tbody>
</table>

**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


References


Thank you for your attention