Helyette: Inflectional Thesaurus for Agglutinative Languages

Gábor Prószéky[1,2] & László Tihanyi[1,3]

[1] MORPHOLOGIC
Fő u. 56-58. I/3
H-1011 Budapest
Hungary

[2] OPKM COMP. CENTRE
Honvéd u. 19.
H-1055 Budapest
Hungary
e-mail:h6109pro@ella.hu

[3] INSTITUTE FOR LINGUISTICS OF H.A.S
Színház u. 5-9.
H-1014 Budapest
Hungary
e-mail:h1243tih@ella.hu

1. Introduction

In the environment of word-processors thesauri serve the user's convenience in choosing the best suitable synonym of a word. Words in text of agglutinative languages occur almost always as inflected forms, thus finding them directly in a stem vocabulary is impossible. **Helyette**, the *inflectional thesaurus* coping with this problem is introduced in the paper.

2. Synonym dictionary with morphological knowledge

The inflectional thesaurus is a tool which (1) first performs the morphological segmentation of the input word-form, then (2) finds its stem's lexical base(s), (3) stores the suffix sequence situated on the right of the actual stem-allomorph, (4) offers the synonyms for the lexical base(s), and (5) generates the new word-form consisting of the adequate allomorph of the chosen stem and the adequate allomorph of the above suffix-sequence.

Both the morphological analysis and synthesis steps are done by the **Humor** (high-speed unification morphology) method described by Prószéky and Tihanyi (1992, 1993). The possible roots and the suffixes following them are temporarily stored, and **Helyette** performs the morphological synthesis on the basis of the new (synonym) root and the internal code of the stored suffix sequence. For more details, see Example 1.

3. Implementation details

The morphological framework behind **Helyette** relies on unification morphology. Both the thesaurus and the morphological/generator (as a stand-alone tool) are fully implemented for Hungarian. The synonym system consists of 40.000 headwords, the stem dictionary of the morphological analyzer/generator contains 80.000 stems, suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries more than 1.000.000.000 well-formed Hungarian word-forms can be analyzed or generated, and approximately 500.000.000 synonyms are handled. The whole software package is written in C programming language. The morphological analyzer based on **Humor** needs 800 KBytes disk space and less than 90 KBytes of core memory. The first version of the inflectional thesaurus **Helyette** needs 1.6 MBytes disk space and runs under MS-Windows.

References

[Prószéky and Tihanyi, 1992] Gábor Prószéky and László Tihanyi. A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. In: Ferenc Kiefer, Gábor Kiss and Júlia Pajzs (eds.) *Papers in Computational Lexicography — COMPLEX-92*, pages 265-278, Linguistics Institute, Budapest, 1992.

[Prószéky and Tihanyi, 1993] Gábor Prószéky and László Tihanyi. Humor: High-speed Unification Morphology and Its Applications for Agglutinative Languages. *La tribune des industries de la langue*, No.10., pages 28-29, OFIL, Paris, 1993.

WORD-FORM TO BE REPLACED:
kupáimra *[onto my drinking $cups_1$]*

MORPHOLOGICAL ANALYSIS:
kupá *+im+ra*

SUFFIX SEQUENCE TO BE STORED:
+PERS-1SG-PL+SUB

BASE-FORM OF ITS STEM:
kupa *[drinking cup_1]*

THE SYNONYM CHOSEN:
kehely *[drinking cup_2]*

TO BE SYNTHESIZED:
kehely *+PERS-1SG-PL+SUB*

ALLOMORPHS OF THE NEW STEM:
{kehely, kelyh}

ALLOMORPHS OF THE SUFFIX ARRAY:
{+im+ra, +im+re, +aim+ra,
+eim+re, +jaim+ra, +jeim+re}

MORPHOLOGICAL SYNTHESIS:
kelyh *+eim+re*

REPLACING WORD-FORM:
kelyheimre *[onto my drinking $cups_2$]*

Example 1.