HUMOR-BASED APPLICATIONS

Gábor Prószéky [1,2] **Miklós Pál** [1] **László Tihanyi** [1,3]

[1] MORPHOLOGIC
Fő u. 56-58. I/3
H-1011 Budapest, Hungary
E-mail: h6109pro@ella.hu

[2] OPKM Comp. Centre
Honvéd u. 19.
H-1055 Budapest ,Hungary
E-mail: h6109pro@ella.hu

[3] INSTITUTE FOR LINGUISTICS
Színház u. 5–9.
H–1014 Budapest, Hungary
E-mail: h1243tih@ella.hu

INTRODUCTION

There are several linguistic phenomena that can be processed by morphological tools of agglutinative and other highly inflectional languages, while processing of the same features need syntactic parsers in case of other languages, like English. There are, however, only a few morphological systems that are both fast enough and linguistically sound. **Humor**, a reversible, string-based, unification approach is introduced in the paper that has been used for creating a variety of lingware applications, like spell-checking, hyphenation, lemmatization, and of course, full morphological analysis and synthesis. Having discussed the philosophical and design decisions we show the above mentioned systems and then survey some competing approaches.

DESIGN PHILOSOPHY OF HUMOR

Several philosophical commitments regarding the NLP systems are summarized in Slocum (1988). **Humor** has been designed according to the Slocum requirements. It is language independent, that is, it allows *multilingual* applications. Besides agglutinative languages (e.g. Hungarian, Turkish) and highly inflectional languages (e.g. Polish, Latin) it has been applied to languages of major economic and demographic significance (e.g. English, German, French). **Humor** overcomes simple orthographic errors and mis-typings, thus it is a *fault-tolerant* system. The morphological analyzer version, for example, is able to analyze Hungarian texts from the 19th century when the orthographic system was not as uniform as nowadays. Word-forms are first "autocorrected" into the standard orthography and then analyzed properly.

Example 1: fault-tolerance in Humor

```
>> hivó
   hiv => hív[V] + ó[PART]
    (calling)
>> galyak
   galy => gally[N] + ak[PL]
    (fences)
```

Humor descriptions are *reversible*. It means that there is an opportunity to input a stem and several suffixes and the system generates every possible word-form satisfying the request.

Example 2: reversibility

Analysis:

```
>> házaddal          (with your house)
   ház[N] + ad[PERS-SG-2]+dal[INS]
>> mondsz                    (you say)
   mond[V] + sz[PERS-SG-2]
>> mondasz                  (you say)
   mond[V] + asz[PERS-SG-2]
```

Synthesis:

```
>> ház[N] + [PERS-SG-2] + [INS]
   házaddal         (with your house)
>> mond[V] + [PERS-SG-2]
   mondsz, mondasz          (you say)
```

The basic strategy of **Humor** is inherently suited to parallel execution. Search in the main dictionary, secondary dictionaries and affix dictionaries can happen at the same time. What is more, a simultaneous processing level (higher than morphology) based on the same strategy is under development.

In real-world applications, number of linguistic rules is an important source of grammatical complexity. In the Humor strategy there is a single rule only that checks unifiability of feature graphs of subsequent substrings in the actual word-form. It is very simple and clear, based on surface-only analyses, no transformations are used; all the complexity of the system is hidden in the graphs describing morpho-graphemic behavior.

Humor is *rigorously tested* on "real" end-users. Root dictionaries of the above mentioned languages contain 25.000--100.000 entries. The Hungarian version (90.000 stems) has been tested in every-day work since 1991 both by researchers of the Institute of Linguistics of the Hungarian Academy of Sciences (Prószéky and Tihanyi 1992) and users of word-processors and DTP systems (Humor-based proofing tools have been licensed by Microsoft, Lotus and other software developers).

MORPHOLOGICAL PROCESSES SUPPORTED BY HUMOR

The *morphological analyzer* is the kernel module of the system: almost all of the applications derived from Humor based on it. Humor has a *guessing* strategy that is based on orthographic, morpho-phonological, morphological and lexical properties of the words. It operates after the analysis module, mostly used in the spelling checkers based on Humor and in the above mentioned 19th century corpus application.

Synthesis is based on analysis, that is, all the possible morphemic combinations built by the core synthesis module are filtered by the analyzer.

Example 3: synthesis steps

```
>>   mond[V] + [PRES-SG-2]
      (you say)
(1)  Concrete morphemes instead
     of abstract morphs:
      mond + {el,ol,öl,sz,asz,esz}
(2)  String concatenation:
      mondel,mondol,mondöl,mondsz,
      mondasz,mondesz
(3)  Analysis, one by one:
      %mondel,%mondol,%mondöl,
      mondsz,mondasz,%mondesz
(4)  Filtering:
      mondsz
      mondasz
```

For internal use we have developed a *defaulting* subsystem that is able to propose the most likely inflectional paradigm(s) for a base word. There are only a few morphologically open word classes in the languages we have studied. Paradigms that are difficult to classify are generally closed; no new words of the language follow their morpho-graphemic patterns. The behavior of existing, productive paradigms is rather easy to describe algorithmically .

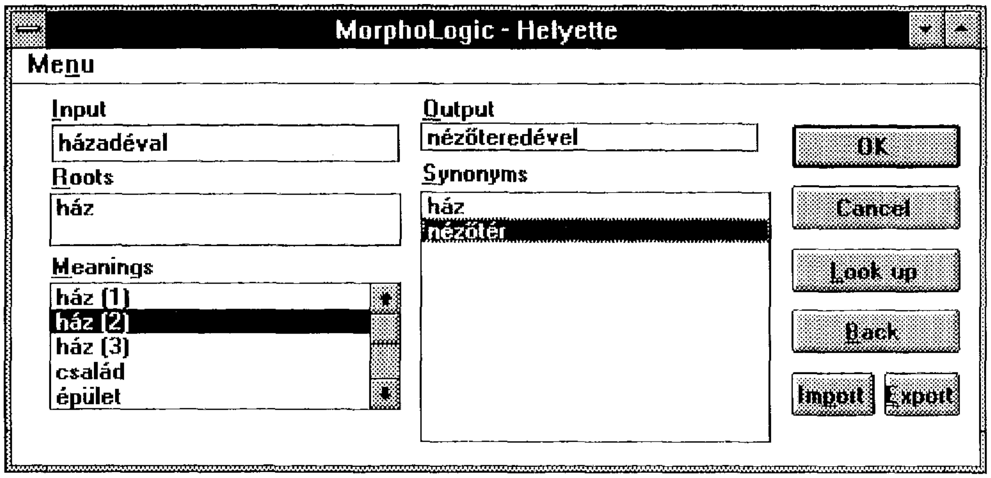
The *coding* subsystem of Slocum (1988) is represented by the so-called paradigm matrix of Humor systems. It is defined for every possible allomorph: it gives information about the potential behavior of the stem allomorph before morphologically relevant affix families.

COMPARISON WITH OTHER METHODS

There are only a few general, reversible morphological systems that can be used for more than a single language. Besides the well-known two-level morphology (Koskenniemi 1983) and its modifications (Karttunen 1985, 1993) we mention the Nabu system (Slocum 1988). Morphological description systems without large implementations (like the paradigmatic morphology of Calder (1989), or Paradigm Description Language of Anick and Artemieff (1992) are not listed here, because their importance is mainly theoretical (at least, for the time being). Two-level morphology is a reversible, orthography-based system that has several advantages from a linguist's point of view. Namely, the morpho-phonemic/graphemic rules can be formalized in a general and very elegant way. It also has computational advantages, but the lexicons must contain entries with diacritics and other sophistications in order to produce the needed surface forms. Non-linguist users need an easy-to-extend dictionary into which words can be inserted (almost) automatically. The lexical basis of Humor contain surface characters only and no transformations are applied.

Compile time of a large Humor dictionary (of 90.000 entries) is 1-2 minutes on an average PC, that is another advantage (at least, for the linguist) if comparing it with the two-level systems' compilers. The result of the compilation is a compressed structure that can be used by any applications derived from Humor. The compression ratio is less than 20%. The size of the dictionary does not influence the speed of the run-time system, because a special paging algorithm of our own is used.

Example 4: **Helyette**, the monolingual thesaurus with morphological knowledge



Input
házadéval

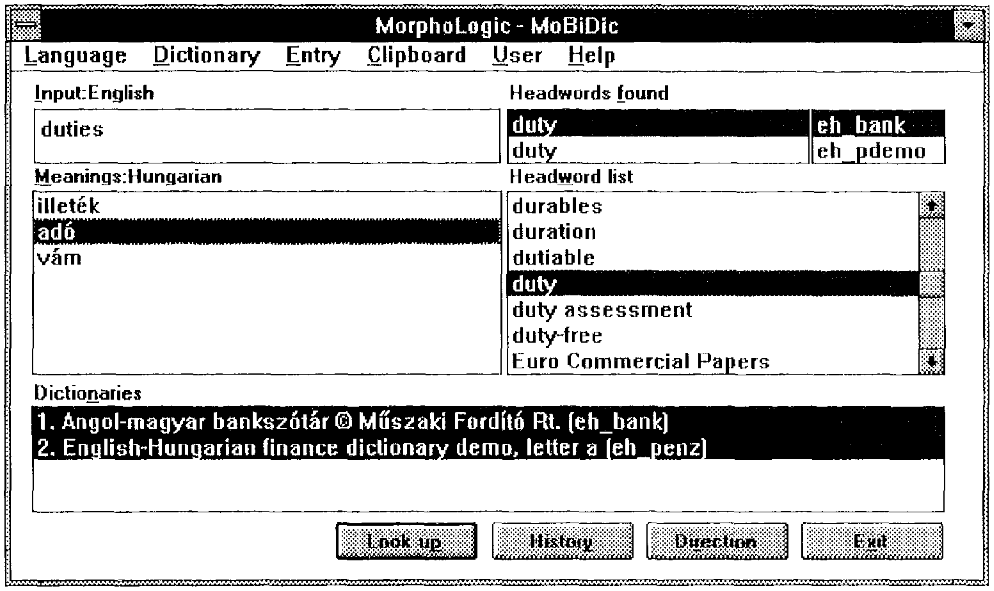
Roots
ház

Meanings
ház (1)
ház (2)
ház (3)
család
épület

Output
nézőteredével

Synonyms
ház
nézőtér

Example 5: **MoBiDic**, the bilingual dictionary with morphological knowledge



Input:English
duties

Headwords found

duty	eh_bank
duty	eh_pdemo

Meanings:Hungarian
illeték
adó
vám

Headword list
durables
duration
dutiable
duty
duty assessment
duty-free
Euro Commercial Papers

Dictionaries
1. Angol-magyar bankszótár © Műszaki Fordító Rt. (eh_bank)
2. English-Hungarian finance dictionary demo, letter a (eh_penz)

HUMOR-BASED IMPLEMENTATIONS

Humor systems have been implemented (at various depth) for English, German, French, Italian, Latin, Ancient Greek, Polish, Turkish, and it is fully implemented for Hungarian. The whole software package is written in standard C using C++ like objects. It runs on any platforms where C compiler can be found[1]. The Hungarian morphological analyzer which is the largest and most precise implementation needs 900 Kbytes disk space and around 100 Kbytes of core memory. The stem dictionary contains more than 90.000 stems which cover all (approx. 70.000) lexemes of the *Concise Explanatory Dictionary of the Hungarian Language*. Suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries Humor is able to analyze and/or generate around 2.000.000.000 well-formed Hungarian word-forms. Its speed is between 50 and 100 words/s on an average 40 MHz 386 machine. The whole system can be tuned[2] according to the speed requirements: the needed RAM size can be between 50 and 900 Kbytes.

There are several Humor subsystems with simplified output: lemmatizers, hyphenators, spelling checkers and correctors. They (called HelyesLem, Helyesel and Helyes-e?, respectively) have been built into several word-processing and full-text retrieval systems' Hungarian versions (Word, Excel, AmiPro, WordPerfect, Topic, etc.).[3]

Besides the above well-known applications there are two new tools based on the same strategy, the inflectional thesaurus called Helyette (Prószéky and Tihanyi 1992) and the series of intelligent bi-lingual dictionaries called MoBiDic. Both are dictionaries with morphological knowledge: Helyette is monolingual, while MoBiDic — as its name suggests — bilingual. Having analyzed the input word the both systems look for the found stem in the main dictionary. The inflectional thesaurus stores the information encoded in the analyzed affixes and adds to the synonym word chosen by the user. The synthesis module of Humor starts to work now, and provides the user with the adequate inflected form of the word in question. This procedure has a great importance in case of highly inflectional languages. The synonym system of Helyette contains 40.000 headwords. The first version of the inflectional thesaurus Helyette needs 1.6 Mbytes disk space and runs under MS-Windows. The size of the MoBiDic dictionary packages vary depending on the applied terminological collection. E.g. the Hungarian—English Business Dictionary (Example 4) needs 1.8 Mbytes space.[4]

Besides the above mentioned products, a Hungarian grammar checker (called Helyesebb) and other syntax-based (and higher level) mono- and multilingual applications derived also from the basic Humor algorithm are under development.

REFERENCES

Anick, P. and S. Artemieff (1992). A High-level Morphological Description Language Exploiting inflectional Paradigms. *Proceedings of COLING-92*, p. 67–73.

Calder, J. (1989). Paradigmatic Morphology. *Proceedings of 4th Conference of EACL*, p. 58-65.

Karp, D., Y. Schabes, M. Zaidel, and D. Egedi (1992). A Freely Available Wide Coverage Morphological Analyzer for English. *Proceedings of COLING-92*, Vol. III. p. 950– 954.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-form Recognition and Production.* Univ. of Helsinki, Dept. of Gen. Ling., Publications No.11.

Prószéky, G. and L. Tihanyi (1992). A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. In: Kiefer, F., G. Kiss, and J. Pajzs (Eds.) *Papers in Computational Lexicography — COMPLEX '92*, Linguistics Institute, Budapest, p. 265–278.

Prószéky, G. and L. Tihanyi (1993). Helyette: Inflectional Thesaurus for Agglutinative Languages. In: *Proceedings of the 6th Conference of the EACL*, p. 173.

Slocum, J. (1983). Morphological Processing in the Nabu System. *Proceedings of the 2nd Applied Natural Language Processing*, p. 228-234.

[1] Up to now, DOS, Windows, OS/2, Unix and Macintosh environments have been tested.

[2] Even by the end-users.

[3] For OEM partners there is a well-defined API to Humor.

[4] MoBiDic's language specific and not application specific parts need not be multiplied because vocabularies of the same languages use a single common morphological knowledge base.