Morphdb.hu: Hungarian lexical database and morphological grammar

Viktor Trón[1], Péter Halácsy[2], Péter Rebrus[3], András Rung[3], Péter Vajda[3], Eszter Simon[4]

[1] International Graduate College Language Technology and Cognitive Systems
University of Edinburgh and Saarland University
v.tron@ed.ac.uk
[2] Budapest University of Technology and Economics
Media Education and Research Center
halacsy@mokk.bme.hu
[3] Hungarian Academy of Sciences
Research Institute for Linguistics, Budapest
{rebrus,vajda,runga}@nytud.hu
[4] Budapest University of Technology and Economics
Cognitive Science Department
esimon@cogsci.bme.hu

Abstract

This paper describes morphdb.hu, a Hungarian lexical database and morphological grammar. Morphdb.hu is the outcome of a several-year collaborative effort and represents the resource with the widest coverage and broadest range of applicability presently available for Hungarian. The grammar resource is the formalization of well-founded theoretical decisions handling inflection and productive derivation. The lexical database was created by merging three independent lexical databases, and the resulting resource was further extended.

1. Introduction

While developing word analysis solutions for Hungarian, we opted for an architecture where the analysis toolkit is language independent and the language-dependent resources needed for the various tasks come from a central primary resource (Németh et al., 2004). Language independent affix-stripping analyzers such as **hunmorph** (Trón et al., 2005) (and its predecessor recognizers such as **ispell**) use a dictionary which stores lexical entries together with affix flags, which license the application of a set of affix rules associated with that flag listed in a so called affix file. Due to their inherently redundant format and linguistically unstructured layout, these resources have proved to be impossible to properly scale up to precise word-analysis tasks. Therefore, we designed and implemented **hunlex** (Trón, 2004) an offline resource compiler which offers a linguistically more motivated morphological description language and allows for principled, flexible maintenance and extension of resources. Hunlex reads a central lexical database and morphological grammar and compiles the dictionary and affix resources used by the word-analysis tools. A central database, with the help of hunmorph and hunlex, provides primary language resources for spell-checking, stemming, morphological analysis and numerous other annotation tasks. Hunlex allows flexible configuration of parameters pertaining to tag choice for various annotation tasks including choice of register standards for adjusting robustness. The rest of the paper is organized as follows. In Section 2. we first discuss the general ideas behind the morphological description, then describe in detail the structure of the morphological grammar. In Section 3. we discuss how our lexicon was obtained by merging three independent preexisting resources. Evaluation and suggestion for further work conclude the paper in Section 4.

2. The principles and structure of the morphological description

Since hunlex gives a great deal of freedom in specifying morphological operations, our choices were based on theoretical considerations of morphology as well as the practical issues of perspicuity and extendibility.

Following lexicographic tradition, our dictionary entries are lexemes (or lemmas) and contain only the unpredictable irregularities in the form of morphophonological and morphosyntactic features, which makes the extension of the lexicon extremely easy, reducing the task to recording a lexeme's standard citation form. A sequence of filters in the grammar creates the stem variants from the the citation forms based on their phonological and orthographic shape. The filters also decorate the stems with features which are used as conditions referred to in affixation rules. This general architecture is elaborated and illustrated below.

2.1. Lexicon and morphological processes

The description language of **hunlex** serves as a formal framework for an Item-and-Process (Hockett, 1954) approach to morphology. A hunlex description of the morphology of a language consists in a lexicon of stems and a grammar formalizing various morphological operations. The operations can be one of two types: *morphosyntactically active* rules, such as the addition of an affix morph to a relative stem, and *filter* rules describing morphophonological processes such as epenthesis or vowel shortening as well as expressing redundancies between features and phonological patterns.

2.2. Affix rules and filters

While designing and implementing the grammar for morphdb.hu we had to make decisions about generalizing some,

```
CAS_INE
        IF: analytic lengthened cas_ine
        TAG: <CAS<INE>>

, +ban IF: back
, +ben IF: front

;
```

Figure 1: The rules separated by commas refer to the allomorphs of the suffix, while the whole rule (CAS_INE) refers to the inessive case morpheme.

but not all, linguistically hypothesized morphological processes. We tried to describe the rules in a way that they would directly reflect the alternations traditionally called allomorphy. This means that nonconcatenative morphological processes were described as an allomorphic rule rather than an abstract one. In the case of the phonological rule governing vowel-harmony, this is illustrated in Figure 1.
The morphosyntactically active rules of the grammar are grouped naturally according to the morphosyntactic features they realize: a set of rules cover the allomorphic variants of what traditionally regarded as an affix morpheme. These sets of rules (CAS_INE in Figure 1.) themselves are thus referred to as affix morphemes, while the individual rules in a set are called affix allomorphs. As a result, the morphological grammar can also be naturally interpreted within an item and arrangement approach lending it more theory-neutral flavour.

2.3. The selection of appropriate allomorphs

Applying certain processes to the input can be set by various conditions, including pattern matching and feature checking. For example, the rules in Figure 1. refer to the features governing vowel harmony. The selection of an allomorph is often based on an idiosyncratic property of a stem (e.g. *hal-ak* ('fish.PLUR'), vs. *dal-ok* ('song.PLUR')), in which case it has to be handled by abstract features. Also, irregular orthography (mainly appearing in foreign proper names) made it necessary to handle some alternations as idiosyncratic and governed by features even if they are phonologically regular (e.g. *Voltaire-rel* 'with Voltaire'). Most of the conditions on allomorphic rules are thus described by means of features.

2.4. Lexemes and underspecification

The lexicon can be extended most efficiently if the entries are lexemes and they are supplied with the minimum necessary information about irregular morphological behaviour. This idea of a lexeme-based lexicon implies that the grammar has to contain special type of processes (so called filter rules) that generate the predictable stem variants of the lexeme as well as provide them with features to be used as conditions in selecting the appropriate affix allomorphs but which are left underspecified in the lexicon. An interesting part of the grammar is therefore the sequence of filters, implementing stem alternations like epenthesis, vowel shortening and final vowel lengthening as well as express certain implications between features and phonological patterns.

2.5. Unary licensing and optionality

A feature is unary if a rule refers to the presence of the feature as a condition of its application. The use of unary features even for binary distinctions makes the treatment of optionality and markedness extremely intuitive. For instance, certain suffixation patterns introduce a linking vowel between the stem-final and suffix-initial consonants, but the height of this vowel (mid or low) depends on the stem. Rules describing the low suffix allomorph with low vowel check the presence of the feature low, while those having a mid vowel reference non_low. These features then are assigned to regular stems and exceptional so called lowering stems, respectively. Optionally lowering stems (e.g. *naptár-ak* or *naptár-ok* 'calendar.PLUR') would be decorated with both features. This approach renders optionality in a particular dimension both notationally transparent and suggestive of markedness. Since lowering stems are exceptional, a filter rule associates stems that are underspecified for lowering in the lexicon with the feature non_low, standing for the unmarked value of this dimension (shown in Figure 2.). Such default feature assignment rules make it possible to leave most features underspecified in the lexicon.

```
NOM_LOWERING_FILTER

    FREE: false
    FILTER: low non_low
    OUT: NOM_KEEP_ALL_FEATURES
    OUT: NOM_ACC_FILTER

,OUT: non_low
;
```

Figure 2: The lowering filter: example of associating a default feature.

Other morphophonological properties where optionality can occur, such as vowel harmony and most of the stem alternations, were also handled by use of unary features.

2.6. Analytic and synthetic suffixes

So called analytic and synthetic suffixation (Rebrus, 2000) governing the selection of stem variants were managed in a similar vein. If a lexeme has alternative stems, each stem variant would get one of the features. Non-alternating stems correspond to optionality, as they allow both analytic and synthetic affixation processes. Features governing analyticity are present only in the grammar, the lexicon contains only the features encoding the type of the stem alternation (i.e., shortening, epenthesis). In the process of generating synthetic and analytic stems, the appropriate features are assigned to the variants, while non-alternating stems automatically receive both features. Defective non-alternating stems (e.g. *siklik* 'glide') can be described by lacking analytic stems (e.g., imperative forms), their exceptional specification of the synthetic feature in the lexicon prevents them from further assignment of default features in the grammar.

2.7. Vowel lengthening

A certain set of affixes trigger vowel lengthening on vowel-final nominal stems yielding stem alternations like with the noun *fa* 'tree': *fá-nak* 'to tree' vs. *fa-ként* 'like tree'. The number of affix allomorphs rules can be kept low and the description simple if this is handled as a rule-governed stem alternation. Lengthening suffixes require a feature `lengthened` present on the stem, whereas non-lengthening ones require `non_lengthened`. Vowel lengthening as a process is thus handled with a rule that generates the stem with a long vowel (*fá-*) and supplements it with the feature `lengthened`, while assigns `non_lengthened` to the basic unaltered stem (*fa-*). Stem types that do not (*mozi* 'cinema') or cannot (*ház* 'house') show final vowel lengthening receive both features `non_lengthened` and `lengthened` by default.

2.8. Suffixation of foreign orthography words

Contrary to the regular phonetic nature of Hungarian orthography, there is a significant number of loan words and proper nouns in the language the spelling of which do not reflect the pronunciation. In such cases, since the choice of affix allomorphs is regular given the pronunciation, (e.g., *Voltaire* /volter/ cf. *Voltaire-rel* 'with Voltaire'), supplying the pronunciation in the lexical entries (*Voltaire/volter*) provides an elegant solution. The hunlex framework allows for specifying rules so that features governing allomorph choice can be based on pronunciation while making sure that it is the orthographic form that combines with the affixes to yield the output. Hunlex allows the specification of rules to determine pronunciation itself in the grammar (loosely corresponding to grapheme to phoneme conversion). Currently, this is only exploited to resolve acronyms (*http*) where the pronunciation consists of spelling out the letters of the orthographic form (*há-té-té-pé*).

2.9. Underspecification and irregularity

As discussed earlier, most of allomorphic conditioning is handled by features. This allows for (i) a concise specification of affix rules and their condition of application (IF: back); (ii) transparent reference to morphological properties familiar from theoretical treatments of Hungarian (backness); (iii) reuse the same set of conditions to several classes of affixes (variants vowel harmony); (iv) allow for marking exceptions (*cél-ok* 'goal.PLUR', front stem vowel with irregular back affixes) and (v) optionality indicated in the lexicon (*fotel-ok* or *fotel-ek*). Since a great deal of features are predictable from the shape of the stem, filters make extensive use of pattern-matching. We tried to achieve maximum economy in lexical specifications by formulating the broadest possible generalizations in the feature assignments, so that exceptional forms (those with features in the lexicon) would constitute small and, for most irregularities, closed classes. The only remaining idiosyncrasies that may need specification when morphdb.hu is extended seem restricted to features governing type possessive suffixation and lowering in case of adjectives.

3. Creating the lexicon of morphdb.hu

3.1. Lexical resources

The lexicon of morphdb.hu is a result of compiling three wide coverage dictionaries. The *Magyarispell dictionary* is the Hungarian resource for ispell-based open-source spell-checker (Németh, 2002), and it is the most up-to-date lexicon of present-day Hungarian, containing 80 thousand entries. Our second source was the *Dictionary of Hungarian Inflections* (Elekfi, 1994).[1] It contains nearly 66 thousand entries of the traditional Hungarian Reference Dictionary classified into paradigm classes. The third source is a dictionary database (Kornai, 1986), which contains 78 thousand entries. Verbs and nominal categories were collected from our sources using mainly automatic methods. As a first step, morphological information present in the three sources were to be transformed into morphophonological features used by the grammar of morphdb.hu. This required three different methods depending on the characteristics of the individual resources.

For the Myspell dictionary, where the various stem types and the irregularities are already grouped, this task was relatively easy. This resource also contained some 30 thousand categorized named entities, which were transferred as well.

Kornai's dictionary database uses a classification scheme, originally developed by Ferenc Papp, where the lexical entry record refer to various morphophonological phenomena, such as the the choice between various allomorphs of a certain suffix and abstract features such as type of stem alternation much in the spirit of morphdb.hu. Since these types of information closely mirror the features as used in morphdb.hu, mapping individual field values to morphdb.hu feature sets was sufficient to transform the resource to a morphdb.hu lexicon.

The Dictionary of Hungarian Inflections uses 1,700 paradigms to group the stems, each paradigm containing words that take exactly the same affixes. The paradigms do not overtly reflect the actual differences between the behavior of the stems, so the morphophonological features used in morphdb.hu had to be specified for each paradigm by semi-automatic methods. This dictionary contains other important grammatical information not encoded in the paradigms such as boundary markers encoding the internal structure of compounds, which are retained in morphdb.hu as well.

As the morphdb.hu lexicon contains only unpredictable irregularities, predictable and regular information of the three resources had to be deleted. In the final lexicon, there are no redundantly specified features that would also be automatically assigned by the filter mechanism of the morphdb.hu grammar.

3.2. Compilation of the lexicon

The second step in creating the lexicon was compiling the transformed dictionaries into one lexicon filtering out multiple occurrences. Each entry received all the features from

[1] digitalized by the Department of Corpus Linguistics at the Research Institute for Linguistics of the Hungarian Academy of Sciences

all three resources, provided they were non-contradictory. A feature that was present in only one of the resources was assigned to the entry as well. This way the whole transformation process could be verified, and in the process a significant number of discrepancies were found. Some of these were simply typos or obvious mistakes in the sources, but many represented the different judgments of their respective authors. Entries having contradictory morphological information were checked and corrected by experts as well as by automatic methods using the Hungarian Webcorpus (Halácsy et al., 2004) in order to best reflect standard present-day Hungarian usage.

All adverbs, conjunctions and interjections present in the three original resources were reviewed and categorized by hand and extended using other sources. To avoid the sometimes controversial classification of our original resources, we created the hugely complex system of pronouns and postpositions from scratch, handling their affixation with rules in the grammar; we only used the resources for evaluation.

The amount of overlap in the three resources turned out to be about a quarter of the overall size of the lexicon, with each of the three contributing at least 10 thousand unique entries. Besides containing rich morphological description, all three resources contain other useful information such as domain information, internal structure, stylistic and usage information, which were all retained in the morphdb.hu database.

4. Conclusion

The evolving morphdb.hu with its 130 thousand entries is the largest lexical database for Hungarian and is freely available under a permissive license. Although morphdb.hu still needs a great deal of lexicographic work in order to reach ultimate gold standard quality, preliminary recall figures obtained from manually corrected, morphologically analyzed Hungarian corpora are already promising.

To measure the coverage of morphdb.hu, we compiled with hunlex the resources for our morphological analyzer and ran the analyzer on two Hungarian Corpora. One of them is the Szeged Corpus (Csendes et al., 2003) which contains 1 million words, and on which the recall of our analyzer is 90%. The missing 10% are mostly proper names and acronyms not analyzed partly due to the difficulty of multi-word named entity tokenization. The other corpus is the 700 million word Hungarian Webcorpus (Halácsy et al., 2004; Kornai et al., 2006) on which the proportion of out-of-vocabulary items is 7%.

The practical use of morphdb.hu for language technology, however, can only be assessed once it is shown how and to what extent it can be exploited for the purposes of automatic morphological tagging of texts. The need for the resolution of ambiguities as well as for fallback to guessing in case of unknown items make it necessary to use statistical methods for this task. Novel methods of enhancing statistical POS tagging with an analyzer are discussed and evaluated in another paper of this volume: lrec06:pos, test various tagging models on Hungarian corpora showing significant increase in performance when using morphdb.hu.

Hungarian is a language with a hugely complex morphology which makes the creation of an extensible central lexical database for word-analysis a rather ambitious task. The fact that this project has been successfully completed and that our database has been extensively used for various tasks demonstrates the power of our word analysis infrastructure, and provides a convincing use case for the hunlex framework which we already use for English and expect to use for other languages as well.

5. References

Dóra Csendes, Csaba Hatvani, Zoltán Alexin, János Csirik, Tibor Gyimóthy, Gábor Prószéky, and Tamás Váradi. 2003. Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 238–245. Szegedi Tudományegyetem.

László Elekfi. 1994. *Magyar ragozási szótár*. MTA Nyelvtudományi Intézet, Budapest.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. 2004. Creating open language resources for Hungarian. In *Proceedings of Language Resources and Evaluation Conference (LREC04)*. European Language Resources Association.

Charles F. Hockett. 1954. Two models of grammatical description. *Word*, 10:210–234.

András Kornai, Péter Halácsy, Viktor Nagy, Csaba Oravecz, Viktor Trón, and Dániel Varga. 2006. Web-based frequency dictionaries for medium density languages. In *Proceedings of the EACL 2006 Workshop on Web as a Corpus*.

András Kornai. 1986. Szótári adatbázis az akadémiai nagyszámitógépen (A dictionary database of Hungarian). *Hungarian Academy of Sciences Institute of Linguistics Working Papers*, II:65–79.

László Németh, Viktor Trón, Péter Halácsy, András Kornai, András Rung, and István Szakadát. 2004. Leveraging the open-source ispell codebase for minority language analysis. In *Proceedings of SALTMIL 2004*. European Language Resources Association.

László Németh. 2002. Magyar Ispell – Válasz a Helyese?-re. In *IV. GNU/Linux szakmai konferencia*, pages 99–107. Linux-felhasználók Magyarországi Egyesülete.

Péter Rebrus. 2000. Morfofonológiai jelenségek [morphophonological phenomena]. In Ferenc Kiefer, editor, *Strukturális magyar nyelvtan. 3. Morfológia. [Hungarian structural grammar. 3. Morphology.]*, pages 763–948. Akadémiai Kiadó, Budapest.

Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*.

Viktor Trón. 2004. Hunlex - morfológiai szótárkezelő rendszer. In *II Magyar Számítógépes Nyelvészeti Konferencia*, pages 177–182. Szegedi Tudományegyetem.