A MODEL OF COMPUTATIONAL MORPHOLOGY AND ITS APPLICATION TO URALIC LANGUAGES

Summary of PhD Thesis

Attila Novák

Roska Tamás Doctoral School of Sciences and Technology

Pázmány Péter Catholic University, Faculty of Information Technology and Bionics

Academic advisor:
Dr. Gábor Prószéky

2015

1

Introduction

Science primarily aims at describing and explaining various aspects of the world as we experience it. But since the second half of the 19th century, science has also grown to be a major contributor to technological knowledge. Nevertheless, language technology (i.e. computer technology applied to everyday language-related tasks) has coped in the past with a number of language-related problems without making much use of linguists' models. Doing morphology (i.e. handling word forms), for example, was not much of a technological problem for some of the commercially most interesting languages, especially for English, because it could be handled either by simple pattern matching techniques or by the enumeration of possible word forms.

Although there have been attempts to handle language technology tasks, such as spell checking, for languages that feature complex morphological structures and phonological alternations avoiding the use of a formal morphological description, these word-list-based attempts failed to produce acceptable results even recently, when corpora of sizes in an order of hundreds of millions of running words are available for languages such as Hungarian. However big the corpus is, even very common forms of not-extremely-frequent words are inevitably missing from it. Moreover, when I analyzed the word forms of the 150-million-word Hungarian National Corpus, I found that 60 percent of the theoretically possible Hungarian inflectional suffix morpheme sequences never occurs in the corpus. This figure does not include any of the numerous productive derivational suffixes. There is nothing odd about these suffix combinations. They are just rare. For a bigger 500-million-word Web corpus, I found the ratio to be 50 percent.

The creation of a formal morphological description is therefore unavoidable for this type of languages. The central theme of my thesis concerns **computational models of morphology that are applicable to such morphologically complex languages**. Some of these languages have also been commercially interesting to some extent: e.g. Turkish, Finnish or Hungarian, just to mention some from Europe. But as soon as there are tools which are readily applicable, what could stop us from applying these to languages that are spoken by less populous or rich speaker communities? So I also explored the task of creating computational morphologies for Uralic minority languages.

Beside the complexity of the morphology of these languages, another factor that makes a data-oriented approach unfeasible in some cases is the lack of electronically available linguistic resources.When working on Uralic minority languages, the corpora I had to do with did not exceed the size of a hundred thousand running words in the case of any of the languages involved, in some cases the size of the corpus did not even reach ten thousand words. In addition to a general lack of such resources concerning these languages, in the case of the most

endangered ones, Nganasan and Mansi, there seems even to be a lack of really competent native speakers. The available linguistic data and their linguistic descriptions proved to be incomplete and contradictory for all of these languages, which also made numerous revisions to the computational models necessary.

The most successful and comprehensive analyzer for Hungarian (called Humor and developed by a Hungarian language technology firm, MorphoLogic) was based on an item-and-arrangement model analyzing words as sequences of allomorphs of morphemes and using allomorph adjacency constraints (Prószéky and Kis, 1999). Although the Humor analyzer itself proved to be an efficient tool, the format of the original database turned out to be problematic. A morphological database for Humor is difficult to create and maintain directly in the format used by the analyzer, because it contains redundant and hard-to-read low-level data structures. To avoid these problems, I created a higher-level morphological description formalism and a development environment that facilitate the creation and maintenance of the morphological databases.

I have created a number of complete computational morphologies using this morphological grammar development framework. The most important and most comprehensive of these is an implementation of Hungarian morphology. Its rule component was created relying mainly on my competence and on research I performed during the preparation of my theoretical linguistics Master's Thesis (Novák, 1999) and later refined during corpus testing and the actual use of the morphology in an English-to-Hungarian and a Hungarian-to-English machine translation system and various corpus annotation projects.

The first version of the analyzer's stem database was based on the original Humor database, from which all redundant (predictable) features were removed and unpredictable properties of words (e.g. category tag) were all manually checked and corrected, and missing properties were added (e.g. morpheme boundaries in morphologically complex entries). Currently, the size of the stem database is several times bigger than that of the original Humor analyzer and also contains specialized (e.g. medical, financial) vocabulary. The underlying grammatical model is much more accurate than that of the original morphology it replaced.

I have also created complete computational morphologies of Spanish and French using the morphological grammar development framework, and also adapted ones for Dutch, Italian and Romanian. In addition, I co-authored morphologies for a number of endangered Finno-Ugric languages (Komi, Udmurt, Mari and Mansi)[i]. In the same project, I created morphologies for two seriously endangered Northern Samoyedic languages: Nganasan and Tundra Nenets. The Uralic project was an attempt at using language technology to assist linguistic research, applying it to languages that can not otherwise be expected to be targets of the application of such technology due to the lack of commercial interest. The Uralic morphologies were further refined in a series follow-up projects, and two Khanty dialects (Synya and Kazym Khanty) were also described.

One aspect of morphological processing not covered by the original Humor implementation is that it does not support a suffix-based analysis of word forms whose stem is not in the stem database of the morphological analyzer, and the system cannot be easily modified to add this

[i]This was done in a project called 'Complex Uralic Linguistic Database' (NKFP 5/135/2001), in which I worked together with László Fejes

feature. Moreover, integration and appropriate usage of frequency information, as would be needed by data-driven statistical approaches to text normalization (e.g. automatic spelling error correction or speech recognition), is not possible within the original Humor system. A third factor that can be mentioned as a drawback of Humor is its closed-source licensing scheme that has been an obstacle to making resources built for morphological analyses widely available. The problems above could be solved by converting the morphological databases to a representation that can be compiled and used by finite-state morphological tools.

The Xerox tools implement a powerful formalism to describe complex types of morphological structures. This suggested that mapping of the morphologies implemented in the Humor formalism to a finite-state representation should have no impediment. However, the Xerox tools (called *xfst*), although made freely available for academic and research use in 2003 with the publication of Beesley and Karttunen (2003), do not differ from Humor in two significant respects: a) they are closed-source and b) cannot handle weighted models. Luckily, a few years later quite a few open-source alternatives to *xfst* were developed. One of these open-source tools, Foma (Huldén, 2009), can be used to compile and use morphologies written using the same formalism. Another tool, OpenFST (Allauzen et al., 2007), is capable of handling weighted transducers, and a third tool, HFST (Lindén et al., 2011), can convert transducers from one format to the other and act as a common interface above the Foma and OpenFST backends.

Creating high-quality computational morphologies is an undertaking that requires a considerable amount of effort, and requires threefold competence: familiarity with the formalism, knowledge of the morphology, phonology and orthography of the language, and extensive lexical knowledge. Thus another interesting subject is the learnability of (aspects of) morphology from corpora or existing lexical databases using automatic methods. Many morphological resources contain no explicit rule component. Such resources are created by converting the information included in some morphological dictionary to simple data structures representing the inflectional behavior of the lexical items included in the lexicon. The representation often contains only base forms and some sort of information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with a few other morphosyntactic features. With no rules, the extension of such resources with new lexical items is not such a straightforward task, as it is in the case of rule-based grammars. However, the application of machine learning methods may be able to make up for the lack of a rule component. Thus, I solved the problem of predicting the appropriate inflectional paradigm of out-of-vocabulary words, which are not included in the morphological lexicon. The method is based on a longest suffix matching model for paradigm identification, and it is showcased with and evaluated against an open-source Russian morphological lexicon.

Since the detailed presentation of each of the morphologies that I developed could itself be the subject of a separate research report, I cannot undertake to present them in full depth within the scope of this thesis. However, fragments of the grammars and lexicons are used in the corresponding chapters to illustrate features of the formalisms. Phenomena from the above-mentioned languages relevant to the subject of my work and the way they are handled in the models are presented and discussed. Emphasis is laid primarily on Hungarian, the language of which the most comprehensive description was created, but I also present some details about the morphologies created for other Uralic languages.

Another group of ‘languages’ for which I describe the method of adaptation of the general morphological analyzer, are some variants of Hungarian. First, the Humor morphological analyzer was extended to be capable of analyzing words containing morphological constructions, suffix allomorphs, suffix morphemes, paradigms or stems that were used in Old and Middle Hungarian but no longer exist in present-day Hungarian. A disambiguation system was also developed that can be used for automatic and manual disambiguation of the morphosyntactic annotation of texts and a corpus manager is described with the help of which the annotated corpora can be searched and maintained. Another ‘language’ for which the analyzer was adapted was the language of Hungarian clinical documents created in clinical settings. This language variant differs from general Hungarian in several respects. In order to process such texts written in a so called notational language, the morphological analyzer had to be adapted to the requirements of the domain by extending its lexicon applying a semi-automated algorithm.

2

New Scientific Results

The better the database of a linguistic program models the language, the better results it can produce. A key module in a linguistic model is the morphological component, which is responsible for the analysis and generation of words in the given language. In this research, I have explored various ways of creating linguistically adequate computational morphologies for morphologically complex languages.

I have created a model for morphological description that has been successfully applied to a number of different languages. The language descriptions and the tools created using the model have been used in various commercial products, such as spell checkers, stemmers, pop-up dictionaries, a rule-based machine translation system, and in various scientific projects.

2.1 A morphological grammar development framework

Creating a morphological analyzer for an agglutinative language is quite a challenge, as the number of morpheme combinations is practically infinite. Thus, the standard methods applied for isolating languages, such as English, do not give satisfactory results for morphologically complex ones. The standard tool used for Hungarian has long been MorphoLogic's Humor ('High speed Unification MORphology') morphological analyzer engine (Prószéky and Kis, 1999). The model this analyzer uses is based on constraints on adjacent morphs. It performs a classical 'item-and-arrangement' (IA)-style analysis. I used this program as the starting point for my research.

Humor analyzes the input word as a sequence of morphs. Each morph is a specific realization (an allomorph) of a morpheme. The word is segmented into parts which have a surface form (that appears as part of the input string, the morph); a lexical form (the 'quotation form' of the morpheme) and a (possibly structured) category label.

The program performs a depth-first search on the input word form for possible analyses. Two kinds of checks are performed at every step: a local compatibility check of the next morph with the previous one and a global word structure check on each locally compatible candidate morph by traversing a deterministic extended finite-state automaton (EFSA) that describes possible word structures. The lexical database of the Humor analyzer consists of an inventory of morpheme allomorphs, the word grammar automaton and two types of data structures used for the local compatibility check of adjacent morphs. One of these are continuation classes and binary continuation matrices describing the compatibility of those

continuation classes. The other are binary properties and requirements vectors. Each morph has a continuation class identifier on both its left and right hand sides, in addition to a right-hand-side binary properties vector and a left-hand-side binary requirements vector.

The database is difficult to create and maintain directly in the format used by the analyzer, because it contains redundant and low-level data structures. To avoid these problems, I designed and implemented a morphological grammar development environment that automatically creates the lexical resources used by the Humor analyzer from a high-level human readable redundancy-free morpheme-based grammatical and lexical representation that only contains idiosyncratic features of morphemes and is easy to maintain and extend. Polymorphemic entries, such as compounds, can also be added to the lexicon, and an inheritance mechanism ensures that these entries inherit idiosyncratic properties from their final element by default, thus minimizing redundancy in the description and enhancing consistency. The system also creates a redundant but still easy-to-read intermediate representation that facilitates the checking of the correctness of allomorph creation rules. The system ensures the consistency of the created lexical resources and automatically checks for possible syntactic errors and contradictions in the source descriptions.

The high-level human-readable description is transformed by the system to the redundant representations of the analyzer by performing the operations described by the rules and converting the features and constraint expressions using an encoding definition description. This defines how each high-level feature should be encoded for the analyzer. Certain features are mapped to binary properties while the rest determine the continuation matrices, which are generated by the system dynamically. The system is not geared to a particular language; it can be effectively used to describe the morphologies of various languages without any modification of the programs.

All Humor morphologies built after the creation of the development environment were developed using this higher-level formalism.

THESIS 1:

> *I designed and implemented a morphological grammar development environment that automatically creates the lexical resources used by the Humor analyzer from a high-level human readable redundancy-free morpheme-based grammatical and lexical representation that only contains idiosyncratic features of morphemes and is easy to maintain and extend.*

Related publications: 13, 14, 54, 55, 56, 61, 63

2.2 Application of the model to various languages

The development environment was used to create computational morphologies for a number of languages, among them agglutinating ones. I created state-of-the-art morphologies for *Hungarian*, *Spanish* and *French*. In addition, I co-authored Humor morphologies for the following Uralic languages: *Komi*, *Udmurt*, *Mari*, *Northern Mansi* and various *Khanty*

dialects. Moreover, based on various morphological descriptions, I also created Humor-compatible morphologies for *Dutch*, *Italian*, *Romanian* and *Russian*, extending the coverage of the original descriptions as required. I also created computational morphologies for two *Samoyedic* languages. Although these agglutinating languages are also members of the Uralic language family, describing their very intricate morpho-phonology using the constraint-based Humor formalism turned out to be too difficult. Thus, these morphologies were implemented using the finite-state formalism of Xerox using *lexc* and *xfst*.

THESIS 2

I created and co-authored computational morphologies for several languages.

THESIS 2a:

I created, co-authored or adapted computational morphologies for the following languages in the formalism I introduced demonstrating the capabilities of the framework: Hungarian, Spanish, French, Dutch, Italian, Romanian, Russian, Komi, Udmurt, Mari, Northern Mansi and the Synya and Kazym Khanty dialects.

THESIS 2b:

I created morphologies for two seriously endangered Northern Samoyedic languages, Nganasan and Tundra Nenets, using the Xerox finite-state formalism.

Related publications: 22, 63, 59, 61, 14, 58, 55, 56, 54, 48, 49, 50

2.3 Adaptation of the Hungarian morphology to special domains

Language use in special domains and language variants may deviate in a significant manner from what one encounters in the standard written dialect of the language. The morphological model needs to be adapted when texts from such a special language variant are to be analyzed. Two examples of such phenomena were described here, demonstrating the adaptibility of the analyzer built using the development framework. The first task for which the analyzer was adapted, was the annotation of Old and Middle Hungarian texts. The adapted analyzer can handle extinct morphological constructions as well as dialectal variants missing from Modern Standard Hungarian. The other example is an adaptation of the Hungarian morphology to the clinical domain, where the domain-specific terminology, which includes a vast amount of word forms of foreign origin, had to be treated in a robust manner.

THESIS 3:

I demonstrated the adaptability of morphologies created using the formalism I introduced by extending the Hungarian morphology I created.

THESIS 3a:

I adapted the Humor morphological analyzer for Hungarian to be capable of analyzing words containing morphological constructions, suffix allomorphs, suffix morphemes, paradigms and stems that were used in Old and Middle Hungarian but no longer exist in present-day Hungarian.

Related publications: 38, 42

THESIS 3b:

I created a method for semi-automatically extending the Humor morphological analyzer to be capable of analyzing words used in the clinical language that contains non-standard word constructions, phrases of foreign origin and a high ratio of abbreviations. I created methods to distinguish words of foreign origin, to predict their pronunciation and to predict the part of speech of words to be added to the lexicon. The resulting analyzer is able to analyze medical language in an appropriate and robust manner.

Related publications: 2, 37, 65, 36, 7, 1

2.4 Finite-state implementation of constraint-based morphologies

Humor's closed-source licensing scheme has been a limitation to making resources made for it widely available. Moreover, there are a few limitations of the rule-based Humor engine: lack of support for morphological guessing and the use of frequency information or other weighting of the models. These problems were solved by converting the databases to a finite-state representation that allows morphological guessing and the addition of weights and has open-source implementations.

The Xerox tools (Beesley and Karttunen, 2003) implement a powerful formalism to describe complex types of morphological structures. This suggested that mapping of the morphologies implemented in the Humor formalism to a finite-state representation should have no impediment.

THESIS 4:

I created a method to convert the Humor databases to a finite-state representation that allows morphological guessing and the addition of weights and has open source implementations.

Related publications: 30, 31

2.5 Extending morphological dictionary-based models without writing a grammar

Most freely available morphological resources contain no rule component. They are usually based on just a morphological lexicon, containing base forms and some information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with some other morphosyntactic features. Resources of this type are much more difficult to extend with new words than rule-based morphologies. However, the application of machine learning methods may be able to make up for the lack of a rule component. I prepared an algorithm that makes the integration of new words into such resources as easy as a rule-based morphology can be extended. This is achieved by predicting the correct paradigm

for words, which are not present in the lexicon. The suffix-trie-based supervised learning algorithm is based on longest matching suffixes and lexical frequency data, and it was used to extend a Russian morphology, on which its performance was evaluated. With minimal adaptation, the tool can be used for any language provided there is a morphological resource available. I assumed that a dictionary with some lexical features is also available, thus such features could be used for disambiguating paradigm candidates. The results showed that the method performs with an accuracy of about 90% in all different setups, achieving the best performance on relatively rare words, which are good candidates of being absent in the original lexicon.

I found that assigning more weight to distributions conditioned on longer suffixes than on shorter ones yields much better prediction performance, not only in terms of the number of exact predicted paradigm matches, but especially when taking into account what sorts of errors the system makes. While the baseline suffix guesser algorithm often proposes paradigms inapplicable to the given lexical item, my algorithm makes errors that arise due to the lack of lexical semantic information. Humans would make similar errors in similar situations.

THESIS 5:

I prepared an algorithm that makes the integration of new words into lexicon-based morphological resources easy by automatically predicting the correct paradigm for words which are not present in the lexicon.

Related publications: 27, 5

2.6 A flexible model of word form generation and lemmatization

The original Humor system lacked the capability of word form generation and the existing lemmatizer was unable to correctly lemmatize certain non-trivial word constructions. I solved these problems by extending the system so that it can also be used as a morphological generator and implementing better lemmatization algorithms.

The generator produces all word forms that could be realizations of a given morpheme sequence. The input for the generator is a lemma followed by a sequence of category labels that express the morphosyntactic features the word form should expose.

The Humor generator is not a simple inverse of the corresponding analyzer: it can generate the inflected and derived forms of any multiply derived and/or compound stem without explicitly referring to compound boundaries and derivational suffixes in the input even if the whole complex stem is not in the lexicon of the analyzer. This is a useful feature in the case of languages where morphologically very complex stems are commonplace. When generating inflected (or derived) forms of a morphologically complex stem, one does not have to be concerned whether the stem is included in the stem database. If the corresponding analyzer can analyze it in any way, the generator will be able to correctly generate its inflected forms.

It is possible to describe preferences for the cases when a certain set of morphosyntactic features may have more than one possible realization. This can be useful for such applications of the generator as text generation in a machine translation system, where the generation of a single preferred word form is required. The Hungarian morphological description was extended with information expressing markedness. Marked forms are automatically removed during compilation from the version of the database that is intended for word form generation in the machine translation system.

Since there is considerable variation in suffix ordering in some of the languages for which I created morphologies (e.g. Komi), I also created a version of the generator that has another useful feature: it does not assume that the morphosyntactic features are properly ordered in the input, rather it considers them a set.

The Humor 'lemmatizer' tool, built around the analyzer core, does more than just identifying lemmas of word forms: it also identifies the exposed morphosyntactic features. In contrast to the more verbose analyses produced by the core analyzer, compound members and derivational suffixes do not appear as independent items in the output of the lemmatizer, so the internal structure of words is not revealed. The analyses produced by the lemmatizer are well suited for such tasks as corpus tagging, indexing and parsing.

There are two implementations of the lemmatizer. Both can properly handle the task of correctly lemmatizing and filtering special word constructions, e.g. ones that are not (only) suffixed at the end of the word thus improving the accuracy of lemmatization.

THESIS 6:

I created new word form generator and lemmatizer tools for the Humor system.

THESIS 6a:

I implemented a word form generator as a Humor module, which can generate the inflected and derived forms of any multiply derived and/or compound stem without explicitly referring to compound boundaries and derivational suffixes in the input even if the whole complex stem is not in the lexicon of the analyzer. The generator was used in various commercial applications and research prototype systems in the domain of information retrieval and machine translation.

THESIS 6b:

I created a lemmatizer tool for Humor, which can properly handle the task of correctly lemmatizing and filtering special word constructions, e.g. ones that are not (only) suffixed at the end of the word. The lemmatizer was used in various corpus annotation projects and it was integrated into information retrieval and machine translation systems.

Related publications (on applications of the tools): 3, 35, 41, 53, 51, 52, 11, 49

2.7 A tool for annotating and searching text corpora

To support the process of manual checking and the initial manual disambiguation of an annotated corpus, I created a web-based interface where disambiguation and normalization errors can be corrected very effectively. The system presents the document to the user using an interlinear annotation format that is easy and natural to read and it supports handling glosses, normalization and translations.

I also created a web-based corpus query tool, which does not only make it possible to search for different grammatical constructions in the texts, but it is also an effective correction tool. Errors discovered in the annotation or the text appearing in the "results" box can immediately be corrected and the corrected text and annotation is recorded in the database. Naturally, this latter functionality of the corpus manager is only available to expert users having the necessary privileges.

A fast and effective way of correcting errors in the annotation is to search for presumably incorrect structures and to correct the truly problematic ones at once. The corrected corpus can be exported after this procedure and the tagger can be retrained on it.

THESIS 7:

I developed a disambiguation system that can be used for automatic and manual disambiguation of the morphosyntactic annotation and glossing of texts and I created a corpus manager appropriate for searching and correcting annotated corpora.

Related publications: 42, 38, 49, 50

3

List of Papers

Journal publications

1 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents, In: *Computer Speech & Language*, Vol. 35, pp. 219-233, ISSN 0885-2308, http://dx.doi.org/10.1016/j.csl.2014.09.001.

2 György Orosz, **Attila Novák**, Gábor Prószéky (2014): Lessons learned from tagging clinical Hungarian. In: *International Journal of Computational Linguistics and Applications*, Vol. 5 no. 2. ISSN 0976-0962

3 László János Laki, **Attila Novák**, Borbála Siklósi, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. In: *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63–78. ISSN 0976-0962

4 István Endrédy, **Attila Novák** (2013): More effective boilerplate removal – the GoldMiner algorithm. In: *Polibits 48.* pp. 79–83. ISSN 1870-9044

Book chapters

5 **Attila Novák** (2015): Making morphologies the 'easy' way, In: A. Gelbukh (ed.) *Lecture Notes in Computer Science Volume 9041: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin–Heidelberg. Part I pp. 127–138. ISBN 978-3-319-18110-3

6 Borbála Siklósi, **Attila Novák** (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (eds.) *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin–Heidelberg. pp. 233–243 ISBN 978-3-319-11396-8

7 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2013): Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In: Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, Bianca Truthe (eds.) *Lecture Notes in Computer Science Volume 7978: Statistical Language and Speech Processing, First International Conference, SLSP 2013.* Springer, Berlin Heidelberg. pp. 248–259 ISBN 978-3-642-39592-5

8 György Orosz, László János Laki, **Attila Novák**, Borbála Siklósi (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue, 16th International Conference, TSD 2013.* Pilsen, Czech Republic. Springer, Berlin–Heidelberg. pp. 280–287. ISBN: 978-3-642-40584-6

9 Nóra Wenszky, **Attila Novák** (2013): The hypercorrect key witness. In: Péter Szigetvári (ed.) *VLlxx: Papers presented to Varga László on his 70th birthday.* Department of English Linguistics, Eötvös Loránd University. ISBN 978-963-284-315-5

10 Borbála Siklósi, **Attila Novák** (2013): Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin Heidelberg. pp. 318–328. ISBN 978-3-642-45114-0

11 László János Laki, György Orosz, **Attila Novák** (2013): HuLaPos 2.0 – Decoding morphology. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin–Heidelberg. pp. 294–305. ISBN 978-3-642-45114-0

12 György Orosz, **Attila Novák**, Gábor Prószéky (2013): Hybrid text segmentation for Hungarian clinical records. In: F. Castro, A. Gelbukh, M.G. Mendoza (eds.) *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin–Heidelberg. pp. 306–317. ISBN 978-3-642-45114-0

13 **Novák Attila**, Wenszky Nóra (2007): Mire jó és hogyan készül egy számítógépes morfológia. In: Alberti Gábor, Fóris Ágota (eds.) *A mai magyar formális nyelvtudomány műhelyei.* Nemzeti Tankönyvkiadó, Budapest. 157–169.

14 Gábor Prószéky, **Attila Novák** (2005): Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.) *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday.* Gummerus Printing, Saarijärvi/CSLI Publications, Stanford. pp. 116–125.

15 **Novák Attila** (2002): Többértelmű vagy homályos? In: Kálmán László, Trón Viktor, Varasdi Károly (eds.) *Lexikalista elméletek a nyelvészetben.* Tinta Könyvkiadó, Budapest. (Segédkönyvek a nyelvészet tanulmányozásához 13.) pp. 277–287.

16 **Novák Attila** (2002): HPSG fonológia. In: Kálmán László, Trón Viktor, Varasdi Károly (eds.) *Lexikalista elméletek a nyelvészetben.* Tinta Könyvkiadó, Budapest. (Segédkönyvek a nyelvészet tanulmányozásához 13.) pp. 99–128.

17 Kálmán László, **Novák Attila** (2001): A magyar egyszerű mondat fajtái. In: Kálmán László (ed.): *Magyar leíró nyelvtan*, Mondattan I. Tinta Könyvkiadó, Budapest, 2001. pp. 10–23.

18 Gyuris Bea, **Novák Attila** (2001): A topik és a kontrasztív topik. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 24–53.

19 **Novák Attila**, Dudás Kálmán, Kálmán László (2001): Igevivők. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 54–75.

20 **Novák Attila** (2001): A kommentelőzmények. In: Kálmán László (ed.): *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 76–91.

21 **Novák Attila** (2001): A hatókör felszíni egyértelműsítése. In: Kálmán László (eds.) *Magyar leíró nyelvtan, Mondattan I.* Tinta Könyvkiadó, Budapest, 2001. pp. 92–97.

22 **Novák Attila** (1999): Inflectional paradigms in Hungarian – The conditioning of suffix- and stem-alternations (Ragozási paradigmák a magyarban – A toldalék- és tőalternációkat kiváltó tényezők), Szakdolgozat, ELTE Elméleti Nyelvészet Szak, Budapest.

23 **Attila Novák** (1998): HPSG Phonology. In: *Lexicon Matters.* ELTE Theoretical Linguistics Programme, Budapest, 1998. pp. 33–48

24 **Attila Novák** (1998): Ambiguity and Vagueness. In: *Lexicon Matters.* ELTE Theoretical Linguistics Programme, Budapest, 1998. 115–120

Conference proceedings

25 **Novák Attila**, Siklósi Borbála (2015): Automatic Diacritics Restoration for Hungarian. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics. pp. 2286–91.

26 Siklósi Borbála, **Novák Attila** (2015): Restoring the intended structure of Hungarian ophthalmology documents. In: *Proceedings of the BioNLP 2015 Workshop at the* 53^{rd} *Annual Meeting of the Association for Computational Linguistics, ACL 2015.* Beijing, China. pp. 152–157

27 **Novák Attila** (2015): "Olcsó" morfológia In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 145–157

28 Siklósi Borbála, **Novák Attila** (2015): Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 237–248

29 Borbála Siklósi, **Attila Novák**, Gábor Prószéky (2014): Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*. Reykjavík. pp. 69–75

30 **Attila Novák** (2014): A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavík. pp. 1068–1073

31 **Novák Attila** (2014): A Humor új Fo(r)mája. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 303–308. ISBN 978-963-306-246-3

32 Siklósi Borbála, **Novák Attila** (2014): Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 167–176. ISBN 978-963-306-246-3

33 Siklósi Borbála, **Novák Attila** (2014): A magyar beteg. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 188–198. ISBN 978-963-306-246-3

34 Orosz György, **Novák Attila** (2014): PurePos 2.0: egy hibrid morfológiai egyértelműsítő rendszer. In: Tanács Attila, Varga Viktor, Vincze Veronika (eds.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 373–377. ISBN 978-963-306-246-3

35 Laki László, **Novák Attila**, Siklósi Borbála (2013): Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 71–82 ISBN 978-963-306-189-3

36 Siklósi Borbála, **Novák Attila**, Prószéky Gábor (2013): Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 148–158 ISBN 978-963-306-189-3

37 Orosz György, **Novák Attila**, Prószéky Gábor (2013): Magyar nyelvű klinikai rekordok morfológiai egyértelműsítése. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 159–169 ISBN 978-963-306-189-3

38 **Novák Attila**, Wenszky Nóra (2013): O & kọzepmaǵar zoalactany elemzọ. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 170–181 ISBN 978-963-306-189-3

39 Endrédy István, **Novák Attila** (2013): Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve. In: Tanács Attila; Vincze Veronika (eds.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 297–301 ISBN 978-963-306-189-3

40 György Orosz, László János Laki, **Attila Novák**, Borbála Siklósi (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (eds.) *2nd Symposium on Languages, Applications and Technologies.* Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249–257 ISBN 978-3-939897-52-1

41 László János Laki, **Attila Novák**, Borbála Siklósi (2013): English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules. In: Marta R. Costa-jussa, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych (eds.) *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra).* Association for Computational Linguistics. pp. 42–50

42 **Attila Novák**, György Orosz, Nóra Wenszky (2013): Morphological annotation of Old and Middle Hungarian corpora. In: Piroska Lendvai, Kalliopi Zervanou (eds.) *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* Association for Computational Linguistics. pp. 43–48

43 György Orosz, **Attila Novák** (2013): Purepos 2.0: a hybrid tool for morphological disambiguation. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.) *Proceedings of the international conference Recent Advances In Natural Language Processing RANLP 2013.* Hissar, Bulgaria. pp. 539–545 ISSN 1313-8502

44 György Orosz, **Attila Novák** (2012): PurePos – an open source morphological disambiguator. In: Bernadette Sharp, Michael Zock (eds.) *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science.* Wrocław, Poland. pp. 53–63

45 Borbála Siklósi, György Orosz, **Attila Novák**, Gábor Prószéky (2012): Automatic structuring and correction suggestion system for Hungarian clinical records. In: *LREC-2012: SALTMIL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages".* Istanbul, Turkey, 2012. pp. 29–34

46 Siklósi Borbála, Orosz György, **Novák Attila** (2011): Magyar nyelvű klinikai dokumentumok előfeldolgozása. In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 143–340

47 **Novák Attila**, Orosz György, Indig Balázs (2011): Javában taggelünk. In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 336–340.

48 Fejes László, **Novák Attila** (2010): Obi-ugor morfológiai elemzők és korpuszok. In: *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010).* Szegedi Tudományegyetem, pp. 284–291

49 Bakró-Nagy Marianne, Endrédy István, Fejes László, **Novák Attila**, Oszkó Beatrix, Prószéky Gábor, Szeverényi Sándor, Várnai Zsuzsa, Wagner-Nagy Beáta (2010): Online morfológiai elemzők és szóalakgenerátorok kisebb uráli nyelvekhez. In: *VII. Magyar*

Számítógépes Nyelvészeti Konferencia (MSZNY 2010). Szegedi Tudományegyetem, pp. 345–348

50 István Endrédy, László Fejes, **Attila Novák** , Beatrix Oszkó, Gábor Prószéky, Sándor Szeverényi, Zsuzsa Várnai, Beáta Wágner-Nagy (2010): Nganasan – Computational Resources of a Language on the Verge of Extinction. In: *Creation and Use of Basic Lexical Resources for Less-Resourced Languages: 7th SaLTMiL Workshop (LREC-2010)*. La Valletta, Malta, pp. 41–44

51 **Novák Attila**, Prószéky Gábor (2009): Kísérletek statisztikai és hibrid magyar–angol és angol– magyar fordítórendszerek megvalósítására. In: **VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2009)**. Szegedi Tudományegyetem, pp. 25–34

52 **Attila Novák** (2009): MorphoLogic's submission for the WMT 2009 Shared Task. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL 2009.* Athens, Greece. pp. 155–159

53 **Attila Novák**, László Tihanyi, Gábor Prószéky (2008): The MetaMorpho translation system. In: *Proceedings of the Third Workshop on Statistical Machine Translation at ACL 2008.* Columbus, Ohio. pp. 111–114

54 **Attila Novák** (2008): Language resources for Uralic minority languages. In: *Proceedings of the SALTMIL Workshop at LREC-2008: Collaboration: interoperability between people in the creation of language resources for less-resourced languages.* Marrakech, pp. 27–32

55 **Novák Attila**, M. Pintér Tibor (2006): Milyen a még jobb Humor. In: *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*. Szegedi Tudományegyetem, pp. 60–69

56 **Attila Novák** (2006): Morphological Tools for Six Small Uralic Languages. In: *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, pp. 925–930

57 **Novák Attila**, Endrédy István (2005): Automatikus ë-jelölő program. In: *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*. Szegedi Tudományegyetem, pp. 453–454

58 **Novák Attila**, Wenszky Nóra (2005): Tundrai nyenyec morfológiai elemző és generátor. In: *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2005)*. Szegedi Tudományegyetem, pp. 200–208

59 **Novák Attila** (2004): Az első nganaszan szóalaktani elemző. In: *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004)*. Szegedi Tudományegyetem, pp. 195–202

60 **Attila Novák**, Viktor Nagy, Csaba Oravecz (2004): Combining symbolic and statistical methods in morphological analysis and unknown word guessing. In: *Proceedings of The Fourth International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon, pp. 1255–1258

61 **Attila Novák** (2004): Creating a Morphological Analyzer and Generator for the Komi language. In: *Proceedings of the SALTMIL Workshop at LREC-2004: First Steps in Language Documentation for Minority Languages.* Lisbon, pp. 64–67.

62 **Novák Attila**, Nagy Viktor, Oravecz Csaba (2003): Magyar ismeretlenszó-elemző program fejlesztése. In: *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003).* Szegedi Tudományegyetem, 45–57

63 **Novák Attila** (2003): Milyen a jó Humor? In: *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003).* Szegedi Tudományegyetem, pp. 138–145

64 **Attila Novák**, Viktor Nagy, Csaba Oravecz (2003): Corpus assisted development of a Hungarian morphological analyser and guesser. In: Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 conference.* UCREL technical paper number 16. UCREL, Lancaster University, pp. 583–590

Research reports

65 Borbála Siklósi, **Attila Novák**, György Orosz, Gábor Prószéky (2014): Processing noisy texts in Hungarian: a showcase from the clinical domain, In: Péter Szolgay (ed.), *Jedlik Laboratories Reports*, Vol. II, no. 3, pp. 5–62 ISSN 2064-3942

Bibliography

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In Holub, J. and Zdárek, J., editors, *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Ventura Hall.

Huldén, M. (2009). Foma: a Finite-State Compiler and Library. In Lascarides, A., Gardent, C., and Nivre, J., editors, *Proceedings of EACL 2009*, pages 29–32, Athens, Greece. The Association for Computer Linguistics.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., and Pirinen, T. (2011). HFST—Framework for Compiling and Applying Morphologies. In Mahlow, C. and Pietrowski, M., editors, *Systems and Frameworks for Computational Morphology*, volume Vol. 100 of *Communications in Computer and Information Science*, pages 67–85.

Novák, A. (1999). Inflectional paradigms in hungarian – the conditioning of suffix and stem alternations. Master's thesis, ELTE Theoretical Linguistics Programme, Budapest.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, College Park, Maryland. Association for Computational Linguistics.