12

Computational Morphologies for Small Uralic Languages

GÁBOR PRÓSZÉKY AND ATTILA NOVÁK

12.1 Introduction

This article presents a set of morphological tools for small Uralic languages. Various Hungarian research groups specialized in Finno-Ugric linguistics and a Hungarian language technology company (MorphoLogic) have initiated a project with the goal of producing annotated electronic corpora for small Uralic languages. The languages described include Mordvin, Udmurt (Votyak), Komi (Zyryan), Mansi (Vogul), Khanty (Ostyak), Tundra Nenets (Yurak) and Nganasan (Tavgi). Most of these languages are endangered, some of them are on the verge of extinction, so their documentation is an urgent scientific task. The most important subgoal of the project was to create morphological analyzers for the languages involved.[1]

In the project, we used the morphological analyzer engine called *Humor* ('High speed Unification MORphology') developed at MorphoLogic (Prószéky and Kis (1999)), which had been first successfully applied to another Uralic (Finno-Ugric) language, Hungarian, and later to various Slavic, Germanic and Romance languages. We supplemented the analyzer with two additional tools: a lemmatizer and a morphological generator. We present the tools through their application to the Komi language, specifically to the standard Komi-Zyryan dialect.

Creating analyzers for the two Samoyed languages involved in the project,

[1] The project was funded by the National Research and Development Programmes of Hungary ('Complex Uralic Lingustic Database', NKFP 5/135/2001).

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

Nenets and Nganasan, turned out to be a great challenge. Nganasan morphology and especially its phonology is very complex and the available linguistic data and their linguistic descriptions proved to be incomplete and partly contradictory, which made numerous revisions to our computational model necessary. Thus using the Humor formalism, which we successfully applied to other languages in and outside the project, was not feasible in the case of Nganasan, as shown in the second part of the present article. We used instead the regular relation calculus based toolset, *xfst* of Xerox to create the analyzer.

12.2 The Humor Tools

12.2.1 Features of the Morphological Analyzer

The Humor analyzer performs a classical 'item-and-arrangement' (IA) style analysis. The input word is analyzed as a sequence of morphs. It is segmented into parts which have (i) a surface form (that appears as part of the input string), (ii) a lexical form (the 'quotation form' of the morpheme) and (iii) a category label (which may contain some structured information or simply be an unstructured label). The lexical form and the category label together more or less well identify the morpheme of which the surface form is an allomorph.

The analyzer produces flat morph lists as possible analyses, since it contains a regular word grammar, which is represented as a finite-state automaton.

The following is a sample output of the Humor analyzer for the Komi word form *kylanly* ('to a listener/listening one').

```
analyzer>kylanly
 kyv[S_V]=kyl+an[D=A_PImpPs]+ly[I_DAT]
 kyv[S_V]=kyl+an[D=N_Tool]+ly[I_DAT]
```

Morphs are separated by + signs from each other. The representation of morphs is `lexical form[category label]=surface form`. A prefix in category labels identifies the morphological category of the morpheme (stem, derivational/inflectional suffix). In the case of derivational affixes, the syntactic category of the derived word is also given.

12.2.2 How the analyzer works

The program performs a search on the input word form for possible analyses. It looks up morphs in the lexicon the surface form of which matches the beginning of the input word (and later the beginning of the yet unanalyzed part of it). The lexicon may contain not only single morphs but also morph sequences. These are ready-made analyses for irregular forms of stems or suffix sequences, which can thus be identified by the analyzer in a single step, which makes its operation more efficient.

In addition to assuring that the requirement that the surface form of the

next morpheme must match the beginning of the yet unanalyzed part of the word (uppercase-lowercase conversions may be possible) is met, two kinds of checks are performed by the analyzer at every step, which make an early pruning of the search space possible.

On the one hand, it is checked whether the morph being considered as the next one is locally compatible with the previous one. On the other hand, it is examined whether the candidate morph is of a category which, together with the already analyzed part, is the beginning of a possible word construction in the given language. Possible word structures are described by an extended finite-state automaton.

12.2.3 The Lemmatizer

Our '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. The output of the lemmatizer and the analyzer is compared in the example below:

```
analyzer>kylanly
 kyv[S_V]=kyl+an[D=A_PImpPs]+ly[I_DAT]
 kyv[S_V]=kyl+an[D=N_Tool]+ly[I_DAT]
lemmatizer>kylanly
 kylan[N][DAT]
 kylan[A][DAT]
```

The lemmatizer identifies the word form *kylanly* as the dative of the noun or adjective (in fact: participle) *kylan* ('listener', 'listening one').

12.2.4 The Generator

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 which express the morphosyntactic features the word form should expose.

The generator is not a simple inverse of the corresponding analyzer, thus 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.

The following examples show how the generator produces an inflected form of the derived nominal stem *kylan*, which is not part of the stem lexicon, and the explicit application of the derivational suffix (and the same inflectional suffix) to the absolute verbal root of the word.

```
generator>kylan[N][DAT]
 kylanly
generator>kyv[V][_Tool][DAT]
 kylanly
```

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 machine translation applications, where the generation of a single word form is required.

12.3 The Morphological Database

Various versions of the Humor morphological analyzer have been in use for over a decade now. Although the analyzer itself proved to be an efficient tool, the format of the original database turned out to be problematic. For the analyzer to work efficiently, the data structures it uses contain redundant data. However, these redundant data structures are hard to read and modify for humans. So we built a morphological description development environment which facilitates the creation of the database.

12.3.1 Creating a Morphological Description

In the environment, the linguist has to create a high level human readable description which contains no redundant information and the system transforms it in a consistent way to the redundant representations which the analyzer uses. The work of the linguist consists of the following tasks:
a. Identification of the relevant morpheme categories in the language to be described (parts of speech, affix categories).
b. Description of stem and suffix alternations: an operation must be described which produces each allomorph from the lexical form of the morpheme for each phonological allomorphy class. The morphs or phonological or phonotactic properties which condition the given alternation must be identified.
c. Identification of features: all features (pertaining to the category or shape of morphemes, or to the idiosyncratic allomorphies triggered) playing a role in the morphology of the language must be identified.
d. Definition of selectional restrictions between adjacent morphs: selectional restrictions are described in terms of requirements that must be satisfied by the set of properties (features) of any morph adjacent to a morph. Each morph has two sets of properties: one can be seen by morphs adjacent to the left and

the other by morphs adjacent to the right. Likewise, any morph can constrain its possible neighbors by defining a formula expressing its requirements on each of its two sides.
e. Identification of implicational relations between properties of allomorphs and morphemes: these implicational relations must be formulated as rules, which either define how redundant properties and requirements of allomorphs can be inferred from their already known (lexically given or previously inferred) properties (including their shape), or define default properties.
f. Creation of stem and affix morpheme lexicons: in contrast to the lexicon used by the morphological analyzer, the lexicons created by the linguist contain the descriptions of morphemes instead of allomorphs. Morphemes are defined by listing their lexical form, category and all unpredictable features and requirements. A simple inheritance mechanism facilitates the consistent treatment of complex lexical entries (primarily compounds).
g. Creation of a word grammar: restrictions on the internal morphological structure of words (including selectional restrictions between nonadjacent morphemes) are described by a regular word grammar.
h. Creation of a suffix grammar (optional): a suffix grammar can be defined by setting up morphotactic classes for the suffixes and creating a directed graph labeled with the name these classes on its arcs. The development environment can produce segmented suffix sequences using this description and the suffix lexicon. Using such preprocessed segmented sequences enhances the performance of the analyzer.

As it can be seen from the description of the tasks above, we encourage the linguist to create a real analysis of the data (within the limits of the model that we provide).

12.3.2 Conversion of the Morphological Database

Using a description that consists of the information described above, the development environment can produce a lexical representation which already explicitly contains all the allomorphs of each morpheme along with all the properties and requirements of each of them. This representation still contains the formulae expressing properties and selectional restrictions in a human-readable form and can thus be easily checked by a linguist.

The readable redundant representation is then transformed to the format used by the analyzer using an encoding definition description, which defines how each of the features should be encoded for the analyzer.

12.4 The Komi Analyzer

In the subproject on Komi, which concentrates on the standard Komi-Zyryan dialect, we created a Komi morphological description using the development environment described in the previous section.

12.4.1 The Language

Komi (or Zyryan, Komi-Zyryan) is a Finno-Ugric language spoken in the northeastern part of Europe, West of the Ural Mountains. The number of speakers is about 300,000.

12.4.2 Creating a Komi Morphological Description

Since the annotated corpora we want to create are intended for linguists, we decided to use a quasi-phonological transcription of Komi based on Latin script instead of the Cyrillic orthography of the language. However, we plan to produce a Cyrillic version of the analyzer as well.

The first piece of description we created in the Komi subproject was a lexicon of suffix morphemes along with a suffix grammar, which describes possible nominal inflectional suffix sequences. One of the most complicated, though quite properly described, aspect of Komi morphology is the intricate interaction between nominal case and possessive suffixes.

A problem we were faced with was the lack of good and thorough modern synchronic grammars on many of the languages involved in the project. This was also the case for Komi, so we had to do a lot of research on the distribution of individual morphemes and allomorphies. In some cases we managed to get some information by producing the forms in question (along with their intended meaning) with the generator and having native speakers judge them.

An initial stem lexicon was created by hand using corpus data and a printed Komi-Russian dictionary (Beznosikova (2000)). Later we managed to acquire the dictionary in an electronic form. It contains about 31,000 stems plus 2800 names. Its conversion to the format used by the development environment is in progress.

There is a number of stem alternations in Komi. They are all triggered by attaching vowel initial suffixes. The alternations themselves are also very simple (there is an *l*–*v* alternation class and a number of epenthetic classes).

In many cases, it is predictable from the (quotation) form of a stem on phonotactic grounds whether it belongs to an alternation class. In other cases, this information must be entered into the stem lexicon. Since the underlying rules had not been described, finding them out was our task, and it is one of the scientific outcomes of the project.

12.5 Creating a morphological description for Nganasan

A formal description of Nganasan was written by fellow linguists taking part in the project (Wagner-Nagy (2002)). They also digitized a Russian-Nganasan dictionary (Kost'erkina et al. (2001)) and converted it to the phonemic transcription based on Latin script used by their team. The dictionary contains approximately 3,650 non-derived roots. The Nganasan team also provided

category labels for each item, which was missing from the original source. Wagner-Nagy (2002) also contains some short texts which we could use as a corpus along with a collection of text from other sources. Later we added another 500 roots encountered when testing the analyzer on this corpus.

During the preparation of the above described root dictionary, we also started to describe the suffixes of Nganasan in a formal manner. The first step of this was the creation of a list of the suffixes that contained the underlying phonological form of each suffix together with its category label, plus a feature that indicates which morphological root form the suffix can attach to. We used the following model to describe Nganasan morphology: we hypothesize that each root morpheme has three morphological stem variants (out of which two or all three might have the same form), and suffixes are sorted into three groups depending on which root allomorph they attach to. We also described the morphotactic restrictions governing the linear order of suffixes by defining a suffix grammar. The underlying phonological representation contains some archiphonemes: harmonic vowels and 'quasi-consonants' which never appear on the surface but condition gradation.

In Nganasan, nominal and verbal roots follow different alternation patterns. Additionally, vowel final and consonant final roots also exhibit different behavior. Some root-final changes are restricted to lexically marked root classes. Each of these roots must have a relevant lexical mark in the root inventory. Other root-final changes occur in each root satisfying the formal requirements of the rule.

12.6 The complexity of Nganasan morphophonology

It was relatively easy to describe root-final sound alternations in the Humor formalism. Those productive phonological processes that are sensitive to local contexts (such as degemination) could be formalized as separate rules. However, the phenomenon of gradation (i.e. the rule-governed alternation of obstruents in syllable onsets) proved to be so complex that we could not describe it satisfactorily. The root of the problem is that the Humor analyzer sees each word as a sequence of allomorphs and during analysis it checks whether the adjacent morphs are locally compatible with each other. Nganasan gradation, however, does not depend on the morphological make-up of the word: the only factor at play is syllable structure. Syllable boundaries and morph boundaries do not usually coincide. In the case of short suffixes (made-up of one segment), it is possible that even non-adjacent morphs belong to the same syllable. Moreover, the rules governing gradation in Nganasan are quite intricate. An obstruent in the onset position is in strong grade (i) in even-numbered open syllables (if not preceded by a long vowel) and (ii) if it is preceded by a non-nasal coda consonant. Otherwise, it is in rhythmical weak

grade (i) if preceded by a long vowel or (ii) if it is in odd-numbered syllable. Otherwise, it is in syllabic weak grade in even-numbered closed syllables. Gradation combines with other alternations in the language: vowel harmony, degemination, root alternations and various morphophonological suffix alternations (as a result of which a monosyllabic suffix can have as many as 20 different allomorphs).

To illustrate the complexity of the above outlined system let us look at the allomorphs of a single verbal suffix (of narrative mood used in the subjective and the non-plural objective conjugations). The underlying representation of the morpheme is `hA2nhV`, and its 12 allomorphs are: *banghu, bjanghy, bambu, bjamby, bahu, bjahy, hwanghu, hjanghy, hwambu, hjamby, hwahu, hjahy*. These allomorphs are produced from the underlying representation by the general phonological processes of the langage, undergoing vowel harmony, *a*-diphthongization and gradation.

While gradation is extremely difficult to formalize as a set of allomorph adjacency restrictions, it is such a productive process in Nganasan that it must be included in a proper morphological analyzer. It seemed, however, that though the formalism of the Humor analyzer proved to be adequate for the description of most phenomena in the language, the rule-formalism of the development environment could not cover all of the essential processes.

12.7 The application of a new formalism

In June 2003, a book was published (Beesley and Karttunen (2003)) with a CD containing a version of the two level morphological toolset of Xerox. This program set is based on finite state transducer technology and the versions published with the book can be freely used for non-commercial purposes. We decided to rewrite our description of Nganasan in the format used by the Xerox programs: lexc (Lexicon Compiler) and xfst (Xerox Finite-State Tool).

Using the xfst formalism, we could create a full description of Nganasan. The calculus implemented by the program makes it possible to ignore irrelevant symbols (such as morpheme boundaries in the case of gradation) in the environment description of re-write rules, therefore environments encompassing non-adjacent morphemes can be easily defined. As during composition the program automatically eliminates intermediate levels of representation created by individual rules producing a single finite-state transducer, generation and analysis can be performed efficiently.

Nganasan gradation was described in xfst as a cascade of rules performing syllabification, the identification of syllable grades, changing the quality of the obstruents in syllable onsets and removing auxiliary symbols. The rule system covers the irregularities of Nganasan syllabification. The whole of the rule system naturally contains several other rules. It describes all pro-

ductive, automatic phonological rules (e.g. the assimilation of nasals to the immediately following obstruent, degemination, vowel harmony, nunnation, palatalization etc.) and morphologically or lexically constrained root and suffix alternations.

We converted our morpheme inventories into the format used by lexc. Some of the feature-based constraints of the Humor description (e.g. the morphological stem selection) were retained in the new formalism: we used the 'flag diacritics' construct of the Xerox tools to implement them.

12.8 Conclusion

In addition to the ones described above, analyzers for Udmurt, Mari and Tundra Nenets have been finished.[2] The former two were prepared using the Humor based formalism, the latter was implemented using xfst and lexc. Additional analyzers for Mansi, Khanty and Mordvin are under construction using the Humor formalism.

A very important result of the project besides creating the programs and annotated corpora using them is that many gaps, uncertainties and inconsistencies were detected and in many cases corrected in the written grammars of these languages. Many details of the description which often remain vague in written grammars (such as the ordering and exact formulation of rewrite rules) must unavoidably made explicit in a computationally implemented grammar. Moreover, the adequacy of the implemented grammar can be very thoroughly tested against a great amount of real lingustic data. Systematic comparison of word forms generated against model paradigms has pinpointed errors not only in the computational implementation (which were then eliminated) but also in the model paradigms or the grammars the computational implementation was based on. We consider it very important to provide feedback to the linguists having prepared the original grammars and to publish the linguistic results of the project. We also hope that the many questions which remained open will induce further field research concerning these endangered languages and that they will be answered before it is too late.

References

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. Ventura Hall: CSLI Publications.

Beznosikova, Ljucija, ed. 2000. *Komi-Roča Kyvčukör (Komi-Russian Dictionary)*. Syktyvkar: Komi kniéznoe izdat.

[2] The individual analyzers were created by Attila Novák in co-operation with László Fejes (Komi, Udmurt, Mari), Beáta Wagner-Nagy and Zsuzsa Várnai (Nganasan) and Nóra Wenszky (Tundra Nenets). The Tundra Nenets analyzer is based on Tapani Salminen's work (Salminen (1997) and Salminen (1998), which he kindly made available to us in a machine readable form) and was created in close on-line co-operation with him.

Kost'erkina, N. T., A. Č. Momd'e, and T. Ju. Ždanova. 2001. *Slovar' nganasansko-russkij i russko-nganasanskij*. Sankt-Pet'erburg: Prosvesčen'ije.

Prószéky, Gábor and Balázs Kis. 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*, pages 261–268. College Park, Maryland, USA.

Salminen, Tapani. 1997. *Tundra Nenets inflection*. Helsinki: Mémoires de la Société Finno-Ougrienne 227.

Salminen, Tapani. 1998. *A morphological dictionary of Tundra Nenets*. Helsinki: Lexica Societatis Fenno-Ugricae 26.

Wagner-Nagy, Beáta, ed. 2002. *Chrestomathia Nganasanica*. Szeged – Budapest: SZTE Finnugor Tanszék – MTA Nyelvtudományi Intézet.