Language resources for Uralic minority languages

Attila Novák
MorphoLogic Ltd.
1126 Budapest Orbánhegyi út 5., Hungary
novak@morphologic.hu

Abstract

Most members of the Uralic language family are small minority languages spoken on the territory of the Russian Federation, which all are endangered. In past and ongoing projects, computational morphologies and annotated corpora have been and are being created for several of these Uralic minority languages: Udmurt, Komi-Zyrian, Eastern Mari, Northern Mansi, and the Kazym and Synya dialects of Khanty, Tundra Nenets and Nganasan. This article presents the morphological analyzers and other annotation tools and the resources developed and used during the projects.

1. Introduction

Besides the national languages spoken by several million speakers: Hungarian, Finnish and Estonian, the Uralic language family includes several minority languages with significantly smaller speaker communities, the majority of which is spoken on the territory of the Russian Federation. In a series of projects[1], computational morphologies and annotated corpora have been and are being created for several of these languages.

2. The projects

One aim of these projects is to make linguistic data concerning these languages available for research to a broader community of linguists, not only the Uralist specialists, and to make corpus-based investigation of these languages possible. Many of these languages exhibit phenomena that would be exciting to explore for a variety of linguists, such as theoreticians specializing in any module of grammar or those interested in language typology. Annotated corpora make it possible to carry out research on various aspects of the language without a long preliminary study of the language itself.

One of the most important lessons that we learned form the first project during which morphologies of six Uralic minority languages (Udmurt, Komi-Zyrian, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan) were created was that since many details of the description which often remain vague in written grammars must unavoidably made explicit in a computationally implemented grammar, the process of creating the implementations as well as the resulting programs themselves shed light on inconsistencies and gaps in the available descriptions of the phonology and morphology of the language, and often help correcting them.

Moreover, while examining linguistic models with regard to exactness and completeness by hand is an impossible task, the computational implementation makes an exhaustive testing of the adequacy of our grammatical models possible against a great amount of real linguistic 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.

Another fact makes a more thorough documentation of these languages urgent is that due to the nature of Russian minority policy, the school system, the great degree of dispersion, the low esteem of the ethnic language and culture and the general lack of an urban culture of their own, all these languages are endangered. On the other hand, there are significant differences among these languages concerning the number of speakers and the exact sociolinguistic situation they are in.

3. Moribund languages

Some of the languages can be categorized as moribund, with virtually no chance of the language still being spoken in another 50 years, not only due to the low number of speakers (some of these languages have existed and developed as the communication medium of small nomadic communities of about a thousand people for thousands of years without an immediate risk of disappearance), but because one generation of speakers has already failed to pass on the language to the next and thus hardly any children speak it. In the case of these languages, an example of which is the Nganasan language of the Northern Samoyedic branch of the Uralic family with about 400 middle-aged and elderly speakers, the most we can do is trying to document as much of the language as possible. Documenting these languages is not a trivial task though, not only because of the extreme complexity of some of them (e.g. in terms of their morpho-phonology), but also because the speaker communities are disintegrating into a small assembly of individuals with more and more uncertain language skills and a heavy influence from their parallel knowledge of the majority language, Russian, that seems to impact not only the syntactic structures they

[1] ('Complex Uralic Linguistic Database', NKFP 5/135/2001), 'Development of Komi and Udmurt morphological analyzers' (OTKA-T 048309) and 'Development of a Ngananasan morphological analyzer' (OTKA / K 60807)

use[2] but even the morpho-phonology[3]. (According to the 2002 census data, there are only 9 monolingual Nganasan speakers, who are all elderly people over 70 living in practically inaccessible spots). The language becomes thus just a collection of idiolects which presumably all differ significantly from both the Nganasan that was spoken by monolingual speakers 60 years ago and from each other. Whose idiolect are we to document? The complexity of the language (e.g. that of the morpho-phonology of Nganasan, or that of the intricate system of verbal moods and evidentiality) might partially account for the fact that no outsiders, including the linguists doing research on the language have managed to master Nganasan. But these languages are not only very difficult to learn for anybody but babies, but they are not very useful to know, either. They have lost much of their function when these nomadic peoples were forced to settle as a minority in settlements inhabited by people speaking another language and to give up their traditional way of life, their rituals and practices. Their tame reindeer herds were collectivized (which subsequently fell victim to epidemics) and they were practically prohibited from reindeer hunting. But the fatal blow on these languages was the schooling of minority children in boarding schools hundreds of kilometres away from their home where the language of education was exclusively Russian. The children had no contact at all with their parents and their home community during the school year, and both their knowledge and their esteem of their mother tongue deteriorated significantly. This was the generation that growing up failed to pass on the language to their children.

There is another factor that makes the documentation of some of these languages difficult. During the Soviet era, making field trips to areas where many of these small minority languages are spoken was only possible for linguists from within then Soviet Union. In the nineties, during the Yeltsin era, an unprecedented freedom of movement made it possible also for foreign linguists to travel freely to the areas previously inaccessible to them and do research there. Fortunately, this is still true for many areas (such as the region of the River Ob, where the Mansi and Khanty live). Certain areas of the northern Arctic regions where some of these minority languages are spoken, however, (the Taymyr Peninsula in particular, where the Nganasans live) have unfortunately been declared divisions of restricted access. Foreign linguists intending to do field work in the region must apply for an entrance permit at the local security authorities which they may fail to issue. This might make it necessary to find alternatives to field trips such as carrying native speakers to places accessible for the researchers as well.

4. Minority languages having a chance of survival

Another group of the languages mentioned do not seem to be threatened by an immediate language death, but even within this group there are significant differences. Although Udmurt and Mari have a similar number of speakers according to the census data, Mari seems to have a different sociolinguistic status than Udmurt due to the native speakers' different attitude toward their mother tongue. While the Mari are proud of their language and their cultural heritage, Udmurts have a rather low esteem of their mother tongue, which they consider inferior to Russian. On the other hand, Maris tend to have more conflicts with the Russian majority than Udmurts for the same reason.

In the case of these languages, the computational tools we create can also be adapted for practical purposes, such as providing the speaker communities with spell checkers and electronic dictionaries in their native language in the hope that the existence of such applications can help to raise the prestige of these languages. In order to be able to create applications of good quality we will need to collaborate with native speakers. Cooperation with publishing houses is vital so that we can obtain corpora that can be used in the process of the development and testing of the tools as well as for linguistic annotation, since on-line resources in these language are rather scarce. On the other hand, there is a stable output of books and newspapers from local publishing houses in all of the languages belonging to this group. The fact that we managed to obtain the manuscript of a 31000 word Komi–Russian dictionary in an electronic form from the company that published it shows that publishers are willing to cooperate. It is important that we make it clear that our goal is to give rather than to take something away from them.

5. Computational morphologies

In our first project, computational morphologies for six languages (Udmurt, Komi, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan) were created and tested on small corpora. These morphologies were based on Latin script based phonological transliterations generally used by linguists dealing with Finno-Ugric and in general with Uralic languages instead of the standard Cyrillic orthographies of the languages, since the tools were intended for linguistic annotation. This also made our lives easier avoiding an inherently non-phonological characteristic of Russian Cyrillic orthography, where palatalized consonants, the *j* phoneme and most vowels are represented by the orthography in a context sensitive manner. Inherently the same system is applied to the palatal consonants of all

[2] A contrast for example between Nenets and Nganasan focus constructions (preverbal vs. postverbal focus) can probably attributed to an adaptation of Russian post verbal focus by the Nganasan.

[3] Among the entries of the Nganasan–Russian dictionary which formed the basis of the stem lexicon of our Ngansan morphological analyzer, we have found about a dozen infinitives that according to our model of Nganasan morpho-phonology cannot be well-formed Nganasan infinitives. All of these 'ungrammatical' forms end in *s'a,* an allomorph of the Nganasan infinitive marker that happens to coincide with the infinitive ending of Russian reflexive verbs instead of some other allomorph that should appear there for the words to be well-formed infinitives.

non-Slavic languages of Russia in their respective Cyrillic orthographies with additional letters or diacritics to represent phonemes that do not exist in Russian.

However, especially in the case of the languages where orthographic texts (newspaper articles, books, etc.) are available, it is desirable that we can directly annotate these, so in a follow-up project, the goal of which was the enhancement of the Komi (Zyrian) analyzer, we created a version of the analyzer that can directly analyze orthographic text. In addition, the stem database of the analyzer was significantly enhanced by incorporating the entries form a 31000 word Komi–Russian dictionary (Beznosikova, 2000). Using standard orthography is of course also a prerequisite if we want to create spell checkers for these languages.

In another follow-up project that has just started this year, we are to create morphologies and annotated and glossed corpora for various dialects of the two Ob-Ugric languages: Khanty[4] and (Northern) Mansi. These analyzers will be based on the Latin script based phonological transliterations generally used in the lingustic works dealing with these languages.

Uralic languages are of the agglutinating type with a high frequency of words containing long suffix sequences and several thousands of possible word forms for each stem in the open word classes. We used two morphological development and analysis toolsets both of which are capable to handle this type of morphologies.

Of the six computational morphologies in our first project, the ones describing Finno-Ugric languages, Komi, Udmurt, Mari and Mansi were created using the formalism of the Humor ('High speed Unification MORphology') morphological analyzer engine of MorphoLogic (Prószéky and Kis, 1999), while the tools for two Samoyed languages, Nganasan and Tundra Nenets were developed using xfst ('Xerox Finite State Tool') of Xerox (Beesley and Karttunen, 2003). We plan to implement the additional Ob-Ugric analyzers using the Humor formalism.

The following table summarizes properties of the morphologies created in our first project and the follow-up Komi analyzer project. The size of the affix lexicons is indicated as a number of morphemes and lexicalized morpheme sequences in the source lexicon.

Language	stem lexicon (lemmas)	affix lexicon (morphemes)
Komi1	2100	156
Komi2	31000+2800 names	156
Udmurt	14100	238
Mari	2200	189
Mansi	1800	270
Nganasan	4150 non-derived	334
Tundra Nenets	19 500	254

[4] We are to create resources for two Northern Khanty dialects: Kazym and Synya Khanty, each named after that tributary of the Ob River along which the dialect is spoken.

5.1 The Humor analyzer

The Humor analyzer performs an 'item-and-arrangement' (IA) style analysis segmenting the input word into a sequence of morphs. The analyzer contains a regular word grammar and it produces flat morph lists as possible analyses. The program performs a search on the input word form for possible analyses looking up morphs in its lexicon that both match the beginning of the yet unanalyzed part of the input and satisfy all morph adjacency constraints of the previous morph. In addition, the candidate morph must form, together with the already analyzed part, the beginning of a possible word construction in the given language. Possible word structures are represented by an extended finite-state automaton in the analyzer.[5]

The morphological database that the Humor engine uses is not directly created and maintained manually, since for the analyzer to work efficiently, the data structures it uses must contain redundant data, which are both hard to read and hard to maintain for humans. The linguistic resources used by the Humor engine explicitly contain allomorphs instead of descriptions of morphemes, along with data structures such as binary vectors and continuation matrices that describe morph adjacency constraints. These resources are created using a morphological description development environment from a feature-based high level human readable description that contains no redundant information and is thus easy to maintain. The system transforms it to the redundant representations that the analyzer uses in two steps.

First, a lexical representation is created that already explicitly contains all the allomorphs of each morpheme along with all their properties and adjacency constraints (using a feature-based formalism) in a human-readable form, which can thus be checked easily by a linguist. This transformation is based on implicational relations, 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. These rules also describe how allomorphs should be created for each morpheme and what properties and constraints the individual allomorphs have (in addition to morpheme level properties and constraints).

The human 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 used in the description should be encoded for the analyzer.

In addition to the analyzer, the toolset contains a lemmatizer and a word form generator.

The lemmatizer, built around the analyzer core, outputputs simplified analyses of word forms consisting of a lemma and morphosyntactic category tags that, in contrast

[5] One can use feature variables in the automaton in to check long distance dependencies a fashion rather similar to flag diacritics in the Xerox tools.

to the more verbose analyses produced by the core analyzer, do not reveal the internal structure of words: compound members and derivational suffixes do not appear as independent items in the output of the lemmatizer.
The output of the lemmatizer and the analyzer is compared in the example below (analyses of the derived Komi word form *kylanly*):

```
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 analyses produced by the lemmatizer are well suited for such tasks as corpus tagging, indexing and parsing.
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 word form 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 listed in the source stem lexicon (like in the case of the Komi derived nominal stem *kylan*):

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

5.2 The Xerox Tools

The two level morphological toolset of Xerox contains various formalisms to create morpheme lexicons and phonological and morpho-phonological rule systems. Morpheme inventories can be created using the *lexc* formalism by defining sublexicons. A sequential phonological rule-system can be defined using the formalism of *xfst* resembling the form used in classical generative phonology as a set of context dependent re-write rules. Using *xfst*, one can compose the rules and the lexicon and during composition the program automatically eliminates intermediate levels of representation created by individual rules. The emerging single two-level finite-state transducer, called a lexical transducer, is a full morpho-phonological description of the language, which can be efficiently used both for analysis and generation. While *xfst* is a compiler for lexical transducers, actual morphological analysis and generation is performed by another program called *lookup*. Lookup may be invoked with either a single transducer, or a script containing an ordered sequence of transducer chains. The chains are applied to the input in order until one produces analyses, so each chain represents a fallback strategy to be applied if all previous strategies have failed. The default strategy is usually simple lookup with the lexical transducer of the language, others may include a chain of a case normalization transducer and the lexical transducer etc. The last fallback strategy can be a guesser, a lexical transducer featuring an extremely underspecified stem lexicon of open word classes besides the normal phonology and suffix grammar of the language. The fact that *lookup* is able to handle chains of transducers as individual strategies instead of just single transducers is important because normally the composition of e.g. a case normalization transducer and a lexical transducer would yield an enormous single transducer.
The two Samoyed languages: Tundra Nenets and Nganasan have a particularly complex phonology with a great abundance of very productive and quite complex phonological and surface phonetical processes. In both of these languages, the combination of phonological and morpho-phonological alternation processes can quite easily result in a single mono- or disyllabic suffix having as many as 20 different allomorphs and stems also tend to have several allomorphs. In the case of these languages, the exact form of a morpheme required by the morpho-phonology of the language cannot in general be determined by considering only local constraints between morphs, because the very intricate well-formedness constraints on syllable structure may involve phonological segments in non-adjacent morphemes. Formalizing these non-local phonological constraints would have been difficult in a formalism based on morph adjacency constraints. Since the descriptions we based our computational morphologies on used a sequential rewrite rule system formalism that was much easier to convert to an *xfst* grammar than to a Humor rule system, we decided to use the Xerox tools for the implementation of these two morphologies.
The feature-based Humor formalism proved to be an efficient means of describing morphological constraints. We also extensively used the corresponding flag diacritics feature of the Xerox tools to describe selectional restrictions between morphemes, such as morphological root selection in Nganasan, suffixes attaching to perfective or imperfective verbal roots; suffixes of verbs requiring an Agent; suffixes attaching to transitive verbs etc. Many of these constraints are local. The flags corresponding to the local constraints can be eliminated from the networks without a size penalty. They are just a convenient way to describe the constraints. The flags constraining long distance dependencies, on the other hand, help to keep the network sizes manageable.

5.3 The two morphological tool sets

Both the Humor analyzer and the Xerox tools are proprietary commercial software. Since Humor was developed by MorphoLogic, it was a natural choice for us to use in these projects. The Xerox tools were published on a CD accompanying Beesley and Karttunen (2003) published in June 2003, accompanied by license that made the version published with the book freely available for non-commercial purposes.
The Xerox tools have an advantage in terms of analysis speed over Humor of a factor of 1.5–4 at an expense of a

significant compile time and runtime memory requirement overhead. Depending on the complexity of the language and the structure of the word grammar, the runtime memory requirement of the Xerox lookup tool may be 10 times as much as that of the Humor analyzer for the same language (even when using Flag diacritics and transducer chains to reduce the memory requirements of the Xerox analyzer). The ratio of compile time memory requirement seems to be at least another order of magnitude higher (i.e. xfst may require more than a hundred times as much memory as the Humor lexicon compiler). 17 years ago, when the Humor analyzer was conceived, the compile time and even the runtime memory requirements of the finite-state tools would have been unfeasibly high. With today's RAM sizes, even a 30 MB analyzer lexicon does not seem to be a serious problem anymore. The Humor analyzer, however, seems to be more applicable in environments with limited memory resources.

The compile time memory requirement of xfst depends significantly on the compilation scenario used. The standard procedure suggested in (Beesley and Karttunen, 2003) of compiling the rule component separately by compiling and composing all the rules using xfst and then composing it with the lexicon compiled by lexc completely failed in a 512 MB machine for lack of memory when first trying to compile our Nganasan morphology. Finally, we managed to tackle this problem by changing the procedure of creating the final transducer: we composed the rules one by one with the lexicon. The lexicon constrained the space of possible underlying representations from the very beginning and thus the size of the network remained manageable throughout the whole compilation process.

6. A web based corpus annotation tool

Although morphological analyzers can be used to rapidly analyse huge amounts of text, they cannot be used alone to create morphosyntactically annotated corpora, because there is always a great degree of morphological ambiguity in the texts. In addition, corpora always contain a number of out of vocabulary word forms that the morphological analyzer is not able to recognize. Usually, some kind of morphological guessing may be used to solve this latter problem, but that usually leads to a disambiguation problem again: that of the possible guessed analyses. The morphological annotation needs to be disambiguated. Although there are standard (statistical) techniques of automatic disambiguated morphosyntactic (part of speech) tagging, these tagging tools must always be trained on manually disambiguated texts. And in fact for the automatic tagging to be of an acceptable accuracy, a huge amount of manually tagged training data is needed (and even then there will be tagging errors). Another problem with standard part of speech taggers is that they do not identify the lemma of words (only the part of speech tag), which is only half of the annotation that we would like to have. Moreover, the word form and the part of speech tag does not always identify the lemma unambiguously, because the paradigms of different lemmas quite often partially overlap at the same paradigm slots[6]. In those cases the lemma cannot be identified fully automatically from the part of speech tagged text. Thus manual disambiguation is inevitable (for at least a subset of the corpus). So a tool is needed that makes the manual disambiguation task as efficient as possible.

We have created a tool that can be used for the morphosyntactic annotation and manual disambiguation of corpora. In order to make the use of this tool efficient, we implemented it as a web application so that it can be concurrently used by linguists/native speakers remotely. It can of course also be installed on and used locally form a local web server.

After tokenizing and morphologically analyzing the text uploaded to the web server, the tool presents individual sentences to the user along with their context clearly indicating ambiguous and unanalyzed words, with the possibility of manually adding analyses of unknown words, removing bogus nonsense analyses (regular expressions can be used to override whole classes of unwanted analyses). The program uses statistical methods to initially rank analyses so that the automatically top ranked analysis of ambiguous words rarely need to be manually overridden. The program learns the decisions of the user. Initial ranking of the analysis candidates can be based on the output of a tagger, the accuracy of which can be incrementally enhanced by adding more and more texts to its training set. In addition to annotating words with their lemmas and morphosyntactic tags, the tool can be configured to add glosses in various languages. When, after making the needed adjustments, the top ranked analysis and glossing candidates are all deemed correct, the user can accept the sentence as correctly analyzed. Manually overridden ranking is always recorded as such. For each disambiguated sentence, the user id of the annotator is logged. Manual correction of typos in the original text is also possible. The user can also mark sentences as problematic. If an update of the database of the morphological analyzer is needed, the corpus can be reanalyzed using the recompiled analyzer without the already disambiguated and accepted sentences being affected.

7. Conclusions

In this paper, we have presented the results of completed projects as well as work in progress the goal of which is to create electronic linguistic resources for several minority languages spoken in Russia belonging to the Uralic language family, also comparing strengths and weaknesses of the two morphological toolsets used in the projects. We have also described a web based corpus annotation workbench that we developed.

A lesson that we learned from the projects is that the need of strict formalization when creating computational grammars may play an important role in creating more adequate grammatical descriptions. We have also found that classical linguistic fieldwork might not be the only

[6] E.g. most forms of the Hungarian verbs '(felül)múl' and 'múlik' coincide. There are many similar lemma pairs.

way to acquire linguistic data in endangered languages. Moreover, we think that further projects with the goal of providing tools such as spell checkers and electronic dictionaries to speaker communities of minority languages (and publishers of books and newspapers) could be a reasonable sequel to these projects.

8. Acknowledgements

The individual analyzers were created by Attila Novák in co-operation with László Fejes (Komi, Udmurt, Mari, Mansi, the grammar being mostly László Fejes's work), with Beáta Wagner-Nagy and Zsuzsa Várnai (Nganasan) and with Nóra Wenszky (Tundra Nenets). The Tundra Nenets analyzer is based on Tapani Salminen's work (his dissertation, Salminen (1997) and his morphological dictionary, 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. The projects were funded by the National Research and Development Programmes of Hungary ('Complex Uralic Linguistic Database', NKFP 5/135/2001) and by OTKA ('Development of Komi and Udmurt morphological analyzers' (OTKA-T 048309) and 'Development of a Ngananasan morphological analyzer' (OTKA / K 60807)).

9. References

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

Ljucija Beznosikova, ed. 2000. *Komi-Roča Kyvčukör*. Syktyvkar.

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

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,* pp. 261–268. College Park, Maryland, USA.

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

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