Morphological Tools for Six Small Uralic Languages

Attila Novák

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

Abstract

This article presents a set of morphological tools for six small endangered minority languages belonging to the Uralic language family, Udmurt, Komi, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan. Following an introduction to the languages, the two sets of tools used in the project (MorphoLogic's Humor tools and the Xerox Finite State Tool) are described and compared. The article is concluded by a comparison of the six computational morphologies.

1. Introduction

This article presents a set of morphological tools for six small Uralic languages. The tools were created in a project that involved various Hungarian research groups in Finno-Ugric linguistics and a Hungarian language technology company (MorphoLogic).

2. The Languages

The languages described were Udmurt, Komi, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan. They are all small minority languages spoken in Russia, and 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 total lack of an urban culture of their own, they all are endangered. Table 1. presents information concerning the alternative names of the languages, their geographical distribution, the estimated number of speakers and the branch to which they belong within the language family.

As is evident even from the number of speakers, two of the languages, Mansi and Nganasan are on the verge of extinction. In their case, the documentation of the language, and any remnants of the oral cultural heritage of these peoples is an urgent scientific task.

2.1. Morphology

These languages, being members of the Uralic language family, are of an agglutinating type. The morphology of agglutinating languages is characterized by the relatively high frequency of words containing long suffix sequences. The following example is from Udmurt.

jaratonoosynyz 'with the sweethearts (ones in love)'

jarat	on	o	os	yny	z
to love	nomen acti = love	having love	plural	instr.	def.

The high number of productive suffixes and possible suffix positions results in a combinatorial explosion of the number of possible word forms (yielding several thousands) for each stem in the open word classes. In some of the languages (e.g. Mari and Nganasan) certain suffixes (clitics) can assume a wide variety of positions within the suffix sequence.

The following corpus examples are from Mari. Both the form (wlak vs. šaməč) and the position of the plural suffix relative to other suffixes (whether it precedes or follows other inflectional endings) exhibit variation:

jeŋ[N]+že[Def]+[NOM]+-wlak[Pl]	'the people'
artist[N]+kə[COM]+že[Def]+-wlak[Pl]	'with the artists'
šydər[N]+-wlak[Pl]+še[Def]+[NOM]	'the stars'
jeŋ[N]+-wlak[Pl]+lan[DAT]	'for people'
pašajeŋ[N]+-šaməč[Pl]+ən[GEN]	'of workers'

Although there have been attempts to handle language technology tasks, such as spell checking, for languages of this type 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 (another member of the Uralic language family). However big the corpus is, even very common forms of not extremely frequent words are inevitably missing from it. Moreover, the analysis of the 150 million words Hungarian National Corpus revealed the fact that 60 percent of the theoretically possible inflectional suffix morpheme sequences never occur 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 words Web corpus, the figure is 50 percent.

The creation of a formal morphological description is therefore unavoidable for this type of languages. There is another factor that made a data oriented approach totally unfeasible in the case of the small Uralic languages project in the first place, namely that there are hardly any electronically available linguistic resources in most of these languages. The corpora we 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 available linguistic 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. In the case of all of the languages, the available linguistic data and their linguistic descriptions proved to be incomplete and contradictory, which made numerous revisions to our computational models necessary.

Language	Other Name	Geographical Distribution	Number of Speakers	Language Branch
Komi	Zyrian	Komi Republic, west of the Urals	262,000	Permian (Finnic)
Udmurt	Votyak	Udmurtia, west of the Urals	500,000	Permian (Finnic)
Eastern (Low) Mari	Cheremis	Mari El Republic, by the Volga	534,000	Volgaic (Finnic)
Northern Mansi	Vogul	west of the Urals, between Urals and Ob River	3,000	Ugric
Nganasan	Tavgi	Taimyr Peninsula, North Siberia	500	Northern Samoyed
Tundra Nenets	Yurak	Northwest Siberia	25,000	Northern Samoyed

Table 1: The six languages

2.2. Phonology and Morphophonology

The rather complex morphological makeup of words is a manifestation of the agglutinating nature of all languages belonging to the Uralic language family. If it were just simple concatenation that happens to morphemes making up a word, creating a formal grammar describing the morphology of these languages would not be a difficult task even in spite of all the variation of suffix ordering that occurs e.g. in the Permic languages or Mari. The following examples from Udmurt, for example, show that the order of possessive and case suffixes is different depending on the case.

kyšno[N]+je[PSS1]+ly[DAT]	‘to my wife’
ares[N]+a[INE]+m[PSS1]	‘in my age’

In Komi, there are even cases where there is free variation, or the order depends on both the case and the possessive suffix:

along my man (transitive case)	
mortöjti	mort[N]+öj[PSS1]+ti[TRA]
morttiym	mort[N]+ti[TRA]+ym[PSS1]
without my/your man (caritive case)	
mortöjtög	mort[N]+öj[PSS1]+tög[CAR]
morttögyd	mort[N]+tög[CAR]+yd[PSS2]
towards my/your/etc. man (approximative case)	
mortöjlań	mort[N]+öj[PSS1]+lań[APP]
mortlańyd	mort[N]+lań[APP]+yd[PSS2]
mortlańys	mort[N]+lań[APP]+ys[PSS3]
mortlańnym	mort[N]+lań[APP]+nym[PSP1]
mortnydlań	mort[N]+nyd[PSP2]+lań[APP]
mortlańnys	mort[N]+lań[APP]+nys[PSP3]

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. This makes not only the implementation of a computational model of the morphology of these languages very difficult, but poses a serious problem even for those trying to do field work and gather linguistic data concerning these languages in a consistent notation, and for the linguists trying to create any acceptable grammar of them. Morphemes in Nganasan and Nenets tend to have numerous surface forms (allomorphs), which hardly resemble each other. Thus what is common in all of the surface forms of a morpheme (its ‘underlying form’) is necessarily something very abstract, which in turn hardly resembles any of the allomorphs.

As an illustration, the following examples demonstrate the purely phonological allomorphy of a single verbal mood suffix (of narrative mood used in the subjective and the non-plural objective conjugations) in Nganasan. Each of the rows of the following table shows a Nganasan word form segmented into a stem, followed by the narrative mood suffix and a subject agreement ending with a gloss added to the end. Each of the word forms contains a different allomorph of the mood suffix. The superscript letters indicate the lexical vowel harmony class of the stems (I: unrounded, U: rounded).

stem	narr. sfx	subj. agr.	‘the rumor
	(hA2nhV)	suffix	is that...’
i^{U}	bahu	[Sg3]	he is
aukum^{U}	hwahu	[Sg3]	he tames sg.
ngungkegimtü^{U}	banghu	[Sg3]	he increases sg.
ngungkegimtü^{U}	bambu	ng[Sg2]	you increase sg.
nguem^{U}	hwanghu	[Sg3]	he attaches to sg.
nguem^{U}	hwambu	ng[Sg2]	you attach to sg.
ngumsyqe^{I}	bjahy	[Sg3]	he answers
ngus1iir^{I}	hjahy	[Sg3]	he moves
ngya’kebty^{I}	bjanghy	[Sg3]	he annoys sy.
ngya’kebty^{I}	bjamby	ng[Sg2]	you annoy sy.
ini’jaim^{I}	hjanghy	[Sg3]	he becomes old
ini’jaim^{I}	hjamby	ng[Sg2]	you become old

The underlying representation of the morpheme is hA2nhV, A2 and V being abstract vowel archiphonemes displaying harmonic behavior. These allomorphs are produced from the underlying representation by the general phonological processes of the language, undergoing vowel harmony, the diphthongization of *a* after *h* and gradation (an extremely complex and rigid system of systematic alternation of obstruents in syllable onsets).

Moreover, the same morpheme has another 12 allomorphs used in the reflexive and the plural objective conjugations, six of which coincide with six of the forms cited above. The underlying form of these is hA2nhA1. The case of the narrative suffix is by no means extreme in Ngansan, the combination of phonological and morphophonological alternation processes can quite easily result in a single monosyllabic suffix having as many as 20 different allomorphs.

Due to their high complexity, it is very difficult to construct an adequate description of the phonology and morphology of the Samoyed languages, and the first such descriptions appeared only very recently (Helimski (1998), Wagner-Nagy (2002), Salminen (1997)).

2.3. Orthography

Linguists dealing with Finno-Ugric and in general with Uralic languages outside Russia tend to use Latin based phonological transcriptions instead of the eventual Cyrillic orthographies of the languages. Since the tools we created were intended for linguists, we decided to use a Latin script based phonological notation in our morphologies instead of the standard Cyrillic orthographies of the languages. This inevitably made our lives easier. On the other hand, the tools we created cannot be directly applied to orthographic input, only with an intermediate converter, which makes the operation of the analyzer less efficient in terms of speed. With the exception of the Tundra Nenets analyzer, the mapping between the orthographic forms and our representation is rather straightforward. In the case of Tundra Nenets, we used Tapani Salminen's phonological notation (Salminen, 1997), which is less phonetic than the standard orthography.

3. The Tools

Of the six computational morphologies, 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).

3.1. The Humor Tools

3.1.1. Features of the Morphological Analyzer

The Humor analyzer performs an '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 indicated.

3.1.2. 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 represented by an extended finite-state automaton in the analyzer.

3.1.3. The Lemmatizer

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.

The lemmatizer identifies the Komi word form *kylanly* as the dative of the noun or adjective (in fact: participle) *kylan*:

```
lemmatizer>kylanly
 kylan[N][DAT]
 kylan[A][DAT]
```

3.1.4. 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 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.

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 preferred word form is required. Since, as we have seen, there is considerable variation in suffix ordering in some of the languages, we 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.

3.2. The Morphological Database

For the analyzer to work efficiently, the data structures it uses contain redundant data. These redundant data structures would be hard to read and modify for humans. We built a morphological description development environment which facilitates the creation of the database (Novák, 2003).

3.2.1. 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 features of any adjacent 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.

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 IA model) instead of assigning each word a cryptically named paradigm ID.

3.2.2. 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.

3.3. The Xerox Tools

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. Since the book gives a very detailed account of the tools, we only give a very brief description here. The program set is based on finite state transducer technology. Although the Xerox tools are commercial products, the authors and the company decided to make the versions published with the book freely available for non-commercial purposes.

The Xerox toolset contains various formalisms to create morpheme lexicons and phonological and morphophonological rule systems. Morpheme inventories can be created using the *lexc* formalism by defining sublexicons. Sublexicons contain morpheme (or morpheme sequence) entries each of which has a lexical form (and optionally a different surface form) and a continuation class. The continuation class is either the name of a sublexicon each member of which may follow the given morpheme, or the word-boundary symbol. 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, a transducer for not properly accented word forms 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 reason for lookup being able to handle chains of transducers as individual strategies instead of just single transducers is that composing a case normalization transducer and a lexical transducer would normally yield an enormous single transducer.

The Xerox program set also contains a construct, *Flag Diacritics*, for the description of feature-value constraints. Flag Diacritics are special labeled epsilon transitions in the transducers the labels of which are interpreted by the lookup tool during analysis performing feature value setting and checking operations thus extending the one dimensional state space of the transducer into a multidimensional state space. (The Humor analyzer utilizes a similar state space extension technique in the implementation of word grammar automata.) The main purpose of this construct is to handle long-distance dependencies (e.g. when certain prefixes are licensed by not necessarily adjacent morphemes), which would otherwise lead to an unmanageable explosion of the size of lexical transducers due to the necessary multiplication of each path in the transducer containing interdependent distant morphemes. But flag diacritics can also be used to implement feature-based constraints on lexical dependencies between neighboring morphemes. In the latter case, the flags can be eliminated from the network using the appropriate command of xfst without an increase of the size of the transducer.

4. Comparison of Humor and xfst

Advantages and drawbacks of the two toolsets follow from the properties of the internal representation of the morphological database of the analyzer/generator and that of the linguistic formalisms used to create the databases.

4.1. Speed and Memory Requirement

In the Xerox tools, morphology is represented by a simple and homogeneous data structure, a set of finite-state transducers. An analysis of a word form is simply the traversal of a path in this homogeneous stream of states and transitions. On the other hand, since transducers are indeterministic with regard to their input side, the traversal of the net usually involves a lot of backtracking. Moreover, in real life situations, normally a synchronized traversal of a chain of transducers is needed, and lookup also has to handle epsilon arcs containing flag diacritics.

The database of the Humor analyzer is less homogeneous and the search for analyses involves more different operations such as lexical lookup, checking of morph adjacency constraints, word grammar automaton traversal and case conversion checks.

Due to the simpler data structure and lookup algorithm, the Xerox analyzer is 1.5–4 times faster. In fact, there is a tradeoff between speed and memory requirement: the Humor analyzer, on the other hand, requires much less memory. We can only estimate the ratio of the size of runtime memory footprints for the two analyzers, since we described different languages using the two sets of tools, but depending on the complexity of the language and the structure of the word grammar, at an extreme even a ratio of 1 to 10 seems to be a realistic estimate (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. 15 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.

We must add though that in the case of Nganasan, the standard procedure 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. 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.

4.2. The Grammar Formalisms

Both grammar formalisms are powerful enough to handle the complex morphology of agglutinating languages without difficulties or compromises. However, the xfst formalism seems to have an advantage when one has to deal with very complex phonology. Context dependent re-write rules of the Generative Phonology tradition have been a popular formalism to describe phonological processes. Such a grammar is obviously easier to translate to the xfst formalism than to a Humor grammar. Moreover, many details of the description often remain vague in written grammars (such as the ordering and exact formulation of rewrite rules). These must unavoidably be made explicit in a computationally implemented grammar. It is also clear that one's first guess at the setting of these parameters is not likely to be totally correct, especially if the model is very abstract, as it was in the case of the Samoyed languages. Rather, one has to experiment with various parameterizations and test them on the available linguistic data to find a model that adequately describes the morpho-phonology of the language. It is obvious that this experimentation requires much less human effort if the computational model which one applies is closer to the formalism used in the original account.

4.3. Lemmatization and Generation

There is a point where the lexicon format geared to the slicing-up approach of the Humor analyzer seems to have a clear advantage over the transducer-based lexicon implementation. The fact that the Humor analyzer returns both the lexical and surface form of each morph allows for an extremely high level of parameterizability when doing lemmatization for word form generation. The key difference between a usual Xerox lexicon created using xfst and a Humor lexicon is that while the 'lexical form' of suffixes is normally an abstract deep phonological representation in the former, in the latter it is the form that the suffix would assume if no further suffixes were attached.

Whether various derivational affixes should be considered to be part of the lemma, and in what constructions, often depends on the actual application. In the Humor lemmatizer, these parameters can be set at runtime without a recompilation of the lexicon. The rich output of the analyzer and the non-abstract lexical forms returned make merging the morphs constituting the lemma very straightforward. The flexibility of the word form generator described in section 3.1.4. (i.e. the fact that, if necessary, the generator can handle non-atomic stems as if they were atomic) is also made possible by the fact that the corresponding analyzer database can be easily converted into a generator lexicon in which stems, derivational affixes and inflections have exactly the kind of representation needed for versatile word form generation.

5. The Morphologies

In this section, we briefly describe the properties of each of the computational morphologies created in the project. With the exception of Tundra Nenets, our morphologies handle all productive word formation processes of the language including inflection, derivation and compounding (if compounds are written as one word). The Tundra Nenets analyzer only handles inflection productively (including gerunds and participles), but its lexicon contains many derived stems manually segmented into morphemes. The size of stem lexicons depends on whether we were able to acquire a dictionary of the language in an electronic form.

Language	stem lexicon (lemmas)	affix lexicon (UR entries)
Komi$_1$	2100	156
Komi$_2$	31000+2800 names	156
Udmurt	14100	238
Mari	2200	189
Mansi	1800	270
Nganasan	4150 non-derived	334
Tundra Nenets	19 500	254

The stem lexicon of the first version of the Komi analyzer 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 (Komi$_2$ in the table above). The stem lexicon of the Udmurt analyzer is based on an Udmurt–Hungarian dictionary (Kozmács, 2002). We did not manage to get an electronically available Mari dictionary, thus the Mari lexicon is entirely corpus-based. Various Mari grammars and printed dictionaries form the basis of the description of closed class words (pronouns, postpositions etc.). The Mansi analyzer was created using Kálmán (1963) and Kálmán (1976) as a source, which were typed in manually. The Nganasan lexicon contains all non-derived stems from a Russian–Nganasan dictionary (Kost'erkina et al., 2001), the vocabulary section of Wagner-Nagy (2002) and words encountered in corpus. The Tundra Nenets stem lexicon contains the vocabulary in Tapani Salminen's morphological dictionary (Salminen, 1998).

6. 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), Beáta Wagner-Nagy and Zsuzsa Várnai (Nganasan) and 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 online co-operation with him. The project was funded by the National Research and Development Programmes of Hungary ('Complex Uralic Linguistic Database', NKFP 5/135/2001).

7. References

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

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

Eugene Helimski. 1998. Nganasan. In Daniel Abondolo, editor, *The Uralic Languages*, pages 480–515. Routledge, London.

Béla Kálmán. 1963. *Chrestomathia Vogulica*. Tankönyvkiadó, Budapest.

Béla Kálmán. 1976. *Wogulische Texte mit einem Glossar*. Budapest.

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

István Kozmács. 2002. *Udmurt-Magyar Szótár*. Savaria University Press, Szombathely.

Attila Novák. 2003. Milyen a jó Humor? (What is good Humor like?). In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*, pages 138–145, Szegedi Tudományegyetem.

Gábor Prószéky 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.

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.

Beáta Wagner-Nagy, editor. 2002. *Chrestomathia Nganasanica*. Studia Uralo-Altaica Supplementum 10. SZTE Finnugor Tanszék – MTA Nyelvtudományi Intézet, Szeged – Budapest.