Nganasan – Computational Resources of a Language on the Verge of Extinction

István Endrédy[1], László Fejes[2], Attila Novák[1], Beatrix Oszkó[2], Gábor Prószéky[1], Sándor Szeverényi[2], Zsuzsa Várnai[2], Beáta Wagner-Nagy[2]

[1] MorphoLogic
5, Kardhegy u., 1116 Budapest, Hungary
{endredy, novak, proszeky}@morphologic.hu

[2] Department of Finno-Ugric and Historical Linguistics, Linguistics Institute, Hungarian Academy of Sciences
33, Benczúr u., 1063 Budapest, Hungary
{fejes, oszko, szeverenyi, varnai, wnbea}@nytud.hu

Abstract

This paper describes the creation and dissemination of computational resources for Nganasan: annotated corpora, morphological analyzer, morphological generator and the development of a website where all of them are available for a wider public. The morphology and especially the phonology of the language are so complex that the implementation of the morphological tools was a real challenge.

1. Introduction

Nganasan belongs to the Northern branch of Samoyedic languages: it is an endangered language spoken in Northern Siberia, in Russia. It is a language on the verge of extinction, namely, it is spoken by less than 500 people most of whom are middle aged or older, and due to the Russian minority policy Russian is the language of teaching in schools for Nganasans. Therefore, it has been an urgent scientific task to provide documentation for the language. Similar work has been done earlier for Sami languages as well (http://giellatekno.uit.no/english.html).

2. Nganasan Root and Suffix Dictionaries

The Nganasan side of the Russian–Nganasan dictionary of Kost'erkina et al. (2001) has been elaborated and converted to the phonemic transcription made up of Roman characters. In the course of building morphological tools further roots have been added to the system, which currently contains approximately 4200 roots. The team also has provided category labels for each item, which was missing from the source, e.g. harmonic features of nouns that cannot be seen on the surface, features of verbal aspect, or irregularity.

During the preparation of the 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 the language: we hypothesize that there are three allomorphs for each root morpheme (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. Some suffixes display ambiguous behavior: they can attach to two of the root allomorphs. There are some vowel symbols in the underlying phonological representation that mark vowels that vary according to the vowel harmony rules of Nganasan: in the case of suffixes the quality of these vowels depends on the harmonic features of the root they are attached to. The first suffix list we compiled contained additional information for derivational suffixes: we gave the category label for the root it attaches to and the category label for the derived form as well.

3. The Complexity of Nganasan Morpho-phonology

This language displays many special phonological and morpho-phonological features, including the phenomena of vowel harmony and two types of consonant gradation. Nganasan gradation does not depend on the morphological make-up of the word: the only factor at play is syllable structure. Syllable boundaries and morph boundaries hardly ever coincide. In the case of short suffixes (made-up of 1 segment), it is possible that even non-adjacent morphs belong to the same syllable. There are additional factors that are needed for the description of gradation: (i) whether the syllable in question is closed (ii) whether the previous syllable is closed (iii) the length of the vowel in the previous syllable (iv) whether the syllable in question is odd or even numbered in the word. Gradation also combines with other alternations in the language: vowel harmony, degemination, root alternations and suffix alternations (as a result of which a one-syllable long suffix can easily have fourteen different allomorphs).

To illustrate the complexity of the above outlined system let us look at the allomorphs of a verbal suffix (Narrative Mood, Nominative–Accusative). Let us see the underlying

Kepa Sarasola, Francis M. Tyers, Mikel L. Forcada (eds.)
7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, Valetta, Malta, 23 May 2010, p. 41–44

representation of the morpheme *hA$_2$nhV*. It has twelve allomorphs: *banghu, bjanghy, bambu, bjamby, bahu, bjahy, hwanghu, hjanghy, hwambu, hjamby, hwahu, hjahy.* These allomorphs are regularly produced from the underlying representation, which undergoes the following phonological processes. The harmonic vowel A_2 surfaces as *a* or *ja* as a result of root dependent roundness harmony, and *a* diphthongizes to *wa* when it follows an *h.* Roots are sorted into lexical classes depending on their harmonic features. This feature must be marked in the lexicon as it is totally arbitrary. Some roots may belong to more than one class, as they display vacillating behavior. The harmonic vowel *V* can surface as *u, y, ü* or *i,* its behavior being regulated by roundness and frontness harmonies (in the suffix being discussed it can only surface as *u* or *y,* however, as there is a back vowel (*a, ja, wa*) in the previous syllable in every case). The consonant *h* appears as *h* in strong grade and as *b* in weak grades. The consonant cluster *nh* surfaces as (i) *ngh* in strong grade or if it undergoes the so-called "nunnation effect"[1], as (ii) *h* in rhythmical weak grade, as (iii) *mb* in syllabic weak grade. (A nasal consonant assimilates in place of articulation to the following consonant, and it disappears in rhythmical weak grade unless there is an immediately preceding nasal on the consonantal tier: this latter phenomenon is called nunnation.) An obstruent in the onset position is in strong grade (i) in even-numbered open syllables 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.

We created morpheme inventories by defining adjacency classes using the program *lexc* (Beesley-Karttunen 2003). The program *xfst* serves to describe a sequenced phonological rule-system by a set of context dependent re-write rules broadly used by generative phonologists. The program set composes the rules and the lexicon and the emerging full morpho-phonological description of the language is a two-level finite-state translating automaton, which can be used both for analysis and generation. 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 a re-write rule; therefore environments encompassing non-adjacent morphemes can also be listed. As the program automatically eliminates intermediate levels of representation created by individual rules, generation and analysis can be performed efficiently.

4. Corpus and Other Tools for Testing of the Nganasan Description

We have elaborated the fairy tale collection of Labanauskas (2001). It consists of 58 texts and more than 17000 running words. Our corpus contains other texts collected by the members of our research group as well: they consist of 4400 words altogether. Unfortunately, almost all of the elements of the collection use inconsistent encoding system, thus their normalization was one of the first tasks. Then we made frequency statistics using the corpus. Word form statistics served as input of the morphological analyzer showing various parses (in the box below the entry in question) according to the recent status of our description of Nganasan morphology[2] (Figure 1).

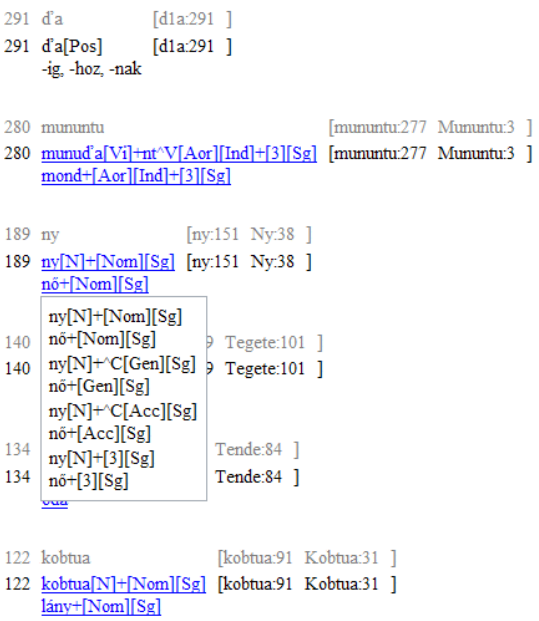

Figure 1: Output of the Nganasan morphological analyzer

The present version of the analyzer leaves 3.7% of the words of the above mentioned corpus unanalyzed. In the case of another 7.9% of the words, analysis is only successful with a version of the analyzer in which some phonological constraints are relaxed.

Although morphological analyzers can be used to rapidly analyze huge amounts of text, they cannot be used alone to create morpho-syntactically 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 morpho-syntactic (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 much larger amount of manually tagged training data is needed

[1] If there is a nasal in the previous syllable, weakening of the nasal+obstruent cluster optionally blocked (and it surface in strong grade).

[2] In the present version glosses are in Hungarian only.

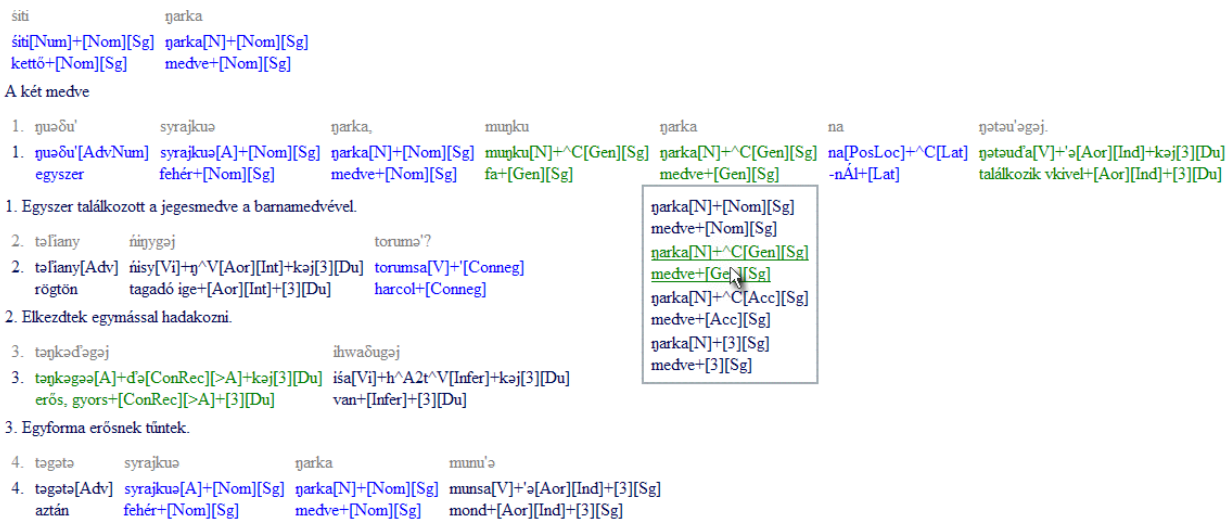

Figure 2: Screen based output of the morphological analysis of Nganasan texts

than available (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 do not always identify the lemma unambiguously, because the paradigms of different lemmas quite often partially overlap at the same paradigm slots. 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 morpho-syntactic 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 analyses of ambiguous words rarely need to be manually overridden. The program learns from 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 morpho-syntactic 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.

A morphological analyzer is not enough for checking the adequacy of the inflectional paradigms. Namely, one cannot detect that a possible alternative form of a certain word form is missing from the word's sample paradigm with the help of analyzer only. A morphological generator is also a very useful tool to track down problems when word forms in the corpus remained unanalyzed. With the help of the generator we could create the form that was adequate according to the grammar. In many cases this strategy led us to the source of the problem in our system.

5. Features of the Web-based System

The Nganasan analyzer and generator have been combined with a web page, and it runs on the web server of MorphoLogic: http://www.morphologic.hu/urali/index.php?lang=english.

As the title of the page suggests, in the future we plan to add further tools for other Uralic languages (Prószéky–Novák 2005).

We have developed an ergonomic way to show the potential analyses of ambiguous words. The parses of the word forms of the input sequence are interlinear. It means that a single analysis is shown on the screen for each input word, but further segmentations can be seen in a pop-up window if the mouse cursor is over the word form in question. The user can chose any of the offered analyses to

replace the original output on the screen (Figure 2). The word form generator is also given in the form of a web service (Figure 3).

We have also developed a soft keyboard the help of which authentic Nganasan texts can be inputted without installing new drivers to the system (Figure 4).

Figure 3: Web-based Nganasan word form generator

Figure 4: Soft keyboard for inputting Nganasan

6. Conclusion

We have developed a computational toolset – morphological analyzer, morphological generator, dictionaries and test corpus – for Nganasan. The language has not yet been morphologically described as thoroughly as with the help of these tools: details of the description which often remain vague in non-computational grammars unavoidably had to be made explicit in the computationally implemented grammar. Many gaps, uncertainties and inconsistencies were detected and in many cases we could correct our grammars and dictionaries. With the help of a corpus we built, the adequacy of the implemented description was very thoroughly tested. It is very important to note here that many questions which remained open should induce further field research. The tools we developed can be used to annotate corpora to facilitate research on other aspects of Nganasan.

7. References

Beesley, K.R., Karttunen, L. (2003) *Finite State Morphology.* CSLI Publications. Stanford: Stanford University.

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

Labanauskas, K.I. (2001). *Nganasanskaya folklornaya khrestomatiya (Нганасанская фольклорная хрестоматия).* Dud'inka: Таймырский окружный центр народного творчества..

Novák, A. (2008) Language Resources for Uralic Minority Languages. *Proceedings of the SALTMIL Workshop at LREC-2008: Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages*. Marrakech: ACL. pp.27–32.

Prószéky, G., Novák, A. (2005). Computational Morphologies for Small Uralic Languages. In: Arppe Antti et al. (eds.) *Inquiries into Word, Constraints and Contexts (Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday).* Stanford: CSLI Publications, Stanford University, pp. 150-157.

Wagner-Nagy, B. (ed.) (2002). *Chrestomathia Nganasanica.* SUA Supplementum 10. Szeged–Budapest: SZTE Finnugor Tanszék – MTA Nyelvtudományi Intézet