Developing electronic lexical resources for Sámi Languages
Ciprian-Virgil Gerstenberger, University of Tromsø, Norway
ciprian.gerstenberger@uit.no

A major task when working with minority languages such as the Sámi languages is to keep them alive or, even more, to revitalize them. Yet language revitalization means first and foremost letting or helping native speakers to keep up with the time and the cultural development of the society, which in turn follows global trends. Fact is that language vitalization cannot take place without deploying a modern society’s universal tool for creating, changing, exchanging, and storing data – the personal computer – as well as various information and communication channels related to it like internet, mobile phones, PDAs, etc.

However, since catching up with the fast-paced evolution of the digital world is not for free, one fairly less expensive but efficient way is to start developing lexical resources. On the one hand, lexical resources are attractive because they exhibit a whole range of complexity: from simple monolingual list of “words” to highly elaborated, both machine- and human-accessible, multi-purpose dictionaries. On the other hand, lexical resources can be regarded as the “bricks” of a language: downwards, they can be morphologically modeled using Finite State technology, upwards they constitute the building material for the work with bigger units (multiword expressions, phrases, clauses, etc.) in syntax and discourse analysis.

At the Center for Sámi Language Technology in Tromsø, preliminary steps have been taken to create a uniform, yet extendable platform for working with lexical resources for all Sámi languages. First, basic problems related to this undertaking are illustrated. Then, requirements for state-of-the-art electronic dictionaries – multifunctionality, detail of description, access and retrieval – will be weighed up against the realities and current development possibilities. Finally, an outline of the proposed environment for maintaining and extending multi-lingual lexical resources will be presented. The overall goal is to create a unified, easily maintainable mono- and multilingual lexical database. The basic idea is that having the lexical knowledge for monolingual applications in the same database as for multilingual applications (1) enables an easier maintenance, (2) facilitates the extension of the existing bilingual resources and (3) offers the possibility of an automated translation check.

Keeping lexical resources for monolingual application (e.g., for morphological analysis) and for each language pair in different lexical databases leads to a multiplication of lexical knowledge (e.g., one for North-Sámi–Norwegian dictionary, one for the Norwegian–North-Sámi dictionary, one for North-Sámi–Lule-Sámi dictionary, etc.). For the task of resource maintenance, this results in time-consuming, yet avoidable operations.

Following the DRY principle¹ of software engineering, I propose an architecture in which each lexical resource is represented only once for each language in a well-structured format so that it can be used both for the generation of input for FST technology and for various other applications. What happens then with the multilingual knowledge? Since this is only a question of lexical knowledge, multilinguality is represented as pointers between a lexeme in the source language and a lexeme (or an ordered group of lexemes) in the target language.

This architecture brings three significant advantages:

1. easier resource maintenance: Updating operations (insertion, deletion, or changing) occur in only one place.

2. automatic extension of individual bilingual resources: Using the transitive closure² of pairwise interlinked monolingual resources, it is possible to infer new links.

¹Don’t Repeat Yourself (DRY) principle: Every piece of knowledge must have a single, unambiguous, authoritative representation within a system. (see, for instance, http://c2.com/cgi/wiki?DontRepeatYourself)

²Transitive closure: An extension or superset of a binary relation such that whenever (a,b) and (b,c) are in the extension, (a,c) is also in the extension.
Given \( \text{viessu}_{\text{sme}} \rightarrow \text{hus}_{\text{nob}} \)\(^3\) and \( \text{hus}_{\text{nob}} \rightarrow \text{gåetie}_{\text{sma}} \): a further pointer \( \text{viessu}_{\text{sme}} \rightarrow \text{gåetie}_{\text{sma}} \) is inferable. Or given \( \text{giele}_{\text{sma}} \rightarrow \text{språk}_{\text{nob}}, \text{giele}_{\text{sma}} \rightarrow \text{giella}_{\text{sme}}, \) and \( \text{giella}_{\text{sme}} \rightarrow \text{snare}_{\text{nob}} \): Since \( \text{giele}_{\text{sma}} \) yields different values between the direct and inferred translations into Norwegian, and moreover, since both translations are correct, respective extensions for both North-Sámi and South-Sámi entries are possible.

3. automatic check of specific bilingual resources: Using the **transitive closure**, it is possible to check whether a lexeme meaning in a source language has the same value(s) both via direct links and via all transitive links to its pendant(s) in a target language.

Given \( \text{giele}_{\text{sma}} \rightarrow \text{språk}_{\text{nob}}, \) \( \text{giele}_{\text{sma}} \rightarrow \text{giella}_{\text{sme}}, \) and \( \text{giella}_{\text{sme}} \rightarrow \text{*sprac}_{\text{nob}} \): The translation of \( \text{giele}_{\text{sma}} \) into Norwegian yields different values: \( \text{språk}_{\text{nob}} \) (directly) and \( \text{*sprac}_{\text{nob}} \) (via North-Sámi). The incorrect spelling can be then rectified.

The concept of pointer naturally extends to an ordered group of pointers for cases in which a single lemma in the source language is translatable by at least two different lemmas in the target language. In addition to the pointer itself, its position in the group and possible morpho-syntactic features or other selectional constraints (needed for a correct rendering) are contained in the entry. This mechanism allows to incorporate any type of multi-word expressions (e.g., collocations, proper nouns, etc.) into the lexicon. Since the pointer is unidirectional, there is no danger to get wrong values when the direction of translation is reverted.

The current work concerns setting up an uniform, easy maintainable infrastructure for the development of mono- and bilingual lexical resources for Sámi languages. The main idea is that interlinking monolingual resources would enable both the extension of bilingual resources and the check of specific translations via transitive closure. Moreover, given the existing lexical resources for English, German, or French, on the one hand, and interlinked Sámi–Norwegian resources, on the other hand, it would be possible to compile, for instance, a reliable Sámi–English or Sámi–German dictionary.

\[^3\text{Following ISO 639-2 code for the representation of names of languages is used: sme for North-Sámi, sua sma for South-Sámi, and nob for Norwegian Bokmål.}\]