Exploiting Similarity in the MT into a Minority Language

Boštjan Dvořák*, **Petr Homola**†, Vladislav Kuboň†

*Zentrum für allgemeine Sprachwissenschaft
Jägerstraße 10–11
10117 Berlin, Germany
dvorak@zas.gwz-berlin.de
† Institute of Formal and Applied Linguistics
Malostranské náměstí 25
110 00 Praha 1, Czech republic
{homola,vk}@ufal.mff.cuni.cz

Abstract

The paper presents a machine translation system from Czech to related languages, including 'small' (minority) languages Lower Sorbian and Macedonian. Lower Sorbian, a West Slavonic language, which is spoken by less than 20,000 people, preserves many acient features, it has supine, dual and some other grammatical forms which disappeared in most Slavonic languages. Macedonian is interesting for us in that its typology is quite different from other Slavonic languages (except **Bulgarian). The paper presents the most important problems encountered during the implementation of the MT system** *Česílko*.

1. Introduction

The problem of preservation of the cultural heritage of the mankind is very closely related to the problem of preservation of minority languages. A complete extinction of a language usually also to a large extent means the loss of most of the cultural heritage of the people speaking that language. The modern era has brought both positive and negative aspects into this problem. Among the negative ones is definitely the globalization bringing a very strong stress on unification, including the unification of languages. The more globalized our world becomes, the higher the number of people capable of expressing themselves in one or more of the "big" languages. This is of course also a positive factor decreasing the number of mutual misunderstandings among the people belonging to different nations. The negative effect is the decreasing number of people capable of speking or reading their own (minority) native language.

The positive effects of the modern world are the scientific and technological achievements which enable to preserve to some extent even the extinct languages in the form of various kinds of corpora of spoken or written language. This is of course the most obvious effect of natural language technology, but not the only one. In this paper we would like to show how a very simple machine translation system can help both to preserve the cultural heritage of a minority language by translating into some "bigger" language and to increase the number of translations in the opposite direction thus increasing the number of texts available in the particular minority language.

Democratic governments usually care about the minority languages, but in many cases a minority language is at the same time much less similar to a majority language than to a language spoken in a neighboring country. The MT from the similar language definitely can complement all other efforts preserving the minority language. This paper describes a particular case of a Lower Sorbian, a minority language spoken in Germany in the area around Cottbus. Lower Sorbian (and its neighbor Upper Sorbian) are Slavonic languages, typologically very different from the majority language, German, but very closely related to the languages spoken in a close geographical proximity, to Czech and Polish. Our paper presents an enrichment of an existing multilingual MT system to exploit the proximity of related languages for Lower Sorbian. The system is not new — it already exists for several language pairs (Czech-Slovak, Czech-Polish, Czech-Lithuanian), cf. (Hajič et al., 2003), and this paper describes the modifications of the original system allowing to insert a new module.

2. Česílko — a multilingual MT system for related languages

The system Česílko has been developed as an answer to a growing need of translation and localization from one source language to many target languages. It is quite clear that the independent translation or localization of the same document into several typologically similar target languages is a waste of effort and money. Our solution proposes to use one language from the target group as a pivot and to perform the translation through this language. It is quite true that applying the pivot language approach has a serious drawback — the translation quality, which needs to be very high, may deteriorate in this two-step process. A negligible shift of the meaning during the translation into a pivot language may be amplified by a subsequent translation from the pivot language to the actual target language. We focus on the 'second step' in the paper.

In order to overcome these problems we have suggested an approach combining the human-made translation

from the source language into a pivot language with a machine translation between a pivot and a (closely related) target language. The reviewer of the target language text may then review the translation against the original source language text and he thus can eliminate any problem caused by the translation from the source into the pivot language.
The system consists of the following steps:

1. Morphological analysis of Czech
2. Morphological disambiguation of Czech by means of a stochastic tagger
3. Search in the domain-related bilingual glossaries
4. Search in the general bilingual dictionary
5. Morphological synthesis of the target language

The need to account for phenomena which cannot be handled by this very simple architecture led us to the inclusion of a additional modules: a shallow parsing module for Czech for some of the language pairs which directly follows the morphological disambiguation of Czech (Homola and Rimkutė, 2003), and a named entity (NE) recognizion module (Homola and Piskorski, 2004).

3. Extending the system to new target languages

In this section, we describe the most significant problems encountered while adapting the system to new target languages — Lower Sorbian and Macedonian. The adaptation itself included, first of all, the creation of a bilingual glossary (with Czech as source language) and the implementation of syntactic and morphological synthesis. The most interesting part has been the modification of transfer. There are about twenty quite complex transfer rules that had to be rewritten according to the grammar of the target language to guarantee correct phrase structure and constituent order.

3.1. MT problems encountered with Lower Sorbian

Sorbian is a West Slavonic minority language spoken in Lusatia in Germany. It splits into many dialects which differ significantly from each other. Two written standards are used in the present, Upper Sorbian in Saxonia and Lower Sorbian in Brandenburgia. We have chosen Lower Sorbian for our experiments, mainly because there exists a morphological tool capable of generating inflected forms from many lemmas obtained as a result of the translation process.[1]
Both morphology and syntax of Lower Sorbian are very similar to Czech, nevertheless the grammar of Lower Sorbian is more complicated than the Czech one since Lower Sorbian is more conservative.[2] In the following text we describe some aspects of Lower Sorbian which are important with respect to MT from Czech.

- Lower Sorbian has *dual*, a special number used instead of plural for the amount 2, e.g. *dub* (1), *duba* (2), *duby* (more than 2) "oak(s)". We ignore this number because the number of persons or objects can only be decided with a proper understanding of the context. This may result in an translation error although the sentence as such is grammatical, but such a strategy is unavoidable if we want to keep the whole system as simple as possible.
- The *supine* is another grammatical form which is not present in Czech. It is an infinite verb form used to express a goal or decisions, usually together with a verb of movement, e.g., *źi spat* "go to sleep" (cf. the infinitive form *spaś*).
- The system of tenses is richer in Lower Sorbian. Whereas Czech only uses one periphrastic past form, Lower Sorbian also has synthetic past forms, *aorist* and *imperfect*. Nevertheless these forms are rarely used in contemporary texts, i.e., one can use the periphrastic form to translate past tense.
- Lower Sorbian does not drop the auxiliary verb *byś* in the third person of the past form (cf. Czech *převzala* "took over" vs. Lower Sorbian *jo pśiwzeła*). We ignore this difference in the current version of the system, since the participle forms are the same for all persons, therefore the shallow parser does not deliver the information about the person at all (in a full parse, the subject would contain the missing information; nevertheless the subject can be dropped as well, so the person may remain underspecified).
- The passive is constructed differently in some cases. There is the specific *bu*-pattern (e.g., *dom bu natwarjony* "the house has been built") and the colloquial *wordowaś* (e.g., *dom worduje twarjony* "the house is being built"), whereas Czech only has one equivalent with *být* "to be". Moreover, the reflexive passive is used more often (e.g., *drjewo se wót nana rubjo* "the tree is being cut by the father"). We use the reflexive pattern if there are more possibilities. The agent is expressed by a prepositional phrase with *wót* "from", whereas Czech uses the instrumental case.

One of the important things which really may substantially decrease the quality of output provided by our system is the word order. Due to the typological similarity of both languages and the fact that both Czech and Lower Sorbian use the word order to express topic-focus distribution, we can preserve the word order of the source (Czech) text.

3.2. MT problems encountered with Macedonian

Macedonian is a South Slavonic language spoken in the Republic of Macedonia and by national minori-

[1] **We are very grateful to Gerat Nagora and Georg Müller who allowed us to use there morphology tool based on (Starosta, 1999).**

[2] The transfer is based mostly on (Janaš, 1976).

ties in Albania, Bulgaria and Greece. It belongs to the South-East Slavonic Bulgarian-Macedonian dialect continuum, the written standard is based on South-West dialects. Macedonian is an interesting target language especially because its typology differs extremely from other Slavonic languages; it has a simplified nominal system with an analytical structure, but on the other hand, its verbal system is very complicated.
Although the vocabulary is similar to Czech, the sentence structure differs in many aspects, since synthetic construction have to be translated analytically in most cases and analytical constructions (e.g., past tense) have different syntactic structure too. For these reasons, the shallow parser and a deeper transfer are more important than for the other implemented langauge pairs as, e.g., for Czech-Polish.
For the first evaluation phase on small texts, we have developed our own morphological synthesizer with a limited word list based on (Koneski, 1952). The comparative analysis is partially based on (Koneski, 1965). In the following list, the most frequent discrepancies between Czech and Macedonian are presented.

- Macedonian has almost no cases except for pronouns, Czech cases have to be translated by analytic (prepositional) phrases, e.g.:

  (1) *hlavní* *město*
  main-**NEUT,SG,NOM** town-**NEUT,SG,NOM**
  Makedonie
  Macedonia-**FEM,SG,GEN**
  "the capital of Macedonia" (Cze)

  (2) главен град на
  main-**MASC,SG** town-**MASC,SG** on
  Македонија
  Macedonia-**FEM,SG**
  "the capital of Macedonia" (Mac)

  The assignment of prepositional cases to grammatical functions is quite straight-forward.

- There is object doubling in Macedonian, i.e., both direct and indirect objects get an additional pronoun in some cases, e.g.:

  (3) Му$_1$ ја$_2$ дадов
  him her-**ACC** gave-**1SG**
  книгата$_2$ на Стојан$_1$
  book-**FEM,SG,DEF** on Stojan
  "I gave the book to Stojan." (Mac)

  The Czech sentence would be:

  (4) *Dal* *jsem*
  gave-**RESPART,MASC,SG** am
  knihu *Stojanovi*
  book-**FEM,SG,ACC** Stojan-**SG,DAT**
  "I gave the book to Stojan." (Cze)

  The solution of this problem involves the decision, whether the enclitic pronoun has to be present in the sentence or not, and eventually the insertion of the pronoun at the right position.

- A complicated problem arises with the past tense. Czech has only one past tense — the compound perfect with a resultative (*l*-)participle, e.g.:

  (5) *On* *byl* *v*
  he was-**RESPART,MASC,SG** in
  Bitole
  Bitola-**FEM,SG,LOC**
  "He was in Bitola." (Cze)

  Unfortunately, an analogical construction cannot be used in Macedonian since these participles are used to express the renarrative (see below), so the English translation of the following example means "reportedly":

  (6) Тој бил во
  he was-**RESPART,MASC,SG** in
  Битола
  Bitola-**FEM,SG**
  "He reportedly was in Bitola." (Mac)

  Instead of that, one can use the compound past tense with има or the concise past tense (aorist or imperfect), e.g.:

  (7) Тој беше во Битола
  he was-**3SG** in Bitola-**FEM,SG**
  "He was in Bitola." (Mac)

  (8) Што имаш речено?
  what have-**PRES,2SG** said-**NEUT,SG**
  "What did you say?" (Mac)

- Verbal nouns are used in Macedonian more often since it has lost the infinive, e.g.:

  (9) Не треба седење
  not needed-**ADV** sitting-**NEUT,SG**
  треба работење
  needed-**ADV** working-**NEUT,SG**
  "One should not sit, one should work."
  (Mac)

  The other way to express the Czech infinitive is the embedded *da*-phrase, e.g.:

  (10) *Chci* *jít* *domů*
  want-**1SG** go-**INF** home
  "I want to go home." (Cze)

  (11) Сакам да одам дома
  want-**PRES,1SG** that go-**PRES,1SG** home
  "I want to go home." (Mac)

- Macedonian has a special verbal category which is not present in any other Slavonic language except Bulgarian, the renarrative. It is used to express facts which the speaker cannot verify, e.g.:

(12) Toj кажа дека Стојан
he says-**PRES,3SG** that Stojan-**SG**
бил в куќи
was-**3SG** in house-**FEM,LOC**
"He says that Stojan was at home." (Mac)

- Existential propositions of the type *there is* are expressed in Macedonian using има+acc., whereas Czech uses *to be*, e.g.:

(13) *V horách jsou*
in mountains-**FEM,PL,LOC** are-**3PL**
medvědi
bears-**MASC,PL,NOM**
"There are bears in the mountains." (Cze)

(14) Има мечки во
has-**PRES,3SG** bears-**FEM,PL** in
планините
mountains-**FEM,PL,DEF**
"There are bears in the mountains." (Mac)

The general pattern in Czech is a sentence with a subject and a locative phrase (with the auxiliar verb), thus the latter has to be transformed to accusative (this change is only relevant for pronouns), e.g.:

(15) Него го нема
Him him-**ENCL** has-not-**3SG**
"He is not here." (Mac)

- The order of clitics is different. Basically, they are attached to the verb, often to the left, in Macedonian, whereas Czech usually places them at the second position in the clause (i.e., they follow the first (accented) phrase), e.g.:

(16) *Nechce se mi číst*
Not-want-**3SG** REFL me-**DAT** read-**INF**
tu knihu
that-**FEM,SG,ACC** book-**FEM,SG,ACC**
"I do not feel like reading the book." (Cze)

(17) Не ми се чита
Not me-**DAT** REFL reads-**3SG**
книгата
book-**FEM,SG,DEF**
"I do not feel like reading the book." (Mac)

As for the topic-focus articulation, we are trying to preserve the word (constituent) order given in the input sentence. Almost all changes concern enclitic elements which usually have a fixed position in the sentence or verbal phrase, e.g.:

(18) *Nemám rád politiku*
not-have-**1SG** like-**ADV** politics-**FEM,SG,ACC**
"I do not like politics." (Cze)

(19) Не ја сакам политиката
not her-**ACC** like-**1SG** politics-**FEM,SG,DEF**
"I do not like politics." (Mac)

(20) *V knize se netvrdí,*
in book-**FEM,SG,LOC** REFL not-says-**3SG**
že...
that
"One does not say in the book that..." (Cze)

(21) Не се тврди во книгата
not REFL say-**3SG** in book-**FEM,SG,DEF**
дека...
that
"One does not say in the book that..." (Mac)

Some changes concern noun phrases with embedded sentences, e.g.:

(22) *dívka, kterou jsem*
girl-**FEM,SG,NOM** which-**FEM,SG,ACC** am
viděl
saw-**RESPART,MASC,SG**
"the girl I have seen" (Cze)

(23) таа чупа што
that-**FEM,SG** girl-**FEM,SG** what
ја видов
her-**FEM,SG,ACC** saw-**1SG**
"the girl I have seen" (Mac)

Obviously, some elements can be dropped or added to the target syntactic structure.

4. Parser and transfer

4.1. Data structures

There are two essential data structures: a *multigraph*, which represents the input sentence and its structural analysis on different stages, and abstract *objects* that are embedded in an object-oriented hierarchy and contain properties and methods. These objects are widely autonomous in the parsing process, i.e. there are almost no global rules any more.

4.1.1. Objects and their instances

Objects are abstract entities which represent elements of the sentence and the result(s). Every object has a predefined template that defines which properties and methods the object has. Concrete realizations of objects are called *instances*.
Let us have a look at the following Czech noun phrase:

(24) *velmi staré auto*
very old-**NEUT,SG,NOM** car-**NEUT,SG,NOM**
"(a/the) very old car"

This NP is an abstract entity consisting of three words, or of other hierarchically organized entities which could be schematically described as follows: (((velmi) staré) auto)$_{NP}$

4.1.2. Properties

Each object can contain static data, called *properties*. Properties can be atomic values (strings, integers etc.) as well as complex entities (instances of other object). All properties of an object are accessible only to its instances or instances of its descendants; this feature is called *encapsulation*. In general, objects appear as black boxes that act autonomously. The behavior of the objects is defined in their methods (pieces of code). Each object defines an interface which is visible for others and allows invoking internal object methods which can manipulate object properties or decompose the action by using other objects' interface.
For example, we would define an object representing nouns. Its instance of the word *auto* from example (24) could be as follows:

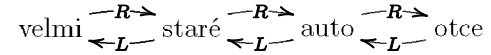
Noun	
LEMMA	**'auto'**
FORM	**'auto'**
NUMBER	**sg**
CASE	**nom**
GENDER	**neut**

4.1.3. Autonomous code

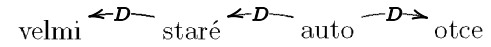
Besides properties, objects may have their own code (organized as methods) which can be invoked by other objects through a shared interface. This code is object specific, i.e. an object representing integers may provide methods for arithmetic functions, whereas an object representing a string or a text document may provide methods for searching for fragments, replacing pieces of text etc. In the described framework, most objects have a linguistic background, of course. Most of their methods are designed to build up the syntactic structure of parts of the sentence in an autonomous manner. For example, objects that represent nouns contain methods which assure building up noun phrases by incorporating adjectives and other attributes etc. The concrete implementation of such an object is, of course, language specific (e.g. there has to be a congruence in some morphological categories between a noun and its attributes, such as in case, gender and number).

4.1.4. Multigraph

The initial state of the multigraph is a chain of morphologically annotated objects. The sentence from example (24) would be represented by the following graph (the labels L/R denote immediate left/right neighbourhood) :



The parser would add dependencies, so the resulting syntactic structure may be:



4.2. Parsing

The parsing process consists in our framework of two phases:

1. setting up hypotheses about relations,
2. transforming hypotheses to tectogrammatical structures (in general, acyclic directed graphs).

The first phase involves recognizing of relations between words and word groups, i.e., between nodes of the graph. We are working with four basic types of relations:

- dependencies,
- coordination,
- (co-)references,
- shackles.

Moreover, each relation (edge of the graph) may be labeled. The second phase involves the recognizing of tectogrammatical patterns in the graph.
The shallow parser that recognizes chunks is implemented as a set of Prolog rules. For example, the rule for combining an adjective with a noun is defined as follows:

```
rule(X1, Y1, X2, Y2, X, Y) :-
subType(X1, adjective), subType(X2, nounPhrase),
splitAvm(Y1, [gender, number, case], A1, B1),
splitAvm(Y2, [gender, number, case], A2, B2),
unifyAvm(A1, A2, A), unifyAvm(A, B2, Y0),
appendAttribute(Y0, attrAdj, B1, Y),
X = nounPhrase.
```

There are several auxiliar predicates. First of all, the type of the objects is checked (**subType**). For this rule to apply, the adjective has to agree with its governor in gender, number and case, thus we have to unify these attributes (**unifyAvm**). Finally, the adjective becomes a feature of its governor (**appendAttribute**).

4.3. Implementation

The code which integrates the independent modules of the system is written in Java (version 1.5), so that it is platform independent. Parser and transfer are written in Prolog.

5. A note on evaluation

The results have been evaluated using Trados Translators' Workbench. The measure gives the work amount of translator necessary to adapt the target sentence so that it would be grammatical (the method is described in more detail in (Hajič et al., 2003)).
Table 1 gives the results for implemented language pairs (the source language is Czech; (**P**) means that a shallow parser has been used).
We have no representative results for Macedonian yet (the preliminary result measured on a short text is about 88%).

target language	accuracy
English	*30%*
Slovak	90%
Polish	71.4%
Lithuanian (**P**)	87.6%
Lower Sorbian (P)	**93%**

Table 1: Evaluation of implemented target languages

6. Conclusions

Machine translation might become a very important tool for increasing the amount of texts available in minority languages. Although the work described in this paper has reached only an experimental stage for some of the language pairs mentioned in the paper, we believe that the experiments we have completed show the advantage of exploiting the language similarity among "smaller" languages may result in an good quality of the translation. Our paper shows that a thorough investigation of linguistic phenomena having a negative influence on the MT quality between two similar languages and the application of relatively simple but adequate means reflecting those phenomena is a relatively effective way how to add new language pairs to an existing simple MT system.

7. Acknowledgements

We are very grateful to Kiril Ribarov for his comments on Macedonian examples. This research was supported by the Ministry of Education of the Czech Republic, project MSM0021620838, and by the grant No. 1ET100300517. We would like to thank the anonymous reviewers for their valuable comments and recommendations.

8. References

Jan Hajič, Petr Homola, and Vladislav Kuboň. 2003. A simple multilingual machine translation system. In *Proceedings of the MT Summit IX*, New Orleans.

Petr Homola and Jakub Piskorski. 2004. How can shallow NLP help a machine translation system. In *Conference Human Language Technologies*, Riga, Latvia.

Petr Homola and Erika Rimkutė. 2003. Shallow machine translation — in between of two extremes. *In: Proceedings of the Tbilisi Symposium, Tbilisi.*

Pětr Janaš. 1976. *Niedersorbische Grammatik.* Domowina-Verlag, Bautzen.

Blaže Koneski. 1952. Граматика на македонскиот литературен јазик *[Grammar of the Macedonian literary language].* Skopje.

Blaže Koneski. 1965. Историја на македонскиот јазик *[History of the Macedonian languages].* Kočo Racin, Skopje.

Manfred Starosta. 1999. *Dolnoserbsko-nimski słownik.* Ludowe nakładnistwo Domowina, Bautzen.