A Translation Model For Languages of Acceding Countries

Petr Homola, Vladislav Kuboň

Inst.of Formal and Applied Linguistics,
Charles University, Prague
{homola|vk}@ufal.mff.cuni.cz

Abstract. . The paper proposes a model for translation between syntactically similar languages of acceding countries. This model is based on the presupposition that the translation of related languages should exploit the relatedness by using as simple methods and tools as possible. In the first part the paper discusses the properties of some "new" languages, the second part describes a simple translation model which has already been tested on several pairs of syntactically similar languages..

1. Introduction

The historical event of EU enlargement scheduled for the 1st May of this year brings many new challenges. Apart from political and economical ones there are also linguistic challenges. The enlargement introduces ten new languages, increasing thus the number of official languages of EU to 21 and the number of language pairs to 210. It is quite clear that such a sudden huge increase of the number of language pairs requires solutions exploiting all possible advantages which might make the enormous translation task a little bit easier.

In this paper we would like to propose a model for translation between those languages of acceding countries which are more or less related. This model is based on the presupposition that the translation of related languages should exploit the relatedness by using as simple methods and tools as possible. It also advocates the idea that although for every acceding country it is extremely important to translate from "big" EU languages (English, German, French), it would be useful to shift the research focus to a machine translation among related "new" languages. It might quickly solve at least some problems of translation inside enlarged EU and to help to bridge a gap until full-fledged MT systems are developed.

2. The classification of relatedness

Natural languages are grouped into language families. Usually, languages from the same language family are more similar than languages from different language families. Although the division of languages into language families is not a perfect criterion for classification, it provides a raw hierarchy. Let us distinguish the following levels of proximity of languages with examples of language pairs from the Central and East Europe:

- Variants of one `underlying' language
  (e.g., Serbian and Croatian, Upper and Lower Sorbian).
- Very closely related languages
  (e.g., Czech and Slovak or Upper Sorbian) are very similar in morphology, syntax and lexics. Semantic ambiguities are rare. We suppose that no syntactic parser is needed in MT systems for such languages.
- Closely related languages
  (e.g., Czech and Polish or Russian) are similar in morphology and lexics, although some semantic ambiguities occur. Syntactic constructions are not fully compatible, e.g., counterparts of analytic constructions are synthetic and vice versa or different lexems are used, cf. Czech *byl jsem* [I was] with Pol. *byłem*, Czech *byl zničen* [has been destroyed] (aux. verb *být*) with Polish *został zniszczony* (aux. verb *zostać* instead of *być*).
  Partial transfer is needed to perform MT among these languages.
- Related languages
  (e.g., Czech and Lithuanian or Latvian) are not as similar as closely related languages, although there are still many similarities (because of a common origin and/or strong mutual influence). The morphological system is similar and there are many one-to-one correspondences in the lexics. Although the syntax is very similar, there are differences (as in the previous case) and, moreover, there are some constructions that do

not have direct counterparts in the other language. For example, Lithuanian half-participles can be expressed by Czech transgressives (e.g., Czech *odešel* ***nerozloučiv se*** [he left without saying good bye]-> *išėjo* ***neatsisveikinęs***). On the other hand, Lithuanian gerunds have no direct counterpart in Czech (the transgressive has the same semantic function, but its use is restricted grammatically). They have to be expressed by other means, usually through nominalization or embedded sentence (e.g., Czech ***při výbuchu*** *bomby zahynul člověk* [a man has died by the bomb explosion] -> ***sprogus*** *bombai žuvo žmogus*).

The remaining two categories in decreasing order of similarity (Languages with common origin and Languages of different origin) are irrelevant from the point of view of MT using simplified methods.

If we look more closely on the set of "new" languages with regard to their mutual relatedness (or linguistic similarity), we must notice that although three languages are more or less isolated in the community (Hungarian (Finno-Ugric), Turkish and Maltese (Semitic)), the remaining languages are either related mutually or related to some of the languages of current EU members. The latter case is Estonian (Finno-Ugric), which is closely related to Finnish. The largest subgroup, namely six "new" languages, belong to the Balto-Slavic family: Czech, Polish, Slovak (West Slavic), Slovenian (South Slavic) and Latvian and Lithuanian (East Baltic). Moreover, Russian, an East Slavic language, is used widely in Latvia and Estonia and should therefore be also taken into account.

3. Typology of language similarity

As the term *similarity of languages* (we intentionally replace the notion of relatedness by this notion due to the fact that from the point of view of our translation model it is more important than relatedness) is very vague, it is necessary to classify the similarity into several categories:

- typological
- morphological
- syntactic
- lexical.

3.1 Typological similarity

The first type of similarity is the most important one. Due to the fact that in the following text we propose to use only shallow syntactic analysis which does not take into account syntactic relationship among larger constituents, it is clear that any difference in the constituent order on the sentence level will have dire consequences for the translation quality. The shallow approach would definitely be unsuccessful for translation between languages of different typology.

Let us take Czech and Lithuanian as an example of the language pair, which supports this claim. These languages have rich inflection and very high degree of word order freedom, thus we can suppose that it won't be necessary to change the word order at the level of inner participants in most cases. On the other hand, both languages differ a lot in the lexics and morphology.

For example, both Example 1 and Example 2 mean approximately *Dad read a/the book*. The difference between these two sentences is in their information structure. Example 1 should be translated as *Dad read a book*, whereas Example 2 means in fact *The book has been read by Dad*[1]. In the first sentence, the noun *book* is not contextually bound (it belongs to the focus), in the latter one it belongs to the topic. The category of voice differs in both sentences due to a strict word order in English, although in both Czech equivalents, active voice is used[2]. We can see that in the Lithuanian translation, the word order is exactly the same.

Example 1:

Czech:	Otec	četl	knihu.
Lith.:	Tėvas	skaitė	knygą.
	Father(nom.)	read(3sg., past)	book(acc.)

Example 2:

Czech:	Knihu	četl	otec.
Lith.:	Knygą	skaitė	tėvas.
	Book(acc.)	read(3sg., past)	father(nom.)

3.2 Lexical similarity

The lexical similarity does not mean that the vocabulary has to have the same origin, i.e., that words have to be created from the same (proto-stem). What is important for shallow MT (and for MT in general), is the semantic correspondence (preferably one-to-one relation between meanings of both words).

Similar morphological systems simplify the transfer. For example, Slavonic languages (except

1 Note that in the first sentence, an indefinite article is used, whereas in the latter one, a definite article stands in front of „book".

2 Passive voice (except of the reflexive one) occurs rarely in Czech (and most other Slavic languages). It can be used if one would like to underline the direct object or if there is no subject at all (for example, *Kniha byla čtena* [The book has been read]).

of Bulgarian and Macedonian) have 6-7 cases. The case system of East Baltic languages is very similar, although it has been reduced formally in Latvian (instrumental forms are equal as dative and accusative and the function of instrumentral is expressed by the preposition *ar* [with], similarly as in Upper Sorbian). (Ambrazas 1996) gives seven cases for Lithuanian, but there are in fact at least eight cases in Lithuanian (or ten cases but only eight of them are productive[3]. Nevertheless the case systems of Slavonic and East Baltic languages are very similar which makes the shallow MT approach possible.

Significant differences occur only in the verbal system, East Baltic languages have a huge amount of participles and half-participles that have no direct counterpart in Czech. The Lithuanian translation of an example from (Gamut 1991) is given in the following example:

Example 3
Gimė vaikas valdysiantis pasaulį..
was-born(3sg.) child(nom.)
ruling(fut.,masc.,sg.,nom.) world(acc.).
[A child was born, which will rule the world.]

The participle *valdysiantis* is used instead of an embedded sentence, because Lithuanian has future participles. These participles have to be expresses by an embedded sentence in Slavic languages.

The discrepancy between Czech and Lithuanian in the building of past and future tenses and subjunctive is at the border of morphological and syntactic differences. Whereas Czech uses analytical constructions with the auxiliary verb *být* [to be], Lithuanian uses synthetic forms, e.g. Czech *byl jsem* [I was] vs. Lith. *buvau*, Czech *budu kupovat* [I will buy] vs. Lith. *pirskiu* (cf. inf. *pirkti*), Czech *tancoval bych* [I would dance] vs. Lith. *šokčiau* (cf. inf. *šokti*).

3.3 Syntactic similarity

Syntactic constructions of related languages are usually very similar. There are often two equal possibilities in both languages. For example, a condition can be expressed by subjunctive or imperative in Russian, as in the example from (Panevová 1980):

Example 4:
Pridi on na čas raňše,…
Come(imp.sg.) he(nom) on hour(acc.) sooner
[If he would have come sooner, …]

In Czech, only one form is possible, namely subjunctive. This case of syntactic ambiguity is no problem for MT from Czech, as the syntactically similar construction can be used.

A well known problem is the inherent semantic ambiguity of transgressives. In Baltic and Slavic languages, the meaning of transgressive is very vague. (Sgall 1974) gives the following example:

Example 5
Tatínek prodav pole vybavil dceru.
Dad(nom) sell(trg,past) field(acc) dower(3sg) daughter(acc)
[After selling the field, Dad dowered the daughter.]

Syntactic similarity is also very important especially on higher levels, in particular the verbal one. The differences in verbal valences have negative influence on the quality of translation due to the fact that our model does not use valency information. Syntactic structure of smaller constituents, such as noun and prepositional phrases, is not that important, because we are able to analyze these constituents syntactically using a shallow syntactic analysis and thus it is possible to adapt locally the syntactic structure of the target.

3.4 Morphological similarity

Morphological similarity means similar structure of morphological hierarchy and paradigms such as case system, verbal system etc. Baltic and Slavic languages (except for Bulgarian and Macedonian) have a similar case system and their verbal system is quite similar as well. Some problems are caused by synthetic forms, which have to be expressed by analytical constructions in other languages (e.g., future tense or conjunctive in Czech and Lithuanian). The differences in morphology can be relatively easily overcomed by the exploitation of full-fledged morphology of both languages (source and target).

Lexical similarity is the least important one from the point of view of MT, because the lexical differences are solved in the glossaries and general dictionaries. For example, even though Polish is lexically much more similar to Czech than Czech to Lithuanian, both modules (Czech to Lithuanian and Czech to Polish) of our MT system mentioned below work with approximately the same quality.

[3] Although some Balticists argue that illative forms are adverbs, it is a fact that this case is productive and used quite often, though it has been widely replaced by prepositional phrases. Allative and adessive are used only in some Lithuanian dialects, except of a few fixed allative forms (e.g., vakaropi)

4. RUSLAN – A Czech-to-Russian MT System

The first attempt to verify the hypothesis that related languages are easier to translate started in mid eighties at Charles University in Prague. The project of the Czech-to-Russian MT system called RUSLAN cf. (Oliva 1989) aimed at the translation of documentation for operating systems for mainframes. It started in 1985 and it was terminated in 1990 for the lack of funding – there was no need for such system after the political changes in 1989, when the Czechoslovak economy struggled to break free from the old ties and to shift its orientation towards the EU countries. The system has never been used commercially, but it is still being used in teaching at the Charles University.

The system was transfer-based, implemented in Colmerauer's Q-systems (Colmerauer 1969). The transfer phase followed a full-fledged morphological and syntactic analysis of Czech. The dictionaries of the system had almost 10 000 basic word forms covering the domain of manuals for operating systems for mainframes. It also contained a syntactic and morphological generation of Russian.

Originally, there was an assumption that due to the similarity of both languages the transfer phase will be minimal. This assumption turned to be wrong and several phenomena were covered by the transfer in the later stage of the project (for example the translation of the Czech verb "být" [to be] into one of the three possible Russian equivalents: an empty form, the future tense form "будет" and the verb "являться"; or the translation of verbal negation).

Although the system has never undergone any serious evaluation of the translation quality, the tests made during grammar implementation showed that roughly 40% of input sentences were translated correctly, about 40% with minor errors correctable by a human post-editor and about 20% of the input required substantial editing or re-translation.

For the illustration of the translation quality, let us introduce a sample translation of the following Czech paragraph into Russian

Pro nastavení pravidel ocenění jsou k dispozici tak zvané odpisové plány. Tyto odpisové plány jsou definovatelné nezávisle na ostatních organizačních jednotkách. Určitý odpisový plán může být například použit pro všechny účetní okruhy jednoho státu.

Translated by RUSLAN:
Для установки правил ocenění являются к диспозиции так называемой odpisové plány. Этот odpisové plány являются определительный независимо на остальных организационных единицах(!сегментах). Определенный odpisový plán может быть например использован для весь бухгалтерские цепи одного стата.

Post-edited MT :
Для установки правил оцененки к диспозиции так называемые планы одписи. Эти планы одписи определены независимо от остальных организационных сегментов. Определенный одписовый план может быть например использован для всех бухгалтерских цепей одного государства.

There were two main factors responsible for the errors. The first factor was the incompleteness of the main dictionary of the system. A word not contained in any of the dictionaries of the system usually did not have only locally negative impact, it may have caused a failure of the whole module of the syntactic analysis of Czech with dire consequences for the result of the translation.

The reason for this instability was the lexically syntactic information accompanying each lexical item in the dictionary (especially verbs). Without that information it was not possible to complete the syntactic analysis (and therefore also the translation) successfully. The rich lexically syntactic information contained in the dictionary had also one very unpleasant side-effect – the process of building the main dictionary of the system was extremely expensive. It required lot of linguistic expertise and, especially, lot of checking, because the content of the dictionary, being created by several linguists, had a tendency to become inconsistent. The testing and debugging of the grammar went hand in hand with enlarging the dictionary and the changes made in the grammar sometimes required also some adjustments of the dictionary information.

The second factor responsible for errors was the incompleteness of the grammar. Even though the grammar was really large, there were still many less frequent phenomena that were not covered by the grammar rules. Apart from those which were simply omitted due to their lower frequency in technical texts there were also the complicated constructions, which were extremely difficult to handle. For example, quite common in Czech are the so-called non-projective constructions, which may appear even in relatively short sentences, such

as *Soubor se nepodařilo otevřít.* [File (Refl.) was_not_possible to_open. – It was not possible to open the file]. The formalism used for the implementation of the system (Q-systems) is not suitable for an efficient handling of non-projective constructions.

The similarity of both the source and the target language did not help with neither of the two factors mentioned above. It only allowed for certain simplification of the translation process, mainly due to the fact that both languages allow a high degree of word-order freedom. This experiment has clearly shown that a transfer-based approach with full-fledged syntactic analysis of the source language is not able to exploit the similarity of closely related languages to a full extent. The bottleneck is the syntactic analysis of the source language. The syntactic analysis of Czech is equally difficult in the Czech-to-Russian as in the Czech-to-English system.

5. Czech-to-Slovak "shallow" MT system Česílko

The message of our first larger scale experiment is clear – if we really want to exploit the syntactic similarity of related languages, we must abandon the full-fledged syntactic analysis in favor of simpler methods. The simplest possible method is probably the word-for-word translation of individual word forms. It in fact means that the transfer phase follows immediately after the morphological analysis (the morphological analysis is necessary due to a rich inflection of Slavic and Baltic languages, it cannot be omitted).

The greatest problem of the word-for-word translation approach is the problem of ambiguity of individual word forms. The type of ambiguity differs slightly between the group of languages with a rich inflection (majority of Slavic languages) and the group of languages that do not have such a wide variety of forms derived from a single lemma. For example, in Czech there are only rare cases of part-of-speech ambiguities (*stát* [to stay/the state], *žena* [woman/chasing] or *tři* [three/rub(imper.)]), the estimation based on the data from the Prague National Corpus is that only slightly more than 10% of word forms have two or more lemmas with different POS tag. However, the ambiguity of gender, number and case is very high (for example, in Czech the form of the adjective *jarní* [spring] is 27-times ambiguous). The main problem is that even though several Slavic languages have the same property as Czech, the ambiguity is not preserved at all or it is preserved only partially, it is distributed in a different manner and the "form-for-form" translation is not applicable.

Without the analysis of at least nominal groups it is often very difficult to solve this problem, because for example the actual morphemic categories of adjectives are in Czech distinguishable only on the basis of gender, number and case agreement between an adjective and its governing noun. An alternative way to the solution of this problem is the application of a stochastically based morphological disambiguator(tagger) for Czech whose success rate exceeds 94%.

The basic problems of automatic translation from Czech may be also demonstrated on the following example, where the target language is Slovak :

Example 6
Source: *Při* zakládání *třídy výkazů se* třídě *nejprve přidělí* označení *a přiřadí se skupině uživatelů.*

Target: *Pri* zakladaní *triedy výkazov sa* triede *najprv pridelí* označenie *a priradí sa skupine užívateľov.*
[When a report class is founded, the class first receives a label and it is assigned to a group of users.]

The sample sentence contains two interesting phenomena – the translation of similar Czech word forms *zakládání* [founding] and *označení* [label] (both are nouns regularly derived from verbs) into Slovak forms *zakladaní* and *označenie* and the translation of the ambiguous Czech word-form *třídě* [class/sorting].

The translation of the pair of similar words illustrates the fact that even though both languages are really very similar, a „full size" bilingual dictionary is necessary. The translation of similar words is irregular to the extent that prevents the use of some simpler mechanism (direct transcription).

The word form *třídě* may be translated into Slovak either as *triede* (if the original word form represents a noun) or as the form *triediac* (if the original form is a transgressive derived from the verb *třídit* [to sort]). This word form is another ilustration the need of a reliable tagger capable of high quality morphological disambiguation of the input.
Taken these facts into account, we came to the following composition of the system:

1. Morphological analysis of Czech
2. Morphological disambiguation of Czech

3. Domain-related bilingual glossaries
4. General bilingual dictionary
5. Morphological synthesis of Slovak

ad 1.
The morphological analysis of Czech is based on the morphological dictionary developed by Jan Hajič and Hana Skoumalová in 1988-99 (for the tagset description, see (Hajič 1998)). The dictionary covers over 700,000 lemmas and it is able to recognize more than 15 mil. word forms. The morphological analysis uses a system of positional tags (each morphological category has a fixed place in the tag) with 15 positions.

The morphological analyzer is written in C and can effectively process about 5000 tokens per second (sustainable rate, including file compression/decompression, network file sharing, etc.).

ad 2.
The module of morphological disambiguation is a key to the success of the translation. It currently gets an average number of 4.29 tags per unit of text (word) on input (it used to be less in the recent past, but the average number of tags per token is growing due to the continuing expansion of the dictionary, the process of which creates new homonyms). The tagging system is based on an exponential probabilistic model (for the model definition and motivation, end evaluation results see (Hajič 1998). The learning is based on a manually tagged corpus of Czech texts, containing roughly 1.2 mil. tokens. The system learns contextual rules (features) automatically and also automatically determines feature weights. The average accuracy of tagging is now over 94% (measured on tokens of running text).

ad 3.
The domain related bilingual glossaries contain pairs of individual words and pairs of multiple-word terms. The glossaries are organized into a hierarchy specified by the user; typically, the glossaries for the most specific domain are applied first. There is one general matching rule for all levels of glossaries – the longest match wins.

The multiple-word terms are a sequence of lemmas (not word forms). This structure has several advantages, among others it allows to minimize the size of the dictionary. However, it entails preprocessing of the terminological dictionary by the same tools (morphology and tagger) since typically words in terminological phrases are inflected, too, and usually there is no external indication which word is the headword. (In fact, this means we have to have a morphological analyzer and a tagger available for the target language as well, or at least an approximation of a tagger suitable for noun phrase handling.) On the other hand, this greatly simplifies the terminological dictionary handling by the end users: in general, it does not require any special involvement on their part – the linguistic experts responsible for terminology simply maintain the terminological dictionaries as if they are to be used by humans. We believe that this approach might prove to be very important part of our system design, since it eliminates the well-known high cost factor for MT dictionary maintenance

ad 4.
The main bilingual dictionary contains data necessary for the translation of both lemmas and tags. The translation of tags (from the Czech into the Slovak system) is necessary, because both systems use close, but slightly different tag sets. Also, the tags do not always correspond exactly: for example, there are some Slovak nouns which have different gender, or tags with variants which do not exist in the other language. Therefore, a Czech tag is not translated into a single tag, but into a priority-ordered list of tags.

ad5.
The morphological synthesis of Slovak is based on a monolingual dictionary of Slovak, developed by J.Hric (1991-99), covering more than 100,000 lemmas. The coverage of the dictionary is still growing. It aims at a similar coverage of Slovak as has currently been achieved for Czech.

6. Additional language modules

The success of the Czech-to-Slovak MT module (the similarity of the translated and post-edited text exceeded 90%) encouraged additional experiments. The next step was quite natural – Polish also belongs to a group of Western Slavic languages, but it is less similar to Czech than the original target language. In general, according to our expectations, with the decreasing similarity level also the quality of results has decreased. The main translation problems we have encountered concerned word-order (in some kinds of nominal groups), agreement, valency frames and some other minor issues. They caused the drop in the translation quality to a level slightly better than 70%.

The third (and currently the last) module is the Lithuanian one. The choice was also natural – after the tests carried on Western Slavic languages it was necessary to cross a borderline between different language groups. Due to the fact that Slavic and Baltic languages are relatively typologically similar (rich morphology, relatively free word order), we have decided to test the limits of the method by developing a Czech-to-Lithuanian module.

The initial comparative study showed that for Czech-to-Lithuanian translation it is necessary to enrich the scheme of the system by creating a shallow parser working with the results of the tagger and preceding the dictionary lookup phase. The combination of a tagger and a shallow parser has been described for example in (Megyesi 2002).

Although we do have a full Czech statistical parser for Czech (Collins 1999), its current accuracy (about 82-84% correct dependencies) was deemed not being sufficient for our task, while we even did not need a full parse. Therefore, the module of a shallow syntactic analysis of Czech is based on the LFG formalism, even though it does not use the complete LFG framework, as described in (Bresnan, 2001). We leave out e.g. the completeness and coherence conditions and anaphoric binding. The grammar consists of a set of phrase structure rules. Constraints (equations) are assigned to every element of the right-hand side of the rules. The application of the phrase structure rules gives the c-structures, whereas the constraints define the associated f-structures.

The main goal of the module is to analyze only the simpler parts (constituents) of the sentence, such as nominal and prepositional phrases. The result of this module is an underspecified dependency tree. The shallow syntactic parser solved some of the translation problems satisfactorily and allowed the overall quality of the translation to achieve higher level as the quality of the Czech-to-Polish module, one of the test showed even more than 80% similarity between the translated and post-edited output.

7. The Translation Model

The encouraging results of experiments with very simple ("shallow") methods of MT support our initial hypothesis that a similarity of (some) languages of acceding countries may be exploited in a multilingual shallow MT system capable of reasonable quality translation from one (pivot) language (Czech in case of our system, but there is no other reason to stick to Czech than the availability of a high quality tagger, which constitutes a key part of the system). This pivot language will serve as some kind of a "secondary" source language for the translation to a whole set of similar languages. Similar approach has been adopted in (Mann 2001) using so-called "bridge" languages.

The model using pivot language has generally several well-known drawbacks, both linguistic and technical ones. To name just the most important ones, let us mention the danger of a shift of the meaning of the original text by a subsequent translation. Such a shift is unavoidable especially if we translate between non-related languages which are typologically different. Probably the most important technical issue is the necessity to make subsequent translations, that means that the translation to a pivot language must always precede the translation from the pivot to the target language. This in fact doubles the translation time and causes delays in delivering the translated output.

Our translation model tries to address these issues, although it fully solves only the linguistic issue of the shift of the meaning of the translation. The technical issue of translation speed is addressed only partially, due to the speed of our translation system there is almost no delay in delivering the translation from the pivot to the target language. The problem, that everything must be translated into the pivot language first, remains. This is a serious drawback of our translation model especially in case that we need translation to several target languages, but the pivot language is not among them. In that case we still must translate everything into the pivot language and only then to the relevant target languages. Nevertheless, this issue may be solved by adding more pivot languages into the system.

The shift of the meaning is addressed in our model through the exploitation of human translation from the source language (one of the "big" EU languages) using translation memories. These memories may than serve as a basis for the pivot-target language translation. At the end of the combined human (source-pivot) and machine translation (pivot-target) we have at our disposal two translation memories, one human made for the source-pivot language pair and one machine made for the pivot-target language pair. Both translation memories can be combined together by the complete omission of the pivot language and by combining the translation segments for the source and target language. The post-editor will then have a chance to compare the machine translation output with the real original.

The exploitation of translation memories in the translation workflow has one more advantage – if

there are already human made translation memories from the source to the target language, partially covering the translated text, they may be very well combined with the memories automatically created by our system.

We hope that the translation model introduced here is suitable not only for the translation and localization of technical texts and manuals, but also for all kinds of documents which are translated in parallel into multiple syntactically similar languages.

8. Conclusion

This paper advocates the idea that the enormous task of translation of important documents into all official languages of the bigger EU can be simplified if the relatedness of a large group of Balto-Slavic languages is taken into account. In order to make a real advantage from the similarity of these language we suggest to use only shallow syntactic analysis of the source language and to rely on the typological and syntactic similarity of all languages from this group. The results of experiments with our multilingual MT system Česílko, which is based exactly on these assumptions, support our claims.

In the future we would like to improve the shallow parser and transfer rules as well as to extend the MT system to other language pairs.

Furthermore, we would like to integrate the system SProUT (Becker (2002))and its linguistic resources for Central and Eastern European languages (Drożdżyński et al. 2003) with Česílko to simplify the development of grammars.

Acknowledgements

The work described in this paper has been supported by the grant of the MŠMT ČR ME642, No. LN00A063 and partially supported by the grant of the GAČR No. 405/03/0914.

References

Ambrazas V. (1996): Dabartinės lietuvių kalbos gramatika. Mokslo ir enciklopedijų leidykla, Vilnius

Becker M., Drożdżyński W., Krieger H.U., Piskorski J., Schaefer U. and Xu F. (2002): SproUT – Shallow Processing with Typed Feature Structures and Unification. In: *Proceedings of ICON 2002*, Mumbai, India

Bresnan J.. (2001). *Lexical-functional syntax*. Blackwell Publishers, Oxford.

Collins M. et al. (1999). A statistical parser of Czech. In *Proceedings of the 37th ACL'99*, University of Maryland, College Park, MD, USA. 505-512.

Colmerauer, A. (1969): Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Mimeo, Montréal

Drożdżyński W., Homola P., Piskorski J. and Zinkevičius V. (2003): Adapting SproUT to processing Baltic and Slavonic languages. In: *Proceedings of Information Extraction for Slavonic and Other Central and Eastern European languages*, Borovec, Bulgaria

Gamut L.T.F. (1991): *Logic., language and meaning 2: Intensional logic and logical grammar*, University of Chicago Press, Chicago

Hajič, Jan ; Homola, Petr; Kuboň, Vladislav (2003): A Simple Multilingual Machine Translation System. *Proceedings of Machine Translation Summit IX*, pp. 157-164, New Orleans, USA.

Hajič, Jan ; Hric, Jan; Kuboň, Vladislav (2000): Machine Translation of Very Close Languages. In: *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, USA, April 2000, pp. 7-12

Homola, Petr; Rimkutė, Erika (2003): Shallow machine translation - in between of two extremes. *Proceedings of The Fifth International Tbilisi Symposium on Language, Logic and Computation*. Tbilisi, Georgia.

Mann, Gideon S., Yarowsky David (2001): Multipath Translation Lexicon Induction via Bridge Languages, *Proceedings of the Second Conference of the North American Association for Computational Linguistics 2001*, Pittsburgh, PA

Megyesi, B. (2002). Shallow Parsing with PoS Taggers and Linguistic Features *Journal of Machine Learning Research: Special Issue on Shallow Parsing*, *JMLR* (2): 639-668. MIT Press

Oliva, Karel (1989): A parser for Czech implemented in Systems Q. *Explizite Beschreibung der Sprache und automatische Textbearbeitung* XVI, MFF UK, Prague.

Panevová J. (1980): Formy a funkce ve stavbě české věty. *Studie a práce lingvistické*, Academia, Praha

Sgall P. (1974): Úvod do algebraické lingvistiky, Lecture notes, Charles University, Prague.