Vladislav Kuboň

Multilingual Aspects of Monolingual Corpora

If someone would collect opinions among the computational linguists what had been the most important trend in linguistics in the last decade, it is highly probable that the majority would answer that it was the massive use of large natural language corpora in many linguistic fields. The concept of collecting large amounts of written or spoken natural language data has become extremely important for several linguistic research fields.

The majority of large corpora used by linguists are monolingual, although there are several examples of bilingual corpora (e.g. Hansard corpus). This paper would like to present evidence that even the monolingual corpora can be useful for multilingual applications.

1 The brief history of major Czech corpora

If we would look approximately ten years back, we would find that the Czech computational linguistics still suffered from the technical gap caused by the long period of isolation from the modern trends in hardware and software development. The first RISC workstations were introduced to Czech universities in the frame of the IBM Academic Initiative in 1991. Also from the point of view of the system of funding the research in our country had undergone a major transformation, the Grant Agency of the Czech Republic (GACR) had been established at the beginning of nineties and it started to organize the research in a standard form of research grants of various forms.

1.1 The Czech National Corpus

The support of the GACR and the efforts of several people from the Charles and Masaryk Universities and from the Academy of Sciences allowed to create a Czech National Corpus project and, subsequently, led to the establishment of the Department of the Czech National Corpus at the Charles University in 1994. Having gradually gained support from the GACR, Ministry of Education, various publishing houses etc. the consortium of cooperating institutions has grown and currently consists of five faculties from three universities and two institutes of the Academy of Sciences.

All these efforts have led to the creation of a large scale corpus of contemporary Czech language called the Czech National Corpus (CNC) and allowed to open the first part of it, a 100 million word corpus called SYN2000, for a general use in 2000 (CNK

2000). Apart from SYN2000 the corpus contains another 400 million words in files not yet publicly available. The structure of the SYN2000 is the following:

- 15 % Literature (11% Fiction)
- 60% Journalism
- 25% Technical and specialized texts

1.1.1 Tagging of the CNC

The CNC is annotated on the morphemic level, therefore the key procedure in its building has been a procedure of morphological tagging. The process of tagging consists in two basic parts – morphological analysis and the disambiguation of ambiguous tags.

1.1.1.1 Czech morphological analysis

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.

Example 1:
tags assigned to the word form “pomocí” (help/by means of)

```
NFP2------A----
NFS7------A----
R--2-----------
```

where:
N – noun; R – preposition
F – feminine gender
S – singular, P – plural
7, 2 – case (7 – instrumental, 2 – genitive)
A – affirmative (non-negative form)

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.).

1.1.1.2 Morphological disambiguation of Czech

The module of morphological analysis 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).

Training of the tagger is based on manually annotated newspaper text. The resulting tagger has over 11 thousand rules total for all morphological categories (feature batches, in the terminology of Hajič 1998) selected and weighted during the training process. These rules are stored in an SGML format. The tagger is reasonably fast, with the full set of 11 thousand rules it can tag at a sustainable rate of 200 tokens per second.

1.2 The Prague Dependency Treebank

The experience gained during the first stage of building the Czech National Corpus opened new directions for all the participating people and institutions. Some of them devoted their efforts towards improvements and enlargements of the CNC, the others, namely the people from the Masaryk University in Brno, have decided to create several small and middle size specialized corpora (e.g. DESAM, Pala 1998 etc.).

Our group at the Faculty of Mathematics and Physics at the Charles University in Prague led by prof. Eva Hajičová undertook a very ambitious project – to create a corpus annotated on multiple levels – morphological, analytical and underlying-syntactic layer (for a description of the tagging scheme of PDT, see e.g. Hajič 1998, Hajič and Hladká 1997, Hajičová 1998, 1999, and the two manuals for tagging published as Technical Reports by UFAL and CKL of the Faculty of Mathematics and Physics, Charles University Prague and available also on the website http://ufal.mff.cuni.cz). The annotation on the underlying syntactic level the result of which are the so-called tectogrammatical tree structures (TGTS in the sequel) is based on the original theoretical framework of Functional Generative Description as proposed by Petr Sgall in the late sixties and developed since then by the members of his research team (Sgall, Hajičová and Panevová 1986).

The work on the Prague Dependency Treebank (PDT) has started in 1996. In the first phase (1996-2000), the morphological and syntactic analytic layers of annotation

have been completed and were published together with the preview of tectogrammatical layer annotation in 2001 by Linguistic Data Consortium as PDT 1.0. In the second phase (2000 - 2004), the work goes on the tectogrammatical layer of annotation and PDT 2.0 will be available in the end.

The corpus uses as the source textual material a subcollection of texts from the Czech National Corpus. From the very start, the build-up of PDT has been conceived as a combination of automatic and manual procedures, the latter being supported by various kinds of user-friendly software tools. As for the annotations on the morphological layer, detecting the discrepancies between the annotators and discrepancies between the annotations and the output of the automatic morphological analysis is carried out fully automatically. Regarding the annotation on the analytic syntactic layer, the situation is different in that the human annotators had at their disposal an automatically preprocessed sentences (using Collins' parser) into a form of dependency syntactic trees and their task was to edit these trees (again with the use of a special software tools) and to attach labels according to their own judgements.

1.2.1 PDT 1.0 Overview

The PDT 1.0 contains 1 725 242 tokens (words, punctuation marks etc.) in 111 175 sentences annotated on morphological level, 1 507 333 tokens in 98 263 sentences annotated on analytical level and a sample of 3 490 tokens in 203 sentences annotated on the tectogrammatical level.

The annotation at the morphological level is an unstructured classification of individual tokens (words and punctuation) of the utterance into morphological classes (morphological tags) and lemmas. The original word forms are preserved, too. In fact, every token has gotten its unique ID within the corpus for reference reasons. Sentence boundaries are preserved and/or corrected if found wrong (the errors in original texts contained in the Czech National Corpus have been preserved in the corpus). The number of tags actually appearing in the PDT is about 1100 out of 4257 theoretically possible. The data has been double annotated fully manually, the annotators selected a correct tag out of a set provided by a module of an automatic morphological analysis (cf. Hajič 2001).

At the analytical level, two additional attributes are being annotated:

- (surface) sentence structure,
- analytical function

A single-rooted *dependency tree* is being built for every sentence as a result of the annotation. Every item (token) from the morphological layer becomes (exactly) one node in the tree, and no nodes (except for the single „technical" root of the tree) are added. The order of nodes in the original sentence is being preserved in an additional attribute, but non-projective constructions are allowed. Analytical functions, despite

being kept in nodes, are in fact names of the dependency relations between a dependent (child) node and its governor (parent) node. Only a single (manually assigned) analytical annotation (dependency tree) is allowed per sentence. There are 24 analytical functions used, such as *Sb* (Subject), *Obj* (Object), *Adv* (Adverbial), *Atr* (Attribute in noun phrases) etc.

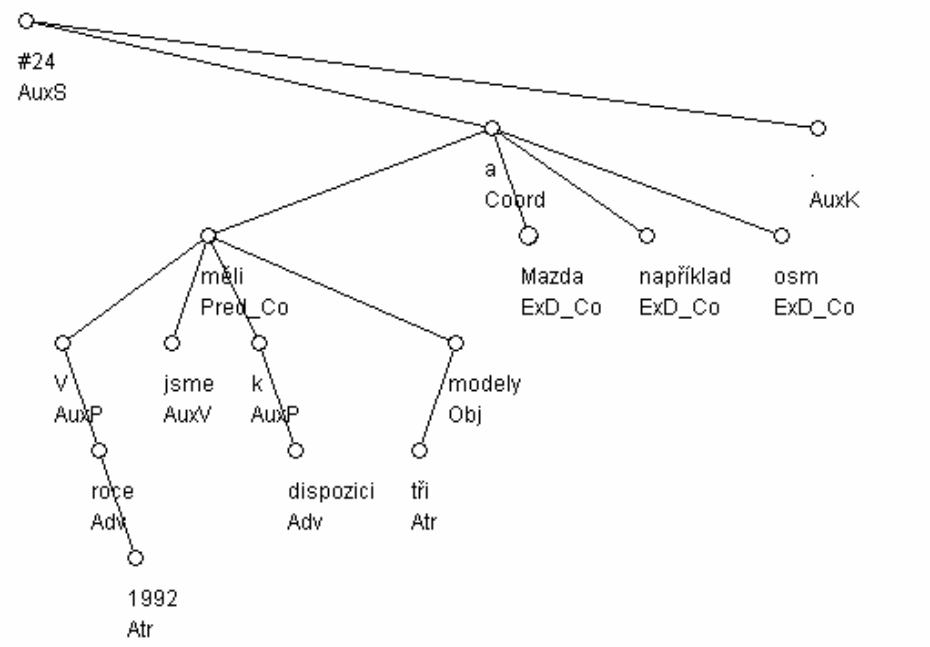

Fig. 1: Analytical annotation of the sentence „V roce 1992 jsme měli k dispozici tři modely a Mazda například osm." [In the year 1992 we had at our disposal three models and Mazda (had) for example eight (models).].

The tectogrammatical level is the most elaborated, complicated but also the most theoretically based layer of syntactico-semantic (or „deep syntactic") representation. The tectogrammatical layer annotation scheme is divided into four sublayers:

- dependencies and functional annotation,
- the topic/focus annotation including reordering according to the deep word order,
- coreference,
- the fully specified tectogrammatical annotation (including the necessary grammatical information).

As an additional data structure we use a syntactic lexicon, mainly capturing the notion of *valency*. The lexicon is not needed for the interpretation of the tectogrammatical representation itself, but it is helpful when working on the annotation since it defines when a particular node should be created that is missing on the surface. In other words, the notion of (valency-based) ellipsis is defined by the dictionary.

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as „subject“ and „object“ by notions like „actor“, „patient“, „addressee“ etc. The representation itself still relies upon the language structure itself rather than on world knowledge. The nodes in the tectogrammatical tree are *autosemantic words* only. Dependencies between nodes represent the relations between the (autosemantic) words in a sentence, for the predicate as well as any other node in the sentence. The dependencies are labeled by *functors*, which describe the dependency relations. Every node of the tree is furthermore annotated by such a set of grammatical features that enables to fully capture the meaning of the sentence.

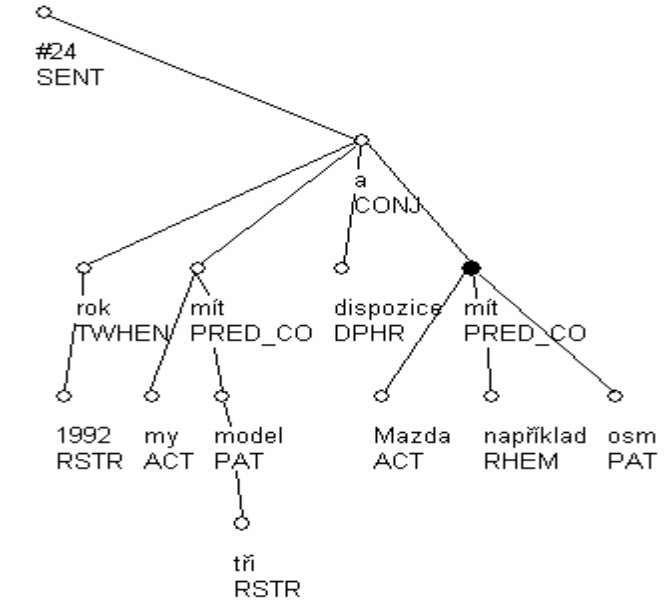

Fig. 2: Tectogrammatical tree for the sentence from the Fig.1.

2 The state-of-the-art of the PDT

Since the PDT 1.0 was published, the work continued especially on the annotation of the tectogrammatical level. In the last year we have annotated the structure, topic/focus articulation and coreference. The valency dictionary has been fully integrated into the annotation for the reason of consistency of the annotation. A fully annotated valency dictionary was being prepared in parallel to the annotation (Straňáková-Lopatková and Žabokrtský 2002).

20 000 sentences had been annotated on the tectogrammatical level, about 1000 have got also the TFA annotation. The adding of coreference markers has just started. Approximately 300 sentences had been fully annotated (including coreference and grammatemes).

Manual annotation of lexico-semantic data had been in progress. The main task of this annotation is to distinguish the polysemy.

The software tools for annotation, checking and processing of data have been improved. Among these tools the most important ones are the tree editor TrEd (the main tool for manual annotations) and the multi-user system Netgraph (at the moment the main user tool for accessing and searching the treebank, based on the client-server architecture).

Since the beginning of the project we are working on the (partial) automatization of the transformation of an analytical tree into the tectogrammatical one. Due to the fact that the analytical tree contains less information than the tectogrammatical one, it is impossible to make the transition fully automatic. The automatic transformation fills in the values of attributes, adds some nodes elidated on the surface and deletes nodes of synsematic words and punctuation (Řezníčková, Veronika (in print)).

It turned out that one of the major obstacles in the tectogrammatical annotation is the addition of nodes elidated on the surface, especially if those nodes belong to a valency frame of a particular word. Even though this activity has been done using the valency dictionary being created in parallel to the annotation, it is extremely difficult. The main problem is the rich variety of verbal valency frames. The majority of Czech verbs has several different variants of valency frames. Thus even if the verb is completely covered by the valency dictionary, it is not easy to assign it the correct valency frame out of the set provided by the dictionary.

The building of the valency frame is organized with respect to the relative frequency of verbs in the corpus, the most frequent verbs take precedence. The dictionary currently contains complete valency frames of approximately 1000 most frequent Czech verbs.

3 Multilingual applications related to Czech corpora

Although both major Czech corpora are monolingual, it is not very difficult to find several examples of multilingual applications at least partially exploiting either the corpora directly or indirectly by sharing the tools or exploiting the data on the theoretical level. Let us briefly mention two examples of machine translation systems related to the CNC and PDT. The first system, Česílko, is built upon the tagger used for tagging the CNC, the second one, an experimental Czech-to-English stochastic MT system uses the concept of tectogrammatical trees as means how to simplify the transfer phase. It also relies on the experience of human annotators trained on the annotation of PDT. These annotators have created a small size parallel Czech, English and Arabic training corpus based on the Wall Street Journal data from the PennTreebank.

3.1 The multilingual MT system Česílko

One of the most widely used techniques of machine-aided human translation of the last decade is without doubts a method of human translation supported by a translation memory. This technique can substantially speed up the translation process especially when it concerns the translation and localization of various kinds of technical documentation.

The main idea of the translation memory is very simple. It takes an advantage from the fact that it is often the case (especially when localizing technical documentation) that for the currently translated document there is at least one document with similar content that had already been translated. Such a document may for example be a part of the previous version of the documentation to a particular software or hardware. The translation memory in fact contains both the source and target text divided into pairs of segments. These segments are typically sentences. When a human translator starts translating a new sentence, the system tries to match the source sentence with sentences already stored in the translation memory. If it is successful, it suggests the translation and the human translator decides whether to use it, to modify it or to reject it.

3.1.1 The use of the translation memory in the system Česílko

It is quite clear that the localization of the same source into several typologically similar target languages individually, one language pair after another, is a waste of money and effort. In the translation process it is necessary to solve very similar problems for each source-target language pair.

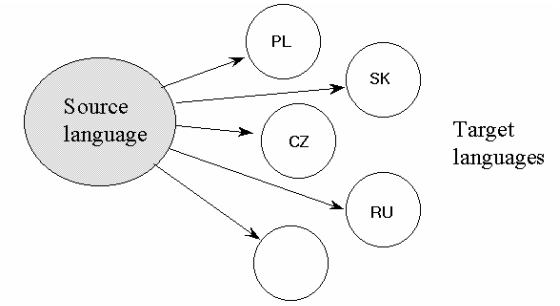

Fig. 3: A traditional model for localization.

The use of one language from the target group as a pivot and to perform the translation and localization through this language seems to be quite natural solution for these problems. It is of course much easier to translate texts from Czech to Polish or from Russian to Bulgarian than from English or German to any of these languages.

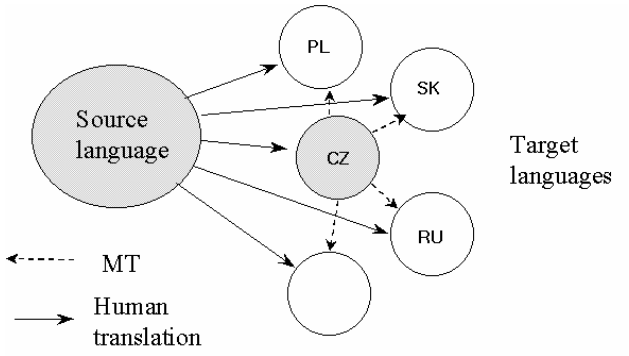

Fig. 4: Our model for localization.

The system Česílko was designed as a tool allowing automatically constructing translation memories between very closely related languages (such Czech and Slovak) for human translators. Such translation "memory" would then be used as if created by humans, but appropriately marked for the human translators.

If we have two translation memories at our disposal – one human made for the source/pivot language pair (say, English/Czech) and the other one created by an MT system for the pivot/target language pair (Czech/Slovak), the substitution of segments of a pivot language (Czech) by the segments of a target language (Slovak) is then only

a routine procedure. The human translator translating from the source (English) to the target language (Slovak) then gets a translation memory for the required pair (source/target).

The system of penalties which is normally applied to results of MT in translation memory based systems guarantees that if there is already a human-made translation present in the memory, it gets higher priority than the translation obtained as a result of MT.

3.1.2 Basic properties of the system

In the group of Slavic languages there are very few languages more closely related than Czech and Slovak. This fact has led us to the idea that the core of the system of a Czech to Slovak MT should use as simple method of analysis and transfer as possible. Our experience from an existing MT system RUSLAN (Czech-to-Russian MT system) aimed at the translation of software manuals for operating systems of mainframes (cf. Oliva 1989) led us to the idea that a full-fledged analysis of Czech is not necessary. According to this experience, full syntactic analysis would be too unreliable. The other reason why the syntactic analysis of the source text was omitted was the fact that such an approach would not profit from the closeness of both languages as much as a less complicated method. The system therefore uses the method of direct word-for-word translation, the use of which is justified by the similarity (even though not identity) of syntactic constructions of both languages.

The system has already been tested on texts from the domain of documentation to corporate information systems. It is, however, not limited to any specific domain. Currently it is being tested on texts of a Czech general encyclopedia. Its primary task is, however, to provide support for translation and localization of various technical texts.

3.1.3 Problems of machine translation between Czech and Slovak

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.)]), however, the ambiguity of gender, number and case is very high (for example, 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 was the application of a stochastically based morphological disambiguator for Czech whose success rate is relatively very high. This is the point where the multilingual MT system touches monolingual corpus – for successful development and training of the stochastic morphological disambiguator (tagger) we need morphologically annotated corpus (although it doesn't need to be as large as CNC).

The necessity of disambiguation may be illustrated on the following example of Czech to Slovak translation:

Example 2:
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 *zakládaní* and *označenie* and the translation of the 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 illustration 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. Import of the source (Czech input) sentence (a segment from an "empty" translation memory")
2. Morphological analysis of Czech
3. Morphological disambiguation of Czech

4. Domain-related bilingual glossaries
5. General bilingual dictionary
6. Morphological synthesis of Slovak
7. Export of the output to the original translation memory (Slovak target sentence)

The role of tagging in a MT system between closely related languages is crucial. The tagger (together with the corresponding morphological analyzer) is used at three different places in the system: in the preprocessing stage, for the tagging of both the source and target dictionaries, and at runtime, for tagging the input (source) sentence. In all three cases, we also use the results of tagging for lemmatization, due to relatively high degree of lexical homonymy in Czech and Slovak (even though the lexical homonymy, as opposed to morphological homonymy, is lower than, say, in English), since lemmatization amounts basically to major part of speech disambiguation.

3.1.4 Evaluation of results

The problem how to evaluate results of automatic translation is very difficult. For the evaluation of our system we have exploited the close connection between our system and the TRADOS Translator's Workbench. The method is simple – the human translator receives the translation memory created by our system and translates the text using this memory. The translator is free to make any changes to the text proposed by the translation memory. The target text created by a human translator is then compared with the text created by the mechanical application of translation memory to the source text. TRADOS then evaluates the percentage of matching in the same manner as it normally evaluates the percentage of matching of source text with sentences in translation memory. In the first testing on relatively large texts (tens of thousands words) the translation created by our system achieved about 90% match (as defined by the TRADOS match module) with the results of human translation.

3.1.5 Experiments with other target languages

The success of the Czech to Slovak module has encouraged further experiments. It is clear that a word-for-word approach to MT as it was described in previous sections is applicable only to languages with high degree of syntactic similarity. An open question is where is the real limit of applicability of our method, which pairs of languages are close enough for our method to provide reasonable quality of translation and which are not. It is therefore quite natural to extend our system to other Slavic languages.

3.1.5.1 Czech-to-Polish experiment

Due to the fact that, as far as we know, no other Slavic language has so many resources for stochastic natural language processing, it is quite natural that we are going to stick to Czech as a source language. The candidate for a new target language was Polish. It is close enough to Czech but it contains several phenomena that are different and thus it provides the natural "next step".

In order to obtain results comparable to the Czech-to-Slovak system we have used the same set of test data and the same evaluation method. The Polish morphological data was kindly provided to us by Morphologic, Inc. (Budapest, Hungary). We converted the data for use with our morphological generator. The comparison of the output from our system with the text post-edited by a Polish native speaker led to following results:

- 25,6% of sentences from the test sample did not require any postediting
- 16,7% of sentences were marked with less than 50% match against the correct post-edited sentences
- 33,3% of sentences achieved a match between 75% and 99%
- 24,4% of translated sentences had a match between 50% and 75%

The weighted average match (the length of a particular sentence was used as a weight) throughout the testing sample reached **71.4%.**

A match lower than 50% does not mean that the sentences were not usable for post-editing. For example, one of the sentences with very low match was the following sentence:

Czech original:
Požadavky starší třiceti dnů se mažou.
[The requests older than 30 days are deleted.]
The result of MT:
Żądania starszy trzydziestu dzieni się smarują.
Post-edited Polish sentence:
Żądania starsze niż trzydzieści dni są wymazywane.

The match between the result of MT and the correct Polish sentence was 32% (according to TRADOS Translator's Workbench standard computation), even though we need only 21 elementary operations to get the correct sentence (50 characters long) from the automatically translated one.

3.1.5.2 Czech-to-Lithuanian experiment

The results of Czech-to-Polish module were relatively good and encouraging with respect to the possible simple improvements – several of the problematic phenomena

responsible for errors can be handled by relatively simple means (e.g. syntactic analysis of nominal groups etc.). On the other hand, they were not surprising. Polish is less similar to Czech than Slovak and the results are therefore worse. In order to achieve scientifically more interesting results we have decided to abandon for the moment the group of Slavic languages and to try to apply our method to a language from a different language group.

The choice was natural –Baltic languages are both geographically close and relatively syntactically similar to Czech. The experiment with Czech-to-Lithuanian MT was performed on a very small sample of text. It indicated some language phenomena responsible for the lower quality of the translation. Some of those phenomena are similar to those already encountered in Czech-to-Polish translation, for example word order variations or an agreement in gender between adjectives and nouns in nominal groups, some will require special treatment (verbal aspect and past tense, reflexive verbs etc.), but generally it seems that the translation quality might be similar to the quality achieved in the Czech-to-Polish experiment. The thorough testing on the data used for the evaluation of Polish and Slovak modules is currently in progress.

3.2 Czech-to-English stochastic MT system

In the fall of 2001 the Center of Computational Linguistics at the Faculty of Mathematics and Physics started a new machine translation project. The main idea of this project came from a generally accepted belief that the deeper the analysis, the smaller the transfer and vice versa. We believe that the tectogrammatical representation is deep enough to allow minimizing even the transfer between typologically different languages.

The project is oriented mainly towards the translation between Czech and English, but it is not limited to this language pair, as there is also a parallel work going on concerning the tectogrammatical annotation of Arabic (Smrž, Zemánek and Šnaidauf 2002). In the first step we have translated about a half of PennTreebank (approx. 500,000 words) by human translators, native speakers of Czech. A small portion of the translated sentences (and their original counterparts from PennTreebank) had been tectogrammatically annotated in the same manner as the sentences from the PDT (Kučerová and Žabokrtský (in print)). The small parallel corpus had then been used as training data for the stochastic procedure transforming Czech tectogrammatical trees into the English ones.

The project (described e.g. in Cuřín, Havelka, Čmejrek (in print) or in Hajič 2002) aims at the fully automatic stochastic translation, including stochastic analysis of the source text to the tectogrammatical level (Honetschläger 2002). Such a procedure, if successful, would not only serve the MT system, but it would speed up the tectogrammatical tagging of PDT substantially, similarly as Collins‘ parser did for the ana-

lytical level. In this respect this project really brings multilingual flavor to the monolingual PDT.

This project really shows that even a monolingual corpus has many multilingual implications. If not the data itself, then the methods, tools and theoretical results connected to the long and complicated process of building a large scale multiple level annotated corpus are reusable even for other languages.

4 Conclusion

Building large scale monolingual corpora is always an effort consuming huge amounts of funding and manpower. Especially at the beginning of the process, when a substantial amount had already been invested and the successful applications exploiting the corpus are still far away, it is very difficult to keep working, to gain more financial support and not to loose a track leading towards the successful end. We hope that our example shows that it is never too late to start building a new corpus according to the requirements of modern linguistics, and that after a painful period of hard and monotonous work it is suddenly possible to use the data in various ways, including multilingual applications.

The future efforts concerning PDT will be concentrated on four domains:

- to reach a solid volume of annotated data on all the three layers,
- to extend the scenario to cover a more subtle representation of coreferential links, and to make it possible to design as a next step a translation of the TGTSs to some kind of formal semantic (logical) representation,
- not to loose sight of applications based on the PDT both in the domain of automatic translation, document retrieval and information extraction, and
- to prepare grounds for a similarly systematic compilation and annotation of a spoken language (speech) corpus.

Acknowledgement

The work presented in this paper has been supported by the grant of MŠMT ČR No. LN00A063.

References

Cuřín, Jan/Havelka, Jiří/Čmejrek, Martin (in print): Czech-English Dependency-based Machine Translation. PBML 78. MFF UK, Prague.

Český národní korpus. Úvod a příručka uživatele, (Czech National Corpus. An Introduction and User's Manual) 2000. Eds. Kocek J., Kopřivová, M., Kučera, K., Filozofická fakulta KU Praha (in Czech).

Hajič, Jan (2001): Disambiguation of Rich Inflection (Computational Morphology of Czech). Prague, Czech Republic: Faculty of Math. and Physics, Charles University, hab. thesis.

– (1998): Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Festschrift for Jarmila Panevová*. Karolinum, Charles University Press, Prague, pages 106-132.

– (2002): Tectogrammatical Representation: Towards a Minimal Transfer in Machine Translation. *Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, 20-23 May 2002, 216-226.

Hajič, Jan/Hladká, Barbora (1997): Probabilistic and Rule-Based Tagger of an Inflective Language – A Comparison. In: Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington D.C., 111-118.

Hajičová, Eva (1998): Prague Dependency Treebank: From Analytic to Tectogrammatical Annotations. In: Text, Speech , Dialogue, ed. by P. Sojka, V. Matoušek and I. Kopeček,Brno: Masaryk University, 45-50.

Hajičová, Eva (1999): The Prague Dependency Treebank: Crossing the Sentence Boundary.. In: Text, Speech and Dialogue, ed. by V. Matoušek, P. Mautner, J. ocelíková and P. Sojka, Berlin: Springer, 20-27.

Honetschläger, Václav (2002): Analytical and Tectogrammatical Syntactical Parsing. In: Proceedings of the WDS 2002, MFF UK, Prague.

Kučerová, Ivona; Žabokrtský, Zdeněk (in print): Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. PBML 78. MFF UK, Prague.

Pala, Karel/Rychlý, P./Smrž, Pavel (1998): DESAM – An Annotated Corpus for Czech, Proceedings of SOFSEM'98, Springer.

Sgall, Petr/Hajičová, Eva/Panevová, Jarmila (1986) The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel.

Oliva, Karel (1989): A Parser for Czech Implemented in Systems Q; Explizite Beschreibung der Sprache und automatische Textbearbeitung XVI, MFF UK Prague

Řezníčková, Veronika (in print): PDT: Two Steps in Tectogrammatical Annotation ... with respect to the issues of deletions. PBML 78. Charles University, Prague.

Smrž, Otakar; Zemánek, Petr; Šnaidauf, Jan (2002): Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. Proceedings of the International Symposium on the Processing of Arabic, pp. 147-155.

Straňáková-Lopatková, Markéta; Žabokrtský, Zdeněk (2002): Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. LREC2002, Proceedings, vol.III. (eds. M. González Rodríguez, C. Paz Suárez Araujo).