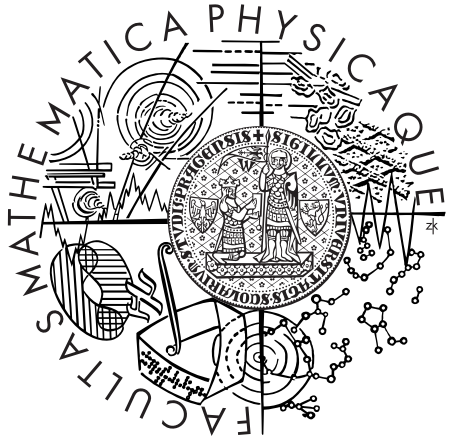
Charles University in Prague
Faculty of Mathematics and Physics

Doctoral Thesis



LINGUISTIC ISSUES IN MACHINE TRANSLATION BETWEEN CZECH AND RUSSIAN
MT between Czech and Russian

Natalia Klyueva

Prague, 2015

Doctoral Thesis

Natalia Klyueva

Supervisor of the doctoral thesis:
Vladislav Kubon

Linguistic Issues in Machine Translation between Czech and Russian
MT between Czech and Russian

Study programme: Computer Science
Specialization: Mathematical Linguistics

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

Prague, 2015

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In . . . . . .

NГғzev prГғce:	Lingvistické Otázky ve Strojovém Překladu mezi Češtinou a Ruštinou.. Strojový překlad mezi češtinou a ruštinou
Autor:	Natalia Klyueva
ГЭstav:	Ústav formální a aplikované lingvistiky
VedoucГн disertaДЊnГн prГғce:	Vladislav Kubon, Ústav formální a aplikované lingvistiky
KlГнДЊovГғ slova:	strojový překlad, slovanské jazyky, blízké jazyky, čeština, ruština, SMT, RBMT, Moses, valence

Abstrakt: Popisujeme strojový překlad z hlediska lingvisty.

Title:	Linguistic Issues in Machine Translation between Czech and Russian. MT between Czech and Russian
Author:	Natalia Klyueva
Department:	Institute of Formal and Applied Linguistics
Supervisor:	Vladislav Kubon, Institute of Formal and Applied Linguistics
Keywords:	machine translation, Slavic languages, related languages, Czech, Russian, SMT, RBMT, Moses, valency

Abstract: In this thesis we analyse Machine Translation from a poin of view of a linguist.

to my family

Contents

1 Introduction

Machine Translation (MT) is considered to be one of the most popular branches of the Natural Language Processing. The work on MT systems generally presented a collaboration between linguists and computer scientists: linguists worked on the initial data - like dictionaries and rules; computer scientists implemented the baseline of the system; linguists analysed the output translations and on their basis suggested further improvements to data and rules and then the cycle was repeated.

The interplay between linguistics and computer science as described above was true for rule-based machine translation (RBMT) systems only before the data-driven (statistical, SMT) approach was adopted in the beginning of 90's. There was no longer a need for a linguist with a knowledge of the source and the target languages: all the necessary information was acquired from data, the evaluation was done either automatically or manually by native speakers, not experts in linguistics.

In our work, we combine observations and findings from both theoretical linguistics and computer science, exploring the performance of several MT systems - RBMT and SMT - through a prism of a linguist.

1.1 Motivation

When the research started in 2006, there was no working MT system between Czech and Russian, so out primary goal was to make an experimental implementation(s) of Czech-Russian MT system within the available frameworks.

Our work was initially supposed to answer a range of questions. First one is which system architecture - rule-based or statistical - is more appropriate for the translation between related languages. Another goal was to spot errors that are typical for each strategy. Our initial hypothesis was that for such related languages as Czech and Russian there is no need to train statistical models to achieve good quality.

The second hypothesis was that under the similar setup, the translation for the related languages is easier to set than for those unrelated.

Finally, we aimed to specify a classification of errors for the MT between Czech and Russian and tie those errors to certain linguistic discrepancies between the two languages; then to compare how SMT and RBMT systems cope with certain linguistic phenomena. As it was virtually impossible to describe all the errors and all points of differences that can cause problems, we concentrated on one of the issues - valency in Czech and Russian languages.

1.2 Outline of the Thesis

The thesis is structured into 5 Chapters.

The first three chapters are of an introductory nature. In Chapter 2, we make a brief overview of Machine Translation - the basic concepts, history of MT, methods and strategies. Chapter 3 presents data and tools that will be used in our experiments. In Chapter 4, we describe the existing MT systems between Czech and Russian. The accent is made on the two systems: a Statisctical one - Moses and a Rule-Based - TectoMT, because those systems are the only ones that we can test, change and adjust[1]. First, we set a baseline for those systems, then we propose some improvements and we take the output of the best experiment to be evaluated further.

Next, in Chapter 5, we explore the output of 4 MT systems, two experimental: TectoMT, Moses and two commercial: PC Translator and Google Translate. Exploring errors in th MT output, we try to answer the question which types of discrepancies between Czech and Russian are successfully processed by an MT and which pose a problem. We also describe which of them result in a system architecture peculiarities a which are tied to the language issues. The evaluation presented here is of a linguistic nature, it is applicable only to the concrete language pair. First, we propose a classification of errors suitable for our language pair. Each error type is analysed and illustrated with examples. Then, we contrast the linguistic phenomenon underlying the error for the two languages and suggest possible reasons why they occurred in the concrete system.

[1] **Collaboration remarks**. Because of the author's linguistic background, some of the experiments described here were carried out with the help of colleagues from ÚFAL. Implementation of Moses Statistical Machine Translation was done in collaboration with Ondřej Bojar, Karel Bílek and David Kolovratník. Zdeněk Žabokrtský and Martin Popel helped to set a baseline for a Rule-Based TectoMT between Czech and Russian. Some of the results presented in Chapter 4 partially intersect with the Master Thesis (Bílek, 2014) done under the same project, but in the latter work the accent is put mainly on technical aspects of the implementation of the MT systems without a deeper linguistic analysis.

In several cases, we did some experiments to fix the error, but for the majority of cases this was not possible in an adequate amount of time.

Out of all the errors, we concentrate on surface valency, especially on theoretical description of valency discrepancies between the two languages (Chapter 6). We conduct some experiments meant to spot the cases of differences and present some observations when the surface valency in Czech and Russian tends to be different.

The summary of our main results and the discussion on what has been done is presented in the concluding Chapter 7.

2

An overview of the Machine Translation

In the first part of the thesis, we discuss the area of Machine Translation. We review the MT architectures and describe the MT systems available for the Czech-Russian language pair, both historical and state-of-the-art ones.

MT is a process of converting a text coded in one language into the other language by means of computer programs. MT is considered to be one of the most popular branches of Natural Language Processing: publications on the topic cover a vast range of problems - as, ex. technical questions of a system development, theoretical research on language aspects, quality estimation, end-user application etc. The most famous web collection of articles about MT[1] counts more than 11,400 items (as of April 2015).

The detailed overview of virtually all aspects of the Machine Translation is provided by (Hutchins, 2007), and the facts presented in this introduction are partially based on this survey.

2.1 History

The first attempts to build the system that substitutes a human translator started shortly after first ascendants of computers appeared. Perhaps the most famous quote that inspired many MT researches comes from Warren Weaver (in 1949):

I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.

Russian and English languages were the first ones for which an MT system was developed by teams in the US and in the former USSR. That was made evidently due to the political reasons, as both countries needed to translate huge amount of texts between the mentioned languages. The first MT system was Russian-to-English, developed by the research team from IBM. Afterwards, many

[1] http://www.mt-archive.info/

other researchers all over the world started to work on their own systems for various languages. Among the early and most successful MT's were, for instance, METEO[2] or Systran[3].

The first systems were rule-based, using dictionaries and implementing linguistic rules (detailed description follows). Only in 1990's with the growth of computer speed and memory capacity it became possible to develop first statistical models, and IBM team again was the first to introduce the first prototype, see the Section 2.4.

The choice of languages for Machine Translation projects had often been politically or geopolitically motivated. For example, in the Czech Republic a Czech-to-Russian MT project Ruslan (Hajic, 1987), (Oliva., 1989) has started in 80's, as Russia had a high influence in this region that time, and it was a high-priority pair that days. At the same time, the experiments in the MT from English into Czech(APAC) had taken place (Kirschner – Rosen, 1989). Ruslan and APAC were both implemented in Q systems.

Nowadays the majority of MT systems, both industrial and research ones, are developed for English (as an international language) and some other language. Still, there exist some MT projects aiming directly at translation between languages other than English (especially MT between related languages) which we will describe in this chapter.

2.2 Approaches to MT

Generally, researchers distinguish two main approaches to the MT systems: **Rule-Based** and **Statistical**, though some other types related to the two main ones can be considered as well - e.g. Example-Based MT and Hybrid. In our work we will stick to the dichotomy Rule-Based vs. Statistical, mentioning further in the texts how they intersect with Hybrid[4]. Many works exist on the topic where the two types of systems were compared - (Thurmair, 2004) for English and German or (Bojar, 2012) for English and Czech, to mention some of them.

[2] An MT between English and French within a domain of weather forecast (Chandioux, 1988)

[3] http://www.systran.co.uk/

[4] We should add a disclaimer that RBMT systems can be regarded as Hybrid, as soon as they exploit statistical modules in analysis, transfer or synthesis. The same is true of SMT, which can be considered to be Hybrid as soon as some linguistic knowledge is being introduced, ex. additional morphological information in a form of morphological dictionaries. In this work when we say RBMT, we actually mean MOSTLY Rule-based that might contain some statistical modules, and under SMT we mean MOSTLY Statistical that might involve some linguistic knowledge.

Moreover, a competition between the most popular MT systems (mostly statistical, but RBMT are also present) is held each year, where the systems are evaluated according to various criteria (WMT - Workshop on Statistical Machine Translation).

Both approaches have their advantages and disadvantages that are well-described in the literature. Statistical approach is language-independent, there is no need in linguist description, so it is cheap and quick to deploy in comparison with RBMT, where rules and electronic dictionaries can be constructed for years in order to reach a sufficient quality. On the other hand, statistical systems require parallel corpora, which are not so easy to obtain for under-resourced languages.

A big advantage of RBMT systems over SMT is that the first ones are more controllable and predictable, the errors produced by RBMT are easy to spot and to fix - only additional rules are needed. SMT functions like a black box, though some issues can be predicted and some errors can be fixed, it is generally not known what output it will produce.

As for linguistic issues, different systems have their own weak and strong points. When speaking about syntax, rule-based systems containing proper syntactic rules generate sentences with better sentence structure, than the statistical. The problem of Word Sense Disambiguation is resolved much better within statistical systems than within rule-based, as WSD is one of the most challenging problems for systems of this type. More detailed description of advantages and disadvantages of the approaches related to the pair Czech-Russian can be found further in the text.

One of the main points that we address in our research is the question which of the approaches - Statistical or Rule-Based is more suitable for our language pair - Czech and Russian. Secondly, we want to find out how phenomena similar in the languages are exploited in distinct MT systems and how the latter cope with discrepancies between Czech and Russian.

2.3 Rule-Based MT systems

As it was mentioned above, the first systems that appeared were **Rule-Based**, they exploited bilingual dictionaries and manually written rules of transfer. They were labor-intensive, involving many linguistic annotation, and sometimes it took years to build a system.

Researchers define three main architectures of RBMT: direct, transfer and interlingua[5] . They are generally pictured in the Machine Translation triangle, or Vauquois triangle, see Picture 6.1.2:

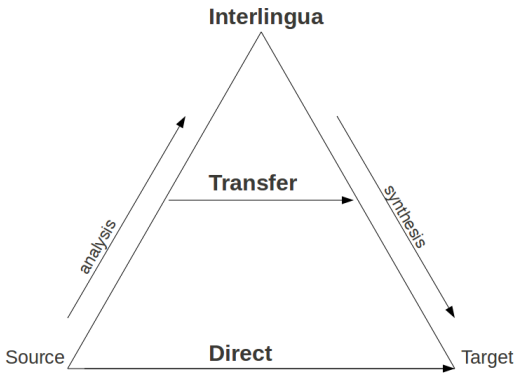


Figure 2.1: Machine Translation Triangle

It should be noted that this distinction, though started initially for the RBMT, is also applicable to SMT as it indicates the level of linguistic annotation integrated into a system.

2.3.1 Direct systems

Direct systems provide a word by word translation from a source to target language. The "pure" RBMT direct systems were used in the early days of the research on MT (1950 – 1960's) and were rather primitive in comparison with the modern ones. In the end of 90's it was believed that for the related languages this architecture might be the best option, as it avoids mistakes originating from the analysis and synthesis modules. The method of ''pure" direct RBMT translationis not really used nowadays[6], as even for very related languages some linguistic

[5] In the scientific literature the notions of **MT architectures** and **MT approaches** are sometimes confused, for instance, some researchers may refer to Direct, Transfer and Interlingua MTs as approaches

[6] However, statistical Phrase-Based Machine Translation systems can be considered as 'direct'.

analysis should be introduced. Almost direct architecture was used in the translation between Czech and Slovak languages Česílko (Hajič et al., 2000a) exploiting only a morphological dictionary.

2.3.2 Transfer translation

Transfer systems rely on the collection of rules aiming to cover morphological, syntactic or semantic incorrespondences between the languages. When the system is declared to be a rule-based, it is mostly probably a transfer system, ex. the already mentioned Systran, METEO, Ruslan, Česílko(Hajič et al., 2000b) for more distant languages, Apertium(Forcada et al., 2011) and many others developed all over the world for various languages. The process of translation generally consists of three phases: text **analysis**, **transfer** and text generation(**synthesis**). The analysis can be made up to different language levels – morphological, shallow syntactic, deep syntactic or shallow semantic. The borders between the levels are often quite deem and depend on the formalism under which the system is developed.

One of the types of the RBMT that we will use in our work is Dependency-Based Machine Translation system. It was first developed for the Czech-English pair (Čmejrek et al., 2003). Functional Generative Description (FGD) theory (Sgall et al., 1986) served as a theoretical platform for this MT. It exploited analytical and tectogrammatical parsers for the analysis of Czech; the transfer was made on the tectogrammatical layer using a bilingual dictionary and a parallel dependency Czech-English treebank; synthesis of a target English text from the tectogrammatical representation was provided by a number of rules. The system has changed a lot since then, so we use a more newer version of this project – **TectoMT** (Popel, 2010) under a platform called Treex, to be described in Section 4.3.

A similar research was conducted in Russia: the Machine Translation system **ETAP** (Boguslavsky, 1995) supports several language pairs with a focus on Russian-English, it is based on another dependency formalism ''Meaning-Text Theory".

As for Česílko, the system adopted transfer architecture when less related language pairs – Czech-Polish, Czech-Lithuanian and Czech-Russian were added. Czech-Russian pair within Česílko will be discussed in the Section 4.2.

A very popular RBMT platform involving many language pairs – Apertium – is based on a shallow syntactic analysis. It exploits the same idea as Česílko, which states that the simple architecture is more suitable for the related languages. Apertium supports MT between unrelated languages as well.

2.3.3 Interlingua

The Machine Translation systems that account for a level of deep semantics are called Interlingua systems. In Interlingua systems no transfer is needed, only synthesis and analysis is made. The core of the Interlingua MT is a Universal language that encodes all possible meanings – semantic primitives – for every natural language. The most famous project on the Interlingua nowadays is UNL – Universal Networking Language[7]. It is based on one formal representation, and each language within it has a stage of analysis from a plain text to the Universal representation and the synthesis from the UNL . The Interlingua architecture is quite complex to build, as it is very laboursome to specify the required language information – all possible universal semantic primitives.

True Interlingua is still considered to be rather a dream than a reality, and the majority of existing RBMTs exploits a transfer architecture.

2.4 Statistical Machine Translation

Statistical Machine Translation nowadays has become one of the easiest to deploy paradigms of the MT systems. Researchers can now use various toolkits to experiment with different language pairs provided the appropriate data exist. It was the IBM research team that pioneered the SMT field introducing the famous IBM models in early 1990's (Brown et al., 1990) and the first SMT Candide system (Berger et al., 1994). The central idea of Statistical MT can be roughly described as follows: we introduce hypothetical translations[8] – $\mathbf{e}$ – of linguistic units (these can be words, phrases, sentences), and define the probability

$$p(e|f) \tag{2.1}$$

– the probability that the unit $\mathbf{e}$ is a good translation of $\mathbf{f}$. Then the best translation $\hat{e}$ is a hypothesis that receives a maximum probability:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f) \tag{2.2}$$

The latter formula presented the ideal conditions of the hypothesis, now comes an approximation on data. The score of a hypothesis $\hat{e}$ is calculated from the two

[7] http://www.unlweb.net/

[8] As the first Statistical MT system was constructed for French to English pair, the source is traditionally denoted as $\mathbf{f}$ and target $\mathbf{e}$

components: the Language model(LM) and the Translation model(TM), which are introduced by the transformation according to the Bayes theorem[9] into:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \frac{p(e) \cdot p(f|e)}{p(f)} = \underset{e}{\operatorname{argmax}}\, p(e) \cdot p(f|e) \tag{2.4}$$

The above formulation presents a combination of a language model and a translation model, it is used in Phrase-Based models as well, more in the Chapter 4.4.

The **language model** $p(e)$ gives us a probability of how likely is some unit e present in an English text. In order to estimate parameters of the **translation model** $p(e|f)$, alignment models from a parallel text are extracted. Given the two models, we can find the best-scoring hypothesis.

The first IBM model was based on a simple word alignment, modern statistical models used more sophisticated techniques that brought better translation results. With the development of Internet technology, the amount of data – monolingual and parallel – has grown rapidly, as well as the performance of SMT.

Currently, the most widely-used statistical models are Phrase-Based Translation models[10]. Nowadays, anyone can implement an MT system for any pair of languages within, for example, Moses SMT toolkit (Koehn et al., 2007) provided that one has parallel data. Our implementation of Moses for the Czech-Russian language pair will be presented in the Section 4.4.

The results of SMT are fascinating in a sense that we can built many MT systems for different languages in a short time and at a low price. The quality of the output highly depends on the amount of parallel data and the match between the training and test data. For some under-resourced languages high translation quality cannot be expected. Let us take the example of Russian and Czech. Czech – as an official language of the European Union – can profit from multilingual texts produced by the European Parliament, this presents roughly 50 million words for each EU language. So it is relatively easy to produce experiments for the Czech and any other language pair within the EU. That is not true for Russian, which has a substantial amount of resources primarily for the English-Russian pair. As for Czech and Russian, some parallel resources can be found online on news sites and quite a few direct translations of fiction in a machine-readable

[9] $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

This equation follows the noisy-channel model, used also in speech recognition, spell checking and other NLP tasks.

[10] In this work, we experiment with Phrase-Based models only, and further in the text, we will refer to them as statistical.

form exist for the languages. Later in the text, we will give more information on data we used in our experiments.

2.5 Hybrid MT systems

Each of the approaches – statistical and rule-based – has its drawbacks that we have mentioned above. In the present more experiments are focused on combination of those two approaches exploiting the advantages of each. Hybrid MT is a rather "fuzzy" and broad term. We can distinguish two main types: originally Rule-Based with some Statistical build-ups (like disambiguation module), or originally a Statistical platform with some rules. The borders are not quite defined, so many existing systems can be considered to be "slightly" Hybrid.

Let us take an example of a couple of systems developed at ÚFAL. Česílko and TectoMT are considered to be Rule-Based systems, though they have some statistical components as parsers, morphological analyzers or WSD modules. Factored model within the system Moses can exploit morphological data, which can be also considered as some brute force learning of language rules. The two examples above are not actually Hybrid. Hybrid MT presupposes that there is some really huge effort to tune the system into more "rule-based or "statistical" direction, to prune the two architectures so as to harvest the advantages of both. Several publications on the Hybrid MT had been presented within the "Workshop on Hybrid Approaches to Translation (HyTra)"[11]. Commercial systems also showed rather promising results in Hybrid MT (Systran [12], AppTek [13] etc.)

2.6 Pivoting in MT

Pivoting is a popular technique in MT that exploits the idea of an intermediate language, and it is used both in the RBMT and SMT. This idea was justified by the fact that some languages have very few language resources, and that the translation from English into "resource rich" Czech and then the direct translation from Czech into some other Slavic language could have brought better results.

Nowadays, when almost each language in the Slavic group has many resources - lexicons, morphological dictionaries, often even treebanks, the idea of a pivot as presented in Česílko, seems to be quite outdated. It might though bring some fruit

[11] http://hytra.barcelonamedia.org/hytra2013/

[12] http://www.systran.co.uk/systran/corporate-profile/translation-technology/systran-hybrid-technology

[13] http://www.speechtechmag.com/Articles/News/News-Feature/AppTek-Launches-Hybrid-Machine-Translation-Software-52871.aspx

for the case of the very under-resourced languages, like Upper or Lower Sorbian (Lusatian). The pivot approach is successfully exploited within the Statistical MT systems, for Slavic, see: (Hartley et al., 2007) and (Galuščáková – Bojar, 2012), where the pivot languages serves as the source of additional data (phrase tables).

We should also mention here that Google Translate can use to some extent pivot English when translated between Czech and Russian, see the next chapter (Examples 5.24 and 5.58).

2.7 MT Evaluation

The problem of evaluation goes hand in hand with the notion of Machine Translation. Developers generally set a baseline system, evaluate it and on the basis of this evaluation introduce respective improvements. Evaluation component is very crucial in this ''MT circle" as it provides the feedback to the developer in which direction the research should go.

2.7.1 Manual evaluation

One of the earlier techniques to estimate translation quality was the **edit distance** technique. A translated sentence(target sentence) was compared against its gold standard translation(reference) in terms of how much insertion/deletions need to be introduced into it to make it fluent and adequate. This reference translation is created manually from a source as close to the target sentence as possible. Manual evaluation tells us a lot about the system, but it takes a lot of time and human resources to construct an ideal reference set.

Another type of human evaluation – **fluency/adequacy test** – is simply to say if the sentence is fluent (forms a correct sentence) or adequate (reflects the sense of the source text). This predicts which system is better, but this evaluation type does not answer the question which steps should we take to improve the performance. Moreover, human evaluators sometimes mix the concepts of fluency and adequacy. Upon the experience of WMT evaluation campaign[14], a more appropriate technique - ranking systems on a scale is now generally used. During the evaluation, the annotators are asked to rank several translations from best to worst.

Another technique which considers types of errors from a linguistic point of view – **error flagging** – will be discussed in the next chapter.

[14] http://www.statmt.org/wmt15/ and earlier

2.7.2 Automatic metrics

Automatic metrics are cheap and fast, they are used mostly by researchers to monitor the progress of system development. They suffer from several drawbacks that we show later. The automatic evaluation techniques generally exploit reference translations which are human translations of the test set. Generally, they come from a parallel corpus, not intersecting with development and training data. Those translations are produced regardless the fact that they should be used for the Machine Translation, so the source can be translated by human and by MT in different ways. The information loss during the translation process also plays its role. Above this, the evaluation between Czech and Russian demonstrates the same problem as the training set: those translations are generally not the direct translations, they present translations from English into Russian and then from English into Czech. So Czech and Russian sides of a test set contain significant structural and semantic discrepancies quite often.
When SMT became popular, some new evaluation metrics suitable especially for this type of MT had been created - BLEU(Papineni et al., 2002), TER(Snover et al., 2006), WER(Tillmann et al., 1997)[15], NIST(Doddington, 2002), Meteor(Lavie – Denkowski, 2009). These techniques are automatic based on a reference corpus. In this work, we will not use most of the metrics though, it might be an interesting idea. We will stick to manual linguistic evaluation and BLEU.

BLEU

In our work, we will provide the most widely-used **BLEU** score, which is generally used to track the progress while developing the MT. BLEU is calculated based on the number of correspondences between translated and reference n-grams according to the formula:

$$\mathrm{BLEU} = \mathrm{BP} \times \{\prod_{i=1}^{4} P(i)\}^{1/4} \tag{2.5}$$

where **i** is a length of an n-gram hypothesis in words and P(i) is the percentage of n-grams that are present both in the hypothesis and in the reference. This value is generally presented as a number between 0 and 1, but it can be also indicated in percentage (range from 0 to 100). BP - brevity penalty - is applied when a hypothesis is shorter than a reference.

[15] TER and WER resembles edit distance metric as they also measure the discrepancies between a hypothesis and a reference in terms of Levenshtein distance, but references come generally from a parallel corpus

For morphologically rich languages with free word order the automatic evaluation method BLEU can not be trustworthy. We will show the BLEU scores in the next chapter for several MT systems, but they actually say very little to us about the translation quality. We can only say that the margin of the SMT(16%) over RBMT(4-6 %) systems in our case is huge and it really indicates that the quality of the Cesilko and TectoMT systems is poor. That is why BLEU score will not be the main criteria of quality in this work.

(Koehn, 2011) has outlined major drawbacks of BLEU: this metrics does not say anything about MT really, it underscores Rule-Based systems and is most unsuitable when translating into morphologically rich languages as it counts the precision of exact n-grams. Also, it is not suitable to evaluate minor improvements of concrete language phenomena as the difference in terms of BLEU will be really insignificant(see Section 6.2.3). Still, as it is the mostly used metric now, we will mention this score when describing the concrete MT systems, but our main evaluation technique[16] will be of a manual nature.

In this chapter we have briefly presented main types of MT and outlined several problematic issues. In the following chapter we will describe data and tools used to create the MT systems for the pair Czech and Russian: dictionaries, parallel corpora, treebanks, morphological taggers.

[16] Described in detail in Section 5.1

3

Data and Tools for MT

In this chapter we will describe the data and tools both external or the ones we have created. We exploited them not only in the MT experiments, but also in a theoretical part where we conducted a contrastive linguistic analysis between Czech and Russian.

3.1 Parallel and monolingual data

Here we overview the corpora used as training data for SMT and for some other experimental comparative studies. We describe in more detail a process of compilation of a parallel Czech-Russian corpus UMC.

3.1.1 UMC

For the needs of our experiments we created the UMC 0.1(UFAL Multilingual Corpus) - a multilingual parallel corpus of texts in Czech, Russian and English languages with automatic pairwise sentence alignments (Klyueva – Bojar, 2008)[1]. UMC is closely related to CzEng[2] , a Czech-English corpus which has been successfully employed in SMT experiments. The primary goal of compiling UMC was statistical machine translation, but it served as data source for the dictionary extraction and some other experiments with valency (see Section 6.3).

We have chosen only one web source(see below) to download our texts and up to now we were able to obtain over 1.7 million words in each of the three languages[3]. We included also the English part of parallel texts into the corpus on purpose, as this served as a platform to compare how the SMT works for the related languages compared to the unrelated.

[1] The scripts to automatically download the corpus were written by Ondřej Bojar

[2] http://ufal.mff.cuni.cz/czeng/

[3] Since the corpus was first compiled in 2008, we have not downloaded new data from the web. So if the update was done now, we would obtain much more data.

Collecting parallel texts meets such challenges as copyright, translation quality and representativeness of the language. The problem of copyright is solved by contacting the page editor asking for a license agreement for educational purposes. It is more complicated with a translation quality, because when downloading automatically huge amount of texts, they can not all be checked, so we look only at the extralinguistic factors. Let us inspect the texts in both Czech and Russian that we can come across in the Internet.

Many of them probably belong to the tourism industry as many hotels, restaurants, tourist sites are advertising their services both in Czech and Russian. The texts are generally short and the translation quality is doubtful.

Technical texts present the second, more reliable and broad group, but their representativeness is low, as they contain lots of technical terminology and general language usage is limited. On the other hand, those types of text are most suitable as a limited domain for MT, as the language is formal and the metaphorical use of language is rare. In most cases the original language is English, and the texts are translations from English into Czech and from English into Russian.

Text of another genre - news and commentaries - are written in a language rich with metaphors sometimes with tricky constructions, which can be translated differently in different languages. However, the language of news covers the most essential part of standard language usage, so we have chosen to use the news articles in the first phase of the experiment.

As it was mentioned, all the texts were downloaded from a single source - The Project Syndicate[4], which contains a huge collection of high-quality news articles and commentaries. We were given the permission to use the texts for research and non-commercial purposes. Texts were downloaded with the help of tools developed under the project CzEng. The total amount of downloaded documents is 2,186 in each of the three languages. The table 3.1 summarizes the statistics of the corpus.

	Czech	Russian	English
Words	1,747,997	1,815,550	1,920,164
Tokens	2,002,990	2,152,326	2,255,901
Sentences	96,335	101,528	97,250

Table 3.1: Summary of corpus size.

[4] http://www.project-syndicate.org/

Corpus processing

Following steps were applied to the downloaded data:

- **Transforming formats.** HTML files are transformed into text documents by extracting text paragraphs from the web pages. The original pages do not include pictures, tables or mathematical formulas, so the process is rather straightforward. Unlike the project CzEng, where the preference was given to the XML storage format, in UMC we use plain text format as this will be enough for our purposes of training models.
- **Segmentation and Tokenization.** In order to make a segmentation and tokenization we used a trainable tokenizer described in (Klyueva – Bojar, 2008). 160 automatically segmented and tokenized sentences were manually annotated with respect to the correctness of segmentation and tokenization. The tokenizer was retrained on this data.
- **Sentence alignment.** In CzEng and in UMC, the texts are aligned only on a sentence level using the hunalign tool[5] . We did not use any additional dictionary, the dictionary was learned automatically by the tool.

3.1.2 Subtitles

The subtitles data were downloaded from the web by Karel Bílek(Bílek, 2014). Texts coming from subtitles are considered to be very unreliable as training data for SMT and for the comparative linguistic purposes as well. The chunks are generally not aligned to each other very well, often they do not form a complete sentence, and they are translated from English into some other languages, and not directly (from Czech into Russian). The main advantage of those data is that they can be obtained easily, and they are quite big.

3.1.3 Intercorp

Intercorp is a collection of parallel corpora in various languages created at the Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University in Prague mainly for linguists to search for specific language phenomena[6]. Though the data are not open-source, but the creators of this resource[7]

[5] http://mokk.bme.hu/resources/hunalign

[6] After our experiments were finished, the data from Project Syndicate and the Subtitles were included into the Intercorp as well.

[7] Alexander Rosen on behalf of ÚTKL(utkl.ff.cuni.cz)

were so kind to provide a Czech-Russian corpus with sentences shuffled in a random order for the purposes of this experiment. The collection contains fiction in Czech and in Russian. The advantage of the data is that they mostly present the direct translation from Russian into Czech or the other way round and that the sentence alignment was checked by linguists making this corpus a very reliable resource.

3.1.4 Parallel data summary

The table 3.2 summarizes the size (number of sentences) of the three corpora:

corpus	sentences	Words		Tokens	
		Czech	Russian	Czech	Russian
UMC	93,395	1,762,325	1,773,616	2,019,683	2,073,102
Subtitles	2,324,373	12,035,512	11,927,075	15,631,855	16,019,077
Intercorp	148,847	1,595,524	1,509,817	2,030,920	1,956,916
Total	2,584,300	15,393,361	15,210,508	19,682,458	20,049,095

Table 3.2: Statistics of Czech-Russian parallel corpora

These corpora will be involved in our experiment with statistical machine translation and in the other experiments concerning corpus-based comparative studies.

3.1.5 Monolingual data

As we experimented with translation from Czech into Russian, we also needed to create a large monolingual Russian corpus to train a language model(LM). The Russian part of the parallel corpora was included into the LM data alongside with other resources in Russian:

- russian side of the parallel corpora described above
- NewsCrawl[8]
- Russian side of a parallel English-Russian corpus from Yandex[9]

[8] NewsCrawl and CommonCrawl are data that were gathered during WMT competitions and available on the web http://www.statmt.org/wmt14/

[9] https://translate.yandex.ru/corpus?lang=en

- CommonCrawl[10]

Totally, those data comprises around 11,665,247 lines of texts. (The texts are not segmented or tokenized). The data comming from CommonCrawl and NewsCrawl are not very reliable as they can contain chunks of text in a foreign language and automatically translated texts, see (Bílek, 2014) for details.

3.2 Czech-Russian dictionary

In the experiment with Machine Translation between Czech and Russian we used a dictionary automatically extracted from a parallel corpus as there is no Czech-Russian dictionary in a plain-text format available online[11]. We will now shortly describe the process of the dictionary extraction.

A very similar work on extracting dictionary entries was done for Chinese and English (Baobao et al., 2002), for Czech and English (Bojar – Prokopová, 2006), for English and Romanian (Tufis, 2002). The tool we used for word alignment process is GIZA++ (Och – Ney, 2003), which is broadly exploited by many researches while generating bilingual dictionaries. Our Rule-Based MT system was supposed to have a morphological analyzer and generator, so the dictionary should include lemmas instead of wordforms. For this purpose we used the taggers: Hajic's tagger for Czech - the same as for the analysis of Czech text and TreeTagger [12] for Russian.

Next, we describe how the dictionary was created. We took the Czech-Russian part of UMC parallel corpus(Section 3.1.1), with only 1-to-1 aligned sentences in order to prevent noise during the alignment process that can be introduced by many-to-many sentence alignment. Then, GIZA++ was run on the lemmatized data. It provided over 406 973 candidate translation word pairs, and we have sorted them according to the frequency of occurrences. Next is the example illustrating how such automatically extracted dictionary looks like, the list is taken from the top of the dictionary, so those are most frequent word pairs. The first comes the number of occurrences of an alignment pair, the second is a Czech word, the third is a Russian one:

```
37,188 a-1      и
25,490 v-1      в
12,269 že       что
```

[10] http://commoncrawl.org/.

[11] We have exploited some commercial dictionaries, ex. PC Translator in order to estimate and improve the quality of our dictionary, but we could not used it directly in our translation

[12] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

```
8,834 být     быть
8,303 na-1    на
5,345 tento   этот
```

The majority of generated translation pairs are obviously wrong. For example there are 509 different translation pairs for the Czech word *aby* – 'to, in order to'. Only the first two with the highest pair frequency are correct:

```
2266 aby     чтобы
517 aby     для
```

and the others, for example those, are wrong:

```
5 aby     восстановление
5 aby     восстанавливать
5 aby     стараться
5 aby     способствовать
```

In order to select the pairs that are mostly probably the correct ones, we used information on frequency of pair occurrence from UMC and searched in the Czech-Russian dictionary of PC Translator[13], released in 2007.

This dictionary was exploited while constructing Rule-Based MT systems. As Česílko or TectoMT(MT systems to described later, Sections 4.2 and 4.3) were not supposed to have any word sense disambiguation module, we took those top frequent translation pair confirmed by the dictionary. After applying these two steps, the dictionary size was reduced by over 91%, so that the cleaned dictionary contained 19,861 translation pairs with the most frequent translation equivalent used.

3.3 Tools for Morphosyntactic Analysis

A morphological tagger analyzes a sentence and assigns lemmas, morphological tags and sometimes syntactic information to each word. In our research we used taggers for Czech and Russian both for RBMT and SMT. The output of the taggers is written in a **form**|**lemma**|**tag** pattern.

3.3.1 Tagger for Czech

Morphological tagging has a long tradition at the Charles University. Firstly, a unique system of a Czech positional tag has been developed.Secondly, due to the

[13] http://www.langsoft.cz/translator.htm

existence of morphologically annotated corpora, the researchers has trained quite a few taggers for Czech, e.g. (Hajič, 2001), (Raab, 2007), (Spoustová et al., 2009) and recently Featurama[14] or MorphoDiTa[15].

In our work, we used the state-of-art tagger MorphoDiTa incorporated into a Czech analysis pipeline. Following is an example of an annotated chunk of text(in English, "Culture of OSN is":

```
Kultura|kultura|NNFS1-----A----
OSN|OSN-1_:B_;K_^(Organizace_spojených_národů)|NNFXX-----A---8
je|být|VB-S---3P-AA--- následující|následující_^(*5ovat)|AGNS1-----A----
```

For example, a tag 'NNFS1-----A-----' denotes a part of speech (Noun), the second position is a specification of a part of speech(N for general noun), the third is a position for gender (Feminine), the fourth is a number(Singular) and the fifth is a case(first case -Nominative), the eleventh is a feature of negation(non-negated in this case), the rest positions are not defined for a noun.

3.3.2 A tagger for Russian

For Russian language, the only available open-source tagger TreeTagger[16] was used. Following is the example of a tagged phrase(In this relies culture):

```
В|в|Sp-l этом|это|P--nsln и|и|C заключается|заключаться|Vmip3s-m-e
культура|культура|Ncfsnn ООН|ООН|Ncfsgn
```

The Russian tag is positional with respect to the part of speech and it does not have a fixed number of positions. Unlike the fully positional Czech tag where each position always stands for a distinct category, positions in TreeTagger's tag can be filled by different features for different part-of-speech categories. E.g., the tag 'Ncfsnn' codes the following features: noun, feminine gender, singular number, nominative case and non-animated, which is more or less the same information as for a Czech word.

3.3.3 SynTagRus

In our work, we will also use another resource of morphological information coming from a Russian Dependency Treebank SynTagRus(Boguslavsky et al., 2000)[17]

[14] http://sourceforge.net/projects/featurama/

[15] http://ufal.mff.cuni.cz/morphodita/

[16] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[17] SynTagRus is not an open-source, but one of the corpus creators - Leonid Iomdin - was kind to provide us some annotated data for our experiments.

This is a valuable resource, as the annotated data are manually hand-checked, so the tagging information is highly reliable. The tag is also semi-positional - a part-of-speech category denotes the sequence of morphological features. For example, a word *шофера* – 'driver.Gen' has a tag "S ЕД МУЖ РОД ОД" (noun, singular, masculine, genitive case and animated). Information on form, lemma and tag from SynTagRus were used in the text generation process in Czech-Russian implementation of TectoMT.

4

MT Systems between Czech and Russian

In this chapter, we present MT systems developed or adjusted at our department - Ruslan, Česílko, Treex, then we will introduce experiments with SMT system between Czech and Russian built with the Moses toolkit. Next, we will take a look at industrial systems Google and PC Translator. In this thesis we will consider only the direction from Czech into Russian, as it was initially made for the systems we can not change - Ruslan and Česílko. Even though those systems were not included into the linguistic evaluation, the translation direction was not changed.

4.1 Ruslan

Experiments in MT between Czech and Russian started in the mid 1980's. The system Ruslan (Hajic, 1987), (Oliva, 1989) of automatic translation between the languages was of a high interest because of extensive cooperation between Russia and the Czech Republic (Soviet Union and Czechoslovakia that time). It was created especially for the domain of manuals for mainframe computers originally written in Czech that were to be translated into Russian.

Implementation of Ruslan required huge amount of manual linguistic annotation. The core of the system was a dictionary, its entries were enhanced with morphosyntactic and semantic information. The system - like a typical RBMT - consisted of several translation steps processing the input in Czech and generating the Russian text. These were Czech morphosyntactic analysis, shallow semantic parser, transfer module, text generating modules to name some of them. While in the beginning of the project it was believed that a transfer module was not needed due to the relatedness of the languages, it became evident that there is a large number of differences to cover, so modules grew larger and the dictionary entries changed as well.

In the 1988 the dictionary contained about 10,000 entries, and the system translated mainframe manuals at a sufficient quality that was worth post-editing.

However, due to the political changes after 1989, there was no need for such MT between Czech and Russian anymore and the project terminated.

Since then, the resources created under the project served for other stand-alone experiments, for example, in (Bojar et al., 2005) authors re-use the module of syntactic analysis of Czech for the Czech-English Machine Translation, the paper (Klyueva – Kuboň, 2010) describes the extraction of morphosyntactic information from the Ruslan dictionary for Czech and Russian valency dictionaries; (Bílek et al., 2013) show the experiments with automatic extraction of semantic features based on those from Ruslan.

Despite the high quality of human annotation of the words in the dictionary, Ruslan also has one drawback - a relatively limited domain of mainframes manual. The computer terminology has changed during the past 25 years, so some of the words contained in the dictionary are slightly outdated. Moreover, while processing the dictionary we have spotted some mistakes both in translation of words and in linguistic information.

We will not go into a detailed description of all the modules and just show the example of two Ruslan dictionary entries which can give some idea of how the system worked:

```
LE2KAR3==MZ(@(*H),!,MA0111,VRAC2).
```

- `LE2KAR3` represents the Czech lemma *lékař* - a doctor; the diacritics is encoded in a rather primitive way corresponding to the time when it was created and when encoding of national characters still constituted a challenge.
- `MZ` represents a declension pattern **muž** (it also determines the part of speech information because this particular declension pattern is used for masculine animate nouns in Czech).
- `@(*H)` represents the semantic category 'human'.
- `MA0111,VRAC2` represents the declension class of the Russian equivalent and the equivalent itself, encoded into basic ASCII

```
VLASTNI==R(5,PRM,?(N(N),A(I)),04,OBLADAT6).
```

- `VLASTNI` represents the Czech stem from the verb *vlastnit* – 'to possess';
- `==R` represents a verb, `5` is a conjugation pattern, `PRM` are morphological characteristics

- `(N(N),A(I))` represents a surface valency pattern: N(N) - an actor(agent) is expressed by a surface Nominative case in Czech and in Russian(in brackets), A(I) - the patient argument is coded with the (A)ccusative case in Czech and with the (I)nstrumental in Russian
- `04,OBLADAT6` represents the respective Russian verb and the pattern according to which the proper morphological form is generated

The major problem in reusing the system is that of a dictionary: a single unknown word can crash the process of analysis, so the two distinct paths (trees) - before and after the unknown word - are generated. The out-of-vocabulary words are rather hard to built in. They have to be transfered into a Ruslan format so that the analysis and synthesis processes can be carried out.

Unfortunately, it was impossible to translate a standard WMT test set that we used for other systems, so Ruslan translation output will not be included into the linguistic analysis.

4.2 Česílko

4.2.1 System description

Česílko(Hajič et al., 2000a) is a Rule-Based MT system for closely related languages. The underlying idea of this project was to exploit the relatedness of languages in the MT. It was believed that for close languages there was no need to build huge linguistic resources and high number of rules. This idea worked pretty nice for the very related Czech and Slovak, but for more distant languages more deep sophisticated linguistic analysis was needed.

In further experiments (Hajič et al., 2003) with Czech-Polish and Czech-Lithuanian pairs a new module of shallow syntactic analysis was introduced. It covered such discrepancies between Czech and other languages, as, e.g. adjective postpositions in noun phrases in Polish or inflective past tense formation in Lithuanian. When a Czech-Russian pair was introduced into Česílko, the shallow syntactic rules were written to covered the most frequent syntactic discrepancies between Czech and Russian. The process of translation consisted of the following modules:

- morphological tagging and lemmatization of Czech
- partial parsing of Czech
- lexical and structural transfer, syntactic synthesis
- morphological synthesis to Russian

4.2.2 Rules of analysis and synthesis

The tagger (Hajič, 2001) provides the morphological information - a lemma and a tag, the partial parser (Homola, 2009) lowers the morphological ambiguity and ensures some of the sentences or sentence structures (like noun phrases) are passed to the transfer module in an appropriate form. The transfer module was joined with the module of syntactic synthesis of Russian. Following are some transfer rules which were implemented in the experiment:

- a rule for a copula to be, which is omitted in Russian and frequently used in Czech.
- usage of reflexives (part of a word in Russian and separate in Czech, though is considered to be a part of a lemma after a morphological analysis)
- negation prefix 'ne', which is a part of a wordform in Czech and is written separately in Russian
- some cases of preposition incorrespondences

The following code illustrates transfer rule for the copula verb 'být' - to be:

```
if ([lemma isEqual: @"быть"]) { // adapting aux 'být'
  if (![dict objectForKey: @"subj"]) {
NSString* person = [dict objectForKey: @"person"];
NSString* number = [dict objectForKey: @"number"];
NSString* lemma;
if ([number isEqual: @"sg"]) {
if ([person isEqual: @"1"]) lemma = @"я";
if ([person isEqual: @"2"]) lemma = @"ты";
if ([person isEqual: @"3"]) lemma = @"он";
} else {
if ([person isEqual: @"1"]) lemma = @"мы";
if ([person isEqual: @"2"]) lemma = @"вы";
if ([person isEqual: @"3"]) lemma = @"они";
}
  }
}
```

4.2.3 Evaluation

The overall quality of the system was evaluated in terms of BLEU score, and it was far from ideal. (Homola, 2009) tested the system on a test set that contained 1000 sentences and BLEU reached only 5%.

Unfortunately, due to some technical reasons, it became impossible to re-use this system on other data or to introduce some other improvements, as the Czech-Russian adopted version of Česílko can not be compiled on more modern systems and, whatsmore, an original morphological module was missing. So this system will not be taken into account in the linguistic analysis as well as Ruslan.

4.3 TectoMT

4.3.1 System description

TectoMT system between Czech and Russian was implemented within the project **Treex** (Popel – Žabokrtský, 2009). Treex is a modular system of NLP tools, such as tokenizers, taggers, parsers and various data - corpora, treebanks and lexicons. The system is multilingual, so above "home-created" tools and data the authors collected some tools from the outside and adjusted them to the system environment.

One of the main projects under Treex is English-Czech Machine Translation system(Popel, 2010). As modules of the system are easily reusable for other languages, the idea emerged to build an experimental system of machine translation from Czech into Russian investing as less effort as possible. Provided that the analysis of Czech already existed, it took only one day for two persons (M.Popel and Z.Žabokrtsky) to adjust TectoMT for the Czech-Russian pair. The process of gathering data (dictionary, morphological data) was not that quick, though. After the baseline was set, there were several improvements introduced by Martin Popel, Karel Bílek and Natalia Klyueva.

Next, we provide a brief description of a system. Each experiment is presented as a scenario consisting of a sequence of blocks, each of which performs some NLP subtask. Each block belongs to one of the four language layers: word (w-layer), morphological (m-layer), analytical (a-layer) and tectogrammatical (t-layer) layers. More precisely, the blocks ensure the transformation between those layers. Let us now briefly describe those layers in detail. This division has its roots in the Functional Grammar Description theory - FGD (Eva et al., 1986), but its implementation in Treex is slightly different from original FGD concepts. The dependency treebank PDT(Hajič et al., 2006) is based on this theory, the

annotation in the PDT is done on the four mentioned language layers, but FGD distinguished more layers.

- **The Word and Morphological Layers.** The word layer is a simple sentence represented as a sequence of words. On the morphological layer, each word in a tree is represented by a node with a lemma and a tag assigned.
- **The Analytical Layer.** Syntactic annotation is presented in a form of a dependency tree, where each morphologically annotated token from the previous level becomes a node with an assigned analytical function. Analytical function (afun) reflects a syntactic relation between a parent and a child node and it is stored as an attribute of the child. Examples of analytical functions: Subject (Sub), Predicate (Pred), Object (Obj) etc.
- **The Tectogrammatical Layer**. The annotation on the tectogrammatical layer (t-layer) goes deeper towards the level of meaning. Function words (prepositions, auxiliary verbs, conjunctions, etc.) are removed from the corresponding analytical tree; they are stored as attributes of autosemantic words, leaving only content words as the nodes on the t-layer.

The picture 4.3.1 shows the sentence from the PDT on each of the three layers[1].

4.3.2 Translation scenario

Now we will describe the translation scenario of Czech-Russian MT itself, it is schematically depicted in the Picture 4.2. Following is the detailed description of what we do on each language layer.

M-layer

On the input we get a text in Czech and apply the following sequence of blocks which provide tokenization, lemmatization and tagging.

A-layer

Morphological layer presents only a flat structure of a sentence. When we make a step from morphological to the analytical layer, the sentence is parsed by the MST parser which generates an analytical (surface syntactic) tree. The nodes are

[1] Sentences presented in this sections are visualized in Tred editor of trees (Pajas – Fabian, 2011).

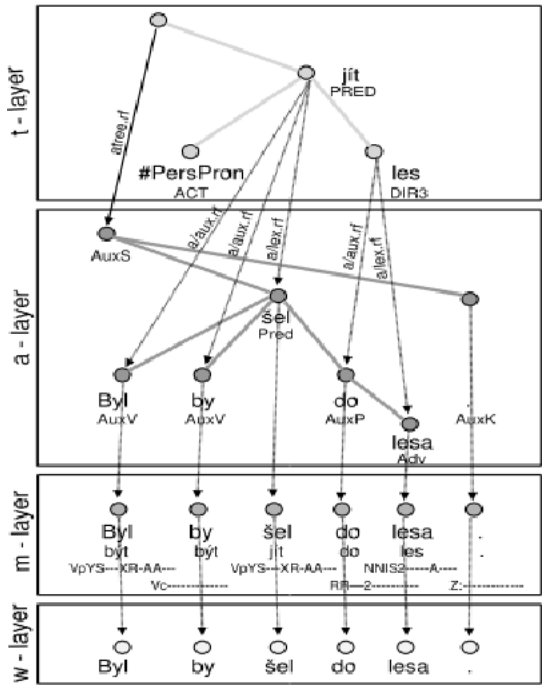


Figure 4.1: PDT sentence in Tred. In English: "He would have gone to the forest".

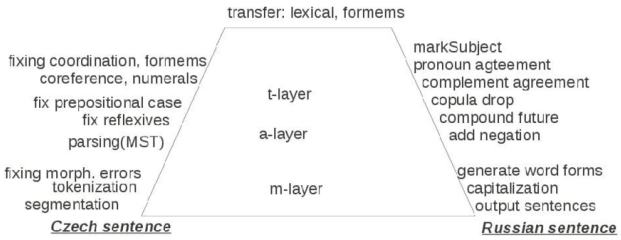


Figure 4.2: Some blocks of the TectoMT translation scenario

marked with analytical functions reflecting dependency relations between nodes of the tree. The blocks that process the sentence are the following:

Picture 4.3 illustrates a parsed Czech sentence ''How to sell negotiations on the global market", the first tree (a-tree) presents the analysis of the sentence up to the analytical layer.

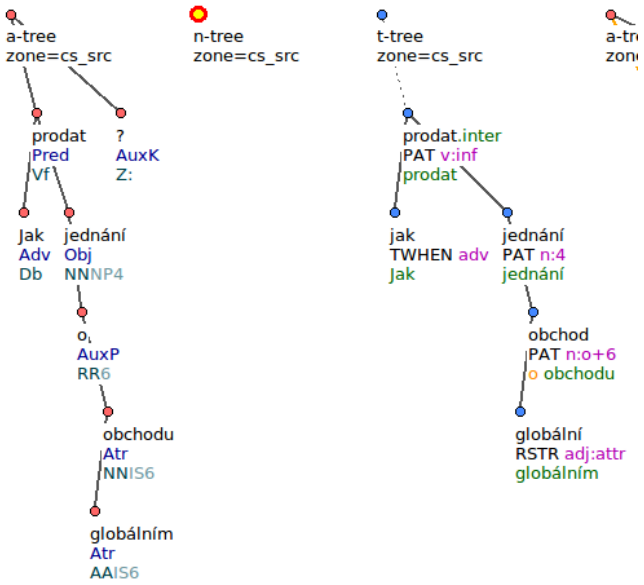


Figure 4.3: Analytical tree

The head of a sentence is a verb "prodat"(to sell) that has an analytical function Pred - Predicate. It has two dependent words - the node with afun Adverbial and the node with afun Object, the latter has child nodes as well.

T-layer

On the tectogrammatical layer some nodes - representing auxiliaries, prepositions, reflexives - are collapsed, fformemes and grammatemes are introduced. In particular, we should mention **formemes** (Dušek et al., 2012) - morphosyntactic properties of the node which were created especially for the MT purposes as a simplification of tectogrammatical attributes. They provide a quick and transparent connection between a tectogrammatical functor and a surface morphological form of a word. Here are few examples of functors:

(4.1) existovat [v:inf] - Infinitive form of a verb
rozdíly [n:1] - Noun in a first - Nominative case.

The second tree (t-tree) in the Picture 4.3 illustrates how analytical functions were converted to tectogrammatical functors; the node received a deeper semantic interpretation - e.g., from afun Object - to functor Patient for the word *jednání* – 'negotiations'. The preposition "o" (about) was transformed into an attribute of a governing word "jednání".

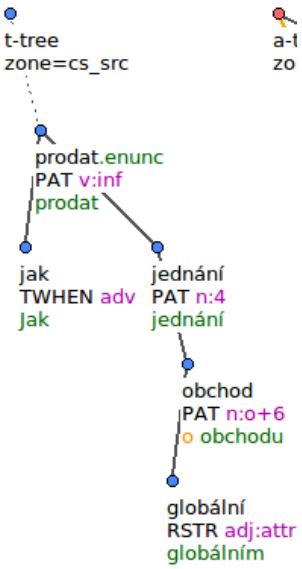


Figure 4.4: Tectogrammatical tree

Transfer and Generation

In the phase of transfer (T2T blocks) Czech lemmas in nodes of trees are substituted with their Russian equivalents. Again, we made use of an automatically generated dictionary that we exploited in Česílko(4.3), though we have fixed some entries with respect to the system architecture. The preliminary transfer of Czech formemes into Russian is made: it covers the most frequent formemes in Czech and Russian. The following example illustrates the transfer of formemes containing a preposition in Czech and in Russian:

(4.2) qw(n:o+6) => qw(n:o+6) - about+6
qw(n:v závislosti na+6) -> qw(n:в зависимости от+2) - 'depending on+X'

In the first case the same preposition is used in Czech and in Russian, and a noun after the preposition requires the Locative case in both languages. The

second example demonstrates a discrepancy: a noun after this multiword preposition requires different case morphemes in Czech and Russian. Valence formemes coincide in the two languages in most cases, the detailed research on differences in surface valency will be presented in the next two chapters.

Introducing a list of formemes resulted in a minor improvement of the translation quality, see Table 4.1.

As for the **T2A blocks**, they ensure proper generation of a Russian sentence, including blocks that fix Future tense, negation, Russian copula constructions and formemes in both languages. The list of discrepancies between Czech and Russian is rather big, even though the languages are related, and we were not capable to cover the majority of them, just the most evident ones.

Picture 4.5 illustrates the Russian tectogrammatical and analytical trees.

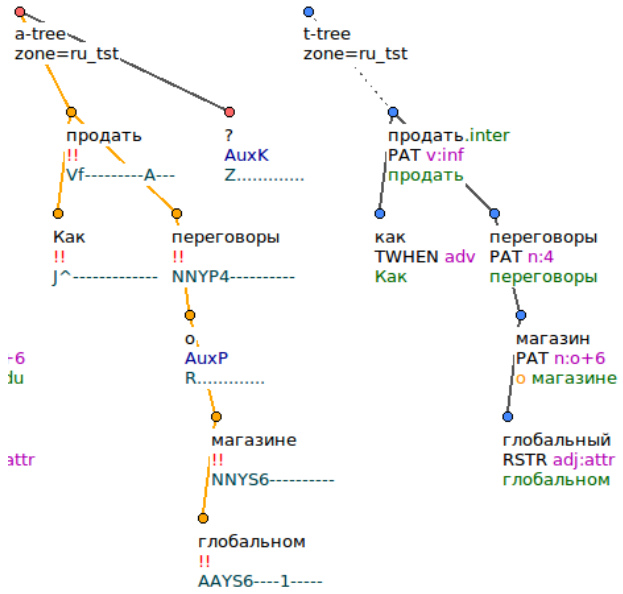


Figure 4.5: Russian tectogrammatical and analytical trees

The nodes mostly do not have any afuns (except for the ones that are relevant under some transfer rule), only the morphological tags of respective Czech equivalents. This means that almost no transfer is made on the tectogrammatical layer and very little on the analytical. Therefore, the mistakes originating from parser inaccuracy do not affect the translation process, because if we do not have a rule for some language issue, it is just processed word-by-word.

All in all, we can say that the translation is made word-by-word handling some discrepancies between the languages. So this experiment can be considered to be very close to Česílko. The advantage was that it could be adjusted, reused and tuned more easily than Česílko.

Another milestone for TectoMT (as well as Česílko) is that the disambiguation module could not be introduced like into English-Czech TectoMT[2] because we do not dispose a parser for Russian language. So, for instance, in the example tree, a word *obchod* – 'trade' was mistranslated into Russian with the word *магазин* – 'shop', more on lexical issues will be discussed in the next chapter in Sections 5.2.15 and 5.2.16.

After words are generated from Russian lemmas and respective morphological tags, the resulting wordforms are concatenated to form a sentence.

4.3.3 Evaluation and improvements

Initially, the system baseline was established with a minimum number of rules handling the most obvious differences between the languages, as, e.g., copula drop, negation particle handling.

The BLEU score of a baseline experiment was poor, almost as Česílko - 4.44%. The lexical transfer was made using the same automatically generated dictionary, and almost the similar lower scores are partially due to the dictionary quality. Some of the errors were introduced by a tagger and a module of wordforms generation multiplied by the incompatibility of the tools and data formats(tag format) for the two languages.

After adding some linguistic information in a form of blocks, the score increased, but not significantly. Some of the new blocks and changes are described in (Bílek, 2014), some of them can be related to change in the analysis of Czech.

Following are some points where the rules/changes were introduced:

- Fixing verb aspect. The wordforms extracted from a SynTagRus have only one aspectual type - imperfect, and the generation of a perfective counterpart is ensured by the tools we do not have at disposition. We bypassed the problem by adding infinitive forms of verbs, where an imperfective lemma was substituted with the wordform.

- The list of formemes with prepositional complements like in Example 4.1 was enlarged that lead to a sufficient improvement.

[2] For English-Czech TectoMT, a maximum entropy classifiers were trained to distinguish different senses of lemmas (Mareček et al., 2010).

- Surface valency frames from Ruslan were added as formemes. The experiment will be described in detail in Section 6.2
- Enlarging the dictionary. The experiment where the new entries were extracted from the parallel corpus was described in (Bílek, 2014).
- Some blocks to fix ceratin linguistic phenomena were added/improved: copula drop, modal verbs, fixing date construction in Russian.

In the Table 4.1 summarizes the performance change in terms of blue score as the specific rules/data were introduces.

experiment and improvements	BLEU score
baseline	4.44%
Fixing verb tenses and aspect	5.09%
Adding preposition formemes	6.62%
Bigger dictionary	7.04%
Fix in Czech analysis (punctuation)	9.04%
Fixes in some rules and valency	9.38%

Table 4.1: Baseline and improvements

Fixes in the Czech analysis were not made for some other language pair, but our system benefited from those improvements as well. The last line - 'Fixes in Rules and valency' concerns mainly the improvements of existing rules and introduction of a module handling verbal valency. Those issues will be discussed in more detail when we will be describing the concrete language phenomena, see Sections 5.2.10 and 6.2.3. The improvement of concrete issues was insignificant in terms of BLEU, but the analysis always showed some improvement in an issue that we aimed to. actually, rules are written to work.

The process of creating a baseline system was not that hard, but introducing improvements that capture problematic issues of differences is a long-lasting and laborious process. Detecting the errors which often reflect this or that discrepancy between the languages or some bug in a system architecture or data can last for several years[3] and it is not at all easier than for Czech and English language pair, no matter how related are the languages.

[3] And sometimes adding new rules resulted in introducing new errors.

4.4 Phrase-Based SMT and Moses

In this section we will look at experiments with Moses, an open-source implementation of phrase-based statistical translation system. Moses[4] is very much language independent since it uses purely data driven methods. The most important property of phrase-based systems is the ability to translate sequences of words (phrases or n-grams) rather than single words.
In the introductory section of this chapter we have briefly mentioned IBM models used for the first SMT systems. Now we will describe the basics of phrase-based models that are very popular nowadays. Their main advantage over the first IBM models is that they support many-to-many alignments, so that they can cover cases when more words in the source language correspond to several target ones.

The equation 2.4 adapted to the phrase-based models combines three components: the phrase translation probability, the language model and the distortion cost:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) * p_{LM}(e) * d(start_i, end_{i-1}) \tag{4.3}$$

- **Phrase Table** Phrase translation probability is stored in a so-called phrase table together with the phrases (n-grams, or combination of words of various length) in both languages. Those n-grams are extracted from the sentences. The translation probability then presents the relative frequency of an -n-gram:

$$score(f|e) = \frac{count(e, f)}{count(e, f_i)} \tag{4.4}$$

  Following is the example of a Czech-Russian phrase table entry[5] indicating an n-gram *to believe in*:

  (4.5) doufat v ||| надеяться на ||| 0.25 0.0128977 0.0588235 0.00630776

  The first and second scores after the n-grams present inverse phrase and lexical probability, the third and the fourth are direct translation and lexical probabilities respectively.

  Generally, two bidirectional tables - p(e|f) and p(f|e) - are considered to maximize the precision of the translation and also a word-based score ("lexical probability") is used to smooth on probability estimates for infrequent

[4] http://www.statmt.org/moses/

[5] We have hidden some probability parameters like word alignment numbers

phrase pairs.

- **Language model** indicates the probability of how good (fluent) a phrase is. It is estimated from the corpus using n-gram modeling which uses the probability of the previous (n-1)-word history to predict the next word n:

$$p(w_1, w_2, ..., w_n) = p(w_1 * p(w_2|w_1))... * p(w_n|w_1, w_2, ..., w_{n-1}) \quad (4.6)$$

  The most widely used model is a trigram language model estimates the probability of a phrase in the target language based on the history of previous two words. Following is an example of phrases from a generated language model with calculated probabilities:

(4.7) -4.584007 надеяться -0.6977018 (*believe*)
-2.196512 на -0.7243898 (*in*)
-0.4852926 надеяться на -0.2465586 (*believe in*)
-3.703978 надеяться на закон (*believe in law*)

- **Distortion parameter** The distortion parameter, or penalty, penalizes a phrase in which a reordering limit (which is set in particular) is exceeded. Reordering limit indicates how phrases neighboring in the source sentence stay far from each other in the target. If we set the reordering limit to 2, it will allow the phrases with 2 switches. This limit reduces the space of hypothesis and does not allow sentences where the phrases stays far from the respective source ones.

Decoding

The various combinations of phrases that constitutes a sentence are therefore being scored using the above formulas and the search algorithm constructs the output sentence (hypothesis). The best scoring hypothesis forms an n-best list, and the final translation is chosen from this list.

4.4.1 Moses toolkit and experiment manager Eman

The Moses toolkit relies on and also includes several components for data preprocessing and MT evaluation[6] which we will further describe in detail.

[6] For example, GIZA++ (http://www.fjoch.com/GIZA++.html) involved in finding word alignment, the SRI Language Modeling or SRILM Toolkit (http://www.speech.sri.com/projects/srilm/), implementation of model optimization (Minimum Error Rate Training, MERT) on a given development set of sentences

Eman

As the development under Moses is a very dynamic and interactive process, where dozens of experiments are carried out, it is always useful to exploit some experiment manager in order not to get lost in the experiments and to make them in a proper order. this purpose several managers have been created, as, for example, EMS(Koehn, 2010). In our work, we will use a system developed at the Institute of Formal and Applied Linguistics - **eman** (Bojar – Tamchyna, 2013). An eman experiment consists of a sequence of steps, each of which executes some specific task:
s.corpus... - ensures test, development and training corpora are provided with all necessary annotation and in an appropriate format
s.mosesgiza... - compiles a specific version of Moses and GIZA++
s.align... - aligns two parallel corpora with GIZA++
s.srilm... - a step for training a language model
s.tm... - generation of the translation model (phrase table)
s.mert... - minimum error rate training, the tuning of the model
s.translate... - translation of a test set
s.evaluator... - automatic evaluation in terms of BLEU and other scores

In our experiment, we have trained two types of models - a simple phrase-based system and a factored model. While the first one is based on plain text data from a parallel corpus, the second one uses linguistic knowledge - morphology in our case - in order to better model fluency of the target side. The latter is especially crucial while translating between morphologically rich languages, as lots of morphological forms will not be present in the training data. In addition to this, we used the so called stemming technique(to be described later), which is frequently used to reduce the out-of-vocabulary rate caused by huge number of word forms in morphologically rich languages.
As we have stated, the goal of our work was not to compete with other systems, but to manually explore MT output so as to find interconnections between language relatedness, peculiarities of Slavic languages and system characteristics. For this purpose, we have chosen the following experiments that will be described in the next sections:

- Setting a baseline: simple models
- Factored Translation and out-of-vocabulary issues
- Language relatedness in SMT: comparing English-to-Russian and Czech-to-Russian translation.

- Discussion on other improvements/possible improvements

4.4.2 First baseline: simple model

The first experiments we made concerned the basic settings of the system - running a baseline experiment for Czech-to-Russian and choosing the best translation model. We trained and tuned the system on the UMC data(see Section 3.1.1) from news commentary as these data are often used within the WMT competition. After finding the optimal setup we trained a system on the whole data that we managed to obtain, see section 4.4.4. As a baseline we take an experiment carried out on the UMC test set without introducing more complex (factored) models.

The two phrase tables were generated, the language model was built from the Russian side of a parallel corpus only and did not contain additional data. We used a specially created development and test set (Kolovratník et al., 2009) for training the models. The BLEU score reached **10.71%**.

The preliminary analysis[7] has shown that the most frequently occurring errors concern morphological endings. Besides this, we have encountered a large amount of untranslated words (out-of-vocabulary, OOV) in the output text. These errors mostly originates from data sparseness which is especially severe in the morphologically rich languages. There are not enough data as the same lemma (basic form of a word) can occur in many various forms (with different affixes), which causes many surface forms not to appear in the training data. Even the language relatedness (both Czech and Russian have very similar wordforms and almost the same number of morphological features) do not help in this case.

Next, we will present the general extension of simple models that take into consideration linguistic information - like lemmas and tags - to handle the OOV words and morphology.

4.4.3 Factored models and OOV rate

Generally, researchers improve the OOV rate and morphology using several techniques. The first one is domain adaptation. The large percentage of the unknown words come from a domain different than training data so the ways to handle out-of-domain words were suggested. The second option is to use **factored models** (Koehn – Hoang, 2007) that are trained on a corpus with linguistic annotation. E.g. (Turchi – Ehrmann, 2011), (Bojar – Tamchyna, 2011) address the problem

[7] The detailed linguistic evaluation of baseline MT and the improved MT system will be given in the next chapter

of how to reduce the OOV rate by introducing morphological information or using additional dictionary resources.

In order to include syntactic information, researchers exploit syntax-based approaches like treelet translation(Quirk et al., 2005), dependency-base translation(Bojar – Hajič, 2008), decoding-as-parsing (Yamada – Knight, 2002) approaches and some others, but all of them do not have some significant impact on the translation performance. Also, they presuppose the existence of parsers for source and target languages, and as we do not have one for Russian we will neglect those approaches and include only more morphologically-oriented ones.

Exploiting the surface form of a word - like division into morphemes, stemming - brought positive results in terms of increasing the percentage of translated words especially when building a translation model from/to morphologically rich languages, see (Popovic – Burchardt, 2011), (Gispert et al., 2005).
Our approach mainly follows the line of the research described above - making use of morphological resources and exploiting simple stemming technique within the factored translation models. Factors can represent virtually any piece of information one need to take into consideration in MT. Generally, factors include information on part of speech, morphological tag; syntactic or semantic components may be also used. We have exploited factors as they were initially suggested in (Koehn – Hoang, 2007). In addition to baseline settings, we have added models that exploit stemming and lemmatization. We can schematically depict our experiments as follows:

```
BASELINE:
        form->form
LEMMA:
        main:    form->form+tag
        backoff: lemma->form+tag
STEM:
        main:  form->form+tag
        backoff:  stem-6->form+tag
```

Let us take an example of one entry from the phrase table[8] and examine it within several models.

- The **Baseline** setup is based on a simple translation model from a wordform to a wordform, following is an example entry from a phrase table, the first n-gram before the delimiter ||| is in Czech, the second a respective n-gram in Russian. The n-gram includes a female surname Bhutto that receives

[8] We will leave aside the probabilities, just showing the data structure

an ending 'ová' in Czech and is declined according to a feminine adjective paradigm, whereas female surnames are indeclinable in Russian.

(4.8) dohodě s bhutto**vou** ||| сделке с бхутто
agreement.Dat with Bhutto

- **Factored model - main**. In the improved setup we used two models - the first one with a form on the source side and a form and a morphological tag on the target side of the phrase table:

  (4.9) dohodě s bhutto**vou** ||| сделке|Ncfsdn с|Sp-i бхутто|Npmsiy

  If a word/or a phrase is not found in this main model, an additional back-off model is applied. The back-off model has the same parameters on the target side, but the source side is different: it is either a lemma or a stem:

- **Lemmatized model - backoff**. Lemmatization of a source side relies on the information coming from a MorphoDiTa tagger 3.3.1 that was described in the previous section. The back-off phrase table thus contains mapping from foreign lemmas into target form+lemma phrases:

  (4.10) dohoda s bhutto**vou** ||| сделке|Ncfsdn с|Sp-i бхутто|Npmsiy

  Note, that 'bhuttovou' was not in the dictionary of lemmas.

- **Stemmed model - backoff**. We define stemming as stripping off a word ending - which often bears some morphological feature. The problem of how much should we cut off a word was not that trivial. Several experiments have shown that leaving six characters in a stem showed better results. So, the back-off table provides a mapping between words with up to six characters derived from a form and a form+lemma on the target side:

  (4.11) dohodě s bhutto ||| сделке|Ncfsdn с|Sp-i бхутто|Npmsiy

  Though very simple at the first sight and not as sophisticated as lemmatization, stemming brought a slightly bigger improvement in terms of BLEU score, but the number of out-of-vocabulary words decreased significantly(see table 4.2). The lead of stemming experiments over lemmatization can be possibly explained by the fact that the morphological dictionary of TreeTagger did not contain lemmas for unknown words (which are rather infrequent) whereas the stemming model guessed the closest translation variant and sometimes even correctly.

Let us demonstrate on the imaginary example how this works. Suppose, we have to translate a word *Bhuttové* – 'Bhutto's' and suppose we do not have a wordform *bhuttové* – ''Genetive case) in our training data[9], but there are words *bhuttová, bhuttovou*. Neither Simple, Factored-main and not even Lemmatized model can provide a translation because the wordform *bhuttové* was not seen in the phrase tables, so it will be an OOV word. In the stemmed back-off model, however, the stem *bhutto* is present and aligned with the respective Russian word *бхутто*. As foreign surnames in Russian are not declined, this word will be translated properly[10].

All in all, we can say that introducing back-off models helped to make texts in Czech or Russian less morphologically rich, reducing to some degree the data sparseness problem.

experiment	BLEU	OOV
simple model	10.71%	6%
factored model + lemmatization	12.80	3%
factored model + stemming	13.73%	1.8%

Table 4.2: Simple and factored models

As the settings with stemming yielded better results, we use the second - factored - setup with a stemmed back-off model as the best one when providing linguistic analysis in Chapter 5.

4.4.4 Data issues: genre

Here we will describe experiments involving different types of training data(genres) and various test sets, comparing the results of Moses translation for the two genres - news (UMC) and fiction (Intercorp). The table below shows the results for our experiments.

It can be seen from the table that under similar experimental setup, a model trained on news data scored better than that trained on fine literature. This can only prove the theory that belletristic texts are less suitable as training data for MT than news, even though some of them are the direct translation between

[9] We will not go into a technical details with uppercase/lowercase tricks in Moses, they are described in detail in (Bílek, 2014).

[10] We admit that it is only an artificial example, because the transliteration can be also applied in this case.

data	BLEU
train, test, dev:**umc**	10.71%
train, test, dev:**fiction**	7.06%
train: **umc**+**fiction**, dev, test - umc	12.90%

Table 4.3: BLEU score for simple model trained on different genres.

Czech and Russian. Combination of the two corpora leads only to an insignificant improvement in terms of BLEU score. We have not trained a model on subtitles only, as the data are very specific and unreliable, but we exploited the data in the overall experiment.

Adding more data Apart from parallel data it is also important to gather monolingual data for the language model, which can improve translation performance significantly. In the baseline system, we just used the Russian side of the respective parallel corpus that served as training data for a translation model - UMC. Further, we have exploited a much bigger monolingual and parallel data as described in 3.1.5 and contrasted this setup to the experiment with little data, see Table 4.4. Then we made an overall experiment involving all the parallel and monolingual data under the best setup with a back-off model.

translation model	language model	experiments setup	BLEU
UMC: 92,233 sent	UMC: 92,233 sent	simple	10.71%
UMC: 92,233 sent	UMC: 92,233 sent	factored	13.73%
mixed: 2,566,615 sent	mixed: 24,261,517 sent	factored	**17.23**%

Table 4.4: Baseline experiment vs. all virtual improvements

If we compare the experiments with factored models from the previous subsection with the experiments that involve data , we can say that adding the bigger language model and larger training data helped the translation more than introducing some linguistic data in a form of factors. These results are in line with the main statistical principle that the bigger the data the better the results.

4.4.5 OOV: Named Entities

It is not so straightforward to introduce linguistic information into the SMT systems as it is for the RBMT systems. Unlike in the Rule-Based TectoMT we can not influence the MT system directly by writing a rule and monitor the specific change.

Another way of introducing linguistic information is adding training data that contain the specific phenomenon. We made one experiment of this kind adding information about named entities. This choice was motivated by the fact that there were quite a few unrecognized named entities (NE) , including multiword NEs[11]. Our approach below was quite standard, it was very similar to the one exploited in (Tan – Pal, 2014).

We have exploited a list of names and phrases extracted from the wikipedia headlines (Bílek, 2014) as this was the only parallel Czech-Russian resource of NEs . In the final experiment these data were used as parallel data and it was not measured how adding this corpus affected the translation.

So using the factored configuration of Moses, we ran two experiments:

- the baseline with models trained on data without the Wikipedia headlines
- model trained on data including the headlines

Above BLEU, we calculated the number of OOV words with the same method as described in 4.4.3 - searching for non-latin characters. In the first experiment, the BLEU score was 17.23% with 1216 OOV words. BLEU score in the second experiment was insignificantly better - 17.90% with 1011 OOV words.

We examined the list of OOV words in the outputs from the two experiments. Among those 205 words/mwe that were recognized in the second experiment, there were NEs from the added resource, like e.g. *Tremblay, Carlo Ancelotti, Lyon, Mursi, Amschel Rothschild, Rouno, alt soprán* etc., but also some other words like *mandat* – 'mandate', *regál* – 'shelf', *vévodkyně* – 'duchess' etc.

We also made a maunal evaluation of the randomly chosen 50 sentences that had some changes with respect to NEs. Table 4.4.5 indicates if the overall quality of a sentence/MWE usage was improved/became worse/stayed the same/ after adding the new data, see Table 4.4.5.

In most cases the overall quality of a sentence was not changed, there was often some improvement in MWE translation, and only in 3 cases the overall quality of a sentence became poorer.

[11] For a detailed linguistic evaluation please refer to Sections 5.2.18 and 5.2.1.

Overall improved/MWE improved	same	better	worse
same	25	3	0
better	2	10	0
worse	1	2	0

Table 4.5: Influence of data with named entities

4.4.6 Impact of language relatedness on SMT

One of our goals was to compare machine translation output across the languages, focusing on the dichotomy of related Czech-Russian vs. unrelated English-Russian pair. We are aware that it is not quite correct to compare results of MT across different language pairs, especially if languages are typologically different. In this experiment we try to compare if the relatedness has some positive effect when using phrase-based statistical models.

Intuitively, we assumed that the translation between related languages should bring better results than MT between those non-related. In the case of statistical machine translation this does not hold. The morphological richness of these two languages implies more severe problem of data sparseness than for the pair English - other language. Translation from/into English obviously scored better than into any morphologically rich language. In order to make the comparison more fair and meaningful, we have trained the two systems on the equally big data from the UMC parallel corpus.

Table 4.6 demonstrates that BLEU score for the two language pairs under the same - improved system setup (factored models, with back-off stem-6 model) is significantly higher for unrelated English and Russian, than for the related Czech and Russian. As for the untranslated words, their number decreased more for the morphologically rich langhuage pair.

We can suggest two reasons for such a gap between the translation quality for the related vs. unrelated language pairs. The first one was already mentioned - the morphological richness of the two languages spreads the number of wordforms leading to data sparseness. The second reason have also been mentioned with respect to other problem: the corpus is better parallel for English-Russian pair than for Russian-Czech, because both Czech and Russian texts are translation from English.

language pair + setup	BLEU	OOV
Czech → Russian - simple model	10.71	6.65 %
Czech → Russian - factored model	13.73	1.88 %
English → Russian - simple model	20.43	4.81%
English → Russian - factored model	23.49	1.52 %

Table 4.6: MT between related and unrelated languages

However, we should also note that introducing morphological elements into SMT decreases the number of unrecognized words especially for Czech and Russian pair. This can be explained by the fact that stemming makes a language less morphologically rich reducing the number of distinct words. It should be stressed that the figures reflecting the percentage of OOV can not be taken as an absolute as-is because some of the "improved" out-of-vocabulary words may be guessed incorrectly.

4.4.7 Discussion

We have described the implementation of Phrase-Based MT Moses for the pair Czech-Russian: the baseline model, the factored model with morphological information, various experiments with data. In the next Chapter while comparing the output of several MT systems, we will use the experiment with the best factored model trained on all data we managed to obtain that reached 17.90% BLEU.

4.5 Commercial MT systems

Let us now briefly describe two external MT systems that we will use in our work just for the sake of comparison - Google and PC Translator. As we have already mentioned, this study is of a theoretical nature, and we do not aim to compete with either of the two systems, but rather explore their performance and compare to the research ones described in the previous sections . These commercial systems have been developed for a number of years using many human resources (as RBMT PC Translator) and all the virtually available data (as SMT Google Translate).

4.5.1 Google

Google Translate is one of the most popular online MT applications in the world nowadays. The system exploits statistical models and since year 2002 Google team constantly improves the system and introduces new language pairs. Generally, as for all SMT systems, the main principle is "The more data the better", and that is why the system is one of the best nowadays, as it exploits tremendous amount of data indexed by a Google search engine.

Parallel data for some under-resourced language pairs are not that easy to obtain, so it is possible that sometimes some sort of pivoting is used. We should say that the system is being constantly developing and improving to the way unknown to us (some system settings can change or new training data added), so the output of google translate we use here might have changed. So all the linguistic issues that we describe for Google Translate might be outdated in some time. One of the observations we made while analysing the output of Google Translate is that some translations are not made directly from Czech into Russian, but via English (Czech->English, English->Russian)[12].

4.5.2 PC Translator

PC Translator is a commercial MT system with a rule-based architecture. We are not aware of methods and modules the PC Translator uses in the translation process, we can only make suggestions analyzing the MT output. In general, the quality of a rule-based MT is, by far and large, determined by the quality of a dictionary. The English-Czech dictionary contains almost million entries [13], and the pair Czech-Russian 650,000. Dictionary entries are not only single words, they can also present multiword expressions, covering above all many idioms.

The PC Translator output for the pair Czech-English was included into the WMT-2012 competition, and though the BLEU score was relatively low (10% and 15th place against the winner 16.8%), the manual evaluation showed that the system took the third place (after Google and Moses implementation exploiting Depfix). We can only suggest that for the English-Czech pair a lot of manually written rules had been introduced into the system. It is evident that as in the case of TectoMT, the less perspective Czech-Russian pair received much less attention and the relative score was lower.

[12] The examples demonstrating this fact will be given in the next Chapter

[13] according to http://www.langsoft.cz/translator.htm

4.6 Conclusions and discussion

In this chapter we have described basic properties of several MT systems, and showed the development procedure and preliminary evaluation. Table 4.7 summarizes different information on the systems involved in our comparison. BLEU scores are (with the exception of Cesilko) computed on 3000 test sentences from WMT2013 test set [14].

	System	reusable	time	best BLEU	Value added
RBMT	Ruslan	-	4 years	n/a	dictionary
	Cesilko	-	4 days	5.12%	-
	Treex	+	1 day	6.64 %	Russian analysis
	PC Translator	-	unk	4.73 %	n/a
SMT	Moses	+	3days	18.57 %	tools, parallel corpus
	Google	-	unk	14.44 %	n/a

Table 4.7: MT systems and various criteria

The **time** shows an approximate time spent on adapting the system to translate a new language pair. Here we should make a disclaimer that this estimate time does not include the huge amount of work done by others - creating platforms for the MT, which was initially done for different language pair (Czech-Slovak for Cesilko, English-Czech for Treex); Moses toolkit, which comes with many language-independent tools like language modeling, GIZA++ etc. It does not inlcude the effort the time spent on creating data (dictionaries, parallel corpora, morphological dictionaries) and the related NLP tools (taggers, parsers, segmenters). We consider only the time to combine all the tools and data to produce a system that translates between Czech and Russian.

Table 4.7 indicates that SMT systems, in general, received considerably higher score in terms of BLEU than RBMT. As we have mentioned in the distinct sections, we do not consider BLEU score to be the absolute indicator of quality, nor it is suitable for the involved RBMT systems. We take it as a relative measure that can indicate the progress within some system. However, it is the most popular metric of automatical evaluation and the WMT competitions showed the score correlates with human judgments. Finally, the last column provides information on tools and data that were created while developing an MT system and that can be possibly reused in some other NLP applications.

[14] http://statmt.org/wmt13/

In (Bojar, 2011), a very similar comparison is being made for English-Czech Machine Translation. The author described commercial and research systems that are the same as we use (excluding Česílko and Ruslan). The conclusions are a bit different, because the language pair is very popular, and all the systems have been developed for English-Czech for years involving many people. The main conclusion of the work was that commercial systems generally outperforms the research ones[15]. It should be noted that there is not such a big gap in quality between RBMT and SMT for Czech-English as this language pair received a lot of attention under PC Translator. On the contrary, if RBMT systems for Czech and Russian would have been tuned for years, we can persume the quality might be also comparable to the SMT.

While comparing MT systems of different types, researchers always point out the relativity of such a comparison. In our case, it is not fair to compare the system developed twenty years ago on old machines, based on a formalism not used nowadays with the state-of-art systems that involve new methods, tremendous amount of data and highly modern efficient machines.

A frequently asked question - whether Rule-Based or Statistical methods lead to better results - is generally answered that the best can be Hybrid MT which combine data driven approach and linguistic knowledge.

In this Chapter, we discussed mainly BLEU score, the metrics designed for SMT under which the RBMT systems are stated to be underscored. Our main goal that we will approach in the next Chapter will be mainly **linguistic analysis** of the MT output of the presented systems.

[15] The situation has changed since the time the article was written, and research systems won over commercial in the WMT competition, e.g. English-Czech developed at UFAL (Bojar et al., 2013)

5

Linguistic evaluation of MT systems between Czech and Russian

> It is our task to figure out how to make use of the insights of linguists
>
> *Frederick Jelinek*

In this chapter we provide a detailed evaluation of the Machine Translation systems described in the previous chapter. We will examine the output of these systems from a perspective of a linguist, focusing on concrete language issues that pose challenge to MT. The chapter aims at answering the central questions of this thesis:

- What kind of mistakes Czech-Russian MT's of various architectures tend to make
- Which of the above are caused by the system settings and which originate from the discrepancies between the two languages, and how the latter two correlate with each other

5.1 Evaluation scheme: Error Flagging

The majority of research articles on Machine Translation includes some kind of evaluation, discussing ways of improving various weak points of the systems, presenting the improvement itself and showing some gain (or sometimes loss) in a translation quality. There are also many papers oriented solely on the evaluation strategies - these are mainly about automatic evaluation techniques that allow researchers to estimate their progress without much human annotation effort.

Evaluation of MT implies three points that should be taken into account:

- evaluation of system architecture itself and evaluation of its settings (e.g., comparing RBMT and SMT);

- evaluation of data used in the experiment (e.g., dictionary/corpus quality and size)
- evaluation of tools that ensure proper grammar construction (e.g. transfer rules for RBMT, factoring techniques in SMT).

The task of MT evaluation is therefore quite challenging because we compare several MT systems of different origin and of complete different nature (industrial vs. experimental). In the introductory chapter, we have described several commonly used evaluation metrics (BLEU, NIST, edit distance). As this work is of a linguistic nature, we have adopted a 'linguistic-oriented' annotation scheme called Error Flagging (Vilar et al., 2006) which is based on attaching labels to mistranslated words. This annotation scheme allows to reveal concrete weak points of a system and make a statistics of the most frequent errors. We believe that this framework is more adequate for our evaluation task from the point of view of a linguist.

This scheme (or a similar one) was exploited by other researchers when a detailed linguistic analysis of MT output was needed, ex. (Bojar, 2011) for Czech-English MT. The automatic error detecting tools were proposed recently, ex. Hjerson (Popovic – Burchardt, 2011) or Addicter (Zeman et al., 2011) that exploit the reference translation in order to spot the troublesome places of the MT systems. (Zhou et al., 2008) proposed a system of pre-defined linguistic checkpoints to test if systems translate those linguistic phenomena correctly. A work which is very close to ours - classifying errors according to the language layers - is presented in (Wisniewski et al., 2014) where authors obtained the labeled data from the students exercises.

5.1.1 Error Taxonomy

Labels present a specification which type of error is being made, ex. **lexical**, **unknown word**, **missing word**, **word order**, **word form**[1]. Following is an example of annotated sentences translated by the four MT systems[2]:

(5.1)

[1] Those errors will be explained further.

[2] Here we just show an example of how the annotation looks like without glossing, but mostly we will gloss and translate the source (src) and the MT output sentences to demonstrate concrete linguistic phenomena. We ignore most of those labels in the examples, as there are generally many error tags and we want to concentrate only on one phenomenon at a time. Also, in the majority of examples, we will show only the relevant chunks, not the whole sentence.

(ref) впервые подобное требование было введено в аризоне.

(src) první takový požadavek zavedla arizona.

(mos) первый такой запрос ввела аризона.

(goo) первый такое требование **missAux-byt::** введено **form::**Аризона.

(pct) первый такой запрос ввела Аризона.

(tmt) Первое такое требование **extra::**она **verbform::**ввести **val::**Аризоне.

Our classification of errors follows the scheme as it was suggested by Vilar, but more specifications are introduced taking into consideration the language specific features and the purpose of our work. We have defined several general error classes[3]:

- 'surface' word issues: missing word, extra word, unknown word
- rather lexical semantic issues: wrong lexical choice, wrong disambiguation, bad word sense, wrong usage of multi-word expressions
- rather morphological issues: agreement, noun valency, genitive constructions, incorrect part-of-speech, other errors in endings
- rather syntactic issues: word order, negation, reflexives, other syntactic constructions

Unlike Vilar's general scheme, we do not classify the word order errors into 'Short range' errors or 'long range' errors as this is hardly relevant to the free word order languages. On the other hand, as we want to describe better some linguistic phenomena, we have made specifications (inspired partially by (Bojar, 2011)) in morphological errors ("form" flag) classifying those errors into more fine-grained categories. The difference from the classification in the cited work is that we aimed first to sort the errors according to language layers.

The whole scheme is shown in the picture 5.1.1.

During the annotation process we have made a few more specifications (mostly, POS) into the classifications indicating more linguistic information, ex. when a wrong form of a word is used we include its part of speech (*verbform::ввести*); when another POS should be used, we specified both the wrong one and potentially correct (*n-iof-v::поддержке* - a noun instead of a verb), or in the cases

[3] We included a word 'rather' in front of each language layer because the classification can be arguable and misleading as some errors can be related to more layers.

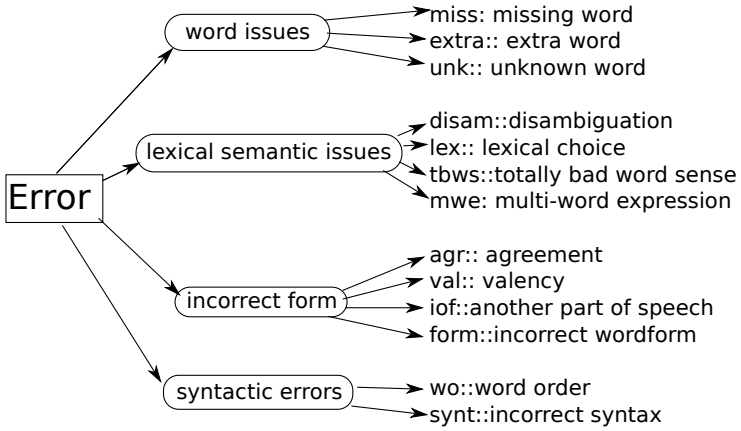


Figure 5.1: Error taxonomy adopted for Czech-Russian language pair

when a word is missing, we specify a POS as well, as in the above example: *missAux-byt::*. Those tags will therefore help to identify concrete reasons that might have caused an error.

5.1.2 Some challenges of the annotation scheme

Before showing straight away the concrete numbers, we will note some weak points and uncertainties of the presented approach and name some challenges we faced during the annotation process.

Annotation 'one tag per one word'

First, we believe that in some cases it is more appropriate to annotate more than one word with one error tag. The latter matters, for instance, when annotating errors in multiword expressions, when two or more words were translated incorrectly. It is not quite clear how to mark a phrase that presents a multiword expression in the source language and is mistranslated in a target. For instance, a multi-word expression "volby v předstihu"(elections in advance) should be translated into Russian as 'досрочное голосование' (pre-term election), but all the systems made some error - either leaving out some component or translating the expression literally:

(5.2) (src) *volby* *v* *předstihu*
elections.Pl in.Prep advance.Loc
'elections in advance'

(mos) *mwe::выборов*
elections.Pl.Gen
'elections'

(goo) *mwe1::выборов* *mwe2::заранее*
elections.Gen beforehand
'(from) elections beforehand'

(pct) *mwe1::выборы* *mwe2::зажечься* *unk::предстиhу*
elections.Pl flare.inf unknown word
'elections to flare (unk)'

(tmt) *выбор* *mwe1::в* *mwe2::упреждение*
election.Sg in.Prep advance.Acc
'election in advance'

So far we have decided to mark each word with the respective tag even if this tag can be related to more words at once, so that the annotation looks more consistent.

Linearity of the approach

Second, we believe that the approach is somewhat "linear" as it is focused on a one word at a time without considering the interconnection between errors. For instance, if some error occurred, it could cause, in turn, incorrectness or mistranslation of another word.

This can be vividly demonstrated on the example when a hypothesis translation misses a verb. As a verb determines morphological properties of the dependent words, we could not exactly answer the question whether the ending of a noun (that should depend on the missing verb) is right or wrong. It is not relevant for English, because if the dependent words are translated properly, there is no need to think about their morphological forms. On the contrary, when a verb in Russian (or Czech) is missing, the question can arise if a morphological form of a dependency is right or wrong.

This problem also occurs when a verb is mistranslated: either it has a totally different word sense or it is disambiguated wrongly, or a different lexical sense is selected. Let us illustrate the problem on the following example:

(5.3) (src) *vystavený* *svou* *školou*
issued.Gerund.Masc.Sg its.Pron.Ins school.Ins.
'issued by its school'

(mos) *tbws::nod* *cвою* *шкoлу*
under.Prep its.Acc school.Acc
'under its school'

Evidently, in Moses translation a gerund is mistranslated as a preposition, but there are some questions about the two dependencies - 'its school'. With the respect to the translated preposition, a morphological form of the NP (the preposition *under* presupposes Accusative case of a noun) is correct, but if we imagine the word "issued" was translated into Russian properly, it would have required another case (Instrumental, as in the source).

One of the possible solutions might be a more deep linguistic analysis, as, e.g. to represent sentences as trees so we can see the dependent nodes and the connections between words. However, syntactic parsers can introduce more mistakes and are rather hard to establish.

Subjectivity of the evaluation

When flagging errors, an annotator processes a sentence having some "right" translation in mind ('target acceptable output'). Though, when the new evidence (word) comes, this ideal translation can change. It should be also noted that researchers claim a very low interannotator agreement when two people flag the errors (less than 50% in average for Czech-English (Bojar, 2011)). So it can be said that judging the MT errors is a very subjective matter and all such analysis (including ours) should be treated with caution.

5.2 Error types in Czech-Russian MT output

As we have mentioned above, we have first tried to classify the errors according to the language layers - morphological, syntactic and semantic. It became evident that such a classification would not be relevant as many errors can be related to more than one class at the same time [4]. So we have decided to define specific error types not relating them to any general class precisely. However, the order in which we will describe the mistakes will go from the superficial language level to the deeper level.

[4] Further in the text, when we say **morphological error** or **syntactic error** we mean that it relates **rather** to morphology or syntax respectively, and it can have some connections to other language layers as well

First we describe a group of 'superficial' errors that are quite easy and unambiguous to annotate and generally do not have some deeper linguistic motivation[5]. They are more or less related to word issues: unknown words, missing words and extra words.

The next larger group of mistakes belongs to **morphological mistakes**: wordform, incorrect part of speech, agreement and valency. Though these mistakes can be related also to ''syntactic" ones as they often originate in improper handling (or, rather, ignoring) of syntactic rules. These are mostly suffixes of words that we take into consideration. When a wrong prefix of a word is encountered, we relate those mistranslations to lexical(dictionary) issues.

Syntactic mistakes are mostly those in word order, and we also describe some concrete syntactic issues that present a challenge to Machine Translation.

Finally, we define 'deep level' mistakes in lexical semantics: disambiguation errors(disam), wrong synonym choice(lex), totally bad word sense(tbws) and multiword expression translation(like idioms etc.).

In the following Table 5.2 the statistics of errors (sometimes with specifications) found in 100 sentences for all the four systems is demonstrated.

In the next subsections we will describe those errors in more detail. This will include the specification of the concrete error type and a short comparative analysis of the phenomena in the two languages if relevant. In the examples, we will not mark every error, because generally there are many of them, we will define one error at a time and accentuate it if needed.

Therefore we will demonstrate examples of errors and discuss possible reasons why this very mistake may have occurred. We will suggest if an error is a result of some discrepancy between the two languages, is it merely a technical issue of the MT system or is it a mixture of both. We should also note that all the conclusions and assumption made in this Chapter are highly connected with the concrete sample of data. The translation output may change for all the systems, because PC Translator is presumably under the development; new training data are constantly added within Google Translate. Moses output is even more tricky to use, because each time the system is trained and tuned, the output is often different.

5.2.1 Unknown words - OOV

The most evident SMT errors that impede understanding of a text are untranslated words (unknown words, out-of-vocabulary - OOV). Some other mistakes -

[5] In spite of being superficial, those mistakes can be considered to be 'serious' as they affect dramatically the general perception of a sentence.

Error/System	MOS	GOO	PCT	TMT
unk	3	0	118	40
missing word	13	13	3	3
extra word	20	15	21	20
form	6	11	26	15
bad verb form	4	13	7	29
incorrect pos	8	10	14	0
valency	9	19	21	16
agreement	7	6	27	21
word order	7	13	10	7
other syntactic errors	7	7	4	3
lex	20	14	60	39
disam	15	13	37	18
multi-word expression	1	2	7	11
totally bad word sense	7	6	41	8
totally	120	136	369	209

Table 5.1: Error types in Czech-Russian Machine Translation

like wrong morphological form of a word or syntax errors - make a text inconvenient to read, but one can still get a sense out of it, whereas unknown words in another language give us no information at all. The source of those mistakes is different for SMT and RBMT systems, we will now discuss both of them in turn.

For the SMT, we have already described this problem in Section 4.4.3 where we explained in detail why the OOV words occur in the Statistical Machine Translation and introduced some obvious improvements to the baseline experiment that helped us to significantly reduce the number of OOV. Another experiment in Section 4.4.5 concerned the lexical part of out-of-vocabulary - true out-of-vocabulary words that were not present in the dictionary in any form. Introducing additional data with Named Entities to the translation model did not bring any significant result in terms of BLEU score, but whenever a word was in the additional data, the translation improved.

Altogether, the reason for the OOV occurrence is that the word has not been seen in the training data. It may be either completely out-of-domain or the word was not found in the training data, though it could have occurred in some other

morphemic form. SMT Google and Moses coped with this problem better: in the testing sample of 100 sentences, google did not have a single untranslated word (due to the excellent coverage of news domain which are often translated into more languages[6]) and there were only 3 errors in Moses output. We have also calculated a number of OOV words in the whole WMT test set just by searching in the translation for a words containing Latin character[7] Google has 315 untranslated words (which makes only 0,006% of all words) and Moses 250(0,004%). So the OOV errors when addressed properly (and in-domain) are not really significant for the statistical systems.

As it is evident from the Table 5.2, the RBMT systems produced much more OOV words than the statistical ones. We could not give an exact percentage of OOV words on the whole test set as unknown words were transliterated by both systems. PC Translator exploits a human-made high-quality dictionary which (according to PC Translator pages) contains quite a large amount of entries (around 650 000). So it is not quite clear why PC Translator made almost three times as many errors as TectoMT though they exploited almost the same dictionary. TectoMT dictionary is based on the automatically extracted lexicon (see Section 3.2) from the in-domain parallel corpus, so the coverage of the test data should be sufficient. We can only suggest that for the commercial system it might not be a matter of a bad dictionary, but rather some problem in a system architecture - most probably in the analysis of a source sentence, and we can not really say which because we can not look inside the system. TectoMT unknown words were mainly named entities, and some of them were reflexive verbs that were not properly handled within the system.

5.2.2 Missing word

Another nasty error that affects the whole perception of an output phrase is that of a missing word, especially of a verb. A verb determines the structure of a sentence. In the dependency tradition it is being viewed as a center of a sentences as it predetermines the arguments, their semantic class and morphological form. The sentence can be more understandable without one of the verb arguments or some other auxiliary part of speech than without a verb. When a verb is missing in a sentence, the problem that was shown previously arises: how to evaluate the arguments of a non-existing verb (ex., whether they have the proper morphological ending required by a verb).

[6] Our test set can be possibly included into the Google's training data as well.

[7] Excluding web addresses which should not be translated.

SMT systems showed more missing words than RBMT, which is natural, since the RBMTs translate mostly word-for-word and only under some circumstances (see further) can they miss out some element of a sentence. Whereas SMT systems are leaving out words on a regular basis because of the nature of the Phrase-Based translation: source and target phrases can have different number of elements.

In the **SMT**, quite often a verb is left out of a sentence, which affects a perception of the whole sentence and sometimes makes it impossible to further evaluate the sentence. It is not evident how to process arguments of the missing verb as it determines morphological features of its dependencies. In the following example all the dependencies stayed 'orphaned' without a verb and they remain only a meaningless set of words:

(5.4)

(SRC)	*přijalo*	*asi*	*dvanáct*	*států*	*zákony*
	adopt.Past.3Pl.Neut	about	twelve	states	laws.Pl.Acc
	'about twelve states have adopted laws '				

(GOO)	*missverb::*	*закон*	*штата*	*десятков*
	law.Sg.Acc	state.Sg.Gen	dozens.Pl.Gen	
	*'law of a state of dozens'			

We have 'found' much less missing words in the **RBMT** than in SMT. In the RBMT systems, those mistakes are all missing prepositions which are on one hand not that serious mistakes and do not affect the perception of a sentence. Still, these are grammatical errors and they can be justified by the difference between Czech and Russian surface valency [8]. So, this error can also belong to the section 'verbal valency' as well.

In the following example, a Czech verb 'to influence' is governed by a noun in an Accusative case, and the system translated a respective noun with an Accusative case as well. However, the surface realization of the argument is different in Russian - the Russian verb requires a prepositional phrase, so the two RBMT produced an error because neither had a rule covering this discrepancy:

(5.5)

(src)	*ovlivnit*	*výsledky*
	influence.inf	outcome.Acc.Pl
	'to influence the outcome'	

[8] The notion of valency and how we understand it will be introduced in the next Chapter.

(ref) *повлиять* ***на*** *результаты*
influence.Inf on.Prep outcome.Acc.Pl
'to influence the outcome'

(pct) *повлиять* ***missprep::*** *исход*
influence missingPreposition outcome.Sg.Acc
*'to influence the outcome'

(tmt) *повлиять* ***missprep::*** *результаты*
influence missingPreposition outcome.Pl.Acc
*'to influence the outcome'

We can conclude that missing words in the SMT output come from the system architecture, whereas unknown words are merely data issues. As for RBMT, a missing word generally indicates some discrepancy between the source and the target languages that is not covered by rules. Again, we should note that the notion of a 'missing' element can be rather a subjective judgment.

5.2.3 Extra words

Extra words also present a substantial number of errors for all the systems, but again the reasons for this are different.

For the SMT they can be justified in the same way as in the case with missing words: within the n-gram based translation model words in phrase tables are not aligned one-to-one. It should be noted that extra words are mostly auxiliaries, not content ones, and most of them are prepositions. It is quite tricky to tag 'extra word' errors in SMT systems because often the whole phrase is mistranslated and more errors are in game. However, because of the linear approach of our annotation, we have to mark one word with some error tag. For instance, the preposition was extra in the following example:

(5.6)

(src) *legislativci* *v* *roce* *2011* ***podpořili*** *zákony*
legislators-N.Pl in year 2011 supported-**Verb**.3Pl.Past law-Pl.Acc
'in 2011 legislators supported laws'

(goo) *законодатели* *в* *2011* *году* ***extra::при*** ***n-iof-v::поддержке***
legislator-N.Pl in 2011 year by-prep support-**Noun**
законы
law-Pl.Acc
'*in 2011 legislators in support laws'

In the above example, for a Czech verb *podpořit* – 'to support' a wrong hypothesis *при поддержке* – 'by support' was chosen. The latter nominal phrase is used typically in other verbal constructions ('make smth in support of'). Ideally, the preposition and the noun should be marked as one mistake, but under our formalism they can not, so the decision was to mark a preposition *при* – 'by' as extra token and a noun *поддержке* – 'support' as 'noun instead of verb'[9]. In several sentences extra punctuation marks were used, but we do not consider those mistakes to be meaningful from the linguistic point of view.

As for RBMT, there can be several reasons for extra words. We can only speculate what can generate so many extra words in PC Translator: they are sometimes completely unrelated to the content of a sentence. TectoMT output does not have any extra content words or even prepositions, all the extra words were personal pronouns, 3rd Person. The roots of this mistake are not that trivial. Those extra pronouns were generated by a rule which covers a pro-drop phenomena[10] in Czech language. In short, whereas Russian (and English) language uses the pronoun, it is in most cases left out of the Czech sentence. So, in order to ensure a proper translation into Russian, a module to cover this discrepancy was written. However, there are the cases where the pronoun should not be used, like in the following sentence:

(5.7)

(src)	*je*	*třeba*	*poznamenat*
	is	necessary	note.Inf
	'It is necessary to note'		

(tmt)	*extra::Он*	*надо*	*заметить*
	He.perspron.3Sg.Masc	necessary	note.inf
	*'He necessary to note'		

Such impersonal constructions with adverbs belong rather to a lexical issue and it is virtually impossible to make an exception to the pro-drop rules for all such troublesome cases[11].

[9] This very error might have also occurred because of an embedded time adverbial 'in year 2011' which interfered into the argument structure. We have tried to translate the phrase without the adverbial by GoogleTranslate (*Legislators supported laws*), and it was translated properly - with a verb. We will discuss the issue of embedded clauses in more detail later.

[10] More on the pro-drop can be found in Section 5.2.11

[11] This is an illustration of how one rule added to a system can improve translation in one case, but can have a negative impact in others.

5.2.4 Agreement

Quite a lot of morphological errors can be related to the **agreement**. Agreement reflects the obligatoriness of a word (e.g, in verb-noun, or adjective-noun pairs) to have a concrete morphological feature (agree in e.g. gender or number) predetermined by another word. Czech and Russian have the same types of agreement, but this fact does not presuppose trouble-free translation even for the Rule-Based systems. Though there exist many types of agreement, we will describe only those most typical cases that are relevant for the machine translation - subject-verb agreement and adjective-noun agreement.

Subject-verb agreement

The relation between a subject and a verb is dual: the noun determines the suffix of the verb on one side (agreement), and on the other side, a verb governs the specific case of the noun (valency). We will set aside the valency issues and describe them in a separate section.

In Czech and Russian, a verb in the past form agrees with a subject in gender and number, and in the present and future forms also in person. The following example demonstrates a verb agreement with a subject in person:

(5.8) (cz) *odejdeš* *vs.* *odejde*
leave.Fut.2Sg vs. leave.Fut.3Sg
'You will leave' vs. 'He/she/it will leave'

(ru) *ты* *уйдешь* *vs.* *он/она/оно* *уйдет*
you leave.Fut.2Sg vs. you leave.Fut.3Sg
'You will leave' vs. 'He/she/it will leave'

Mistakes in agreement can be justified differently for different types of systems. SMT can not consider the connection between a subject and a verb, but as soon as the two words stand not far from each other, the number/gender morpheme has more chance to be chosen properly because the respective n-gram was seen in this phrase table. If this does not happen, an error in agreement can occur. In the example below a verb (should be Plural) does not agree with the subject in number in the output of both SMTs:

(5.9) (src) *advokáti* *za* *posledních* *deset* *let* *zaznamenali*
advocates-**PL** for last ten years noticed-**PL**
'advocates for last ten years noticed'

(goo/mos) *юристы* *за* *последние* *десять* *лет* *записал*
advocates-**PL** for last ten years noticed-**SG**
*'advocates for last ten years noticed'

In theory, RBMTs have better chances to cope with agreement issues as the latter are determined by rules. However, even if those rules are present in the system, the latter can fail to determine which words should agree with each other (due to errors in parsing). The second case is when a reference word with which a verb agrees is not translated properly or is not translated at all. It is therefore not evident how to mark or whether to mark a mistake altogether (see Example 5.10). There was nothing to agree with in PC Translator output and the wrong usage of infinitive form made it impossible to express an agreement in number for TectoMT.

(5.10) (src) *napomohly* *kampaně* *k*
helped-Verb.Past.3**Pl**.Fem campaigns-N.**Pl**.Fem for.Prep
zaregistrování
registration
'Campaigns helped for registration'

(pct) *посодействовать* *кампанский* *к* *реестровый*
help-Verb.**Inf** campaign-**Adj** to registered-Adj
*'help campaign to registered'

(tmt) *помочь* *кампании* *к* *зарегистровани*
help-Verb.Inf campaigns-N.**Pl** to unknown*registration
*'to help campaigns to register'

The possible reason for so many mistakes in the previous example is a non-standard VSO (verb-subject-object) constituent order in a source sentence. It should be noted that all the systems except for Moses translated this sentence with an error in agreement.

Noun-adjective agreement

In Czech and Russian, an adjective agrees with a noun/pronoun in gender and in number. An adjective in a text generally stands before (or not far from) to the governing noun, so the SMT systems generally produce the correct hypothesis. From the table 5.2 it is evident that noun-adjective agreement errors in SMT are really rear whereas they are very frequent in the RBMT output for the same reason as verb-noun agreement: failure to find a connection between words possibly due to errors in parsing. For instance, PC translator uses the original (Czech) gender morphemes for a noun and a pronoun, which results in an error when the gender of a Czech noun is different from the Russian one. In the below example, the possessive adjective *naše* – 'ours' agrees with the noun *obec* – 'village' in gender (Fem.) in the source phrase. The respective Russian word *населенный*

пункт – 'village' is Masculine, and the possessive adjective *наша* – 'ours' has a feminine gender morpheme as in the source.

(5.11)

(src)	*naše*	*obec*	*hlasovala*	*proti*
	our-**Fem**	village-**Fem**	voted	agains
	'Our village voted against'			

(pct)	*наша*	*населенный пункт*	*проголосовала*	*против*
	our-**Fem**	village-**Masc**	voted	against
	*'Our village voted against'			

Moreover, in the following example (Example 5.12) it is evident that PC Translator does not take into account the morphological ambiguity: the phrase *tato nová* can be also analyzed as *this.Sg.Fem new.Sg.Fem* as those two forms - plural neutrum and femininum singular are morphologically ambiguous in Czech, but not in Russian. The system had chosen the wrong variant of a morpheme and a wrong form - femininum singular - was generated in Russian.

(5.12)

(src)	*tato*	*nová*	*ustanovení*
	this-**Pl.Neut**	new-**Pl.Neut**	regulation-Pl.Neut
	'these new regulations'		

(pct)	*эта*	*новая*	*указание*
	this-**Fem.Sg**	new-**Fem.Sg**	regulation-Sg.Neut
	*'this new regulations'		

5.2.5 Incorrect part of speech

We made a special category for mistakes when the word sense is chosen appropriately, but the part of speech is wrong, ex. a noun is used instead of a verb, or an adjective instead of an adverb etc. As it can be seen from the table of errors, those mistakes are typical rather of SMT than RBMT. The errors were annotated specifying the two part of speech tags that got confused.

Statistical systems may produce this type of mistakes because the mistranslated word occurred in the used part of speech in the respective context more frequently than in the appropriate part of speech:

(5.13)

(src) *škody* *se* *jim* *zčásti* *podařilo*
damages-AccPl refl.part they-Dat.Pl partly managed-3SgPast
omezit
limit-**Verb**.inf
'They managed to partly limit damages'

(goo) *ущерб,* *который* *они* *частично* *удалось*
damage, which they-Nom.Pl partly managed-3SgPast.refl
предел
limit-**Noun**.Sg
'damage, which they partly managed a limit'

We have also noticed that it is often a gerund or a transgressive that is confused with some other part of speech:

(5.14)

(src) *státy* *představují* *171* *z* *270* *hlasů*
states represent-**verb**-3PlPres 170 from 270 votes
'states represent 171 from 270 votes'

(goo) *государства* *представляющие* *171* *из* *270* *голосов*
states representing-**gerund** 171 from 270 votes
'states representing 171 from 270 votes'

More detailed analysis of non-finite constructions will be presented in Section 5.2.13.

5.2.6 Genitive of negation

Genitive of negation[12] is a construction typical for some Slavic languages, where a subject (otherwise Nominative in affirmative constructions) or a direct object (Accusative) are marked by Genitive case in phrases with negation. The problem of Genitive of negation is a well-studied, e.g. (Mustajoki – Heino, 1991) made an extensive study of this phenomena and collected a large bibliography on this subject. As for the contrastive work on this phenomenon, (Skwarska, 2002) made a comparative analysis of genitive of negation in three Slavic languages – Czech, Russian and Polish. Here, we will not go into a detailed description of this construction, but only point the errors that are made in this sentences in sentences

[12] Again we should note that the phenomenon is related to the syntactic and to the semantic layer as well, but we put it here because the surface error is a morphological one - using a wrong ending for a noun

containing genitive of negation. Object marked as an Accusative instead of a Genitive case in the context of negation (ex. 5.15) can be considered as an error in a wordform of a noun, but we will treat this issue separately from other morphological mistakes.

(5.15)

(src) *evropané* *nemají* *koherentní* *a* *jednotnou*
Europeans have-not.neg coherent-**Acc** and consequent-**Acc**
politiku
politics-**Acc**
'Europeans does not have a coherent and united politics'

(goo) *европейцы* *не* *имеют* *согласованную-**Acc*** *и*
Europeans not have coherent-**Acc** and
*последовательную-**Acc*** *политику-**Acc***
consequent-**Acc** politcs-**Acc**
'*Europeans do not have a coherent and united politics'

(pct) *европейцы* *не* *имеют* *когерентный* *и* *единый*
Europeans not have coherent-**Nom** and consequent-**Nom**
политику
politcs-**Acc**
'*Europeans do not have a coherent and united politics'

(ref) *европейцы* *не* *имеют* *согласованной* *и* *последовательной*
Europeans not have coherent-**Gen** and consequent-**Gen**
политики
politics-**Gen**
'Europeans do not have a coherent and united politics'

In the above example in a source sentence, a negated verb has accusative dependencies, and in the output all the systems produced objects in Accusative (or other improper cases), whereas the right variant was Genitive (see ref example).

It should be noted that statistical systems do this type of an error from time to time (e.g., when a construction above contains coordinating members), and in some cases the proper - Genitive - case is used after a negated verb (when a construction is not extended).

The negated **possessive constructions** pose challenge to all the MT's, both rule-based and statistical, as modern Czech language does not have this construction anymore[13], and Accusative is used in this case:

[13] It was used in Old Czech, the remainings of this construction are some phrases with the particle **ani**: *nemám ani potuchy* – 'I don't have the slightest idea'.

(5.16)

(1cz) *Nemám* *doklady.*
Not-have-1.Sg documents-**Acc**
'I don't have documents'

(1ru) *У* *меня* *нет* *документов.*
For me not documents-**Gen**
'I don't have documents'

More syntax-oriented problems of the possessive constructions will be discussed further in the subsection 5.2.12.

Let us now demonstrate a difference between a construction with genitive of negation (Example 5.17(1ru)) and one close to the Czech or English (2ru) that can be used in the same context in Russian language:

(5.17)

(1ru) *В* *кино* *не* *было* *зрителей.*
In cinema not was-Sg.Neut.Past viewers-**Gen**.Pl
'There were no viewers in the cinema'

(2ru) *зрители* *не* *были* *в* *кино*
viewers-**Nom** not were in cinema
'Viewers were not in the cinema'

The syntactic construction (1ru) with genitive of negation is not possible in Czech or English, the one equivalent to (2ru) is used instead.[14]. (Barbara H. Partee – Testelets, 2011) and (Babby, 1980) showed that the two constructions can be distinguished with respect to their semantic properties: the genitive of negation (1ru) are claimed to be "existential" constructions whereas (2ru) are "predicative" (affirmative). (1ru) means that the object did not exist in the predefined conditions, whereas the second sentence (2ru) supposes their existence, but not in this concrete place (The viewers may be standing near the cinema).

The genitive of negation is an example of one evident discrepancy between Czech and Russian languages, and sometimes it can cause mistakes in noun endings. At least in our test data, this mistake is not very frequent in SMT as soon as a dependent noun stay close to a verb. It can be properly handled by RBMT with the help of rules (substituting for genitive in the context of negation), but in order to ensure this rule works there should be a good analysis module (a syntactic parser).

[14] For now we will leave aside a problem of word order, as here we focus only on morphological features

5.2.7 Valency

In this work, we have decided to examine in detail one particular issue that we believe is crucial from the point of view of Machine Translation and language comparison as well.

As in previous cases, valency errors can be related to both morphological, syntactic and semantic errors. As we will show further, the notion of valency itself is not very straightforward and can be understand differently by different researchers. Traditionally, in general linguistics, the notion is used to indicate that the verb requires some number of complements of the certain semantic type. However, in more applied works like in (Rosa et al., 2013), the author refers to valency with respect to its surface realization - morphemic endings of nouns or preposition required by a verb (some theoretical linguists use a term 'rection' for surface valency). In this work, for the sake of shortness, we will stick to the second understanding of valency, though this may seem to be confusing from the point of view of the traditional terminology.[15]

It was challenging to set distinct rules to distinguish valency errors from just errors in word forms. In short, we attach a label 'valency' when dependencies of a predicate are used in a wrong form On the surface, these errors look like morphological, but they provoke syntactic and semantic improperness as well. So, next in the text, we use the word 'valency' in sense of surface valency, and under the notion 'valency frames' we will understand mainly surface forms of frame elements.

The origin of these errors (actually, like any error) is different for the SMT's and RBMT's. The most evident case is when Russian and Czech valency have some discrepancies, and the Czech structure is used in a Russian output. The following example demonstrates a verb "to take something from someone".

(5.18)

(src) *odnímají* *volební* *právo* *občanům*
take-3PlPres vote right citizens-**Dat**
'They detach voting rights from citizens'

(goo) *отнять* *право* *голоса* *людям*
take right vote-gen people-**Dat**
*'to detatch voting rights to people'

[15] More sophisticated and broad definition of valency, theoretical and practical aspects of this linguistic phenomena will be given in the distinct chapter. Here we will just present examples of errors not going deep into the comparative analysis of valency in the two languages.

(ref) *отнимают* *право* *голоса* *у* *граждан*
take-3Pl right vote-gen **from** citizens-**Gen**
'detach voting rights from citizens'

In Czech language the verb *odejmout* – 'to take' requires an object in Dative case whereas in Russian a verb *отнимать* – 'to take' with a preposition *у* – 'from' governing Genitive case should be used instead. The systems often use a Czech valency structure of a verb which sometimes results in a mistake like in the example above. This very case - the verbs with the semantics of 'taking something from somebody' is very tricky in Czech. We have tried GoogleTranslate from Czech into English, and the same error was made: the sentence 5.18(src) was translated wrongly as *Take away the citizens right to vote.*

Generally, when a dependent word stands directly near (or not far from) a governing verb, the statistical systems cope with the discrepancies in valency much better than the Rule-Based - in case a respective rule is not added to the RBMT. Let us consider an example of a verb "to influence", which governs a noun in Accusative in Czech and a prepositional phrase with Accusative in Russian - 'to influence on smth.'

(5.19)

(src) *ovlivnit* *výsledky* *voleb*
influence results-Acc.Pl elections-Gen
'To influence results of the elections'

(ref/goo/mos) *повлиять* ***на*** *результаты* *выборов*
influence **on** results-Acc.Pl elections-Gen
'To influence results of the elections'

(pct) *повлиять* *исход* *выборов*
influence result-Acc.Pl elections-Gen
*'To influence results of the elections'

(tmt) *повлиять* *результаты* *выборов*
influence results-Acc.Pl elections-Gen
*'To influence results elections'

In this case, the two Rule-Based systems failed to produce a proper surface form of the argument because a translation was made directly without applying a rule on this discrepancy. Both statistic systems produced the proper translation like in the reference. However, for SMT systems, if some adverbial is introduced between a verb and its argument, other mistakes may arise for the same verb. In

the following example the verb *влиять* – 'to influence' in GoogleTranslate does not have an obligatory preposition *на* – 'on' probably because the verb *dotýkat se* – 'concern' was separated from the depending noun by an adverbial *hlavně* – 'in general':

(5.20)

(src) *nová omezení se dotýkají hlavně mladých*
new restrictions refl concerns mainly young
'Those restrictions concern mainly young people'

(goo) *новые ограничения влияют в основном молодых людей*
new restrictions influence in general young people
*'New restrictions influence mainly young people'

We account valency errors not only in cases when Russian and Czech valency frames differ. Quite often statistical systems produce those errors when a clause is complicated (ex. with a non-standard word order or extended complicated structure of a sentence), but the valency frames in the two languages are the same. Another frequent error lies in confusing deep semantic roles, like subject and object in the following example:

(5.21)

(src) *proti schválení zákonů, jež ..., se postavili*
against approval laws-Gen, which ..., refl stand(against)-*Verb*
demokratičtí zákonodárci
democratic-Adj.**Nom** legislators-Noun.**Nom**
'Democratic legislators stand against the approval of laws, which ...'

(goo) *против законов, которые ..., против демократических*
against laws, which ..., against democratic-Adj.**Gen**
законодателей
legislators-**Gen**
*'against laws about passes, which ..., against democratic legislators'

The source sentence has Object-Verb-Subject structure, the object and the verb are separated by an embedded clause. In the translation of google, a verb **stood against** is missing (or, it can be viewed as mistranslated because the preposition *против* – 'against' is often used with this verb) and the subject of a sentence has adopted a morphological marker of an object dependent on the preposition - Genitive case.

The SMT systems generally produce errors in valency when a verb is separated from its argument by some other phrase or even one word. Otherwise, when a respective n-gram without an embedded element is found in the phrase table, the translation is generally correct. The similar observation on valency in SMT was also made in (Rosa, 2013) for Czech-English MT.

RBMT errors in valency are only partially related to some discrepancy in Czech and Russian. The systems also made errors in cases when the valency structure of a verb in Czech and Russian was similar. Those errors might be a result of an improper analysis of the source phrase or some error in transfer or generation phases.

In the next Chapter, we will go deep into the definition of valency. We will also present a theoretical study on discrepancies between Czech and Russian surface verb frames.

5.2.8 Syntactic issues in MT

The following Sections (5.2.9, 5.2.10, 5.2.11, 5.2.12, 5.2.13 and 5.2.14) describe the errors that can be generally referred to as syntactic ones with strong ties to morphology and lexical semantics.

Syntax is claimed to be a real problem for the Phrase-Based Statistical Systems because they do not form a 'core' scheme of a sentence according to a given source pattern, which might be an advantage when translating between free word order related languages. On the contrary, RBMT rarely make an error here because they take this pattern from a source language and due to the similarity in syntax between the two languages, the trick of just copying the source syntactic structure works out. While describing the syntactic issues we will try to answer the following questions:

- to what extent Czech and Russian syntax is similar and what are the major points of discrepancy.
- do the similarities in languages benefit SMT and RBMT translation, or may the discrepancies result in some regular mistakes for SMT and RBMT.

First, we will have a look on the syntactic mistakes (or, rather, mistakes of a more syntactic nature) and try to find possible explanation for them with respect to the two major points stated above.

Again, we should put a disclaimer that the statement 'SMT handles syntax badly' is not quite correct because SMT does not take into account any linguistics. What we consider to be 'syntactic issues' are some concrete syntactic constructions in the source language that can lead to various mistakes in the target (these

can be extra words, morphological or disambiguation errors). That is why there are not so many errors explicitly marked as 'syntactic', and the actual number of errors that syntax can possibly evoke can be much higher.

5.2.9 Word order issues

Russian and Czech are both free word order languages, and we automatically assume that this similarity may help a machine translation and a word order should not be a problem when translating between them. Except for some syntactic constructions that are different in Czech and Russian, the order of sentence units is quite similar.[16] The tag 'word order mistakes' does not indicate that if we change the order of translated words, the error will diminish. So, **word order issues that we describe here indicate rather a cause of a mistake than a mistake itself.**

A marked word order

While analyzing the output, we have found a correlation between a non-standard order[17] of elements in the source sentence and the amount of mistakes in the sentence.

The basic (non-marked) word order in Czech and Russian is 'SVO' - (subject-verb-object). However, especially in the Czech news texts (which is the genre of our test set) the verb quite often occurs in the first position (VSO order). This is not true of Russian which is more inclined to the standard SVO order, so this difference might probably caused a kind of divergence in the phrase tables. This sometimes results in mistakes of different types, see Example 5.22(goo) - the verb is incorrect and its dependencies do not have proper morphological endings. The RBMTs preserved the source order of elements and produced a relatively correct output (Example 5.22(tmt), except for one error that is related to the Genitive of negation, Section 5.2.6):

(5.22)

[16] Consider an example of French adjective postposition which seems to be an ideal candidate for the word order discrepancy relevant for the MT. So, for instance, when some English-to-French system produced an adjective-noun sequence following the source pattern, this may indicate a mistake. For Czech and Russian no such evident order discrepancy exists, but a lot of minor syntactic constructions connected more to lexical issues are different.

[17] Here, we will use the notion 'marked', also non-standard word order for every constituent order other than SVO. More on theoretical aspects of word order and its connection to information structure of a sentence can be found in the Section 5.2.9 (Word order-theory).

(src) *nemají* *američtí* *občané* *průkaz*
not-have-3Plneg american-adj citizens id
'American citizens do not have an ID'

(goo) *не являются гражданами США удостоверение*
not are citizens-Pl USA ID
*'Are not the US citizens ID'

(tmt) *не имеют американские граждане удостоверение*
not have american citizens *ID-Nom/Acc
'American citizens do not have an ID'

Long-distance dependencies

Another challenging issue is when a sentence contains **long-distance dependencies**: the main arguments of a verb - subject and object - are separated by some adjuncts or embedded clauses and thus stand relatively far from each other and a verb[18]:

(5.23)

(src) *nevlastní* *11%* *amerických* *občanů,* *tj.* *21* *mln* *osob* *ve*
not-possess 11% american citizens, resp. 21 mln persons in
věku *umožňujícím* *volit,* *žádný* *průkaz* *totožnosti* *s*
age allowing vote-inf, no id personality with
fotografií
photograph
lit. '11% of American citizens, or 21 mln people in a voting age, do not possess an ID card with a photograph'

(goo) *принадлежит 11% граждан США, то есть 21 млн человек*
belong 11% citizens USA, resp. 21 mln people
в возрасте не может быть выбран, нет удостоверения
in age not can be elected, not-have id
личности с фотографией
personality with photograph
*'belong to 11% US citizens, or 21 mln people in age that can not be elected, not possess an ID with a photograph

[18] In this example we will hide the tags of other mistakes - like disambiguation or morphological errors, just to demonstrate what a mess can be caused be a complex word order. We believe, though, that those mistakes resulted because of the too complex structure as well

In the example above, a label 'word order mistake' can not be attached to some distinct word, rather a whole sentence (phrase) should be marked as incorrect, as almost all the words are confused, and the output presents a meaningless 'bag of words'. In this very case, Rule-Based systems, though with its typical mistakes like disambiguation or unknown words, produced more meaningful output as they preserved the source syntactic structure.

Embedded clauses

Embedded clauses - like relative clauses or transgressive constructions - present a challenge to machine translation as they generally interfere into the predicate-argument structure of a sentence. Embedded clauses are closely connected to the previous point as they are generally the reason of long-distance dependencies.

We consider a clause to be embedded when it separates the main arguments of the sentences or a verb, e.g. when it is situated between a subject of a sentence and a predicate like in the Example 5.24(src).

(5.24)

(src) *prvním* *státem,* *který* *tento* *požadavek* *zavedl,*
first-Ins state-Ins , which this demand
byla *indiana*
introduced-**finite.Past**, was indiana
'Indiana was a first state to introduce this demand'

(mos) *первым* *государством,* *который* *привел* *к,* *была* *индиана*
first-Ins country-Ins, which came to, was indiana
*'First state which came to was Indiana'

(goo) *первым* *государством,* *ввести* *это* *требование* *было*
first-Ins state-Ins, introduce-**inf** this demand was-**neut**
Индиана
Indiana-fem
*'First state, to introduce this demand was Indiana'

(pct) *первым* *страной,* *какой* *этот* *запрос* *ввести,* *была*
first country, which this demand introduce-**inf**, was
индий - *Диана*
india - Diana
*'First country which this demand to introduce was Indy-Diana'

(tmt) *Первым государством, которое это требование ввести,*
first state, which this demand introduce-**inf**,
она была индиана
she was Indiana
*‘First country which this demand to introduce, she was Indiana’

In this example[19], the finite Czech verb *zavedl* – ‘introduced’ from a relative clause was translated in its infinitive form in three systems (goo, tmt and pct), but generally, the syntactic structure produced by the RBMT was more correct than that from SMT. Note also that the structure of (goo) resembles an English one, as google translates via this pivot language.

However, mistakes generally do not occur when a relative clause depends on a last element of the main clause.

Adversative clauses

Adversative constructions containing words like ‘however’, ‘but’ etc. have tendency to be mistranslated, which can be justified by a discrepancy between Czech and Russian. In Czech, the word **ale** - ‘but’ as a connector is not fixed in a sentence in the first position like in Russian or English, its function is to topicalize a subsequent word/phrase.

Translation of adversative constructions depends on a sentence structure. If it is not very extended, the conjunction stands not far from the beginning of a clause and the surroundings of the conjunction are frequently used words, statistical systems translate the phrase properly - with ‘but’ in the first place, and the RBMTs failed to do this as a respective rule was not present or properly applied:

(5.25)

(src) *když ale vidím něco*
when but see-1Sg something
‘But when I see something...’

(mos) *но когда я вижу то*
but when I see this
+‘But when I see this...’

[19] Just to mention, all the systems failed to properly disambiguate the word ‘stát’ - state which is ambiguous in Russian - it can refer both to a *country* and to one of a state - *province* in Czech and English, but Russian has two distinct words for it - *государство/страна* – ‘country’ and *штат* – ‘province’.

(goo) *но когда я вижу что-то*
but when I see something
+'But when I see something..'

(pct) *когда а вижу что*
when and see what
*'When and I see what'

(tmt) *Когда но я вижу нечто*
When but I see something
*'When but I see something'

If a sentence is extended and a conjunction is situated between a subject and a predicate, or if it stands far from the beginning of the clause as in the Example 5.26, the statistical systems fail to translate a construction properly (as well as the Rule-Based that do not have the specific rules) because the respective n-gram with a conjunction *ale způsobila* – 'but caused' was, evidently, not seen in the training data.

(5.26)

(src) *opatření přijatá po roce 2009 ale způsobila pokles..*
measures taken after year 2009 **but** caused decrease..
'the measures taken after 2009 caused a decrease'

(mos/goo) *меры, принятые в 2009 году, но вызвало падение*
measures, taken in 2009 year, **but** caused decrease
*'Measures taken in 2009 but caused a decrease'

(pct) *осторожное прийата спустя год 2009 а зпусобила*
*careful *unk after *year 2009 and caused
опускание
lowering
*'Careful prijata after year 2009 **and** caused lowering'

(tmt) *Меры принятый по году 2009 но, привести падение*
measures *taken after *year 2009 **but**, *cause decrease
*'Measures taken after year 2009 but, to cause a decrease'

We have checked the whole test set for adversative clauses. In 13 test sentences (out of 3000) a word *ale* – 'but' is not first in the clause (sentence). While RBMTs failed in all cases, Moses made only one mistake, and Google three mistakes - putting 'but' in other than the first position. The same mistakes occur in the translations of the adversitive conjunction *však* - 'however', but it has a bit different usage and behaves like a clitic in Czech, see the next Section.

Sentential clitics

We can say that word order in Russian is less strict than in Czech especially because of the clitic[20] position in a sentence: Czech obeys 'a law of second position', or Wackernagel's law which does not apply in Russian language. In Czech, some clitics are often required to move to the position after a first word/phrase in a sentence. Following are the most frequent classes of clitics subjected to the law:

- **Weak pronouns**, e.g. mi (for me), ti (for you), mu (for him).
- **Reflexives (pronopuns or particles)** se, si
- **Auxiliary clitics** Auxiliary verb 'to be' (jsem, jsi...jsou), conditional auxiliary (bych,bys...by)
- **-li**

In (Hana, 2004), it is illustrated how clitics in Czech have a fixed position not only in the sentence, but also in relation to each other. If more clitics occur in the same cluster[21], they will have a predefined fixed ordering: 1. -li 2. auxiliary 3. reflexives 4. weak pronouns. This is even more complex, because there is also a fixed order for weak pronouns in different cases. Another complication may be the attachment of an auxiliary 'to be' in second person singular (**jsi** contracted to **s**) to a verb (*Přišel**s** pozdě* – 'You came late'). Or, if if a verb is reflexive, to a reflexive particle *Umyl si**s** ruce?* – 'Have you washed hands?' .

All of the four items presented in the list above demonstrate some difference in Czech and Russian. Russian pronouns do not have a weak form like Czech ones. Reflexive particles are incorporated into verbs in Russian, which makes a really huge difference when translating sentences where the reflexive particle stands far from the governing verb. The particle '-li' is presented in Russian as a distinct lexeme - 'если'. Many forms of auxiliary clitical verbs are not used in Russian, and conditional auxiliary 'to be' has only one form. If more than one clitic occurs in a Czech sentence (such called clitic cluster), then the structure of a Russian sentence often looks very different. Following is an example from the training data for Moses:

(5.27)

(cs)	*Podařilo*	*by*	*se*	*mu*	*odejít*
	succeeded	would	refl	him	leave

[20] Here we will talk about sentential clitics only.

[21] Multiple clitic clusters can also occur in a sentence, but we are interested in second position clitics.

'He would succeed to leave'

(ru) *Ему удалось бы уйти*
him succeeded-refl would leave
'He would have succeeded to leave'

As the translation of clitics is done in a different way, the Czech n-gram *by se mu* – 'would refl him' will be associated with various n-grams in Russian in the training data for **SMT**.

In the example below, we show that the translation of the clitic cluster from the above example was not satisfactory in any of the four systems:

(5.28)

(src) *to by se mu mohlo vymstít*
this be-conditional refl him could backfire
'This could backfire on him.'

(mos) *это могло бы ему табло*
this could be-conditional him board
*'This could him board'

(goo) *что он может иметь неприятные последствия*
that he can have unplesant consequences
*'That he can have unpleasant consequences'

(pct) *то чтобы с ему могло поплатиться*
this that with him could pay
*'This in order to with him could payed'

(tmt) *To ему вымстит*
This him *unknown-word
*'This to him вымстит'

GoogleTranslate was the closest to the right variant, whereas Moses, PC-Translator and TMT generated completely inaccurate translations.

Translation of reflexive verbs alone without other clitics is also quite challenging and often results in an error:

(5.29)

(src) *které se mohou stavět*
which se-refl can build
'which can be built'

(mos) *которые могут строить*
which can build
*'which can build'

(goo) *которые могут построить*
which can build
*'which can build'

(pct) *которой с могут сооружений*
which with can building
*'which with can bildings'

(tmt) *которые ставет-се*
which *unknown-word
*'which stavet-se'

As the reflexive particle *se* and the verb are separated by another word in this sentence, the SMT systems translated the reflexive predicate (reflexive passive voice) as a non-reflexive(active voice) which changed the sense of the sentence. **RBMT** systems that we consider have rules to cope with some clitics, but they do not work sometimes even for the single standing clitics, possibly due to problems in analysis/parsing. In Example (5.29), PC Translator misleaded the reflexive pronoun *se* with a preposition *se* – 'with'[22]. TectoMT recognized the reflexive verb properly, but failed to find a translation equivalent.

As with other mistakes, it is impossible to predict where an SMT will make a mistake with clitics and when it will cope with it. Generally, the more frequent the reflexive verb is (or if it is used in collocations), the more it has a chance to be translated properly. RBMT systems should have deeper and more sophisticated rules to handle clitics. We can also suppose that clitics will be more of a challenge when translating into Czech because of these many rules that they are subjected to.

Other mistakes related to word order

Actually, all the mistakes that were marked as 'word order' in Statistical systems can be attributed to a wrong choice of an n-gram phrase, see Example 5.30. In most cases, they are not connected to some real discrepancy in Czech and Russian word order. Such illogical mistakes do not occur in the RBMT just because the basic order of elements is preserved.

(5.30)

[22] This mistake occurs always!

(src) *tato opatření částečně podkopou americký*
this measures partially undermine **american democratic system**
demokratický systém

'This measures will partially undermine the US democratic system'

(mos) *эти меры частично сша подорвали демократическую*
this measures partially **USA** undermined **democratic system**
систему

*'This measures partially USA undermined democratic system'

As for the RBMT, all the word order errors that we came across reflected one very specific construction : in Russian language a phrase 'year xxxx' is used in a reversed order in comparison with Czech or English:

(5.31)

(ru) *до 2004 года*
to 2004 year
'up to year 2004'

(cz) *do roku 2004*
to year 2004
'up to year 2004'

Both Statistical systems translated it properly - This discrepancy, however, can be easily introduced as a rule as in the case with contradictive constructions. We have fixed it in TectoMT system with a block FixDateTime.pm. The order of the two constituents - 'year' and a digit was changed, so this temporal construction is now translated correctly. Like always, the BLEU score was only a bit higher - it increased from 9.4% to 9.55%.

Word order - theory

Having shown some syntactic constructions where errors can most probably occur, we will now discuss the theoretical question of word order and provide a comparative analysis of the phenomenon for the two languages.

In Slavic languages, the word order is stated to be connected with the information structure of a sentence. The latter was extensively described under the Functional Generative Description theory developed under the Prague School (Hajičová et al., 1998). The information that is known (thema, or topic) generally precedes the new information (rhema, or focus). We will not give here exacts

definition of topic, focus and sentence boundness as we believe this is not relevant for our task of a rather shallow overview.

The basic participant order in Slavic languages (except for Lower and Upper Sorbian) is considered to be SVO, it is an unmarked and the most frequent order. Next, we will demonstrate that constituent order in Czech and Russian can be interpreted in the same way in terms of information structure.

Consider the following examples, where we give some interpretation to a pragmatic structure of sentences. As more readings are possible for each constituent order, and it is impossible to interpret the latter without a broader context, we suggest a sentence being an answer to a concrete question. (other readings are possible). Topic is in italics, focus is in bold and uppercased:

(5.32) *Сегодня Петр идет в* **КИНО**
Dneska Petr jde do **KINA**
'Today Petr go to the **cinema**'.

This example illustrates that Petr has some plans for today, and the focus of the sentence reveals those plans. This sentence can be an answer to the question *Where does Petr go today?*

(5.33) *Петр идет в кино* **СЕГОДНЯ**
Petr jde do kina **DNESKA**
'Petr go to cinema **today**'.

Here the known information is that Petr will go to the cinema, the last part gives us information when exactly he will do so - the focus is on an adjunct 'today'. The respective question is *When does Peter go to the cinema?*

(5.34) *Сегодня идет в кино* **ПЕТР**
Dneska jde do kina **PETR**
'Today go to cinema **Petr**'.

This example can be interpreted that going out to the cinema is scheduled, and today is Petr's turn ('Petr' is a focus), the possible question that evokes this answer can be *Who goes to the cinema today?*

Examples above suggest that generally the constituent order in Czech and Russian is very similar. Word-for-word translation is therefore quite acceptable for the RBMT, taking into account the syntactic issues described in the Sections above. Still, other problems may arise if Czech and Russian use different syntactic constructions that are connected with concrete lexical issues, and it is virtually impossible to take into account all those minor discrepancies.

5.2.10 Constructions with a verb 'to be'

The verb 'to be' has many meanings (copulative, existential, auxiliary, modal etc.) and has many translation equivalents in the languages. Here, we will discuss those functions of the verb 'to be' which, according to our data, impact the MT output most.

Copulative meaning

Zero copula is such a striking characteristic of Russian language in contrast to Czech and many other languages, that it was the very first thing to write a rule for when constructing the RBMT. Naturally, we want to test how MT systems handle this construction. Such called copula drop, or zero copula is a phenomenon when a verb 'to be' in its copulative meaning[23] is left out of the sentence. It does not exist in Czech, English, and, actually, in most European languages.

Copulative constructions can have several realizations in Czech and Russian, the basic one is presented in the example 5.35 (cz-1) and (ru-1), but there are more translation variants. The form with instrumental case (cz-2) can be sound more formal than (cz-1) in Czech, and the construction with Instrumental (ru-2) is definitely formal in Russian. This discrepancy is true only for a copula in Present Tense, in the Past Tense the verb 'to be' - 'byl(a/o)' is present in both languages and expressed in the same way with the same cases.

(5.35)

(cz-1) *Jsem* *student.*
be-1SgPres student-Nom.
'I am a student'

(cz-2) *Jsem* *studentem*
be-1SgPres student-Ins.
'I am a student'

(ru-1) *Я (-)* *студент.*
I-1Sg student-Nom.
'I am a student'

(ru-2) *Я* *являюсь* *студентом.*
I be(very official) student-Ins.
'I am a student' (very officially)

[23] 'to be' in its auxiliary function is also dropped from the sentence, but not 'to be' in its existential meaning, see the discussion further in the text.

We will now make some remarks on diachronic linguistics. Old Russian language (or, earlier, Old Church Slavonic) had a copulative construction as well as other Slavic, but it disappeared from the language. Some researchers (like (Clancy, 2010)) claim this phenomena to be a structural calque from Finno-Ugric, especially Hungarian or Hungarian-like. A tight contact with such different non-Indo-European languages can be the reason that Russian has too many grammatical discrepancies with other Slavic languages. The latter kept more or less in contact with each other[24] so there are less dissimilarities between them.

Next, we will show several cases when MT systems made some mistakes while translating a sentence with a copula 'to be'. Statistical systems sometimes translate copula constructions properly and sometimes not. This presumably can be attributed to the fact that there is no unified translation pattern and different variants (see examples from 5.35) can occur in various combinations in the training data. In the following sentence, one mistake in a case (usage of Instrumental instead of Nominative) occured when a verb was not present:

(5.36)

(src)	*toto*	*higgsovo*	*pole*	*je*	*mnohem,*	*mnohem*	*stabilnější*
	this	higgs	field	is	more,	more	stable-**Nom**
	'This Higgs field is more more stable'						

(goo)	*поле*	*Хиггса*	*гораздо,*	*гораздо*	*более*	*стабильным*
	field	Higgs-Gen	much,	much	more	stable-**Ins**
	'Higgs field is much more stable'					

In this example, the usage of Instrumental can be connected to a word *являться* – 'to be - formal' (Example 5.35 (ru-2)) which is not used in the translation, but its dependent (either noun or adjective) should be in Instrumental case. The same mistake in case was made by a TectoMT system. This can be justified only by the Instrumental case from the source Czech and the absence of a rule to cope this discrepancy. PC Translator, on the contrary, used the word *являться* with a proper Instrumental case, but a pronoun 'they' is absent (also a calque from the pro-drop Czech):

(5.37)

(src)	*že*	*jsou*	*americkými*	*občany*
	that	are	american	citizens
	'that they are American citizens'			

[24] Bulgarian presents the second exception since it also contacted much with non-Indo-European languages like Turkish

(pct) *что являются американскими *общаны*
that be-official american-Ins citizens-Ins
'That are american citizens'

(tmt) *что они американскими гражданами*
that they american-Ins citizens-Ins
'That they american citizens'

Again, if the respective n-grams are frequent, SMT systems usually translate copula properly whereas RBMT made comparatively more mistakes.

The **negation of copula** is also a big challenge for the MT systems that have to deal with two discrepancies at once. First, several ways of translating copulative constructions interferes with different surface realization of negation in Czech and Russian leading to rather frequent (according to our data) mistakes in all the four systems:

(5.38)

(src) *od roku 2008 aids není rozsudkem smrti*
from year 2008 aids not-be sentence-**Ins** death-**Noun.Gen**
'Since 2008, AIDS is not a sentence of death'

(mos) *с 2008 спида не смертным приговором*
from 2008 aids-Gen not death-**Adj.Ins** sentence-Noun.Ins
*'Since 2008 AIDS not a death sentence'

(goo) *с 2008 года, СПИД является не смертный*
from 2008 year, AIDS is-official not death-**Adj.Nom**
приговор
sentence-**Noun.Nom**
*'Since 2008, AIDS is not a death sentence'

(pct) *от году 2008 аидс нет смертный приговор*
from *year-Loc 2008 aids there-is-not death sentence
*'Since year 2008, AIDS there is not death sentence'

(tmt) *от года 2008 аидс он не приговор смерти*
from *year 2008 aids *he not death sentence
*'Since year 2008 aids he is not death sentence'

In the latter example all the systems made different mistakes: Moses evidently used an Instrumental case from the source sentence with a copula which will be otherwise correct with the verb 'являться'.

Google has it opposite to Moses: the verb *являться* was used, but the Instrumental case of a dependent that is required with the verb was not. Moreover, the ordering of a verb *являться* and a negation particle is reversed (negation particle should come before a verb). Those errors can be justified by the fact that Google uses English as a pivot language to translate between Czech and Russian. The mistakes where a noun phrase is in its base form, the negation particle stands after a verb 'to be' (**is not**) and a comma comes after an adjunct phrase proves this fact.

PC Translator wrongly used a negative predicate construction with *нет* – 'there is not' that is generally used in existential sentences, but not copular ones where a negation particle 'не' should be used instead.

TectoMT output contained a proper negation particle because the respective rule was applied. On the other hand, it has an extra pronoun *он* – 'he' that was inserted because the parser did not recognized and identified a subject of a sentence and the sentence was treated like a pro-drop.

So, we can see that the discrepancy in copular constructions between the two languages presents a challenge both for SMT and RBMT because several factors are here at play. In affirmative sentences, depending on a translation variant for a copula (either zero copula or 'являться'), different case should be chosen.

As for RBMT, the existing rules are not enough to cover this phenomenon, so the new more complex ones should be written. The most simple decision would be to stick to one 'basic' copula-drop style in the target Russian and to translate every copular construction as the sentence 5.35 (ru-1) always changing a case to a Nominative. Again, a parser should recognize the 'to be' as a copular verb and not mix it up with the existential 'to be' which will be described further.

After this error analysis, we tried to include a respective rule into the TectoMT system, see Section 4.3. In baseline system, the copula verb was dropped, but the case of a predicative noun stayed the same as in the source - Instrumental(Example 5.37(2)). First, we tried to fix this error according to the sentence from Example 5.35(ru-2) - substituting the Czech copula with a verb *являться* without changing a case. The bleu score was not changed, and the translation were grammatical, but sometimes not fitting the language style (too official). So we decided to fix it according to the pattern 5.35(ru-1) - without a verb, just changing the case from Instrumental to Nominative, again the BLEU score changed only 0.001%.

As for the SMT systems, such a variety of surface realization and exceptions on both sides can in some cases lead to the improper handling of the phenomena. One of the possible decisions to this concrete problem can be an introduction of a post-editing rule on the same principle as for the RBMT.

Auxiliary 'to be'

The similar problem concerns the Past Tense formed by the means of the auxiliary 'to be' in Czech which is dropped in Russian (again, it was used in Old Russian and Old Church Slavonic).

(5.39)

(cz) *Včera* *jsem* *byl* *doma*
Yesterday be-**1Sg**Aux be-PastParticipleSg home
'I was at home yesterday'

(ru) *Вчера* *я* *был* *дома*
Yesterday I-**1Sg** be-SgPastParticiple home
'I was at home yeaterday'

Following is an example of how the Past Tense is translated:

(5.40)

(src) *všichni* *jsme* *bojovali* *pro* *tým*
all be-Verb.Aux.2Pl fight-Past for team
'We have all fought as a team'

(mos) *все* *мы* *сражались* *для* *команды*
all we fought-Past for team
?'We all fought for a team'

(goo) *мы* *все* *боремся* *за* *команду*
we all *fight-Verb.**Pres** for team
*'We all fight for a team'

(pct) *мы* *все* *воевали* *для* *команда*
we all fought for *team-noun.**Nom**
*'We all fought for a team'

(tmt) *Все* *боролись* *для* *команды*
all fought for team
?'All fought for a a team'

From the example, we can see that Statistical systems can handle the Past Tense properly (Moses), but not always (Google). Rule-Based translated the verb in a correct tense, but either the preposition or a case of the following noun phrase was wrong.

'to be' - existential meaning

'To be' in its existential function (like in the sentence *There are bears on Red Square*) is mostly present in Russian sentence as well as in Czech. (Barbara H. Partee, 2002), argue that on the surface, the border between existential and subject-predicate (copulative) sentences in Russian language is very vague. It is connected to the Theme-Rheme structure which, to the best of our knowledge, is not at all on the agenda of MT systems as a very non-trivial problem, especially for free word order languages.

Let us have a look at the translation of an existential sentence 5.41 (*There are good consultants in Moscow today*). It was translated as a subject-predicate sentence (*Today in Moscow are good consultants*) by TectoMT, Google and Moses and this translation can be viewed as grammatically acceptable.

That made more trouble for PC Translator that had chosen an autosemantic verb *являться* – 'to be -formal' as translation of constructions for the verb 'to be'. The latter verb has a very limited usage, an it does not have an existential component in its meaning. So using this verb in this context is completely wrong. On the contrary, just leaving out a verb - what was done in the TectoMT - brings more good-looking results, ex. 5.41(tmt). Statistical systems produced the same (almost proper) translation as TectoMT.

(5.41)

(src) *v* *moskvě* *jsou* *dnes* *dobří* *konzultanti*
in Moscow are today good consultants
'There are good consultants in Moscow today'

(mos/goo/tmt) *В* *москве* *сегодня* *хорошие* *консультанты*
in Moscow today good consultants
'Today in Moscow are good consultants'

(pct) *зажечься* *москве* *являются* *сегодня* *добра* *консультанты*
*fire *Moscow are-formal today *good consultants
'Fire Moscow are today good consultants'

However, to properly transfer the sense of the source sentence, a verb *есть* – 'to be' should be used (*В москве сегодня* ***есть*** *хорошие консультанты* – 'Today in Moscow **thare are** good consultants').

5.2.11 Pronoun usage

Another discrepancy between Czech and Russian that has an effect on MT quality concerns pronoun usage. It is closely connected to the copula drop phenomena

described above: the morphological categories of gender and person should be obligatorily expressed in both Czech and Russian[25], but the languages chose different means to encode it. Russian is inclined to use a pronoun, like in the example 5.35(1-ru) whereas in Czech this information is encoded in an auxiliary verb 'to be', see ex. 5.35(1-cz)[26].

To illustrate this discrepancy we have calculated the statistics of pronoun usage in the parallel Czech-Russian corpus (Klyueva – Bojar, 2008). In the Table 5.2.11 we show that for the same sentences the usage of personal pronouns in Russian language is far more frequent than in Czech.

	Russian	Czech
ja(I)	5433	143
ty(you-singular)	24	8
on/ona/ono(they)	5102	264
my(we)	2368	462
vy(you-plural)	334	18
oni/ony(they)	4131	167

Table 5.2: Pronoun usage in Czech and Russian

Due to such a disproportion of pronoun usage, statistical systems often fail to suggest a hypothesis with a pronoun, but it depends on the frequency of the respective n-gram and the chosen translation paths. For example, the following sentence 5.42 was translated by Moses properly, with a pronoun, but not by Google. TectoMT has a special rule for it, so the translation is also proper with respect to the pronoun. PCTranslator's output does not contain the pronoun at all.

(5.42)

(src)	*tvořili*	*pouhých*	*11%*	*ze*	*všech*	*voličů*	
	constituted-Past3Pl	only	11	%	from	all	electors
	'They constituted only 11% of all the electors'						

[25] This information is applicable to copular sentences and to the sentences with a verb in the Past Tense

[26] In the third person, an auxiliary is not used in Czech as it is not needed there - when in the sentence the past form of a verb (past participle) is present without the 'to be', this signifies the third person either singular or plural

(mos) ***они*** *составляли* *всего* *11* *%* *из* *всех* *избирателей*
they-3Pl constituted-PastPl only 11 % from all electors
'They constituted only 11% of all the electors'

(goo) *приходилось* *лишь* *11* *%* *всех* *избирателей*
constituted only 11 % all electors
*'They constituted only 11 % of electors'

(pct) *создание* *один* *11* *%* *из* *всех* *волицу*
creation one 11 % of all *unknown
*'Creation one 11 % of all *unknown'

(tmt) ***Они*** *составили* *только* *11* *%* *из* *всех* *избирателей*
They constituted only 11% of all electors.
'They constituted only 11% of all the electors'

It is evident that in this case, the best system to deal with this phenomena will be TectoMT which has a rule for it. However, this rule produces 'false positives' and inserts pronouns when they are not necessary. This happens especially when a sentence is not parsed correctly or a subject is not recognized at all. In the following example, evidently due to the parsing error, a subject was confused with an object, and the subject pronoun was wrongly inserted into the sentence[27]. This error can also belong to the word order/valency or parsing issues.

(5.43)

(src) *v* *neděli* *ráno* *začíná* *pracovní* *týden*
on Sunday morning starts working week
'On Sunday morning a new working week starts'

(tmt) *в* *неделю* *утром* ***он*** *начинает* *рабочий* *неделя*
on Sunday morning he starts new week
*'On Sunday morning he starts a new week'

Having analyzed the output of the statistical systems with the regard to the pronoun usage, we have not found any regularity where the pronoun is or is not present - so that is more or less the matter of a chance and frequency, see example 5.42 (goo) vs. (mos). So we do not really have a suggestion how to address pronoun drops for this type of systems. As for the RBMT, more sophisticated rules should be written in order to avoid inserting pronouns in the sentences where they are not needed.

[27] The most frequent extra words in TectoMT output are pronouns.

5.2.12 Constructions with 'to have'

The verb 'to have' is also an ambiguous verb as well as 'to be' and it challenges the MT systems to the same degree. In some languages, the verb 'to have' can be substituted with the verb 'to be' when expressing possession. ('somebody has something' vs. 'to somebody is something'). Languages thus can be classified into 'to be' languages and 'to have' languages. Russian belongs to the 'to be' group of languages and stands out of the Slavic and Indo-European 'to have' languages.

Generally, the Czech *mít* – 'to have' construction is translated into Russian as *У кого-л. есть* – 'To smb have' (see Example 5.44). The other variant - *иметь* – 'to have' - is also acceptable in collocations or light verb constructions *иметь право* – 'have the right', sometimes it is also used in very formal written language (e.g. *Он имеет высшее образование* – 'He has a University degree'.

(5.44)

(cz) *Mám* *doklady.*
Have.1Sg documents
'I have documents'

(ru-1) *У* *меня* *есть* *документы.*
For-Prep me-GenSg is documents
'I have documents'

(ru-2) *?Я* *имею* *документы.*
I have documents
'I have documents'

Let us have a look on how statistical systems handle this phenomena (Example 5.45). Again, as with the verb 'to be', the variety of translation equivalents leads to more hypotheses from which the decoder have to choose. Mostly, SMT systems tend to translate Czech possessive constructions with 'mít' as 'иметь' which is is sometimes acceptable in formal contexts. This can be attributed to the fact that the majority of training data come from the domain of news, and news articles are written in a formal language. As our test data are also from the same domain, in the majority of cases this translation is (almost) acceptable.

As for the RBMT systems (Example 5.45), TectoMT does not have a rule to transfer the verb 'mít' into 'у меня есть', so the translation of a possessive phrase mostly corresponds to the Statistical output and is relatively acceptable. PC Translator, on the contrary, has some rule for handling 'to have' constructions, but in many cases it confuses a possessive (have smth./smb.) and a modal (have

to do smth.) functions of a verb 'to have' translating a possessive verb in its modal meaning:

(5.45)

(src) *kteří mají legální postavení.*
who have-Pl legal status
'who have legal status'

(mos|goo|tmt) *которые имеют законный статус.*
who have legal status
'who have legal status'

(pct) *которая должны законный должность.*
who must-Pl legal position
*'Who must legal position'

However, there are more variants that were used by the systems to translate the possessive construction. As the respective n-gram is frequent, the Czech 'mít' was translated by the appropriate 'to be' variant (Example 5.46) by Moses, though Google made a minor mistake leaving out a subject-possessor, so the sentence became impersonal:

(5.46)

(src) *máte majetek za 60 miliónů*
have-2Pl property for 60 millions
'You have a property for 60 millions'

(mos) *у вас есть имущество за 60 миллионов*
for you is property for 60 millions
'You have a property for 60 millions'

(goo) *есть активы на сумму 60 миллионов*
is property for sum 60 millions
'There are assets for a sum'

PC Translator, evidently, also has a rule to transfer *mít* into *у ... есть*, but in most cases it is not applied properly, see Example 5.47. The transformation (mít -> 'у него есть'), though it was in the right place, is presumably very shallow and does not handle the subject-possessor properly. In couple with the marked word order (VOS), unrecognized words and negative polarity, this made the phrase constituents change their semantic roles, so the cases were confused and the verb has a spare actant 'него'.

(5.47)

(src) *nemá* *průkaz* *totožnosti* *pět* *milionů* *nových* *voličů*
not-have-Pl card identity-gen 5 mln new-gen electors-gen
'5 mln of electors does not have an ID card'

(pct) *у* *него* *нет* *пруказ* *тождественность* *5* *млн* *волицу*
for him not *card *id 5 mln *electors
*'He does not have card id 5 mln electors'

So, we can see that statistical systems are unpredictable, they can handle the construction properly, but may not; those mistakes are not easy to identify and to fix automatically. As for RBMT, the rules should be written more carefully taking into account many other factors - like auxiliary, modal usage or polarity and some other word order aspects.

5.2.13 Transgressives

Transgressive is a non-finite form of a verb[28] that expresses an action done simultaneously with/or right after the action of the main verb. We have encountered several mistakes concerning transgressives in SMT systems, e.g. when a transgressive was used instead of an appropriate part of speech:

(5.48)

(src) *zákony* *vyžadující* *předložení*
laws demanding-ger presentation-**noun**
'Laws demanding demonstration'

(mos) *законы,* *требующие* **продемонстрировав*
laws, demanding-ger *presenting-**transgressive.Past**
*'Laws demanding demonstrating'

Possible explanation can be the fact that sentences with transgressives show differences in Czech and Russian which is then reflected in the training data and, consequently, in the phrase tables.

In Czech, transgressives are considered to be archaic, whereas in some other Balto-Slavic languages, like Russian, Polish or Lithuanian, they are used rather frequently especially in the official style and in news.

[28] The terminology on transgressives is not unified: in Slavic linguistics they are generally addressed as transgressives, the closest English term for it is **gerund**, and English gerund is very close to what a Slavic transgressive is. Here, we will operate with both notions.

Formation of transgressives in Czech is more morphologically complex, as the form agrees with the actor of a main clause in number and in gender. Polish and Russian gerunds are not that complex, they have only one form for all numbers and genders, so there are 3 forms for Czech gerund and only one in Russian for either past or present tense. The system of Czech transgressives is more complicated than that in Russian also due to the fact that there are two paradigms of gerund declension (a/ouc/ouce vs. e/-íc/-íce)[29].

As we have shown in (Klyueva, 2013), transgressives are used 40 times as much in Russian than in Czech. The most frequent equivalent constructions to Russian gerunds in Czech are dependent clauses (Example 5.49 or coordination clauses (Example 5.50)[30]:

(5.49)

(cz) *učinil* *nejsmělejší* *krok* , *když* *odvolal*
(he) made boldest step , when (he) removed-**Verb**-Past-3Sg
'took his boldest step , when he removed ...'

(ru) *сделал* *смелый* *шаг* , *отстранив*
made bold step , removing-**Gerund**-Past
'took his boldest step , removing'

(5.50)

(cz) *Mozart* *se* *vzdálil* *a* *ponechal* *Nicholase* *o* *samotě.*
Mozart refl gone and left-**verb-fin** Nicholas about loneliness
'Mozart went away and left Nicholas alone'

(ru) *Моцарт* *удалился,* *оставив* *Николаса* *наедине.*
Mozart gone, leaving-**gerund** Nicholas alone.
'Mozart went away, leaving Nicholas alone'

The fact that gerunds are translated in various ways results in some uncertainty in the phrase table which, in turn, can sometimes lead to a wrong translation hypothesis.

[29] This complexity might be a reason why they have disappeared from the language - native speakers just stopped to use them

[30] English language also uses much more transgressives in comparison with Czech. Actually, the news commentaries were originally translated from English into Russian, and a gerundive English construction is almost always translated into Russian as a gerund (transgressive)

We have not noticed the gerund mistranslations in the output of the RBMT just because the source language - Czech - almost never uses them apart from the lexicalized transgressives (ex. *takříkajíc* – 'so speaking').

We can suggest that for the opposite direction - from Russian into Czech - the translation of transgressives will be a more challenging task.

5.2.14 Some other syntactic constructions

POZNAMKA: Tou kapitolou jsem si nejsem uplne jista, mozna by stalo za to ji vynechat?

In the few points above, we have demonstrated the most frequent, evident or regular mistakes that are justified by some syntactic discrepancies between the source and the target languages. However, many mistakes that we have marked as *syntactic* can not be related to any category as they are very specific to a concrete language and are generally related to some lexical issue. Let us show just a few concrete constructions that are often translated wrongly.

The Czech construction from the Example 5.51 is specific for Czech but not for Russian (or English), which makes a word-for-word translation syntactically incorrect. The Czech connector *s tím, že* has several functions in the sentence, and, thus, several translations, which presents a challenge for both SMT and RBMT.

(5.51)

(src) *bouchli* *dveřmi* *s* *tím,* *že* *nemají* *možnost* *prosadit*
slammed-Pl doors with that, that have-not possibility enforce
své *názory*
their views
'They slammed the door as they did not have an opportunity to establish their views'

(mos) *хлопнула* *дверью* *с* *тем* , *что* *у* *них* *нет*
slam-Sg.Fem doors with that, that for them not possibility
возможности *отстаивать* *свои* *взгляды*
maintain their views
*'(She) slammed the door with that, that they do not have a possibility to maintain their views'

(goo) *хлопнул* *дверью* *с* *ним,* *они* *не* *в* *состоянии*
slammed-Sg.Msc doors with him, they not able defend
защищать *свои* *взгляды*
their views

*‘(He) slammed the door **with him, they** were not able to defend their views’

(pct) *барабанить дверми с тем, что не имеют возможности*
beat doors with that, that not have possibility
продвинуться свой взгляд
move-ahead his view
*‘To beat the doors with that do not have possibility to move ahead his view’

(tmt) *они бухнуть дверьми с тем, что они не имеют*
they swelled doors with that, that they not have
возможность просадит его мнения
possibility *unrecognized his views
‘They swelled doors with that they do not have a possibility to *unk his views ‘

The construction *u ... tomu bylo stejně* – ‘for ... this was the same’, also caused errors in the output of all the four systems. This construction does not have a direct translation equivalent, it generally corresponds to some phrase like, e.g. *то же самое происходило с* – ‘the same happened to’.

(5.52)

(src) *u hispánců tomu bylo stejně*
for hispanic that-Dat was same
‘It was the same for the latin people’

(mos) *у латиноамериканцев не было*
for latin-american not was
*‘Latin Americans did not have’

(goo) *для выходцев из Латинской Америки это было то же*
for immigrants from Latin America it was that very
самое.
same.
?‘For Latin Americans it was the same’

(pct) *у hиспанцу тому было одинаковая*
for unknown-word that-Dat was same
‘for *unk that was the same’

(tmt) *У гиспанцу той оно было также*
for unknown that it was same
*‘For hispanic that it was same’

Above, we have shown only two of many syntactic constructions that are different in the languages and, consequently, may become a source of mistakes. We believe that it is virtually impossible to name all the discrepancies and introduce the rules to cover them all (when speaking about Rule-Based MT). SMT systems will cope with syntactic discrepancies unpredictably and inconsequently: the syntactic structure might be proper, but in a slightly different context it can be totally wrong.

5.2.15 Disambiguation

In the following subsections we will describe errors related to lexical semantics: disambiguation, wrong lexical choice, choice of a completely bad word sense and multiword expressions (idioms, light verbs, phraseological units).

When a word in a sentence is used in a wrong sense, we consider it as a disambiguation error. Languages, even related, demonstrate discrepancies in the lexical structure of words. We can consider several types of polysemic differences between the two languages:

- the first one is when a source (Czech) lexeme is polysemoius and the target (Russian) lexical equivalent does not have some sense that a Czech one has and those senses are expressed by different lexemes (see Example 5.53, word *stát*);
- there are more sense components in the target (Russian) lexeme than in the source Czech. Ex, a word *diplomat* can indicate only a human of a concrete occupation or quality in Czech, whereas the corresponding Russian word *дипломат* has one more non-human sense - *aktovka* – 'briefcase'. The ambiguity on the target side will, most probably, not cause such problems for the MT as the ambiguity of the source.
- Finally, some words may not even have translation equivalents in the other language. Ex. a Czech word *sourozenci* – 'siblings' - can not be translated with one word into Russian (*братья и сестры* – 'brothers and sisters' should be used instead).

The researchers often point out (and it is quite logical) that SMT systems win over RBMT when coping with ambiguous words, and the statistics from the table 5.2 proves this fact. The choice of a proper sense depends on a context of a word, and as statistical systems are based on phrases, the context is often 'taken into account'. For the RBMT systems, a special disambiguation and lexical selection modules have to be created, which requires additional lexicographical resources

and parallel data. Also, it can be introduced in a form of statistical post-editing, as, ex. presented in (Galuščáková et al., 2013).

In the paper (Oliveira et al., 2005) the authors used an approach quite close to the one from the system Ruslan, but modern techniques (like statistical modeling on parallel data) were exploited. The main idea consists in classifying ambiguous words into sense categories like 'human', 'place', 'emotion' and then choose the proper sense according to the context window.

RBMT systems that we research both lack disambiguation or lexical selection modules. The TectoMT transfer was designed to select the most frequent variant from a dictionary. Obviously, this often fails because Czech and Russian are distant enough and have a lot of differences in semantic structure. The same may be true for PC Translator, though we can not say anything for sure about the architecture of this commercial system. On the basis of the analysed data we have only concluded that PC Translator made far too many mistakes in lexical semantic issues in comparison with other systems, even with TectoMT.

Let us show an example where all the systems chose a wrong variant systematically. The word *stát* – 'state' is ambiguous in Czech as well as in English: above all other meanings, it can refer to a state as a country and a state as a province (esp. in the articles about USA). In Russian, a two distinct words are used for these two notions: *страна* – 'a country' and *штат* – 'a state' (e.g., in America). The source Czech word has a meaning of a state as a part of the USA, but all the systems chose the most frequent variant (country):

(5.53)

(src) *v dalších pěti státech*
in other five states
'In the other five states'

(mos) *еще в пяти disam::странах*
more in five countries
'in five more countries'

(goo) *в пяти других disam::государств*
in five other countries
'in five other countries'

(pct) *зажечься еще петиция disam::странах*
fire more petition countries
'fire more petition countries '

(tmt) *в других пять disam::государствах*
in other five countries

'in other five countries'

We have examined several sentences where this word occurred and sometimes either Google or Moses translated it properly in specific contexts. Also, Google (but not Moses) quite often translates this word in a sense *condition* because it translates from Czech into Russian via English language. In Czech, the word **stát** - 'state' - has the two meanings as in the above example (province and country), but 'state' as 'condition' is evidently a middle-step error of Czech-to-English translation:

(5.54)

(src) *nevyžadoval* *žádný* *stát*
demand-not-Past1Sg no state
'No state demanded...'

(goo) *не* *требует* *состояние*
not demand condition
*'the condition does not demand'

In some cases, a homonymy of morphological forms of different lemmas[31] can occur, see ex. 5.55. A Czech word *let* can be translated either as a *peřc* – 'flight', it can also be a genitive form (supletive) of a word *years* in, e.g. a context *10 years*:

(5.55)

(src) *10* *let* *poté*
10 years after
'10 years after'

(pct) *10* *peřc* *затем*
10 flight afterwards
*'ten flight after'

This error evidently comes from the morphological analysis, during which the word *let* was not recognized as it should be - as an indirect genitive plural from a lemma *rok* – 'years'. In TectoMT, the construction with age receives a special treatment, so it was translated correctly. Statistical systems also cope with this

[31] When a wordform has the same surface realization as a concrete morphological form of another word

phenomena because the corresponding n-gram is seen quite often in the phrase table.

All in all, we can say that SMTs are generally better when guessing the word sense, be it for related or non-related (for Czech and English, see the Table 5 in (Bojar, 2011)) languages. The only significant difference in the table of Czech-Russian errors in comparison with that Czech-English is that PC Translator scored significantly worse with respect to all lexical-semantic issues than other systems. Just for comparison, PC Translator for Czech-English pair has less errors in disambiguation than TectoMT. Again, we can attribute this to the fact that PC Translator was developed mainly for Czech-English for many years[32], so it copes better with word sense issues for this language pair.

5.2.16 Lexical choice

There is quite a fuzzy border between *disambiguation* and *lexical choice* errors, and we can say that it is another side of one coin. Both notions indicate that a wrong word was used. **Disam::** errors specify completely different sense of a word. Words marked with **lex::** bare mostly a proper sense, but some very close synonymous word was chosen instead.

Wrong lexical choice is a problem that is really hard to detect automatically because the difference between the two near synonymous words often lies in nuances, style, connotation, polarity or usage in concrete contexts. Different languages cluster close synonyms usually in a very different, unique way which makes the translation challenging, not only for the machines, but for human translators as well. Very often in Machine Translation near-synonymous words are used and, in contrast to other errors that make sentence unreadable/ungrammatical, they do not affect the whole perception/sense.

Following are the examples of the instances marked as lexical errors[33]

(5.56)

(src) *voliči* *předkládali* *průkazy*
voters demonstrated id's
'The voters demonstrated ids'

(mos) *избиратели* *выдвигали* *удостоверения*
voters proposed id's

[32] Czech-Russian pair was not at all a popular direction, which reflects the translation quality

[33] The glossing into English will be extremely difficult as it concerns slightest shades of meaning that can be completely different in Czech, Russian and English, but we tried to find the most close synonyms.

'Voters put forward their ids'

The statistical system finds the most probable n-gram, and the phrase *избиратели выдвигали* – 'voters proposed' (as a candidate), evidently, occurred many times in the training data, more than *voters demonstrated*[34].

Following is an example of an error made by the RBMT system. In Czech, a word *osoba* may refer to 'personality' or it can be more of a general sense - 'people', and the latter sense was used in the sentence 5.57 (src). In Russian, however, a special notion for a sense *personality* – 'личность' is used and in the Czech-Russian dictionary of PC Translator this very variant was the most frequent one, which lead to a mistake.

(5.57)

(src) *21 milionů osob*
21 million people-gen
'21 million people'

(pct) *21 млн. *личностей*
21 mln. personalities
'21 mln. personalities'

Generally, Statistical systems scored better with respect to the lexical choice (as well as disambiguation) because the context is taken into consideration. On the contrary, RBMT systems are not very accurate in the lexical choice; more sophisticated techniques (like adding statistical post-editing, or preprocessing of the source) should be used which makes RBMT more of a hybrid system.

The last note is that the border between 'disambiguation', 'lexical choice' and 'no mistake' is very vague and even subjective. Some annotators will consider a word to be disambiguated wrongly, another can find some difference and put a label **lex::**, others may even not even tag a word as an error. In (Bojar, 2011) the agreement rate between two annotators on these two error types was around 10% when treated separately, and when **disam::** and **lex::** errors were united into one class, it was still around 30%.

5.2.17 Totally bad word sense

We attached a tag **tbws::**(totally bad word sense) when a word in an output has a sense that has nothing to do with a source. This problem can not be motivated

[34] This may seem as a disambiguation error from the point of view of English, but for Russian the two senses are more close synonyms

linguistically by some discrepancy between the languages, but we will try to explain why such cases might possibly occur. Let us have a look at Statistical systems first. As for google, we can justify those errors again by the fact of pivoting through English: all three languages are ambiguous in their own way which can lead to a 'chain' of disambiguation errors:

(5.58)

(src) *absolvovat* *vyšetření* *nebo* *ne*
complete-**verb**.inf examination or not
'to undergo (medical) examination or not'

(goo) *полное* *обследование* *или* *нет*
complete-**adj** examination or not
*'full examination or not'

In the above example, the polysemous word *absolvovat* – 'undergo' was wrongly disambiguated in the context and, evidently, was translated into English as *complete* rather than *undergo*. While translating from English into Russian, the part of speech was confused and a wrong sense was chosen again, which resulted in a word with a totally different sense.

Using of a completely improper word in RBMT systems is evidently a result of a wrong dictionary entry. We have encounterED many words with different sense in the PC Translator. This may be due not only to the dictionary[35], but also to some 'core' error in the PC Translator. One of the most frequent words - a preposition *v* – 'in' is very often translated for some unknown reason as a verb *зажигаться* – 'light' which makes the whole sentence look very clumsy, see Example 5.59. Some other words - mainly named entities were translated by very strange equivalents (like Indiana as 'Hindi- Diana') as well.

(5.59)

(src) *mezery* *v* *modelu*
gaps in model
'gaps in the model'

(pct) *пробелы* *зажечься* *модели*
gaps light-inf model
*'Gaps to light a model'

[35] TectoMT uses an automatically generated dictionary 'checked' by the dictionary from PC Translator, so the quality of PC Translator's dictionary(and number of dictionary entries) should be higher than that from TectoMT

5.2.18 Multi-word units

A multiword expression(MWE) presents a sequence of words with non-compositional meaning - where the meaning of a phrase can not be derived from the meaning of its parts. Handling MWEs is a challenging problem in various areas of NLP, in (Sag et al., 2002) MWEs were called 'A Pain in the Neck for NLP'. Many papers exist on how multiword expressions are identified in the text, aligned with their equivalents in the other language, how they are therefore processed and incorporated into MT systems, e.g. (Anastasiou, 2010), (Bouamor et al., 2012). MWEs are annotated within the Prague Dependency Treebank (Bejček – Straňák, 2010), so we have a resource for Czech MWEs, but there is no such for Russian.

MWEs differ from language to language and are highly idiosyncratic. Even for the related Czech and Russian we can not be sure about their similar structure. Both approaches - Rule-based and Statistical - experience difficulties when processing those units. The meaning of multiword expressions is not compositional, so the RBMT-based systems without an appropriate information will translate the units word for word, which can lead to a mistake. SMT systems generally cope with multiword expressions better, as they consider the n-grams, but it is not always the case that the n-gram will be translated correctly.

We will distinguish several types of the multiword expressions based on their part of speech and function in a sentence.

- Noun multiword expressions
- Auxiliary multiword expressions
- Light verbs
- Idioms

Next, we will show several examples of how the MT systems handle the multiword expressions.

Noun multiword expressions

Multi-word expressions in our test set are mainly named entities or belong to domain specific terminology. They generally contain a noun and some other part of speech. According to the table of errors, SMT systems almost did not make such errors, at least in our evaluated test set.

Following is an example where both SMT and RBMT systems made the error while translating an MWE *návrh zákona* – 'bill' word by word:

(5.60)

(src) *180 návrhů zákona*
180 suggestions law-Gen
'180 bills'

(mos) *180 работ закона*
180 work law-Gen
*'180 works of law'

(goo) *180 предложений по законов*
180 suggestions for laws
'180 suggestions for laws'

(tmt) *180 предложения законов*
180 suggestions law
*'180 Suggestion of laws'

In this example, the two-word expression *návrh zákona* should be translated into Russian as a compound word - *законопроект* – 'lawproject', and three systems made an error in this case. Russian language has a tendency (but not to such an extent as German) to form multiword compounds more often than Czech, so the cases where two or more Russian words correspond to one Czech are not so infrequent, e.g. *теракт* – 'act of terrorism' vs. *teroristický útok* – 'terrorist attack'.

Auxiliary multiword expressions

Auxiliary MWEs present mainly multiword prepositions (e.g., *в течение* – 'during') and they are generally reflected in a dictionary of RBMTs; SMTs also do not have a problem to handle them properly because their co-occurence in the data is quite frequent and parts of an expression are not separated by other words. However, sometimes a MWE is not present in the dictionary, which can result in an error. We have already provided this example in the general introduction of errors, see Example 5.2. A complex Czech preposition *v předstihu* – 'in advance' should have been translated as a one-word preposition *заранее* – 'in advance' in Russian. However, the problem here is more sophisticated as this complex preposition forms a part of a multiword expression itself - *volby v předstihu* – 'elections in advance' and it should be translated as *досрочные выборы* – 'advance elections' into Russian.

Light verbs

Light verb constructions (LVC) are generally formed by a verb and a noun where a verb does not bare its initial meaning, so that the whole construction takes the

semantics of the noun.[36] Following are several examples of Czech LVC in contrast to Russian. They can be equal for both (using the same verb):

(5.61) cz:*hrát úlohu* vs. ru:*играть роль* – 'to play role'
cz:*lámat hlavu* vs. ru:*ломать голову* – 'lit. to broke head'

Or the languages can use different light verb:

(5.62) *brát zřetel* – 'lit. take consideration' vs. *принимать во внимание* – 'lit. accept into attention'
dát smysl – 'give sense' vs. *иметь смысл* – 'have sense' - in English 'to make sense'

The examples above illustrate that some multiword verbs have identical component words in the two languages, and some not. Generally, multiword expressions are translated properly within SMT when an LVC presents a bigram, but when a verb is separated from a noun, this LVC is generally mistranslated, see Example 5.63. The LVC from this example has the same structure in Czech and Russian, so the error is not justified by discrepancies between the languages.

(5.63)

(src) ***neměli*** *průkopníci laseru v bellových laboratořích ani*
had-not.Past.Neg pioneers laser-gen in Bell Labs any
tušení *o revoluci*
idea about revolution
'Laser pioneers in Bell Labs did not have any idea about revolution'

(mos) *не пионеры технике за радиацию* ***понятия***
not pioners machines radiation-Dat idea-gen about
о революции
revolution
*'Not pioners for techiques for radiation idea about revolution '

(goo) ***не было*** *пионеров лазер на Bell Labs ни малейшего*
not was pioneers laser for Bell Labs any idea
представления

*'There was no pioneer laser on Bell Labs any idea'

[36] The researchers do not have a unified definition of light verbs, sometimes it is quite hard to distinguish LV from idioms.

(pct) *прукопници лазер зажечься белловэцһ лабораторный даже*
unk-word laser set-light unk-word laboratory even
предчувствие
presentiment
*'unk laser to set fire unk laboratory even presentiment'

(tmt) ***не имели*** *пионеры лазера в белловых лабораториях ни*
not have pioneers laser-gen in bell laboratories not
предчувствие
presentiment
*'Pioners of laser did not have no presentiment in Bell Labs'

In this example Google and Moses used some fragments from the proper construction (*не имели понятия/представления* – 'not have any idea'), but none of the systems used the proper light verb. On the contrary, TectoMT generated an almost good structure using a proper verb, but the predicate noun was not quite correct (though, understandable).

Idioms

Idioms are MWEs that can include words of any part of speech and they generally bear a meaning that has very little to do with any component of MWE. Idiomatic constructions often present a challenge to MT systems. RBMTs tend to translate them word-for-word unless the idiom is present in the lexicon.

Let us have a look on some concrete examples of idioms. The expression might be equal in both languages, see Example 5.64(cs vs. ru), but that is not always the case, see Example 5.65 (cs vs. ru).

As our data belong to the domain of news, we have not found any idioms in the test set. Just for the sake of the experiment, we translated the two mentioned idioms with Google Translate tool in both directions, and the results were quite interesting. When translating the first - identical in Czech and Russian - idiomatic expression 5.64 from Czech into Russian, the idiom was translated properly as 5.64(ru), however, the reverse translation from Russian into Czech stayed the same as a pivot English - 'rain cats and dogs'. We can speculate that this might be due to the low co-occurrence of the corresponding idiom in Czech (*It rains like from a bucket*) and English (*It rains cats and dogs*) whereas the Russian-English pair evidently has it.

(5.64)

(cs) *Leje jako z konve*
Rains like from bucket

'It rains cats and dogs'

(ru) *Льет как из ведра.*
Rains like from bucket
'It rains cats and dogs'

(goo:cs-ru) *Льет как из ведра.*
Rains like from bucket
'It rains cats and dogs'

(goo:ru-cs) *Prší kočky a psy.*
Rains cats and dogs
*'It rains cats and dogs'

As for the next example, the idioms in Czech and Russian have different structure and the translation is only half correct for both translation directions:

(5.65)

(cs) *Dělá z komára velblouda.*
Makes-he from mosquito camel
'He makes mountains out of molehills.'

(ru) *Он делает из мухи слона.*
He makes from fly elephant
'He makes mountains out of molehills.'

(goo:cs-ru) *Он делает нечто из ничего.*
He makes something from nothing
?'He makes something from nothing'

(goo:ru-cs) *Dělá hory z komára.*
Makes-he mountains from mosquito
*'He makes mountains out of mosquito'

All in all we can conclude that multiword expressions present a problem mostly for RBMT systems that need to have the bilingual lexicon of MWEs. Statistical systems cope with them as soon as a multiword unit fits into a respective bigram which is relatively frequent in the training data.

5.3 Discussion

In this chapter, we have described the most typical errors that the four MT systems make, classifying them from the linguistic point of view. This taxonomy and annotation schema are the most used ones for the task of manual error analysis, but we have made a more fine-grained classification of errors with the regard to our language pair.

For each linguistic problem, we have provided a detailed analysis and explanation of why an error occurred and in some cases outlined possible directions of how an error might be fixed (especially for the Rule-Based Systems). However, we did not make an attempt to fix those errors as this is a task for a team of specialists for several years and, still, it is virtually impossible to fix all of them. For example, PC Translator have been developing Czech<->English pair for years, and there are still many errors in it, and even it did not receive the best score among other systems in the WMT competition. As for the SMT, a lot of research is carried out in this area in order to improve BLEU score, or the system performance for some concrete phenomena.

Quality of translation also highly depends on the source: if a source sentence is complex or contains certain linguistic constructions, the chance that it would be translated properly is rather low. The error analysis has revealed several types of constructions that tend to be mistranslated, and they correspond with what other researchers have written about Rule-Based and Statistical MT.

As it is generally stated, the major cornerstone for Statistical systems is syntax, whereas semantics poses a challenge mainly for Rule-Based MT. According to our error analysis, both systems show a huge number of errors in morphology, though they are generally connected to syntax as well.

Following are some observations that we found to be interesting for each type of MT:

SMT : Discrepancies between the languages do not have much impact on the MT as soon as the elements that constitutes the concrete phenomenon stand not far from one another. If dependent words stay far from each other in the source sentence, the respective n-gram most probably will not be found in the training data and thus the translation might be incorrect.

RBMT : Language discrepancies have much higher impact on the RBMT systems than on SMT. Performance of an RBMT system depends on the rules that are written to capture differences between languages. Each rule, in turn, has to be properly implemented into the process of text analysis and synthesis to generate the expected output. Sometimes the rules are not sophisticated

enough or are not applied properly. Another reason for a mistranslation is often an error in analysis(parsing) or synthesis modules.

To sum up, what helps RBMT are years of hard manual work on rules and the language processing modules. What helps SMT is mainly data (like adding bigger translation and language models, domain adaptation etc.). Also, the translation quality highly depends on the source sentence complexity (the constituent oder, embedded clauses).

6

Valency in Czech and Russian

In this chapter, we will discuss theoretical aspects of valency and focus on the comparative analysis of Czech and Russian surface valency using various linguistic resources.

The results of manual evaluation of MT output revealed that valency errors[1] occurred in the output in all the systems. Our initial suggestion that errors in valency would occur only when there is some discrepancy in Czech and Russian valency structures turned out to be false. Many words were marked as a valency error even though Czech and Russian verbs had the same frame with the same morphological cases. The concrete examples were demonstrated in Section 5.2.7. Those errors are not always directly connected to the discrepancy in valency. The source of those errors are different for the Rule-Based and Statistical MT systems:

- In case of the Rule-Based systems, errors always occur when there is some discrepancy in valency - most often in prepositions and cases - unless this discrepancy is present in the system in a form of a rule or a dictionary entry. On the other hand, due to the low performance of analysis or synthesis modules of the system, the wrong case/preposition can be used even when the valency patterns for Czech and Russian are identical. So our assumption that 'relatedness helps' is overridden by some troublesome technical issues.

- Phrase-based systems are hard to evaluate in terms of linguistics. Generally, a system will generate correct valency connections as soon as a hypothesis contains a proper n-gram no matter whether a valency frame is different or equal in the source and the target languages. When a verb and its dependent noun are separated by one or more words, it is more likely that the noun will have improper case, again irrelevant of valency discrepancies/equality.

[1] As it was mentioned in the previous chapter, we will use the term 'valency' in a sense of 'surface valency'.

Though these errors seem to be less serious for a simple sense gisting than e.g. disambiguation errors or unrecognized words, they may complicate the analysis of a sentence structure and can sometimes change the meaning of a phrase. This is especially true of Slavic languages where words can take almost any position in the sentence, but if used in an incorrect form, they can make the whole text hard to understand.

Our main objective here is to identify the main points of difference between Czech and Russian valency, aiming at building a Czech-Russian valency lexicon and trying to apply it in the MT system. The only resource containing the data with valency information for the two languages was Ruslan. Though this source is quite outdated and not that reliable, we decided to make an experiment on automatical extraction of Czech and Russian surface valency frames from parallel data, it was not successful though. Then, we explored the nature of the verbs that have different valency structures in Czech and Russian. The idea was that if some verb in a semantic class has different surface valency in the two languages, the probability is that semantically related verbs will have this discrepancy as well.

This chapter will be structured as follows. First, we will define what we understand under the term 'valency', show the existing valency resources (Section 6.1). Then we extracted the surface valency frames from Ruslan dictionary, examined differences in valency and implementes the extracted list of verbs + frames into the TectoMT system (Section 6.2). As Ruslan is a rather limited source of information, we also attempted to automatically build a lexicon with surface frames using the large-scale resource Vallex, a bilingual dictionary and a parallel corpus (Section 6.3). In Section 6.4 we try to answer the question which verbs tend to have different valency frames in the two languages. The valuable result of this research is a parallel Czech-Russian valency dictionary extracted from Ruslan.

Our experiments share similar ideas with many other works on valency within Machine Translation, next we will name those that work with either Czech or Russian. In (Bojar – Šindlerová, 2010) authors collect valency translation equivalents for Czech and English verbs exploiting the parallel treebank. (Rosa, 2013) built a simple probabilistic valency model for Czech and English and exploited this information to correct valency errors in the machine translation output. As for the theoretical research, (Hladná, 2012) presents a comparative study of Czech and Russian valency based on a small sample of text, so we can compare the results with ours.

6.1 Notion of valency

6.1.1 Theoretical aspects of Valency

Valency is understood differently by various researchers, and this phenomenon is also known under different names. In the English tradition, the notion *subcategorisation frame* is very close to valency and typically denotes the surface(morphosyntactic) valency, whereas *Predicate-argument structure* refers to more deep, semantic valency[2].

We have already made a disclaimer on the term 'valency' in the previous Chapter. The way how we operate with it here can be quite misleading for many theoretical researchers, because, generally, 'valency' is defined as a capability of a verb/word to bind a specific amount and types of arguments. Valency can be also seen in a broader sense - it can be either deep (concerning such notions as thematic roles or deep cases) and it can be also viewed as surface valency that operates on syntactic and morphological level. In this study, we focus on this second aspect of valency, namely on surface realizations of verb arguments. For the sake of shortness, we will call this information 'valency', as it was done in other computationally-oriented works like (Rosa, 2013). We focus mainly on the differences and as soon as the 'left-hand side' actants (Subjects) almost never show discrepancies in Czech and Russian (they are almost always in Nominative case), the emphasis will be put on the 'right-hand side' valency. Also, we will narrow our research on the noun phrase realizations only.

Valency

For a particular word - mostly a verb - valency presents the number of dependent words in a sentence that a verb must have (obligatory) or that a verb may have (facultative). The Encyclopedic dictionary of Czech (Petr Karlík – Pleskalová, 2002) defines the valency as "a number and the features of valent places (arguments) which is attracted by a verb (or other part of speech) in obligatory or facultative way".

The term 'valency' was adopted to the linguistic terminology from chemistry by Lucien Tesnière (Tesnière, 1959) in association with an atom (a verb) which can attract molecules (complements).

Since Tesnière introduced his theory, many other linguistic schools based their theories on Tesnièr's. His original classification of actants - first, second, third actants - was therefore substituted by a more meaningful terms - Subject, Direct/Indirect object. Afterwards, a new theory with a more semanticalized and

[2] It is not virtually possible to describe all the valency theories, and in this work we will present only those most relevant to our research.

fine-grained classification was proposed by Fillmore(Fillmore, 1968), who introduced a notion of case frames.

In Prague, a valency theory (Panevová, 1974) was developed within the already mentioned FGD framework (see Section 4.3 for detailed description of language layers). We can spot valency on all the three language layers - morphological, analytical and tectogrammatical. On the tectogrammatical layer the valency is more semantic, it is represented by a sequence of **functors** that form a core of a valency frame of a verb[3]. Each of the functors is labled with the respective surface realization (ex. case, preposition+case, relative clause etc.). A frame slot also contains information on the obligatoriness. As we have mentioned, the arguments of a verb might be obligatory or facultative. One of the contributions of Jarmila Panevová to the linguistic theory of valency was introducing criteria for distinguishing between obligatory and facultative complementsbased on a dialog test (Sgall et al., 1986). The wh-question on each complement is being put to the speaker and if the answer ''I don't know" in a coherent dialogue is possible, the complement is then facultative (optional), and if it is not - then it is obligatory.

As for the Russian linguistic school, the **Meaning-Text theory (MTT)** (Mel'čuk, 1988) accounts for the valency also in a semantic and a syntactic sense as the FGD theory[4]. MTT also structures the language into several layers, though those layers do not always match one-to-one with thoe from PDT. For instance, a deep syntactic layer from MTT is very close to the tectogrammatical layer from FGD, and a shallow syntactic level from MTT is close to analytical level from FGD. An extensive lexico-semantic resource - Explanatory Combinatorial Dictionary (ECD) - was developed based on the MTT theory, which will be desribed in the next section valency resources.

Prepositional vs. non-prepositional complements

In our work, we will pay particular attention to the dichotomy of prepositional vs. non-prepositional complements. It should be noted that the status of prepositions in the phrase is a very disputable issue. Some researchers, like (Trask, 1944), claim that a preposition governs its object. And it does actually determine the case of the following noun. According to other theories (Kuryłowicz, 1960), a

[3] Though, a noun can also have a valency, but in this work we concentrate on verbal valency only.

[4] MTT and FGD theories were developed at roughly the same years and they share a lot of common features: division into layers, creation of a treebank based on the theory and application of this theory in the MT system ETAP. More on the similarities and differences between the two theories can be found in (Žabokrtský, 2005).

preposition does not govern a noun/pronoun, it is considered to be a kind of a morpheme itself which is subordinated to a noun.

This theoretical dichotomy is also projected in the treebanks. For example, in the Prague Dependency Treebank a preposition is a parent whereas a noun is a child - but only on the analytical (shallow syntactic) layer. This supports the fact that a preposition really determines the surface form of an argument. However, on the tectogrammatical (more semantic) layer, a preposition becomes only an attribute to the respective noun, and that means that its function in a sentence is really more close to morphological. According to the most recent studies (Universal Dependencies format)[5], a preposition is repsented as a child of the noun.

As we study mainly the surface valency, we will often use phrases like 'a preposition entails a case of a noun', but we are aware that on the deeper language layers the preposition does not play a big role. This decision is also caused by the nature of Machine Translation architecture. In both Rule-Based and Statistical systems, it does matter if a verb has a prepositional or non-prepositional complement, because in the first case (prepositional valency) there is one or more tokens to be processed by the system. Also, this makes a difference for the statistical experiments - when we search a corpus for prepositional complements, we search for at least three tokens (a verb, a preposition and a noun in a certain case), whereas we search only for two tokens (a verb plus a noun) in case of non-prepositional complement.

6.1.2 Valency Resources

Next, we will name the most famous valency resources or those resources relevant to our work[6].

FrameNet

FrameNet (Baker et al., 1998) presents a freely available lexicon of words organized into a semantic hierarchy. Wors in FrameNet are assigned with a semantic frame reflecting roles of main actants of the word. Each lexical unit in a sentence is assigned with a semantic role - or frame element; frame elements, in turn, form a semantic frame of a lexical unit. Following is an example[7] of a frame element of a verb *to fry*:

[5] http://universaldependencies.github.io/docs/

[6] A dictionary from Ruslan (Oliva., 1989) that we base our work on will be described in Section 6.2

[7] Borrowed from https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

(6.1) [Matilde]$_{\text{Cook}}$ fried [the catfish]$_{\text{Food}}$ [in a heavy iron skillet]$_{\text{HeatingInstrument}}$

(Benešová et al., 2008) mapped semantic information from FrameNet into Vallex, but we will not use this resource in our work.

PropBank, PropBank-Lexicon

PropBank (Kingsbury – Palmer, 2002) - a shortcut from the Propositional Bank - is a corpus of texts in which verbs are annotated with predicate frames containing main arguments. In comparison to FrameNet, the PropBank is focused on verbal valency only. It sticks more to the syntactic layer, and the semantic roles are not that deep and granulated in comparison with the FrameNet, see the following example[8]:

(6.2) [John]$_{\text{ARG0}}$ **broke** [the window]$_{\text{ARG1}}$

The arguments are assigned with the concrete numbers. Also, the modifiers are tagged with more semantic specifications like Manner, Time or Locative. PropBank was developed to serve as a source for training data for the semantic role labeling.

VerbNet

VerbNet (Kipper et al., 2007) is a lexicon of verbs based on the Levin's classification (Levin, 1993) and each argument of a verb is assigned with a thematic role (theta-role), ex. Agent, Beneficiary, Cause, Experiencer, Patient. The verbs are grouped into the Verb classes that share typical frame patterns, semantic restrictions on the thematic roles (e.g., concrete, abstract, location).

Verbalex

Verbalex (Hlaváčková, 2005) is a lexicon of Czech verbs which is very similar to Vallex, but the verbs in Verbalex are organized into synsets - a sets of synonyms sharing the same subcategorisational pattern - or valency frame. Verbalex is organized more like an hierarchy of verb classes whereas Vallex semantic classification is just an additional feature. Verbalex is not an open source as Vallex, so we can not exploit this lexicon in our experiments.

[8] From the annotation manual http://verbs.colorado.edu/ mpalmer/projects/ace/PBguidelines.pdf

Vallex

Vallex[9] is a manually created Valency Lexicon of Czech Verbs based on the valency theory of Functional Generative Description. It provides the information on valency frames of the most frequent verbs (in version Vallex 2.5 there are over 2,700 lexeme multiplied by aspectual counterparts). The original valency entry from Vallex contains complex linguistic information:

- a lemma - the basic form of a verb;
- A frame:
  - a deep semantic role called functor (Actor, Patient, Addressee etc.,);
  - a surface realization of the functor;
- a semantic class of the verb;
- examples of using the verb in a real context;
- an information on reflexivity, aspect, idioms and some other.

Let us take as an example a verb *dotknout se* – 'touch', following is the Vallex entry for this lexeme:

The entry for a lexeme *dotknout se* consists of 4 lexical units. Let us examine the first lexical unit (first sense of a lexeme). This verb is valent for three complements: Actor, Patient and Means. Actor is expressed in Nominative case (1)[10], Patient is expressed as a direct object in Genitive case (2), Means is an indirect object in Instrumental case (7).

As we have mentioned, we will be primarily interested in a surface valency, so we will exploit information on morphemic forms of complements.

PDT-Vallex

In our work, we will use data from Vallex and partially PDT-Vallex (Urešová, 2012) (more in Section 6.3), a dictionary that contains annotated valency frames in the Prague Dependency Treebank. The lexicon itself is different from Vallex as it contains less linguistic information (e.g. there is no information about word class, reflexivity, reciprocity, etc.), but there are far more verbs in it - more than 7,000 verbs with over 11,000 valency frames.

As with the case of Vallex, we will use only morphemic forms of the complements.

[9] http://ufal.mff.cuni.cz/vallex/2.5/

[10] The cases in Czech are traditionally encoded with numbers

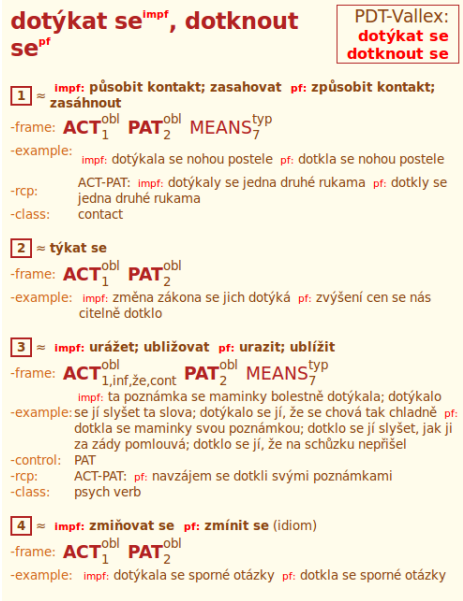


Figure 6.1: Example of a Vallex entry

TKS

As we have already mentioned, information on valency for Russian verbs is included into the TKS (Melčuk – Zholkovsky., 1984) - the dictionary based on the MTT theory. Each entry (called vocable) in the dictionary contains a number of lexical units that define sense(s), morphological, syntactic and semantic characteristics of a word. The most famuous notation from MTT adopted in TKS is that of lexical functions that denote semantic relations between two words within collocations. The information on valency we are interested in is represented in a form of a **government pattern**[11]. Unlike in Vallex, the information on deep semantic roles of complements is not included, only a shallow syntactic function - **X** - subject and **Y** - object. Picture 6.2 depicts a government pattern of the verb *восхищаться* – 'to admire'.

A modern ancestor of the TKS dictionary is *Active dictionary of Russian language* (Apresjan, 2011) that also contains the information on a government pattern of a verb is still under development. We could have used both of the

[11] In Russian, *модель управления*.

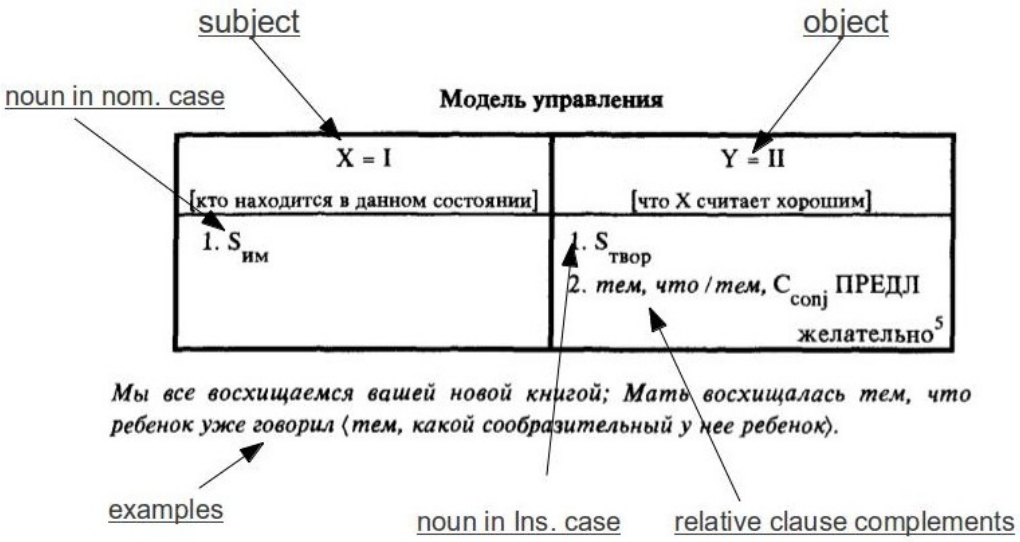


Figure 6.2: Government pattern of a verb *восхищаться* – 'admire' in the TKS dictionary

resources in our comparative analysis and tried to combine them into a bilingual valency resource, but, unfortunately, both TKS and the Active dictionary are not available online.

6.2 Valency information extracted from the Ruslan lexicon

In the first stage of comparison of Czech and Russian valency we will exploit the MT dictionary Ruslan[12] (preliminarily described in Section 4.1). Within the Ruslan MT system, verbs in the lexicon were assigned with their valency frames in Czech and the corresponding frames in Russian with the specification of a semantic class of all verb complements. The following Example 6.3 demonstrates

[12] This section arised from the work (Klyueva – Kuboň, 2010), where the author of the thesis made all the experiments and most of the writing. Some of the passages here may contain the same formulations as the cited paper.

an entry from Ruslan for the verb *vystačit* – 'to be enough', the explanatory notes are given further:

(6.3) VYSTAC3==R(5,PRP,?(N(D),S(I,G)),39,CHVATIT6):

- *VYSTAC3* presents a root of a verb,
- *==R* denotes a verb,
- *5, PRP* - a verb conjugation pattern in Czech,
- *N(D),S(I,G))* is a valency frame that we will further describe in detail,
- *39* is a Russian declination pattern,
- CHVATIT6 is the Russian translation of a lexeme, coded in latin

As the original format of Ruslan entry was written for Q-systems and is very incomprehensible, we transformed Ruslan entries into a more user-friendly format. First, we lowercased entries and transfer Ruslan coding of letters with diacritics (coded in numbers) into common letters and transformed the cyrillic letters for a Russian translation. Then we selected verbs and substituted the verb stem and the morphological information coded in special symbols with an appropriate verb ending. We will not use the semantic feature of a complement as we believe this semantic information would not be necessary in our comparison. The valency frame also contains a passive valency slot, which we will ignore as well because the passivization pattern is quite similar in Czech and in Russian.

Here is an example of a transformed entry from the Example 6.3:

(6.4) vystačit (n(d) s(i,g)) хватить - *to be enough*

Following is a short explanation of the frames:

- **n(d)** means that Czech Nominative case corresponds to Russian Dative.
- **s(i,g)** means that the preposition s (with) governs Instrumental case in Czech whereas in Russian a non-prepositional case - Genitive(g) is used.

Here we will work with several valency resources that mark morphemic cases in different way - with letters in Ruslan or numbers in Vallex. In order to make examples more comprehensible, we rewrite each example into a form with a contracted name of a case[13], Example 6.2 will be therefore depicted as:

(6.5) Nom + vystačit + s + Ins -> Dat хватить + Gen- *to be enough*

[13] Technically, the data that we use will have the original format.

6.2.1 The comparison of valency frames

Out of the 2080 verbal dictionary entries from Ruslan we have analyzed 1856 unique verbs[14]. We examinated how Czech valency frames correspond to Russian ones. We have sorted verbs on the basis whether the verb requires the prepositional case or the non-prepositional one. For shortness, we will cal the non-prepositional case the **simple case**.

This dichotomy is not motivated by some meaningful difference between simple and prepositional valency frames (see a discussion in the introductory Section), it was just more convenient due to the structure of an entry. Then, for each of the types we calculated the percentage of the verbs for which the surface forms in Czech and Russian match.

Due to simplicity, further in the text we will call the non-prepositional complements of a verb **simple complements** and those complements with prepositional cases - **prepositional complements**.

In most cases, we do not take into account a left-hand valency (generally, a Subject), because it is almost always the same in Czech and Russian (Nominative case in both languages).

Simple complements

Next, we will describe the verbs that require both in Czech and Russian a frame complement without a preposition, ex.:

(6.6) Nom vyzývat + Acc -> Nom вызвать + Acc - *to call*

The most typical sequence of frame patterns is *n(n) a(a)*(as in the example above) , which represents simple transitive verbs. 1317 (70 % of all verbs) have this structure. The fact that Czech and Russian have practically the same number of cases[15] makes the comparison easier and it apparently also influences the number of identical frames. Also, because for the majority of verbs the Actor is realized by the Nominative case in both languages, we will ignore the n(n) forms in our examples.

There are not so many verbs that govern simple (non-prepositional) cases and those cases are different in Czech and Russian (see the overall Table 6.3) in comparison with prepositional cases. Some examples:

[14] The reason for this difference is the fact that the original dictionary contains a number of verbal pairs with identical valency frames, usually two variants of a Czech lemma in the present and past tense.

[15] Vocative case is not used in modern Russian unlike in Czech, and it is not relevant for our study of verb complements.

(6.7)

(1) povšimnout si + Gen -> заметить + Acc - *to notice*

(2) vyhýbat se + Dat -> избегать + Gen - *to avoid*

Table 6.1 presents the statistics of simple frame patterns giving a picture of how simple cases in Czech and Russian mutually correspond. Locative case is not included as it is governed by a preposition in both languages.

		Czech				
		Nominative	Genitive	Dative	Accusative	Instrumental
Russian	Nominative	**3070**	8	10	6	3
	Genitive	0	**25**	0	4	0
	Dative	0	3	**178**	7	0
	Accusative	3	19	12	**1388**	7
	Instrumental	5	0	0	3	**1355**
Different surface frames:						90 (1.47%)
Total number of surface frames:						6160 (100%)
Number of verbs with different frames:						68 (3.66%)
Total number of analysed verbs:						1856 (100%)

Table 6.1: Co-occurence of the same cases in Czech and Russian based on Ruslan

As we can see from the table, Czech and Russian non-prepositional valency slots have usually identical cases, the list of verbs exhibiting this difference is not so big (68 verbs out of all the lexicon).

Prepositional complements

Next, we will describe verbs that govern complements with prepositional phrases. We consider the surface frames to be equal in a case when prepositions are translated straightforwardly or typically from Czech into Russian according to the dictionary default translation[16]. For example, the surface form with prepositions

[16] We have taken default translations from a list of formemes from the TectoMT block https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/RuleBasedFormemes.pm. We have linked those formemes to the Ruslan prepositional complements, but we have not transformed them into a human-readable format as this is a technical issue to calculate the differences. Some of the similar prepositions:

na + Acc -> на + Acc) is equal prepositional constituent, it means that in Czech and Russian the same preposition (cz) *na* – 'on' and (ru) *на* – 'on' is used and that the following noun is in the same case - Accusative. As Czech and Russian are very related, the function words , like original prepositions, often are the same or similar. Though, there are cases when translation of default forms does not match the surface forms, ex. a Czech preposition *do* – 'to' corresponds to Russian *в* – 'to'.

Following is an example of a verb with equal prepositional complements:

(6.8) působit na + Acc -> воздействовать на + Acc *to influence*

To select verbs that have different prepositional frames we just excluded verbs with similar surface frames. According to the results, 104 (5.6 %) of verbs have different surface frames containing prepositions. Following is an example of a verb *narazit* – 'come across'.

(6.9) narazit na + Acc -> столкнуться с + Ins - *to rush into*

We sorted the list of 'preposition plus case' pairs from Ruslan, Table 6.2 represents the top of the list with the frequencies of how often this frame occurred in Ruslan[17], different prepositional cases are in bold.

The differences in prepositional frames are not that obvious as in simple cases. Here we should take into account which frames are free modifiers and thus can be covered by the default translation of a dictionary. Unfortunately, in Ruslan, there is no distinction between the actants and free modifiers, so we can not say for sure if this prepositional phrase belongs to a verb or is it more independent of it.

6.2.2 Lexicon and a list of differences

The main output of this transformation is a small bilingual lexicon and a list of verbs that have different valency structure in Czech and Russian. Both resources can be exploited in the rule-based machine translation systems in order to cover such mistakes as in the Examples 5.18 or 5.19.

Table 6.3 shows the statistics of those verbs with regard to our classification on simple and prepositional case frames.

na(a,na(a)), s(i,s(i)), k(d,k(d)), z(g,iz(g)), od(g,ot(g)), v(l,v(l)), o(l,o(l)), do(g,do(g)), na(a, k(d)), o(a,na(a)), z(g,z(g)), na(l,na(l))

[17] As in some other examples, we did not transliterate the Russian prepositions so that the similarity is more apparent.

Czech frame	Russian frame	freq
na+Acc	na+Acc	82
do+Gen	v+Acc	80
z+Gen	iz+Gen	76
k+Dat	k+Dat	58
s+Ins	s+Ins	57
od+Gen	ot+Gen	29
v+Loc	v+Loc	26
o+Loc	o+Loc	22
do+Gen	do+Gen	19
k+Dat	**dlja+Gen**	16
na+Acc	**o+Loc**	15
na+Acc	**k+Dat**	14
před+Ins	**ot+Gen**	12
o+Acc	**na+Acc**	10
na+Loc	na+Loc	9
z+Gen	z+Gen	8
za+Acc	za+Acc	7
od+Gen	od+Gen	7
z+Gen	s+Gen	6
od+Gen	**u+Gen**	6
k+Dat	**na+Acc**	6
nad+Ins	nad+Ins	5

Table 6.2: Prepositional case correspondence - Ruslan

According to the Ruslan data, the number of different verbal valency frames between Czech and Russian is relatively low. However, we admit that the coverage of the dictionary used is rather limited. In further experiments, we will provide a surface valency analysis for the two languages exploiting more large-scale language resources.

Type of difference	Number of verbs	Percentage
Simple case	68	3.6%
Prepositional case	104	5.6%
Total	1856	100%

Table 6.3: Types of valency frames incorrespondences

6.2.3 Exploiting valency information from Ruslan in Machine Translation

We have also tried to exploit the entries from the Ruslan lexicon within the TectoMT (Section 4.3) system in order to see if there is some improvement in the translation. In order to integrate the dictionary into the system, we have transformed the entries into the special format verb+formeme[18]:

(6.10) **narazit n:na+4 => столкнуться n:с+7** - *to run into smb*

The list was incorporated into a system in a form of a block - FixValency.pm[19]. After implementing this block, some sentences with troubleshoot verbs (verbs with different surface valency) were translated with proper surface form. In examples below, (1TMT) is a test translation before applying the rules and (2TMT) after applying the rules.

In the following example, a Czech verb *využívat* – 'use' governs a complement in the Dative case, and in the baseline (1TMT) system, the complement received the same formeme as a default. However, in Russian the Accusative case should be used instead. This discrepancy was covered by the Ruslan entry (využívat + Dat -> использовать + Acc)[20] in the improved system (2TMT).

(6.11)

(SRC) *využívali* *obrovských* *amerických* *zakázek*
used-3Pl huge-**Dat** american-**Dat** contracts-**Dat**
' they made use of huge American contracts'

[18] The notion 'formeme' was introduced in the Section 4.3.2

[19] https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/FixValency.pm

[20] využívat n:2 => использовать n:4 in the block FixValency.pm

(1TMT) *они использовали огромных американских заказов*
*they used huge-**Dat** american-**Dat** contracts-**Dat**
‘ they made use of huge American contracts’

(2TMT) *они использовали огромные американские заказы*
they used huge-**Acc** american-**Acc** contracts-**Acc**
‘ they made use of huge American contracts’

However, there were cases when this rule worsened the translation. In Exampe 6.12, the prepositional complement was translated properly by (1TMT) because a rule for preposition transfer from another module[21] was applied (**n:pro+4 -> n:для+2** - n:for+Acc -> n:for+Gen). In the version with the lexicon, this rule was overridden by the rule from a new FixValency.pm module ("připravit n:pro+4" => "готовить n:про+4"). The latter verb-formeme Russian equivalent is a mistake in a Ruslan lexicon[22].

(6.12)

(SRC) *v kuchyni se pro hosty připravuje čaj.*
in kitchen refl **for** guests-**Acc** prepare tea
‘In the kitchen the tea for the guests is preparing’

(1TMT) *В кухне для гостей готовится чай.*
in kitchen for guests prepare-refl tea
‘In the kitchen the tea **for** the guests-**Gen** is preparing’

(2TMT) **В кухне про гости готовится чай.*
*in kitchen for guests prepare-refl tea
‘In the kitchen the tea **about** the guests-**Acc** is preparing’

In some sentences, both translations were incorrect due to various reasons. In Example 6.13, the light verb phrase *nabývá účinnosti(Gen) vs. вступит в силу(в + Acc)* – ‘takes effect’ is different in Czech and Russian; it should have been translated with another verb and another noun. The rule has no effect in this case, as the translation is wrong all the same.

(6.13)

[21] https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/RuleBasedFormemes.pm

[22] As the dictionary was compiled by non-native Russian speakers, there are a few errors in the lexicon and this one illustrates how people automatically assign a surface frame from their native Czech language to the verb in Russian.

(SRC) *zákon nabývá účinnosti 6 prosince*
law gains effect 6 December
'The law takes effect on 6 December'

(1TMT) *закон приобретать эффективности 6 декабря*
law *gains *effect-**Gen** 6 December
'The law gains effect on 6 December'

(2TMT) *закон *приобретать *эффективность 6 декабря*
law *gains *effect-**Acc** 6 December
'The law takes effect on December 6'

The above examples show that using the valency resource helps in some cases and harms in some others. Also, there was no significant influence on the BLEU score: **9.40%** without valency fix and **9.37%** with the module FixValency.pm. For such a small experiment, the BLEU score can not necessarily indicate if this valency module helped or not - we evaluated the experiment only on one reference example. So we evaluated manually the cases where a valency frame was changed according to the lexicon.

We have marked a list of changes between the (1TMT) and (2TMT) outputs indicating whether the introduction of a new rule:

- lead to some improvement like in Example 6.11
- worsened the translation like in Example 6.12
- did not have any effect as both variants were incorrect - Example 6.13

Effect	number of differences	Percentage
improved	28	58.3 %
worsened	3	6.2%
no effect	17	35.4%
Total	48	100%

Table 6.4: Manual evaluation of changes after adding FixValency.pm

From the table we can see that in the majority of cases the verbal valency will improve, or it will have no effect on the translation which is wrong this way or that. However, such a little fix did not bring any sufficient gain or loss when considering automatic evaluation metrics.

6.2.4 Discussion on Ruslan

Here in this section we worked with and extracted valuable information from the linguistic resource that was created more than 20 years ago and was forgotten ever since. This information contributed to identification of discrepancies in surface valency in Czech and Russian. Also, the extracted lexicon was used in the Rule-Based MT system where the manual evaluation showed that valency errors were corrected in more than 50 % of cases.

We should also note that this information might be as well contained in a textbook of Russian language addressed to Czech native speakers. We doubt that this list can be found in educational resources in a format sufficient for language processing.

6.3 Automatic valency extraction based on Vallex

In this section, we exploit several existing data resources and tools (a parallel corpus, the valency lexicon Vallex, morphological taggers and a bilingual dictionary) in a task of automatic extraction of surface valency frames[23].

Some experiments on automatic valency extraction related to our languages can be found for example in (Bojar – Šindlerová, 2010), (Zeman – Sarkar, 2000), (Pala – Ševeček, 1997). The authors rely upon different methods, formats and language resources, their resulting lexicons represent either surface or deep valency information and vary in sizes.

Our experiment is restricted only to nominal constituents in both simple and prepositional cases. We are aiming to extract surface frames similar to those from Ruslan.

Building a large scale valency lexicon - like Ruslan or later Vallex - is a costly and time-consuming effort which requires years of linguistic work. The automatization of this process is challenging, especially for some types of natural languages, as, e.g., the morphologically languages with free word order like Czech and Russian. In the sentences of free word order languages it is impossible to rely on the order of individual complements of a verb and thus their identification constitutes a complicated problem.

Another problem in automated valency frame extraction is the difficulty of classifying obligatory and optional complements. Quite often obligatory com-

[23] Some passages from the section come from the related paper (Klyueva – Kuboň, 2014) where the author of the thesis conducted all the experiments.

plements are omitted at the surface level because they can be identified in the previous context. This makes the classification practically impossible.

6.3.1 The Setup of the Experiment

We are aiming at using the simplest possible means in our experiment. It is highly desirable to use syntactic parsers to identify dependencies in sentences and syntactic types of the nominal groups (Subject, Object etc.). As we had some problems obtaining a parser for the Russian language, we will not use any parser in this experiment.

Our experiment consists of the following stages:

- adapting valency frames extracting information on a verb + surface frame
- corpus lookup - searching for the Czech verb+frame
- dictionary search for a Russian equivalent of the Czech verb and the complement
- Russian frame extraction from the Russian side of a parallel corpus

Next, we will describe each step in detail[24].

Processing Vallex Frames

We are exploiting only surface realization of verb complements, typically having the form of a case or a combination of a preposition and a case. For the moment we are leaving out the subject complements, assuming that the subject is mostly realized in Nominative in both Czech and Russian, thus it can be included into the Russian valency frame automatically.

The frames are transformed into a formeme-like format: a verb plus a case of an argument without a functor.
Following is an example of a Vallex frame slot in original format representing a functor Patient with various types of surface realizations - direct case realized by either Genitive or Accusative; infinitive or subordinate clause.

```
<slot functor='PAT' type='obl'>
  <form type="direct_case" case="2" />
  <form type="direct_case" case="4" />
  <form type="infinitive" />
```

[24] The script implementing the algorithm can be found here: https://github.com/natalink/CzeRuValency/blob/master/valency.pl

```
    <form type="subord_conj" subord_conj_lemma="aby" />
    <form type="subord_conj" subord_conj_lemma="at'" />
    <form type="subord_conj" subord_conj_lemma="že" />
  </slot>
```

The transformed entry[25]: **vyžadovat+2**, **vyžadovat+4** (to demand + Genitive, to demand + Accusative), the information on subordinate and infinitive clauses was ignored.

For prepositional valency, the format is a **verb+preposition+case** of an argument:
Original Vallex format:

```
<slot functor='PAT' type='obl'>
  <form type="prepos_case" prepos_lemma="na" case="6" />
  <form type="prepos_case" prepos_lemma="od" case="2" />
</slot>
```

Transformed: **záviset na+6**, **záviset od+2** (to depend on + Locative, depend from + Genitive).

Dictionary Lookup

For each Czech lemma from Vallex we search for the Russian translation equivalent in the Czech-Russian commercial dictionary[26] where the translations can be multiple. The Russian equivalents are then searched for in the parallel corpus in the next stage.

Parallel Corpus Lookup

The search is performed in the Czech-Russian part of the corpus UMC (see Section 3.1.1 for corpus description), containing 242,242 pairs of sentences aligned one-to-one. The texts are morphologically tagged, the tags contain a lemma, part-of-speech tag and other morphological characteristics as described in 3.3.2. The labels are assigned to each word in each sentence in the format **form|lemma|tag**.

In the first step of our algorithm, the corpus is searched sentence by sentence, until we identify a verb with surface valency frame matching to Vallex. Vallex then provides its valency pattern - Czech lemma and the surface realization of the nominal dependents - a noun or a pronoun within the same clause.

[25] Here we will again present cases as numbers because of the format of data involved in the experiment.

[26] http://www.langsoft.cz/

The bilingual dictionary then provides lemma(s) translation(s) which are looked up in the corresponding Russian sentence. In case of success (the verb corresponds to one of the lexical equivalents found in the translation dictionary), the respective case of a valency candidate (noun/pronoun)[27] is extracted and stored in the hypothesis set. Following is a chunk from the tagged sentence we used and an illustration of how we process it. The Czech tagger outputs the following information:

(6.14) mír|mír|NNIS1-----A---- vyžaduje|vyžadovat|VB-S---3P-AA---
komplexní|komplexní|AAIS4----1A-- přístup|přístup|NNIS4-----A--
' ... peace requires a complex approach ...'

The bilingual dictionary then provides the translation of the Czech verb *vyžadovat* – 'demand' into the corresponding Russian lemma *требовать*.

This lemma is then identified in the tagged Russian sentence:

(6.15) мир|мир|Ncmsnn требует|требовать|Vmip3s-a-e
всестороннего|всесторонний|Afpns-g-f подхода|подход|Ncmsgn

According to Vallex, the verb *vyžadovat* has two complements apart from the Actor in nominative. The dependent noun should be either in Genitive or in Accusative case, so we search[28] for a noun or a pronoun in Genitive or Accusative case. Genitive case is not found, so the only possible candidate to fill the valency slot of this verb is the noun *přístup* – 'approach'.

With the Czech complement identified, we get its lemma and search for its Russian equivalent from the dictionary. The translation of the Czech noun *přístup* is highly ambiguous, so we have to search for one the following Russian equivalents:

подход, подступ, право входа, допуск, приступ, обращение, доступ.

The only candidate present in the tagged Russian sentence is the noun *подход*. Its morphological tag **Ncmsgn** tells us that the corresponding case in Russian is Genitive (the **g** tag on the 5th position).

The algorithm applied on this clause therefore provides a frame hypothesis:

(6.16) (cz)*vyžadovat+Acc* => (ru)*требовать+Gen*

[27] We account always only for one complement per cycle.

[28] Here we will suppress some details such as optimization of search range (5 words around the verb) and a restriction within a clause.

The above hypothesis means that the Accusative case in the Czech valency frame (probably) corresponds to the Genitive case in Russian. Verbs requiring prepositional surface valency are processed in a similar manner, it is only necessary to identify both the preposition and the case in the Czech text and to take into account that a prepositional case in Czech may correspond to a non-prepositional in Russian and vice versa.

Russian surface frames and Statistics

Finally, we collect all hypotheses (like in Example 6.16) established in the preceding phases for a particular Russian verb and choose **the most frequent** Russian valency frame from this set[29].

The main statistics concerning the total number of patterns identified in the corpus and the number of the extracted patterns are presented in Table 6.5. The fact that we have been able to find equivalent frames for almost one third of verbs and their constituents on the basis of only slightly more than 240 000 sentences seems to be promising. Many patterns were simply not present in the data.

verb + surface form from Vallex	16561
"verb + surface form" matched in the corpus	14046
extracted patterns for Cz and Ru	**4286**

Table 6.5: Statistics of the experiment

The last line in Table 6.5 shows only the identified patterns, it does not reflect whether these patterns are correct or not. The errors we have discovered are discussed further.

The second interesting observation was made while we compared the obtained results with those from the manually created Ruslan. In order to draw a parallel, we splitted the set of frames into two parts - those with simple(non-prepositional) case and those with prepositional ones, like we did with Ruslan.

Simple case The results for the simple case correspondences are presented in Table 6.6. According to this table, out of the total of 1727 surface cases, 343 are different. This represents 19.86% of the total. This number is 6 times higher than the respective figure from Ruslan (Table 6.1). A frequent co-occurence cz:Acc vs. ru:Gen (196 times) reflects the most error-prone frame of our algorithm which we will discuss later in more detail.

[29] The output of the search algorithm can be found here: https://github.com/natalink/CzeRuValency/blob/master/valency_06112013.out.

		Czech			
		Genitive	Dative	Accusative	Instrumental
Russian	Genitive	**21**	20	196(!)	15
	Dative	1	**159**	12	2
	Accusative	8	23	**1026**	22
	Instrumental	4	6	34	**178**
Different surface frames					343 (19.8%)
Total number of surface frames					1727 (100%)

Table 6.6: Co-occurence of the same simple case in Czech and Russian

Prepositional frames

The results for prepositional valency are presented in Table 6.7, Russian prepositions are transliterated. Due to a large number of very rare (and thus unreliable) correspondences we included only those from the table (i.e. those which occurred more than 10 times). The top of the Table is quite similar to the one from manually created Ruslan, see Table 6.3 for comparison.

The pairs with very low frequency are very unreliable, so it would be very doubtful to perform the comparison on all of them. We calculated different prepositional frames only on those most frequent pairs from the table just for the sake of completeness. As in the case with Ruslan, we consider a preposition in Czech as an equivalent of a Russian one if they are typical translation of each other (see Section 6.2.1 for examples) and cases of complements are the same.

Out of the total of 841 prepositional pairs from Table 6.7 there were 154 different unique pairs. That represents 18.3% of the total, slightly more than in the case of non-prepositional cases. However, the 'comparison' from Tables 6.6 and 6.7 should be taken with caution because it is made on automatically extracted and highly erroneous data.

6.3.2 Error Analysis

Manual Error Analysis

As expected, the frequency of errors in our automatically extracted lexicon was quite high. In order to detect errors, to discover a reason why they occurred, we have performed a manual evaluation of a small sample of valency frames. Out

Czech	Russian	freq
na+Acc	na+Acc	159
k+Dat	k+Dat	82
s+Ins	s+Ins	78
z+Gen	iz+Gen	58
v+Loc	v+Loc	56
za+Acc	za+Acc	52
do+Gen	v+Acc	50
od+Gen	ot+Gen	42
o+Loc	o+Loc	35
na+Loc	na+Loc	33
na+Acc	**v+Acc**	32
na+Acc	**na+Loc**	22
z+Gen	**s+Gen**	20
v+Acc	v+Acc	18
na+Acc	**k+Dat**	18
k+Dat	**na+Acc**	16
na+Acc	**v+Loc**	14
před+Ins	ot+Gen	12
proti+Dat	protiv+Gen	12
za+Acc	**na+Acc**	11
na+Acc	**o+Loc**	11
k+Dat	**v+Acc**	10

Table 6.7: Prepositional case correspondence.

of the set of 4,286 extracted frames (Table 6.5) we have manually evaluated 200 verb+frame pairs. Among those, 24 frames, i.e., 12% of the sample, were marked as incorrect. Some errors were caused by tagging inaccuracy, others resulted from an erroneous match of Czech and Russian nouns, and the rest can be attributed to other factors, as, e.g., bilingual dictionary issues.

After simple marking the erroneous entries, we have tried to predict which pairs of frames in Czech and Russian are most likely to cause an error.

- **Tagger inaccuracy** The most frequent error (196 times) has a pattern:

  Czech: Verb+Acc => Russian: Verb+Gen.

This error pattern has its roots in the tagger inaccuracy due to the morphological ambiguity. In Russian, an animated masculine noun has the same form in Genitive and Accusative cases[30], and the tagger often confuses them. So even if the algorithm matches all the dependencies correctly, the extracted case of the Russian noun is incorrect. Let us present an example:

(6.17) ERR: *najímat*+Acc => *нанимать*+Gen (to hire smb.)

The Russian morphological case should be also Accusative. Probably, because the complement of this verb is always animated and is often wrongly tagged as Genitive in the Russian corpus, it was the most frequent hypothesis, so it was selected

- **Experiment setup** As we have already mentioned, our approach is rather shallow and does not take into account syntactic functions or functors. While examining some 'suspicious' cases which contain a prepositional valency frame in one language and a simple one in the other. Let us illustrate this on the following entry:

  (6.18) *odebrat*+Acc (take smth.) => *отобрать*+у+Gen (take from smb.)

  In this example, the Czech morphemic case Accusative is a surface realization of the functor Patient (PAT), whereas the Russian surface form *у+Gen* is a realization of the functor Experiencer (EXP). On the both sides there should be a complement with either PAT or EXP case/functor for both Czech and Russian. However, due to our very shallow approach and lack of syntactic or semantic parsers, those two roles were confused that led to an error.

  Let us look at this case more closely and examine a sentence from our corpus containing this example[31]:

  (6.19)

  (cz) *Simeonovovi* *odebrali* *dort*
  Simeonov.**Dat**-EXP took.3Pl cake.**Acc**-PAT
  'They took a cake from Simeonov'

  (ru) *торт* *отобрали* *у* *Симеонова*
  cake.**Acc**-PAT take.3Pl **from** Simeonov.**Gen**
  'They took a cake from Simeonov'

[30] *Он пришел без друга.Gen* – 'He came without his friend' vs. *Я вижу друга.Acc* – 'I see a friend'. This ambiguity also holds in Czech, but it is not relevant here because the respective case from a frame comes from the manually written Vallex.

[31] In order to simplify the text we leave only the relevant morphological tags.

Although our algorithm has identified both dependencies - object and indirect object, the latter has got mixed up because of the reversed word order in Russian. The same situation was observed in many sentences - when the algorithm chose the most frequent variant, it turned out that it was an incorrect one for that particular verb. It should be noted that the correct valency frame for the indirect object was generated as well, but it was not the most frequent hypothesis:

(6.20) *odebrat*+Dat => *отобрать*+у+Gen(take from smb.)

This mistake is beyond the abilities of our simple algorithm, a possible solution of this problem is to use some deeper parsing strategy which would be able to identify the type of the noun phrases involved.

Comparison with Ruslan frames

It would be a natural thing to compare the generated pairs [verb+frame] with 'golden' data from Ruslan. We can not name it *evaluation against the manually created frames*, because of the nature of the dictionary. Many verbs from Ruslan are not present in the parallel corpus and vice versa, so the evaluation is of a very approximate nature.

We have selected the verbs from Ruslan that also occurred in the automatically generated lexicon (695 verbs). For each [verb+frame] pair from Ruslan and the lexicon we calculated the number when the frame matches - in 309 cases (44%).

Following are the examples of comparing verb+frame pairs from Ruslan and the lexicon:

Czech	Ruslan	Lexicon	match
patřit do+Gen	принадлежать к+Dat	принадлежать к+Dat	+
uškodit+Dat	навредить+Dat	повредить+Dat	**almost**
skládat do+Gen	складывать **из+Gen**	складывать **до+Gen**	–

Table 6.8: Examples of compring frames from the lexicon to Ruslan frames

This result can not tell us much about the quality because we compare incomparable, but at least we know that our algorithm generated a sufficient percentage of correct frames.

Discussion

To conclude, the percentage of correctly identified frames suggests that even such a naive and simplistic approach may lead to a relatively fast method of creating large scale valency lexicon for another language (Russian) from the resources of a related language. The automatically extracted lexicon will not be used in any experiments as the results are not reliable enough. The method introduced here might be improved through the exploitation of a syntactic parser. This would enable us to include also the complements which cannot be easily identified in a sentence, as, e.g., the long-distance dependencies.

6.4 Surface frame discrepancies and verb classes in Vallex

This section describes another experiment with Vallex that concerns more semantic level of representation. It includes theoretical observations on discrepancies in surface frames without a chance of being used in any application. The two previous sections suggest that about 10-15 % of verbs in Czech and Russian exhibit surface valency discrepancies. The question arises which verbs tend to have different valency frame in the two languages.

One of the obvious suggestions is that the discrepancies can be connected to semantic classes - for instance, if valency frames of verbs in Czech and Russian are different, the probability is high that verbs close in meaning will also show the same discrepancy. This hypothesis goes in hand with (Levin, 1993) in which it was stated that verbs from the same verb classes exhibit similar syntactic behavior.

We have mentioned that Vallex entries contain information on a semantic class of a verb. Vallex distinguishes 22 verb classes, among them are communication, exchange, motion, perception, transport, psych verbs to name some of them. Naturally, words that belong to the same semantic field or share some component of meaning will have similar valency frames. Unlike the previous two experiments, here we will also make use of functors (semantic roles) as the surface discrepancies are connected to certain functors.

6.4.1 Frame comparison

We made a comparison of Czech and Russian frames with respect to word classes in the following way. For each word in a semantic class we took a Czech verb and check if the surface frame (for nominal complements only) fits the one of a

Russian equivalent verb[32]. It was impossible to evaluate a big amount of verb frames (in total 2,903 lexical units have a verb class assigned), so we have taken only several verb classes: motion, communication, change, exchange and mental action[33].

Example 6.21 shows the Czech and Russian verbs with the same valency frames:

(6.21)

(cz) *obhajovat ACT(Nom) PAT(Acc)* – 'to defend'

(ru) *защищать ACT(Nom) PAT(Acc)* – 'to defend' - **matches**

The verb in Example 6.22 has two incorrespondences in it:

(6.22)

(cz) *blahopřát ACT(Nom) PAT(Dat) CAUS(k+3)* – 'to congratulate'[34]

(ru) *поздравлять ACT(Nom) PAT(Acc) CAUS(с+7)* – 'to congratulate' - **does not match**

Example 6.22 illustrates that in Czech and Russian different prepositions and different cases are used to express the semantic roles Patient and Cause.

Thus we consider the Russian surface frame to be different from the Czech one if there exists some frame slot with some distinct semantic role for which the Czech and Russian surface realizations are different. As we tried to find the most typical translation equivalent in Russian, the semantic roles were the same in both languages. If a surface form is represented by a preposition with some case, we judge the default translation of prepositions (see the discussion in the Section 6.2) as the similar realization. For example, the Czech frame element PAT(**před**+**Ins**) corresponds to the Russian PAT(**перед**+**Ins**) and is equal with respect to surface valency. Another sense of the same Czech prepositional phrase PAT(**před**+**Ins**) has different surface form in Russian PAT(**от**+**Gen**) (from+Gen).

While analyzing Czech and Russian frames it became evident that differences in valency frames can be either regular or occasional within the individual class. Next, we will present the description of differences in semantic classes of verbs. The overall table 6.9 will be introduced after the description classes.

[32] As in the two previous sections, we will refer to the 'surface frame of nominal complements' just as **frame** for the sake of shortness.

[33] We wanted to test which subclass of verbs - those of physical or mental activity has more verbs with surface frame incorrespondences in Czech and Russian.

[34] The German frame: *gratulieren zum +3* matches the Czech one.

Also, after examining the whole list of differences, it became evident that when there is some discrepancy in Czech and Russian surface valency, the surface frame of a Czech verb quite often matches the one in the German language. As the main scope of this work is the comparison between Czech and Russian, we will mention corresponding German frames only as additional observations. It is possible that this can be a consequence of a language contact (influence of German on Czech surface valency was described e.g. in (Berger, 2008)). Further, we will indicate when Czech and German frame slots match while Czech and Russian do not.

6.4.2 Class of Change

Verbs of the class **Change** often have the complement **DIFF** that characterizes the rate of change, and we observed that it often has different surface realizations in Czech and Russian. For example, the Czech slot DIFF*o+Acc* – 'about+Acc' generally corresponds to *na+Acc* – 'on+Acc' in Russian (other variations are possible), see examples Examples 6.23 and 6.24.

(6.23)

(cz) *ceny klesly* ***o*** *20%* 'prices fall **by** 20%'[35] - functor Diff

(ru) *цены упали* ***на*** *20%* 'prices fall **on** 20%' - functor Diff

(6.24)

(cz-o) *Administrace zkrátila dovolenou* ***o*** *2 dny*[36]
'administration shortened the holiday **by** 2 days'

(cz-na) *Administrace zkrátila dovolenou* ***na*** *2 dny*
'administration cut off the holiday **to** 2 days'

(ru) *Администрация сократила отпуск* ***на*** *2 дня*
'administration cut off the holiday **on**/**to** 2 days'

In Example 6.24 we can spot the ambiguity of Russian surface form *на+Acc* with the functor DIFF for verbs of change. The sentence 6.24(ru) can be interpreted both as (cz-o) and (cz-na): that the administration shortened holidays by two days or to two days.

For the functor DIFF, the surface form *o+Acc* is typical for Czech while Russian language uses the preposition *о* – 'about' mainly with mental predicates

[35] The German surface frame: *sinken um + Acc* matches the Czech

[36] German surface frame: *kürzen um + Acc* matches the Czech

(e.g. *забыть о+Loc* – 'forget about') or communication verbs (e.g. *рассказать о+Loc* – 'tell about'). It does not occur with the Accusative case in Russian at all.

Verbs from this class are troublesome especially for the Rule-Based systems as the preposition is translated into Russian as *о* – 'about' by default, which makes the translation completely wrong. Some of the cases were covered by verbs from the Ruslan converted lexicon (Section 6.2.3) for the TectoMT system, but there are more of them not covered by the lexicon. Here we can suggest some improvements that may be made in future in case we have a parser. If the verb is identified as belonging to this class and the complement realization of DIFF is *о+Acc*, we can apply the rule to set the surface form of the complement in Russian to *на+Acc*.

6.4.3 Class of Motion

We have not found many dissimilarities in Czech and Russian valency frames within the class of **Motion** verbs. One most evident is that verbs with the semantic component of 'going away from somewhere' in Czech have the surface realization of PAT as *před+Ins* – 'before+Ins', but are translated into Russian with the respective verb plus the prepositional phrase *от+Gen* – 'from+Gen'. Just to name some of these verbs: *prchat* – 'be on the run', *ujíždět* – 'speed off', *unikat* – 'escape'.

(6.25)

(cz) *prchat před policii* – 'run before police'[37]

(ru) *убегать от policii* – 'run from police'

Roughly speaking, Russian prefers the preposition *от* – 'from' whereas Czech uses *před* – 'before' in this context. Verbs of other semantic classes with the similar component of meaning, as, e.g. the Location class, share this rule as well (cz:*schovat před+Ins* – 'to hide before' vs. ru:*спрятать от+Gen* – 'to hide from'.

The following example illustrates a coincidental difference in a verb frame, when only one verb from the class has got it[38]:

(6.26)

[37] Only prepositions in the Czech and German frame match, but not the case: *fliehen vor + Dat*

[38] the Czech frame of a verb *trefit+Acc* – 'hit' corresponds to the German one *treffen+Acc*

(cz) *trefit PAT(Acc)* – 'to hit smth '

(ru) *попасть PAT(в+Acc)* – 'to hit into+Acc'

6.4.4 Verbs of Exchange

One of a regular and rather evident discrepancies in Czech and Russian with the respect to surface frames was described in (Lopatková – Panevová, 2004). Some Czech verbs of exchange with the meaning of removing something from someone, ex. *sebrat* – 'take away', *krást* – 'steal', *brát* – 'take' etc. exhibits the regular difference in valency in contrast to the respective Russian equivalent verbs. The addresse (ADDR) functor here is a person or an object from whom something is taken, and in Czech it has surface realization with simple case [Dat][39], whereas in Russian it is a prepositional case with the Genitive case [*у+Gen* – 'from+Gen']:

(6.27)

(cz) *Bere dítěti hračku*
takes baby.**Dat** toy
'(He) takes away a toy from a baby'

(ru) *Он берет у ребенка игрушку*
He takes **from** baby.**Gen** toy
'He takes away a toy from a baby'

If the sentence from Example 6.27(cz) was translated into Russian according to the Czech valency pattern, it would have the reverse meaning in Russian, because Dative case of a noun in this context is understand as Benefactor (BEN) (taken TO someone), not Addressee (taken FROM someone). However, the sentence in Czech is ambiguous as it can have a meaning lit. *He takes a toy to a baby.* In this case, the functor of a Czech complement is BEN, not ADDR.

Especially this causes big problems to learners of foreign language: they project the known pattern from their native language on the phrase to the foreign language and often make a mistake.

The same is true for the metaphorical sense of such verbs, e.g. the verb *zabírat ADDR(**Dat**)* – 'take(time)+Dat':

(6.28)

(cz) *studium mi zabírá hodně času*
study me.**Dat** takes many time
'Study takes me a lot of time.'

[39] The same holds in German: *nehmen + Dat* – 'take+Dat'

(ru) *учеба отнимает у меня много времени*
study takes **from** me.**Gen** many time
'Study takes me a lot of time.'

On the example of this verb class we can see that some semantically related verbs have different surface realization of a functor (ADDR in this case) in Czech and Russian. Only two words of this class with different valency frame (*odpírat* – 'deny' and *opatrovat* – 'look after') does not have the pattern described[40], and we consider them to be occasional incorrespondences.

6.4.5 Class of Communication

Czech and Russian verbs of this class have many differences with respect to valency. Here we could not observe some leading difference as in the previous classes. Incorrespondences concern several functors and several surface forms. They may be considered coincidental, but we can identify several functors for which surface forms can be different in Czech and Russian:

1. **ADDR(na+Acc)** The functor Addressee with the surface form ADDR: **(na+Acc)**-(on+Acc) in Czech is presented differently in Russian - with another preposition and another case:

   (6.29)

   (cz) *mluvil na bratra*
   spoke **on** brother.**Acc**
   'He spoke to his brother'

   (ru) *он говорил с братом*
   He spoke with brother.**Dat**
   'He spoke to his brother'

   The Russian sentence in Example 6.29 can also be translated into Czech as *mluvil s bratrem* – 'speak with brother'. So, there are two meanings of the verb *mluvit* in Czech and only one in Russian.

   The ADDR functor of another Czech verb - *zavolat na + Acc* – 'to call smb' in Russian has got a surface realization without a preposition:*позвать+Acc* – 'to call smb'. As with the verb *mluvit*, the Russian equivalent *позвать+Acc* can be tranlated into Czech as *zavolat+Acc* in a slightly different sense which is not distinguished in Russian.

[40] We counted in only the verbs from the Vallex class, there can be more if we consider all the verbs of this class.

2. **PAT(na+Loc)** The functor Patient with the surface form **(na+Loc)**-(on+Loc) in Czech has another surface realization in Russian, generally the morphemic form is **(o+Loc)**-(about+Loc) for such verbs of 'asking question' as *ptát se* – 'ask', *tázat se* – 'to question' etc.:

   (6.30)

   (cz) *ptát se* *na* *zdraví*
   ask **on** health.**Acc**
   'to ask about health'

   (ru) *спрашивать* *о* *здоровье*
   ask about health
   'to ask **about** health.**Loc**'

   Other verbs of 'speaking' with a frame slot **PAT(na+Loc)**-(on+Loc) are also very similar to the above sample, e.g. Czech *domlouvat se* PAT(**na+Loc**) vs. Russian *договориться* PAT(**o+Loc**) 'to agree on/about'. In Czech, a surface realization *domlouvat se* PAT(**o+Loc**) is also possible.

3. **ADDR(Dat)** Addressee in Dative case for the following verbs corresponds to Accusative in Russian:

   (6.31)

   (cz) *poblahopřát* ADDR(**Dat**) 'congratulate +Dat'[41]

   (ru) *поздравить* ADDR(**Acc**) 'congratulate +Acc'

   (6.32)

   (cz) *děkovat* ADDR(**Dat**) 'thank + Dat'[42]

   (ru) *благодарить* ADDR(**Acc**) 'to thank + Acc'

4. **PAT(o+Acc)** Similar to Change class (Example 6.23), some complements of Czech communication verbs with surface form **(o+Acc)**-(about+Acc) have another surface realization in Russian due to the fact that the Russian preposition *o* – 'about' does not combine with Accusative case at all:

   (6.33)

   (cz) *hlásí se* *o* *slovo*
   asks **about** word.**Acc**
   'She asks for a word'

[41] Compare with German *gratulieren + Dat*

[42] German frame slot: *danken + Dat*

(ru) *она просит слова*
She asks word.**Gen**
'She asks for a word'

5. There are several coincidental differences occurring only once or twice that do not go to any scheme, e.g.:

(6.34)

(cz) *doznávat se* PAT(**k**+**Dat**) 'confess to smth'[43]

(ru) *признаваться* PAT(**в**+**Loc**) 'to confess in smth'

6.4.6 Class of Mental Action

Verbs of this class often have differences in valency frames but they are rather coincidental and there is no regular pattern of difference. The most frequent discrepancy was in PATIENT functor with the surface form **(na+Acc)**. In Czech, it is also different for verbs of class Communication (see Example 6.30) but fo that class it was regularly translated as *o+Loc* – 'about+Loc'. For the class of Mental Action no common frame translation equivalent exists and **PAT(na+Acc)** corresponds to several surface forms in Russian:

(6.35)

(cz) *pamatovat* PAT(**na**+**Acc**) 'to remember on'

(ru) *помнить* PAT(**про**+**Acc**) 'to remember about'

(6.36)

(cz) *myslet* PAT(**na**+**Acc**) 'to think on'[44]

(ru) *думать* PAT(**о**+**Loc**/**про**+**Acc**) 'to think about'

(6.37)

(cz) *zvykat si* PAT(**na**+**Acc**) 'get used on'[45]

(ru) *привыкать* PAT(**к**+**Dat**) 'to get used to'

[43] Gemran frame: *sich bekennen zu + Dat*

[44] German slot: *denken an + Acc*

[45] German slot: *sich gewöhnen an + Acc*

The structure of the following verbs coincides a lot with that from Examples 6.31 and 6.32, though the functor here is PAT, not ADDR:

(6.38)

(cz) *rozumět* PAT(**Dat**) 'understand'

(ru) *понимать* PAT(**Acc**) 'understand'

Following is another example of the coincidental difference in surface frames in cases:

(6.39)

(cz) *pohrdat* PAT(**Ins**) - 'to despice'

(ru) *презирать* PAT(**Acc**) 'to despice'

6.4.7 Overall results on verb class differences

In this experiment, we have compared Czech and Russian surface frames of verbs from 5 semantic classes, totally 1473 lexical entries and examined the connection between surface valency discrepancies and verb classes. The discrepancies between surface verb frames in the two languages can be either regular or coincidental (no semantically related word exhibits the same discrepancy in valency). We made the following observations:

- Most incorrespondences occur in prepositional phrases.
- Within a verb class we can often find some typical (regular) surface valency patterns of Czech verbs which correspond to certain Russian patterns.
- Quite a few Czech surface frames different from Russian follow the same pattern as German. We did not study this issue in detail, but, obviously, this can be attributed to the language contact between Czech and German[46].

The table 6.9 presents the distribution of verbs with different frames according to the verb classes.

[46] It is mostly the vocabulary, not grammar that is generally borrowed from another language, but here we can see that surface valency can be a calque from another language. Also, some observation with the respect to language acquisition. A child of the author (5 years) often uses a Czech morphemic case with a Russian verb, ex. *подожди *на меня* – 'wait for me' is a 'valency calque' from a Czech *počkej na mě* – 'wait for me'. The correct surface form of a Russian complement is the Accusative case without a preposition - *подожди меня* – 'wait me.Gen'.

Verb class	same frame	different frame	# of verbs
Change	309(95%)	14(5%)	323
Exchange	166(92%)	13(8%)	179
Motion	305(99%)	3(1%)	308
Communication	312(88%)	42(12%)	354
Mental Action	270(87%)	39(13%)	309
Total	1362(92%)	111(8%)	1473

Table 6.9: Differences according to the verb classes

From this table we can see that verbs of physical activity (change, motion, exchange) are less likely to have some incorrespondences in surface frames than verbs of mental activity (communication, mental action).

The observations presented here are only of a descriptive nature and does not contain results that can be integrated into the Machine Translation system.

6.5 Discussion on Valency issues

In this chapter, we have discussed some practical aspects of valency and concentrated on the discrepancies in surface valency for Czech and Russian verbs.

The main contribution of this research is the lexicon of Czech and Russian verbs plus their surface frames[47] that was extracted from Ruslan dictionary encoded in Q-Systems format. Then, we have automatically identified the discrepancies in Czech and Russian surface frames.

We have tried to incorporate this lexicon into the Rule-Based system TectoMT from Czech into Russian. The automatic evaluation metric BLEU showed an insignificant decrease, but the manual evaluation of the changes that were introduced by the rule showed that in almost 60% the incorrect valency was improved.

The next step of our research on valency was an attempt to automatically construct a lexicon of surface frames from a parallel corpus exploiting a simple algorithm. The manual evaluation of a small sample revealed the major drawbacks of our approach: more sophisticated information is needed for a proper frame extraction - mainly, a syntactic parser would help.

[47] https://github.com/natalink/CzeRuValency/blob/master/python/lexicon

Finally, we presented more detailed linguistic observations on which verbs tend to have some discrepancy in surface form of arguments. We have explored 5 verb classes and proved our initial hypothesis that some groups of verbs with similar meaning tend to have the same 'discrepancy pattern' in surface valency. Also, the final figure indicating number of verbs with discrepancies in surface frames (8%) agrees with the number of such calculated on Ruslan surface frames.

7

Conclusion

The results presented in our research are theoretical and practical, both contributing to comparative linguistics in a form of analysis of discrepancies between Czech and Russian and computational linguistics in a form of new data.

Data

One of the main output of our research are data that we collected/created for the purpose of MT and that can be exploited in other experiments:

- **Czech-Russian parallel corpus** We have automatically downloaded a parallel corpus UMC that was therefore morphologically tagged. It served mainly as a part of training data for SMT. The corpus was exploited to extract Czech-Russian dictionary and surface valency frames. Also, some linguistic phenomena like usage of pronouns, transgressives were explored using this parallel corpus.
- **Czech-Russian dictionary** was automatically extracted from the parallel corpus, cleaned and checked against manually written dictionary. The quality of the dictionary is not so good, but to the best of our knowledge there is no freely available Czech-Russian dictionary. The dictionary was used for implementing Czech-to-Russian machine translation within the RBMT systems TectoMT and Česílko.
- **A list of surface valency frames** was extracted from the outdated dictionary Ruslan. Though the original resource contained many obsolete constructions, the majority of verbs extracted still belong to the common language. A list was therefore was used in the comparative analysis of Czech and Russian surface valency frames.

MT systems

The main objective of this work was to built experimental MT systems - rule-based and statistical, and then to judge the performance of both types of the

systems from a linguistic perspective. We worked with the rule-based MT system TectoMT and the statistical Moses for Czech-Russian. Also, we compared the performance with the two commercial systems - RBMT PC Translator and Google Translate for Czech-to-Russian. As for the BLEU score, SMT systems (Google and Moses) scored almost 3 times as better as RBMT (TectoMT and PC Translator).

The manual evaluation of errors in the output of the four systems suggested that BLEU correlates with manual evaluation as there were much many words marked as errors in the RBMT output than for those of SMT.

One of the initial hypothesis was that MT between the related languages under the similar settings could bring better results than when constructing an MT between unrelated turned out to be false. The morphological complexity of the two languages hurts more than the relatedness helps.

Differences between Czech and Russian and MT

Our work contributes to the comparative studies of the two languages by exploring the MT output and tying the errors to concrete language discrepancies. We proposed a classification of errors relevant for the language pair under consideration. We concluded that discrepancies between the two languages has more impact on RBMT systems than on SMT.

RBMT and SMT strategy of language acquisition

We want to conclude this thesis with another vision of RBMT and SMT systems that was cultivated while working on the output of the two types of systems. The metaphor we want to present is based on a parallel between machine translation and language acquisition.

Rule-Based MT systems resembles learners who started to learn a language exploiting traditional method of language acquisition: learning words (a dictionary) and a set of rules to combine the words (rules in RBMT). This learning strategy brings reasonable results only after investing a huge amount of efforts and time, which is true both for RBMT and for a typical second-language learner. The learners and RBMT system produce errors of the same nature as they transfer the features from the native/source language into the target and the only way to fix the problem is to learn the rule/add a new word into the dictionary.

SMT systems can be thus compared to either children learning their language from their surroundings or learners who live in the target language environment. They do not need any linguistic rules at all. The only strategy that works for both SMT and the learners is obtaining as much data as possible; the acquisition of data is done in a 'black box' that is not subjected to any control. We can

only speculate whether a human brain can store language data in a phrase-based manner (as n-grams), but, evidently, the statistical 'approach' to language learning is more efficient than the rule-based. It allows us to acquire a language in a short amoun of time without implicit knowledge of abstract rules.

Both statistical and rule-based approaches benefit from borrowing some features from each other, both for the MT systems and for the language learners.

List of Figures

List of Tables

A

Appendix

Links to data:

- Parallel Czech-Russian Corpus UMC with morphological annotation: http://hdl.handle.net/11858/00-097C-0000-0001-4909-7
- Czech-Russian dictionary
- Czech-Russian surface frames extracted from Ruslan https://github.com/natalink/CzeRuValency/blob/master/python/lexicon

Bibliography

Anastasiou, D. *Idiom Treatment Experiments in Machine Translation.* Cambridge Scholars Publishing, 2010.

Apresjan, V. Active Dictionary of the Russian Language: Theory and Practice. 2011. Proceedings of the 5th International Conference on Meaning-Text Theory. Barcelona.

Babby, L. H. *Existential Sentences and Negation in Russian.* 1980.

Baker, C. F. – Fillmore, C. J. – Lowe, J. B. The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, s. 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

Baobao, C. – Danielsson, P. – Teubert, W. Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing*, 2002.

Barbara H. Partee, V. B. Existential Sentences, BE, and the Genitive of Negation in Russian. *Conference on Existence: Semantics and Syntax.* 2002.

Barbara H. Partee, V. B. E. V. P. – Testelets, Y. Russian Genitive of Negation Alternations: The Role of Verb Semantics. *Scando-Slavica, 57:2, 135-159.* 2011.

Bejček, E. – Straňák, P. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation.* 2010, 44, 1-2, s. 7–21.

Benešová, V. – Lopatková, M. – Hrstková, K. Enhancing Czech Valency Lexicon with Semantic Information from FrameNet: The Case of Communication Verbs. In *ICGL 2008 Proceedings of the First International Conference on Global Interoperability for Language Resources*, s. 18–25, Hong Kong, China, 2008. City University of Hong Kong.

Berger, A. L. et al. The Candide system for machine translation. In *In Proceedings of the ARPA Conference on Human Language Technology*, s. 157–162, 1994.

Berger, T. Deutsche Einflüsse auf das grammatische System des Tschechischen. In *Studien zur historischen Grammatik des Tschechischen*, Bohemistische Beiträge zur Kontaktlinguistik. 2008. s. 57–69.

Bílek, K. – Klyueva, N. – Kuboň, V. Exploiting Maching Learning for Automatic Semantic Feature Assignment. In Boonthum-Denecke, C. – Youngblood, M. (Ed.) *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2013*, s. 297–302, Palo Alto, California, 2013. FLAIRS, AAAI Press. ISBN 978-1-57735-605-9.

Boguslavsky, I. A bi-directional Russian-to-English machine translation system (ETAP-3). In *Proceedings of the Machine Translation Summit V. Luxembourg*, 1995.

Boguslavsky, I. et al. Dependency treebank for Russian: Concept, tools, types of information. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, s. 987–991. Association for Computational Linguistics Morristown, NJ, USA, 2000.

Bojar, O. – Homola, P. – Kuboň, V. An MT System Recycled. In *Proceedings of MT Summit X*, s. 380–387, September 2005.

Bojar, O. *Čeština a strojový překlad*. 11 / *Studies in Computational and Theoretical Linguistics*. ÚFAL, 2012. ISBN 978-80-904571-4-0.

Bojar, O. Analyzing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*. 2011, 95, s. 63–76. ISSN 0032-6585.

Bojar, O. – Šindlerová, J. Building a Bilingual ValLex Using Treebank Token Alignment: First Observations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, s. 304–309, Valletta, Malta, 2010. ELRA, European Language Resources Association.

Bojar, O. – Tamchyna, A. Forms Wanted: Training SMT on Monolingual Data. In *Proceedings Research Workshop of the Israel Science Foundation University of Haifa, Israel*, 2011.

Bojar, O. – Tamchyna, A. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*. 2013, 99, s. 39–58. ISSN 0032-6585.

Bojar, O. – Rosa, R. – Tamchyna, A. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, s. 92–98, Sofija, Bulgaria, 2013. Bălgarska akademija na naukite, Association for Computational Linguistics.

Bojar, O. – Hajič, J. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, s. 143–146, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Bojar, O. – Prokopová, M. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, s. 1236–1239, 2006.

Bouamor, D. – Semmar, N. – Zweigenbaum, P. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In Chair), N. C. C. et al. (Ed.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

Brown, P. F. et al. A Statistical Approach to Machine Translation. *Comput. Linguist.* 1990, 16, 2, s. 79–85.

Bílek, K. A Comparison of Methods of Czech-to-Russian Machine Translation. Master's thesis, 2014.

Chandioux, J. *Meteo (tm), an operational translation system.* 1988.

Clancy, S. J. The Chain of Being and Having in Slavic (= Studies in Language Companion Series 122). 2010.

Doddington, G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, s. 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Dušek, O. et al. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, s. 267–274, Montréal, Canada, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6.

Eva, H. – Sgall, P. – Panevová, J. *The Meaning of the Sentence in Its Pragmatic Aspects, Reidel.* 1986.

Fillmore, C. J. The Case for Case. In *Universals in Linguistic Theory.* 1968.

Forcada, M. L. et al. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation.* 2011, 25, 2, s. 127–144. Special Issue: Free/Open-Source Machine Translation.

Galuščáková, P. – Popel, M. – Bojar, O. PhraseFix: Statistical Post-Editing of TectoMT. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, s. 141–147, Sofija, Bulgaria, 2013. Bălgarska akademija na naukite, Association for Computational Linguistics.

Galuščáková, P. – Bojar, O. Improving SMT by Using Parallel Data of a Closely Related Language. In *Human Language Technologies – The Baltic Perspective – Proceedings of the Fifth International Conference Baltic HLT 2012*, 247 / *Frontiers*

in AI and Applications, s. 58–65, Amsterdam, Netherlands, October 2012. IOS Press. ISBN 978-1-61499-132-8.

Gispert, A. – Mariño, J. B. – Crego, J. M. Improving statistical machine translation by classifying and generalizing inflected verb forms. 2005, s. 3185–3188.

Hajic, J. RUSLAN: An MT System Between Closely Related Languages. In *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics*, EACL '87, s. 113–117, Stroudsburg, PA, USA, 1987. Association for Computational Linguistics. doi: 10.3115/976858.976879. Dostupné z: <http://dx.doi.org/10.3115/976858.976879>.

Hajič, J. – Hric, J. – Kuboň, V. Machine Translation of Very Close Languages booktitle = In Proceedings of the 6th Applied Natural Language Processing Conference . 2000a.

Hajič, J. *Disambiguation of Rich Inflection - Computational Morphology of Czech.* I. Prague Karolinum, Charles University Press, 2001. 334 pp.

Hajič, J. – Kuboň, V. – Hric, J. Česílko - an MT system for closely related languages. In *ACL2000, Tutorial Abstracts and Demonstration Notes*, s. 7–8. ACL, ISBN 1-55860-730-7, 2000b.

Hajič, J. – Homola, P. – Kuboň, V. A Simple Multilingual Machine Translation System. In Hovy, E. – Macklovitch, E. (Ed.) *Proceedings of Machine Translation Summit IX*, s. 157–164, New Orleans, USA, 2003.

Hajič, J. et al. Prague Dependency Treebank 2.0. LDC2006T01, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, Jul 2006, 2006. Dostupné z: <http://ufal.mff.cuni.cz/pdt2.0/>.

Hajičová, E. – Partee, B. – Sgall, P. *Topic-focus articulation, tripartite structures, and semantic content.* Kluwer Academic Publishers, 1998.

Hana, J. Czech clitics in Higher Order Grammar. In *Working Papers in Slavic Studies.* Columbus, Ohio: Department of Slavic and East European Languages and Literatures, 2004.

Hartley, A. – Babych, B. – Sharoff, S. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *In Proceedings of the MT Summit XI, pp. 412–418, Copenhagen*, 2007.

Hladná, V. Valence sloves v českých a ruských publicistických textech [online]. Diplomová práce, UNIVERZITA PALACKÉHO V OLOMOUCI, Filozofická fakulta, 2012.

Hlaváčková, D. Verbalex – new comprehensive lexicon of verb valencies for czech. In *In Proceedings of the Slovko Conference*, 2005.

HOMOLA, P. *Syntactic Analysis in Machine Translation.* 6 / *Studies in Computational and Theoretical Linguistics.* Institute of Formal and Applied Linguistics, 2009. ISBN 978-80-904175-7-1.

HUTCHINS, W. J. Machine translation: A concise history, 2007.

KINGSBURY, P. – PALMER, M. From treebank to propbank. In *In Language Resources and Evaluation*, 2002.

KIPPER, K. et al. A large-scale classification of English verbs. *Language Resources and Evaluation.* 2007.

KIRSCHNER, Z. – ROSEN, A. APAC - An experiment in machine translation. *Machine Translation.* 1989, 4, 3, s. 177–193.

KLYUEVA, N. – KUBOŇ, V. Verbal Valency in the MT Between Related Languages. In *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*, Pisa, Italy, 2010.

KLYUEVA, N. Usage of some non-finite constructions in Czech and Russian. In *6th Annual International Conference on Languages & Linguistics*, s. 5–12. Athens Institute for Education and Research, Atiner, 2013.

KLYUEVA, N. – BOJAR, O. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *Proceedings of the Conference Korpusnaja lingvistika - 2008*, s. 188–195, 2008.

KLYUEVA, N. – KUBOŇ, V. Automatic Valency Derivation for Related Languages. In BOONTHUM-DENECKE, C. – EBERLE, W. (Ed.) *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014*, s. 437–442, Palo Alto, California, 2014.

KOEHN, P. What is a better translation? Reflections on six years of running evaluation campaigns. *Tralogy 2011.* 2011.

KOEHN, P. An Experimental Management System. *Prague Bull. Math. Linguistics.* 2010, 94, s. 87–96.

KOEHN, P. – HOANG, H. Factored Translation Models. In *EMNLP-CoNLL*, s. 868–876. ACL, 2007.

KOEHN, P. et al. Moses: Open Source Toolkit for Statistical Machine Translation. 2007, s. 177–180.

KOLOVRATNÍK, D. – KLYUEVA, N. – BOJAR, O. UMC003: Czech-English-Russian Tri-parallel Test Set for MT. Institute of Formal and Applied Linguistics, 2009.

KURYŁOWICZ, J. *Esquisses linguistiques.* 1960.

Lavie, A. – Denkowski, M. J. The Meteor Metric for Automatic Evaluation of Machine Translation. *Machine Translation*. 2009, 23, 2-3, s. 105–115.

Levin, B. *English verb classes and alternations : a preliminary investigation*. 1993.

Lopatková, M. – Panevová, J. Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi). 2004, 5, s. 348–356.

Mareček, D. – Popel, M. – Žabokrtský, Z. Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, s. 201–201, Uppsala, Sweden, 2010. Uppsala Universitet, Association for Computational Linguistics.

Mel'čuk, I. *Dependency Syntax: Theory and Practice, State University of New York Press*. 1988.

Melčuk, I. – Zholkovsky., A. Explanatory Combinatorial Dictionary of Modern Russian. 1984. Vienna: Wiener Slawistischer Almanach.

Mustajoki, A. – Heino, H. *Case selection for the direct object in Russian negative clauses*. University of Helsinki, 1991.

Och, F. J. – Ney, H. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics 29*, s. 19–51, 2003.

Oliva, K. *A Parser for Czech Implemented in Systems Q*. Explizite Beschreibung der Sprache und automatische Textbearbeitung. Matematicko-fyzikální fakulta UK, 1989.

Oliva., K. A Parser for Czech Implemented in Systems Q. 1989. Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague, .

Oliveira, F. et al. Unsupervised word sense disambiguation and rules extraction using non-aligned bilingual corpus. *Natural Language Processing and Knowledge Engineering*. 2005.

Pajas, P. – Fabian, P. TrEd 2.0 - newly refactored tree editor, 2011.

Pala, K. – Ševeček, P. Valence českých sloves. In *Sborník prací FFUB*, 1997.

Panevová, J. On verbal frames in Functional generative description I. *Prague Bulletin of Mathematical Linguistics*. 1974, , 22, s. 3–40.

Papineni, K. et al. BLEU: a Method for Automatic Evaluation of Machine Translation. s. 311–318, 2002.

Petr Karlík, M. N. – Pleskalová, J. *Encyklopedický slovník češtiny*. Praha, 2002.

Popel, M. English-Czech Machine Translation Using TectoMT. In Šafránková, J. – Pavlů, J. (Ed.) *WDS 2010 Proceedings of Contributed Papers*, s. 88–93, Praha, Czechia, 2010. Univerzita Karlova v Praze, Matfyzpress, Charles University. ISBN 978-80-7378-139-2.

Popel, M. – Žabokrtský, Z. Improving English-Czech Tectogrammatical MT. *The Prague Bulletin of Mathematical Linguistics*. 2009, , 92, s. 1–20. ISSN 0032-6585.

Popovic, M. – Burchardt, A. From human to automatic error classification for machine translation output. In *15th International Conference of the European Association for Machine Translation (EAMT 11). Annual Conference of the European Association for Machine Translation (EAMT-11), 15th, May 30-31, Leuven, Belgium*. European Association for Machine Translation, 5 2011.

Quirk, C. – Menezes, A. – Cherry, C. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, s. 271–279, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Raab, J. Morče - Czech morphological tagger, 2007.

Rosa, R. Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia, 2013.

Rosa, R. – Mareček, D. – Tamchyna, A. Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, s. 172–179, Sofija, Bulgaria, 2013. Bălgarska akademija na naukite, Association for Computational Linguistics.

Sag, I. et al. Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing*, 2276 / *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2002. s. 1–15.

Sgall, P. et al. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Springer, 1986.

Skwarska, K. Záporový genitiv v současné češtině, ruštině, polštině a slovinštině. In *IX. Zborník materiálov z IX. kolokvia mladých jazykovedcov*. 2002.

Snover, M. et al. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, s. 223–231, 2006.

Spoustová, D. et al. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, s. 763–771, Athina, Greece, 2009. Association for Computational Linguistics. ISBN 978-1-932432-16-9.

Tan, L. – Pal, S. Manawi: Using multi-word expressions and named entities to improve machine translation. *ACL 2014*. 2014, s. 201.

Tesnière, L. *Eléments de syntaxe structurale*. Klincksieck Paris, 1959.

Thurmair, G. Comparing rule-based and statistical MT output. Workshop on the amazing utility of parallel and comparable corpora, 2004.

Tillmann, C. et al. Accelerated Dp Based Search For Statistical Translation. In *In European Conf. on Speech Communication and Technology*, s. 2667–2670, 1997.

Trask, R. L. *Language and linguistics : the key concepts*. Abingdon England ; New York: Routledge, 2nd ed. edition, 1944.

Tufis, D. A cheap and fast way to build useful translation lexicons. In *In Proceedings of the 19th international Conference on Computational Linguistics - Volume 1*, s. 1236–1239, 2002.

Turchi, M. – Ehrmann, M. Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources. In *Polibits 43*, 2011.

Urešová, Z. Building the PDT-VALLEX valency lexicon. In *Proceedings of the fifth Corpus Linguistics Conference*, s. 1–18, Liverpool, UK, 2012. University of Liverpool, University of Liverpool.

Čmejrek, M. – Čurin, J. – Havelka, J. Czech-English Dependency-based Machine Translation, 2003.

Vilar, D. et al. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, s. 697–702, Genoa, Italy, may 2006.

Wisniewski, G. – Kübler, N. – Yvon, F. A Corpus of Machine Translation Errors Extracted from Translation Students Exercises. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, may 2014.

Yamada, K. – Knight, K. A Decoder for Syntax-based Statistical MT. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, s. 303–310, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

ŽABOKRTSKÝ, Z. Resemblances between Meaning-Text Theory and Functional Generative Description. In IOMDIN, L. (Ed.) *Proceedings of the 2nd International Conference of Meaning-Text Theory*, s. 549–557, Moskva, Russia, 2005. Slavic Culture Languages Publishers House. ISBN 5-9551-0094-6.

ZEMAN, D. – SARKAR, A. Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, s. 227–233, Athens, Greece, 2000. European Language Resources Association.

ZEMAN, D. et al. Addicter: What Is Wrong with My Translations? *Prague Bull. Math. Linguistics*. 2011, 96, s. 79–88.

ZHOU, M. et al. Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-points. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, s. 1121–1128, 2008.