Charles University in Prague
Faculty of Mathematics and Physics

Doctoral Thesis

LINGUISTIC ISSUES IN MACHINE TRANSLATION BETWEEN CZECH AND RUSSIAN

Natalia Klyueva

Prague, 2015

Doctoral Thesis

Natalia Klyueva

Supervisor of the doctoral thesis:
assoc. prof. (doc.) RNDr. Vladislav Kuboň, Ph.D.

Linguistic Issues in Machine Translation between Czech and Russian

Study programme: Computer Science
Specialization: Mathematical Linguistics

ÚFAL
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 . . . .

Název práce:	Lingvistické otázky ve strojovém překladu mezi češtinou a ruštinou.
Autor:	Natalia Klyueva
Ústav:	Ústav formální a aplikované lingvistiky
Vedoucí disertační práce:	assoc. prof. (doc.) RNDr. Vladislav Kuboň, Ph.D., Ústav formální a aplikované lingvistiky
Klíčová slova:	strojový překlad, slovanské jazyky, blízké jazyky, čeština, ruština, SMT, RBMT, Moses, valence

Abstrakt: V této disertační práci zkoumáme strojový překlad mezi češtinou a ruštinou z hlediska lingvisty. Pracujeme s několika pravidlovými a statistickými překladovými systémy a pomocí změn v jejích nastavení se snážíme dosáhnout co nejlepších výsledků překladu. Jedna z otázek, které řešíme v naší práci, je nakolik příbuznost obou jazyků pomáhá strojovému překladu.

Hlavním cílem práce je lingvistický rozbor chyb ve výstupu čtyř systémů strojového překladu, dvou experimentálních – TectoMT, Moses – a dvou komerčních – PC Translator a Google Translate. Analyzujeme každý typ chyb a řešíme, zda daná chyba souvisí s rozdílem mezi češtinou a ruštinou nebo zda je zapříčiněná architecturou jednotlivých systémů. Pro některé chyby nabízíme cesty, jak je opravit.

Ve zvláštní kapitole se zaměřujeme na chyby v povrchové valenci sloves. Zkoumáme rozdíly v české a ruské povrchové valenci, popisujeme extrakci slovníku povrchových forem a tento slovník integrujeme do systému TectoMT. Dále nabízíme souhrn lingvistických pozorování o povaze rozdílů v české a ruské valenci.

Title: Linguistic Issues in Machine Translation between Czech and Russian

Author: Natalia Klyueva

Department: Institute of Formal and Applied Linguistics

Supervisor: assoc. prof. (doc.) RNDr. Vladislav Kuboň, Ph.D., Institute of Formal and Applied Linguistics

Keywords: machine translation, Slavic languages, related languages, Czech, Russian, SMT, RBMT, Moses, valency

Abstract:

In this thesis we analyze machine translation between Czech and Russian languages from the perspective of a linguist. We work with two types of Machine Translation systems – rule-based (TectoMT) and statistical (Moses). We experiment with different setups of these two systems in order to achieve the best possible quality. One of the questions we address in our work is whether relatedness of the discussed languages has some impact on machine translation.

We explore the output of our two experimental systems and two commercial systems: PC Translator and Google Translate. We make a linguistically-motivated classification of errors for the language pair and describe each type of error in detail, analyzing whether it occurred due to some difference between Czech and Russian or is it caused by the system architecture. We then compare the usage of some specific linguistic phenomena in the two languages and state how the individual systems cope with mismatches. For some errors, we suggest ways to improve them and in several cases we implement those suggestions.

In particular, we focus on one specific error type – surface valency. We research the mismatches between Czech and Russian valency, extract a lexicon of surface valency frames, incorporate the lexicon into the TectoMT translation pipeline and present observations on which verbs tend to have different valency frames.

Contents

Acknowledgements

I want to express my gratitude to my supervisor Vladislav Kuboň who supported me throughout my studies, gave me a tremendous amount of invaluable advice and feedback on my writing.

The second person I'm most obliged to is Ondřej Bojar who helped me to make my first steps from a theoretical to a computational linguist, showing me what wonderful things can be done with the command line and a text. The second Ondra I want to thank is Ondřej Dušek who was always willing to answer hundreds of programming and linguistic questions I had.

I want to praise other colleagues for helping me with technical issues related to Machine Translation systems and thesis formatting: Eduard Bejček, Karel Bílek, Martin Popel and Zdeněk Žabokrtský.

Many thanks go to my colleagues whom I asked to read parts of my thesis, firstly, Alexandr Rosen for reading the whole thing, and other colleagues for reading small parts of it: Jiří Mírovský, Markéta Lopatková, Ahmad Ebrahimian, Martin Holub, Katka Rysová, Jiří Hana, Vendula Kettnerová, Anna Nedoluzhko, David Mareček, Rudolf Rosa, Pavel Straňák, Shadi Saleh, Petya Osenova, Lukaš Žilka, Elena Pascual Aliaga and Karolína Skwarská.

Also, I thank Jan Hajič for being a supporter of such a fantastic department. And I want to thank every single person at ÚFAL, their inspiring attitude to life and the whole atmosphere where your colleagues are your friends.

Last but not least I am indebted to my mother who always believed in me and encouraged me to continue when I was about to give up and to my husband for tolerating me being overwhelmed with my studies.

I dedicate this thesis to my kids – Jaroušek and Maruška – who were born during the thesis writing and so patiently shared their mother with computational linguistics.

1

Introduction

Machine translation (MT) is a popular branch of Natural Language Processing. The work on MT systems is traditionally presented as a collaboration between linguists and computer scientists: linguists prepare data (e.g. dictionaries and transfer rules), and computer scientists implement the baseline of the system. Linguists analyze the output translations and on the basis of these translations suggest further improvements and then the cycle is 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 largely adopted in the beginning of 1990's. There was no longer a need for a linguist with the 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 rather than by 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 the prism of a linguist.

1.1 Objectives

When our research started in 2006, the primary goal was to make an experimental implementation of a Czech-to-Russian MT system within the available frameworks.

Our work was initially supposed to answer a range of questions. The 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 related languages such as Czech and Russian there is no need to train statistical models to achieve good quality. The second hypothesis was that under a similar setup, a translation system for related languages is easier to build than for those unrelated.

We aimed to specify a classification of errors for the MT between Czech and Russian and link those errors with 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 the points of differences that can cause problems, we concentrated on one of the issues – surface valency in the Czech and Russian languages.

1.2 Outline of the Thesis

The thesis is structured in 6 Chapters.

In Chapter 2, we give a brief overview of machine translation – the basic concepts, history of MT, methods and strategies; then we present data and tools that will be used in our experiments and name the MT systems that exist for the pair Czech-Russian. Then, in Chapter 3, we focus on two systems of those systems that we will evaluate: a statistical one (Moses) and a rule-based one (TectoMT). We set a baseline for those systems, propose some improvements and take the output of the best experiment to be evaluated further. Also, we describe two commercial systems – PC Translator and Google.

Next, in Chapter 4, we explore the output of the four MT systems. Exploring errors in the MT output, we try to answer the question which types of discrepancies between Czech and Russian are successfully processed by the MT system and which pose a problem. We propose a classification of errors suitable for our language pair. Each error type is analyzed 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 system. In several cases, we did some experiments to fix the error.

Out of all the errors, we focus on surface valency, especially on a theoretical description of surface valency discrepancies between the two languages (Chapter 5). We conduct some experiments designed to spot the cases of differences and present some observations on when the surface valency in Czech and Russian tends to be different.

The summary of our main results and the discussion is presented in the concluding Chapter 6.

Collaboration remarks. Because of the author's linguistic background, some of the experiments described here were carried out with the help of colleagues from Institute of Formal and Applied Linguistics (ÚFAL). Implementation of the Moses statistical machine translation system 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 system TectoMT between

Czech and Russian. Some of the results presented in Chapter 2.3 partially intersect with a Master Thesis (Bílek, 2014), done under the same project (GAUK 639012), but in the latter work the stress is put mainly on technical aspects of the implementation of the MT systems without a deeper linguistic analysis.

2

Machine Translation systems between Czech and Russian

2.1 An overview of Machine Translation

In the first part of the thesis, we discuss the field of machine translation. We review MT architectures and describe historical and state-of-the-art MT systems available for the Czech-Russian language pair.

MT is a process of converting a text coded in one language into another language by a computer. 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, such as technical issues of system development, theoretical research on linguistic aspects, quality evaluation etc. The most renowned web collection of articles about MT[1] counts more than 11,400 items (as of April 2015).

A detailed overview of many aspects of machine translation is provided by (Hutchins, 1986), and facts presented in this introduction are partially based on this survey. Here, we will concentrate only on those approaches and MT systems relevant to our language pair.

2.1.1 History

The first attempts to build a system to substitute human translators started shortly after the first ascendants of computers appeared. Perhaps the most famous quote which has inspired many MT researchers 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.

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

Russian and English were the first languages for which an MT system was developed by teams in the US and in the former USSR. That was obviously due to the political reasons, as both countries needed to translate huge amounts of texts between the two languages. The first MT system was Russian-to-English, developed by the IBM research team. Afterwards, many other researchers all over the world started to develop their own systems for various languages. METEO[2] and Systran[3] were among the first and most successful MT systems.

The first systems were rule-based. They used dictionaries and implemented linguistic rules (detailed description follows). In 1990's an enormous increase in processing and memory capacity led to possibility of statistical MT development. The IBM team was the first to introduce the first prototype, see the Section 2.1.4.

The choice of languages for machine translation projects had often been politically or geopolitically motivated. For example, in former Czechoslovakia, Ruslan, a Czech-to-Russian MT project (Hajic, 1987), (Oliva, 1989) started in 1980's, as Russia had a strong influence in this region, and it was a high-priority pair in those days. At the same time, experiments on MT from English into Czech (APAC) had taken place (Kirschner and Rosen, 1989). Ruslan and APAC were both implemented in Q systems (Colmerauer, 1970).

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

2.1.2 Approaches to MT

Generally, two main approaches to building MT systems can be distinguished: **rule-based** and **statistical**, though some other types related to the two main ones can be considered as well (e.g. example-based MT or hybrid MT). In our work we will stick to the dichotomy rule-based vs. statistical, mentioning further in the texts how they intersect with Hybrid method.[4] There is an extensive pool

[2] An MT system between English and French for the 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, whenever they exploit statistical modules in a rule-based architecture. The same is true of SMT, which can be considered to be Hybrid as soon as some linguistic knowledge is being introduced, e.g., additional morphological information in the form of morphological dictionaries. In this thesis when we use RBMT, we actually mean MOSTLY rule-based that might contain

of work concerning comparison of the two approaches, such as (Thurmair, 2004) for English and German or (Bojar, 2012) for English and Czech.

Both approaches have their advantages and disadvantages that are well-described in the literature. On the one hand, statistical approach is language-independent, there is no need for a linguistic description, hence it is cheap and quick to deploy. In comparison, in RBMT, compiling rules and electronic dictionaries can take years in order to reach a sufficient quality of translation. 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 former are more controllable and predictable. The errors produced by RBMT are easy to spot and it is often obvious how to fix them (but not always easy) – by only some additional rules. In contrast, SMT works like a black box. Although 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 with proper syntactic rules generate sentences with better structure, than statistical systems. On the other hand, statistical systems are much better at resolving the problem of word sense disambiguation (WSD) than rule-based systems.

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 MT systems cope with mismatches between the languages.

2.1.3 Rule-based MT systems

As mentioned above, the first systems that appeared were **rule-based** MT systems. They use bilingual dictionaries and manually written rules of transfer. They are labor-intensive and involve extensive linguistic knowledge. It often takes years to build such a system.

Researchers define three main architectures of RBMT: direct, transfer and interlingua.[5] They are usually represented within the machine translation triangle, or Vauquois triangle (Vauquois, 1968), see Figure 5.1.2:

some statistical modules, and under SMT we mean MOSTLY statistical that might involve some linguistic knowledge.

[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.

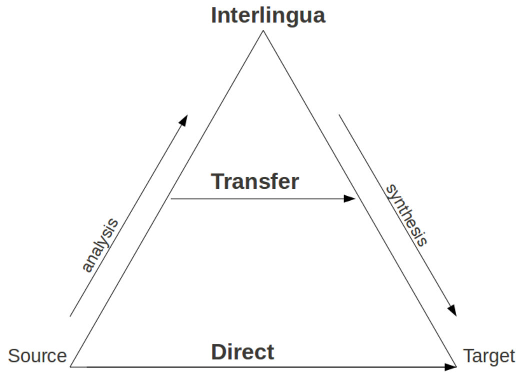


Figure 2.1: Machine translation triangle

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.

Direct systems

Direct systems provide a word by word translation from a source to a target language. “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 1990’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. Method for “pure” direct RBMT translations is not really used nowadays,[6] as even for very related languages some linguistic analysis should be introduced. Almost direct architecture was used in Česílko (Hajič et al., 2000b), a translation system between Czech and Slovak languages, using only a morphological dictionary.

[6] However, statistical phrase-based MT systems can be considered as ‘direct’.

Transfer translation

Transfer systems rely on the collection of rules aiming to cover morphological, syntactic or semantic mismatches between the languages. When the system is declared to be rule-based, it is almost always a transfer system, such as the already mentioned Systran, METEO, Ruslan (to be described in detail in Section 2.3.1), Česílko (to be described in detail in Section 2.3.2), 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**). Analysis can proceed up to different language levels – morphological, shallow syntactic, deep syntactic or shallow semantic. The borders between the levels are often quite vague and depend on the formalism under which the system is developed.

One of the types of RBMT systems that we will use in our work is a dependency-based machine translation system. It was first developed for the Czech-English pair (Čmejrek et al., 2003) and was based on the Functional Generative Description (FGD) theory (Sgall et al., 1986) as its theoretical platform. 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; the synthesis of the target English text from the tectogrammatical representation was provided by a number of rules. Since then, the system has changed, so we use a more recent version of this system – **TectoMT** (Popel, 2010) on a platform called Treex, to be described in Section 3.1.

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" (Mel'čuk, 1988).

As for Česílko, the system adopted the transfer architecture for less closely related language pairs – Czech-Polish, Czech-Lithuanian and Czech-Russian. The Czech-Russian pair within Česílko will be discussed in the Section 2.3.2.

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. However, Apertium supports MT between unrelated languages as well.

Interlingua

Machine translation systems that account for the level of deep semantics are called Interlingua systems. The core of Interlingua MT is a universal language that encodes all possible meanings – semantic primitives – for every natural language. The most renowned project of Interlingua nowadays is UNL – Universal Networking Language.[7] It is based on a single formal representation, where each language has analysis from plain text to the UNL and the synthesis from the UNL . The Interlingua architecture is quite complex to build, as it is very demanding to specify the required language information that can be language-independent – 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.

2.1.4 Statistical Machine Translation

Statistical machine translation has become one of the easiest MT paradigm to deploy. 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 SMT field introducing IBM models in the 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] – *e* – of linguistic units (these can be words, phrases, sentences), and define the probability

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

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

$$\hat{e} = \operatorname*{argmax}_{e} \mathsf{p}(\mathsf{e}|\mathsf{f}) \tag{2.2}$$

The latter formula presents the ideal conditions of the hypothesis. As an approximation on data, the score of a hypothesis $\hat{e}$ is calculated from two components: the language model (LM) and the translation model (TM), which are introduced by the transformation according to the Bayes theorem[9] into:

[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 *f* and target *e*.

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

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

$$\hat{e} = \operatorname*{argmax}_{e} \frac{p(e) \cdot p(f|e)}{p(f)} = \operatorname*{argmax}_{e} p(e) \cdot p(f|e) \qquad (2.4)$$

The above formula, representing a combination of a language model and a translation model, is used in phrase-based models.

The **language model** p(e) gives the probability of how likely a unit e is 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.

While the first IBM model was based on simple word alignment, modern statistical models used more sophisticated techniques bringing better translation results.

Phrase-Based 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, using toolkits such as Moses SMT (Koehn et al., 2007) provided that parallel data are available. Our implementation of Moses for the Czech-Russian language pair will be presented in the Section 3.2. The most important property of phrase-based systems is the ability to translate sequences of words (phrases or n-grams) rather than single words.

Phrase-based MT systems 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:[11]

$$\hat{e} = \operatorname*{argmax}_{e} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) * p_{LM}(e) * d(start_i, end_{i-1}) \qquad (2.5)$$

- **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 probabilities are extracted from a parallel corpus. The translation probability then presents the relative frequency of

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

[11] The formulas from this section are taken from (Koehn, 2010b).

a phrase:

$$\phi(\bar{f}|\bar{e}) = \frac{\mathrm{count}(\bar{e},\bar{f})}{\sum_{\bar{f}}\mathrm{count}(\bar{e},\bar{f})} \tag{2.6}$$

Following is an example of a Czech-Russian phrase table entry[12] indicating a phrase *to believe in*:

(2.7) doufat v ||| надеяться на ||| 0.25 ...

The score 0.25 denotes the probability that the Czech phrase *doufat v* will be translated into Russian as *надеяться на*.

- **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:

  $$p(w_1, w_2, ..., w_n) = p(w_1)*p(w_2|w_1)*p(w_3|w_1w_2)...*p(w_n|w_1, w_2, ..., w_{n-1}) \tag{2.8}$$

  The most widely used model is a trigram language model. It 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:[13]

  (2.9) -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** Distortion parameter, or penalty, penalizes a phrase in which a reordering limit 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 phrases with 2 reorderings. This limit reduces the space of hypothesis and does not allow sentences with the phrases that are far from the respective source ones.

Decoding

The various combinations of phrases that constitute 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.

[12] We hide some other probability parameters, such as lexical probability or word alignment numbers.

[13] The negative numbers are log-probabilities.

2.1.5 Hybrid MT systems

Each of the approaches – statistical and rule-based – has its drawbacks that we have mentioned above. At present, more experiments are focused on a combination of these 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 as "slightly" Hybrid.

Let us take an example of some systems developed at ÚFAL. Česílko and TectoMT are considered to be rule-based systems, even though they have some statistical components as parsers, morphological analyzers or WSD modules. Factored model within the Moses system can exploit morphological data, which can be also seen as a kind of brute force learning of language rules. The two examples above are actually not Hybrid. Hybrid MT presupposes some substantial effort to tune the system into a 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 were presented within the "Workshop on Hybrid Approaches to Translation (HyTra)".[14] Commercial systems also showed rather promising results (Systran,[15] AppTek[16] etc.)

2.1.6 Pivoting in MT

Pivoting is a popular technique in MT that exploits the idea of an intermediate language, and it is used in both RBMT and SMT. This idea is justified by the fact that some languages have very few linguistic resources, and that the human translation from English into "resource rich" Czech and then the automatic translation from Czech into some other Slavic language could bring good results in a relatively short amount of time.

Nowadays, with almost each language in the Slavic group having many resources like lexicons, morphological dictionaries, and even treebanks, the idea of a pivot as presented in Česílko,[17] is not so widespread. It might still bring

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

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

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

[17] Originally, Česílko's idea was that the first translation – from source into the pivot, e.g., from English into Czech – should be a human translation and translation from the pivot language into other (related) languages should be automatized.

some fruit in the case of very under-resourced languages, such as 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 and Bojar, 2012), where the pivot languages serve as the source of additional data (phrase tables).

2.1.7 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.

Manual evaluation

One of the earlier techniques to estimate translation quality was **edit distance** technique. A translated sentence (target sentence) was compared against its gold standard translation (reference) in terms of how many insertions/deletions must be introduced 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 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.

In WMT evaluation campaign,[18] the ranking of systems on a scale is used. During the evaluation, the annotators are asked to rank several translations from best to worst.

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

Automatic metrics

Automatic metrics are cheap and fast. They are used mostly by researchers to monitor the progress of system development, even though they suffer from

[18] WMT (Workshop on Statistical Machine Translation) is held each year - `http://www.statmt.org/wmt15/` and earlier. It aims at evaluating the state-of-the-art MT systems, both statistical and rule-based.

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 of the fact that they should be used for the machine translation, so the source can be translated by human and by the MT system in different ways.

The test set used for evaluation between Czech and Russian demonstrates the same problem as the training set: such translations are generally not direct. They present translations from English into Russian and from English into Czech. Thus, Czech and Russian sides of a test set quite often contain significant structural and semantic discrepancies.

When SMT became popular, some new evaluation metrics suitable especially for this type of MT had been created, e.g. BLEU (Papineni et al., 2002), TER (Snover et al., 2006), WER (Tillmann et al., 1997),[19] NIST (Doddington, 2002), Meteor (Lavie and Denkowski, 2009). These techniques measure the performance automatically 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 evaluation metric – **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:

$$\mathsf{BLEU} = \mathsf{BP} \times \{\prod_{i=1}^{4} \mathsf{P}(i)\}^{1/4} \tag{2.10}$$

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, or 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.

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

[19] TER and WER resemble 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

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 Česílko and TectoMT systems is poor. That is why the 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 underrates RBMTs in favour of SMTs and is not very suitable when translating into free word order languages as it counts the precision of exact n-grams. Also, it is not suitable to evaluate minor improvements of specific language phenomena as the difference in terms of BLEU will be really insignificant (see Section 5.2.3). 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[20] will be of a manual nature.

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

2.2 Data and Tools for MT

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

2.2.1 Parallel and monolingual data

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

UMC

For the needs of our experiments we built UMC 0.1 (ÚFAL Multilingual Corpus) – a multilingual parallel corpus of texts in Czech, Russian and English with

[20] Described in detail in Section 4.1

automatic pairwise sentence alignments (Klyueva and Bojar, 2008).[21] UMC is closely related to CzEng,[22] a Czech-English corpus which has been successfully employed in SMT experiments. The primary goal of compiling UMC was statistical machine translation, but it also served as data source for the dictionary extraction and some other experiments with valency (see Section 5.3).

We have chosen only one web source (see below) to download our texts and so far we were able to obtain over 1.7 million words in each of the three languages.[23] We included also the English part of the parallel texts into the corpus on purpose, as this served as a platform to compare how SMT works for related languages in comparison to those unrelated.

Collecting parallel texts meets challenges such as copyright, translation quality and representativeness of the language. The problem of copyright is solved by contacting the site editor, asking for a license agreement for educational purposes. It is more complicated with translation quality, because when downloading automatically huge amount of texts, they cannot 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 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 a substantial share of technical terminology and the use of general language is limited. On the other hand, those types of text are most suitable for MT in restricted domains, 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 the news covers the most essential part of standard language usage, so we have chosen to use the news articles in the baseline experiment.

As mentioned already, all these texts were downloaded from a single source – Project Syndicate site,[24] which contains a large collection of high-quality news

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

[22] `http://ufal.mff.cuni.cz/czeng/`

[23] Since the corpus was first compiled in 2008, we have not downloaded new data from the web. Should we do an update now, we would obtain much more data.

[24] `http://www.project-syndicate.org/`

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 the downloaded documents is 2,186 in each of the three languages. Table 2.1 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 2.1: Summary of corpus size

Corpus processing

The following steps were applied to the downloaded data:

- **Converting formats.** HTML files are converted 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 CzEng project, where the preference was given to the XML storage format, in UMC we use plain text format as this will be enough for the purposes of training our models.

- **Segmentation and Tokenization.** In order to segment and tokenize our texts we used a trainable tokenizer described in (Klyueva and 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 the sentence level using the hunalign tool.[25] We did not use any additional dictionary for the alignment, the dictionary was extracted automatically by the tool.

Subtitles

The subtitles data were downloaded from the web (http://opensubtitles.org) by Karel Bílek (Bílek, 2014). Texts coming from subtitles are considered to be very

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

unreliable as training data for SMT and for 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 the data is that they can be obtained easily, and they are quite large.

Intercorp

Intercorp (Čermák and Rosen, 2012) 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.[26] We obtained 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.

Parallel data summary

Table 2.2 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 2.2: 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.

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

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 corpus was included into the LM data alongside with other resources in Russian:

- Russian side of the parallel corpora described above,
- NewsCrawl,[27]
- Russian side of a parallel English-Russian corpus from Yandex,[28]
- CommonCrawl.[29]

Totally, those data include 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.

2.2.2 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. We will now briefly describe the process of 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 and Prokopová, 2006), and for English and Romanian (Tufis, 2002). The tool we used for word alignment process is GIZA++ (Och and Ney, 2003), which is widely used by many researchers 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 word forms. For this purpose we used the taggers: Hajič's tagger for Czech – the same as for the analysis of Czech text and TreeTagger[30] for Russian.

Next, we describe how the dictionary was created. We took the Czech-Russian part of the UMC parallel corpus (Section 2.2.1), with only 1-to-1 aligned sentences

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

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

[29] `http://commoncrawl.org/`

[30] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

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. The example below shows a few entries from the extracted dictionary. The list is taken from the top of the dictionary, so those are the most frequent word pairs. The first column gives the number of occurrences of the alignment pair, the second column shows a Czech word, the third column the Russian word:

37,188	a	и
25,490	v	в
12,269	že	что
8,834	být	быть
8,303	na	на
5,345	tento	этот

Most of generated the 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

while the others are wrong, e.g.:

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

In order to select the pairs that are most probably correct, we used the information on the frequency of the pair occurrence from UMC and selected the most frequent translation equivalents.

This dictionary was used while constructing rule-based MT systems. As Česílko or TectoMT (MT systems described later in Sections 2.3.2 and 3.1) were not supposed to have any word sense disambiguation module, we took those top frequent translation pair. 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.

2.2.3 Tools for Morphosyntactic Analysis

A morphological tagger analyzes a sentence and assigns morphological tags and sometimes lemmas 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.

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.[31] Secondly, due to the existence of morphologically annotated corpora, researchers have trained a number of taggers for Czech, see e.g. (Hajič, 2001), (Raab, 2007), (Spoustová et al., 2009) and, more recently, Featurama[32] or MorphoDiTa.[33]

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 UNO is the following’):

```
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, the 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.

A tagger for Russian

For the Russian language, the only available open-source tagger TreeTagger[34] was used. Following is an example of a tagged phrase (lit. ‘In this relies culture’):

```
В|в|Sp-l
этом|это|P--nsln
и|и|C
```

[31] https://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html

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

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

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

```
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.

SynTagRus

In our work, we will also use another resource of morphological information, coming from the Russian Dependency Treebank SynTagRus (Boguslavsky et al., 2000).[35] 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 the Czech-Russian implementation of TectoMT.

2.3 Pioneering MT systems between Czech and Russian

There exist at least six MT system between Czech and Russian that we are aware about. Rule-based are: Ruslan, Česílko, TectoMT, PC Translator. The statistical systems are Moses and Google. In this section, we present MT systems developed at our department – Ruslan and Česílko. We could not use them in our linguistic evaluation due to the reasons explained further. The other systems will be described in Chapter 3 as we will use the output of these systems in the linguistic evaluation.

[35] 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.

2.3.1 Ruslan

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

The implementation of Ruslan required huge amount of manual linguistic work. The core of the system was a dictionary, its entries were enhanced with morphosyntactic and semantic information.

In 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 was terminated.

Since then, the resources created under the project served for other standalone 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 and 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 the translation of individual words and in linguistic information.

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

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

- The left-hand side of a dictionary item – `LE2KAR3` – represents the stem of the Czech noun *lékař* – a doctor; the diacritics is encoded in a 'letter + a digit symbol' as the time when the dictionary was created the encoding of national characters constituted a challenge.

- `MZ` represents the 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 pattern of the Russian equivalent and the translation equivalent itself.

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

- `VLASTNI` represents the stem of the Czech verb *vlastnit* – 'to possess';
- *R* – a root of a tree;
- `5` stands for a verb, `PRM` is the conjugation pattern of the verb;
- `(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 pattern according to which the proper morphological form is generated and the lemma of the respective Russian verb.

The major problem in running Ruslan on the standard test set is that of a dictionary: a single unknown word can crash the process of analysis, so the two distinct paths (trees) – preceding and following the unknown word – are generated. The out-of-vocabulary words are rather hard to include as new entries. 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[36] that we used for other systems, so Ruslan translation output is not included into the linguistic analysis.

2.3.2 Česílko

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 large linguistic resources with a high number of rules. This idea worked

[36] The coverage of the Ruslan dictionary – that is domain-specific and outdated – is rather low, so it would have been necessary to include many new translation pairs in a Ruslan format, which is very labor-intensive task.

well for the very closely related Czech and Slovak, but for more distant languages a deeper and more sophisticated linguistic analysis was needed.

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

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 that 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 the copula 'to be', which is omitted in Russian and is used in Czech.
- the usage of reflexives (part of a word in Russian and separate in Czech, though considered to be a part of a lemma after morphological analysis)
- negation prefix 'ne', which is a part of the word form in Czech and is written separately in Russian
- some cases of prepositional mismatches

The following code illustrates the transfer rule for the copula verb *být* – 'to be'. It reflects the rule to transfer the Czech copulative construction like *Jsem student* – 'I am a student' into the Russian *Я студент* – 'I am a student' which is done by substituting the Czech auxiliary verb with the respective Russian pronoun:[37]

[37] More on linguistic aspects of this problem see Section 4.2.9.

```
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";
    }
 }
}
```

Evaluation

The overall quality of the system was evaluated in terms of the BLEU score, and it was far from ideal. (Homola, 2009) tested the system on a test set that contained 1000 sentences and the 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 modified version of Česílko cannot be compiled on more modern systems and, moreover, the original morphological module was missing. So this system is not subjected to the linguistic analysis, similarly to Ruslan.

2.4 Discussion

In this introductory chapter we made an overview of MT systems and system architectures, concentrating only on those that are relevant to our language pair (either rule-based or statistical MT systems). We then presented the data and tools that we will use while implementing our MT systems. The author of the thesis contributed to the data collection (the parallel corpus for Czech-Russian, Section 2.2.1 and the automatically extracted Czech-Russian dictionary, Section 2.2.2.). In the next chapter, we will present the four MT systems between Czech and Russian that we will use in a linguistic evaluation in Chapter 4.

3

MT systems under study

In this chapter, we present the four MT systems between Czech and Russian that we will use in our linguistic evaluation – TectoMT, Moses, Google and PC Translator. The author of this thesis has a major contribution to the following experiments that will be described in detail in respective sections:

- TectoMT:
    - data for a transfer block - an automatically extracted dictionary
    - adopting data from SynTagRus (see Section 2.2.3) to the morphological synthesis of Russian
    - blocks for handling prepositions, verbal valency and some other minor mismatches in Czech and Russian
- Moses
    - collecting the training data (corpus UMC)
    - major ideas in experiments in reducing OOV rate
    - investigating impact of text genre on the SMT performance
    - investigating impact of adding a resource with Named Entities
    - comparing SMT performance between related and unrelated languages

In the last section, we briefly describe the two commercial MT systems – Google and PC Translator – that we will also use in our comparative linguistic evaluation.

3.1 TectoMT

3.1.1 System description

The TectoMT system between Czech and Russian was implemented within the framework **Treex** (Popel and Žabokrtský, 2009). Treex is a modular system of

NLP tools, such as tokenizers, taggers and parsers that were created to process corpora and treebanks in multiple languages.

One of the main projects under Treex is the 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 machine translation system from Czech into Russian, investing least possible effort. Provided that the analysis of Czech already existed, it took only one day for two persons (M. Popel and Z. Žabokrtský) to adjust TectoMT for the Czech-Russian pair. The process of gathering data (dictionary, morphological data) was not that quick, though.

Next, we provide a brief description of the system. Each experiment is presented as a scenario consisting of a sequence of **blocks**,[1] each of which performs some NLP subtask. The blocks ensure the transformation between the four language layers: word (w-layer), morphological (m-layer), analytical (a-layer) and tectogrammatical (t-layer) layers.

This division has its roots in the Functional Grammar Description theory – FGD (Sgall et al., 1986), but its implementation in Treex is slightly different from original FGD concepts. The PDT dependency treebank (Hajič et al., 2006) is based on this theory, the annotation in the PDT is done on the four mentioned language layers while FGD distinguished more layers. Below, we will provide a description of layers how they are used in Treex/TectoMT.

- **The Word and Morphological Layer.** The word layer is a sentence represented as a sequence of tokens. On the morphological layer, each token in the sequence is represented as a word form with a lemma and a tag assigned.

- **The Analytical Layer.** Syntactic annotation is presented in the form of a dependency tree, where each morphologically annotated token from the previous level becomes a node with an assigned *analytical function* (**afun**). Analytical function reflects the syntactic relation between a parent and a child node and 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 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.

[1] Blocks in TectoMT are Perl modules.

On the t-layer, nodes are labeled with **tectogrammatical functors** - attributes which represent deep syntactic relation of a word to its parent. The notion 'functor' is very close to the notion 'semantic role', but is not yet the same. Some examples of functors: "Actor", "Patient", "Addressee", "Effect" etc. In experiments with TectoMT we do not make use of functors, only in Section 5.4 we will exploit functors in valency comparison.

3.1.2 Translation scenario

Now we will describe the translation scenario of the Czech-Russian MT itself.

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

The morphological layer presents only a flat structure of a sentence. Then, we make a step from the morphological to the analytical layer ensured by a sequence of blocks **M2A** (morphological to analytical). The sentence is parsed by the MST parser (McDonald et al., 2006) which generates an analytical (surface syntactic) tree. The nodes are marked with analytical functions reflecting dependency relations between nodes of the tree.

Figure 3.1 illustrates a parsed Czech sentence *Jak prodat jednání o globálním obchodu* – 'How to sell negotiations on the global market', the first tree (a-tree) presents the analysis of the sentence up to the analytical layer.

The head of the sentence is the verb *prodat* – 'to sell' that has an afun 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

Analysis up to tectogrammatical layer is made by blocks **A2T** (analytical to tectogrammatical). On the tectogrammatical layer, nodes representing auxiliaries, prepositions, reflexives are collapsed, formemes and grammatemes[2] 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 the tectogrammatical attributes. They provide a quick

[2] Grammatemes are tectogrammatical counterparts of the morphological categories.

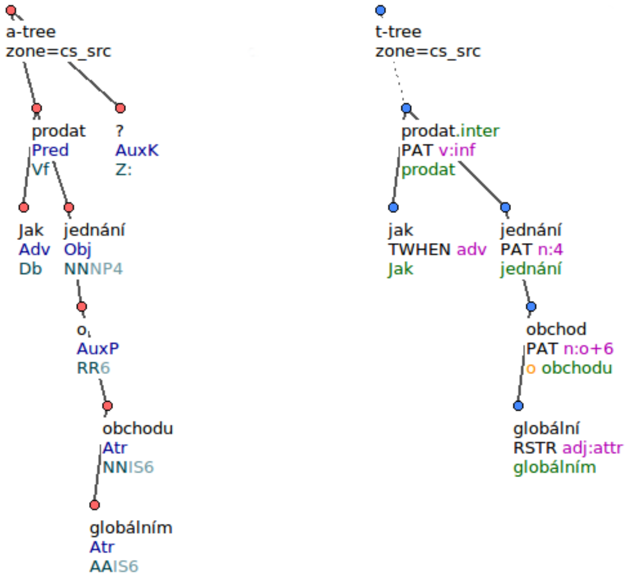


Figure 3.1: Analytical representation of the sentence *Jak prodat jednání o globálním obchodu* – 'How to sell negotiations on the global market'.

and transparent connection between a tectogrammatical functor and a surface morphological form of a word. Here are few examples of formemes:

(3.1) existovat [v:inf] – The infinitive form of a verb
rozdíly [n:1] – Noun in the Nominative case.

The second tree (t-tree) in Figure 3.1 illustrates how analytical functions are converted to tectogrammatical functors; the node receives 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".

Transfer and Generation

In the phase of transfer (**T2T blocks**), Czech lemmas in the tree nodes are substituted with their Russian equivalents. Again, we made use of the automatically generated dictionary that we exploited in Česílko (Section 3.1). 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:

(3.2) 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 (*o*) is used in Czech and in Russian, and a noun (*n*) after the preposition requires the Locative (*6*) case in both languages. The second example demonstrates a discrepancy: a noun after this multiword preposition requires different case morphemes in Czech and Russian. Valency 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 3.1.

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 able to cover the majority of them, just the most frequent ones.

Figure 3.2 illustrates the Russian tectogrammatical and analytical trees.

The nodes mostly do not have any afuns (except for the ones that are relevant under some transfer rule), they only carry the equivalents of source Czech morphological tags. This means that almost no transfer is made on the tectogrammatical layer and very little on the analytical.

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 issue for TectoMT (as well as Česílko) is that the disambiguation module cannot be introduced for our language pair as in English-Czech TectoMT[3] because we do not have a parser for the Russian language. So, in Example 3.2, 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 4.2.13 and 4.2.14.

After words are generated from Russian lemmas and respective morphological tags, the resulting word forms are jointed to form a sentence.

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

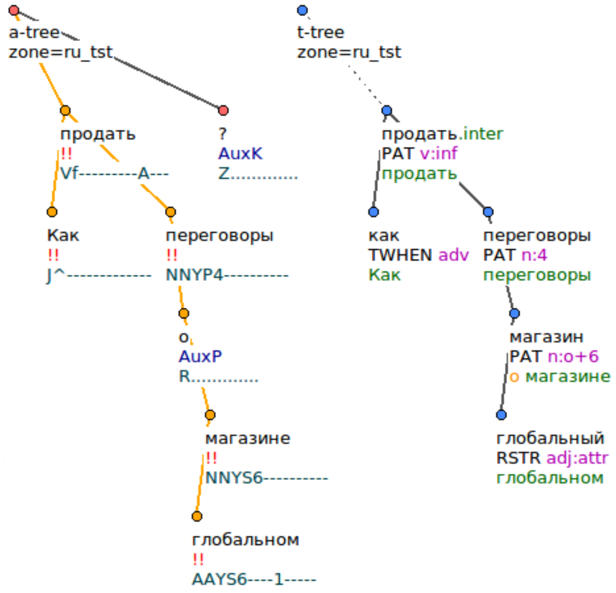


Figure 3.2: Russian tectogrammatical and analytical trees: *Как продать переговоры о глобальном магазине.* – 'How to sell negotiations about global shop.'

3.1.3 Evaluation and improvements

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

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

After adding some linguistic information in the form of blocks, the score increased only a little bit. Some of the new blocks and changes are described in

(Bílek, 2014), some of them are related to the modified analysis of Czech. Following are some points where the rules/changes were introduced:[4]

- Fixing verbal aspect. Word forms extracted from SynTagRus have only one aspectual type – imperfect, and the generation of the perfective counterpart is ensured by tools unavailable to us. We bypassed the problem by adding infinitive forms of verbs, where the imperfective lemma was substituted with the word form.
- Enlarging the dictionary. The experiment where new entries were extracted from a parallel corpus was described in (Bílek, 2014).
- The list of formemes with prepositional complements like in Example 3.1 was enlarged.[5]
- Surface valency frames from Ruslan dictionary were added as formemes. The experiment is described in detail in Section 5.2.3.
- Some blocks to fix certain linguistic phenomena were added/improved: copula drop (more analysis see Section 4.2.9),[6] modal verbs,[7] fixing year[8] construction in Russian (Section 4.2.8).

Table 3.1 summarizes the performance changes in terms of the BLEU score as the specific rules/data were introduced.

Fixes in the Czech analysis (punctuation handling) were made for the Czech-English pair, but our system benefited from those improvements as well. The last line – 'Fixes in rules and valency' concerns mainly improvements of existing rules (see the above list) and introduction of a module handling verbal valency. Those issues will be discussed in more detail when we will be describing specific linguistic phenomena, see Sections 4.2.8, 4.2.9, and 5.2.3. The improvement of specific issues in terms of BLEU was very little, but the analysis always showed some improvement in an issue that we aimed for.

[4] The last three blocks from the list were either written or sufficiently improved by the author of this thesis, the links to the respective blocks in TectoMT are provided in footnotes.

[5] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/RuleBasedFormemes.pm`

[6] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2A/RU/DropCopula.pm`

[7] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2A/RU/AddAuxVerbModal.pm`

[8] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/FixDateTime.pm`

Experiment and improvements	BLEU score
Baseline	4.44%
Fixing verb tenses and aspect	5.09%
Adding preposition formemes	6.62%
Larger dictionary	7.04%
Fixes in Czech analysis (punctuation)	9.04%
Fixes in rules and valency	9.38%

Table 3.1: Baseline and improvements

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

3.2 Moses toolkit and experiment manager Eman

In this section we will look at experiments with Moses, an open-source implementation of a phrase-based statistical translation system. The main principles of phrase-based MT systems were described in Section 2.1.4. Moses is very much language independent since it uses purely data driven methods.

The Moses toolkit[9] relies on and also includes several components for data preprocessing and MT evaluation[10] which we will further describe in detail.

[9] `http://www.statmt.org/moses/`

[10] 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.

3.2.1 Experiment manager 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 an experiment manager in order not to get lost in the experiments. For this purpose several management systems have been created, as, for example, EMS (Koehn, 2010a). In our work, we will use a system developed at the Institute of Formal and Applied Linguistics – **eman** (Bojar and Tamchyna, 2013). An eman experiment consists of a sequence of steps, each of which executes some specific task:
s.corpus... – ensures that 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 former is based on plain text data from a parallel corpus, the latter uses linguistic knowledge – morphology. The latter is especially crucial while translating between morphologically rich languages, as many 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 a large 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 links 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
- Data issues: impact of genre of training data on MT quality.
- Enhancing training data with Named Entities.
- Language relatedness in SMT: comparing English-to-Russian and Czech-to-Russian translation.

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 2.2.1) 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. 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. The data for training were lowercased, as it is standardly done in Moses. 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[11] 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 originate 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 word forms 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.

3.2.2 Factored models and OOV rate

Generally, researchers improve the OOV rate and morphology using several techniques. The first one is domain adaptation. A large percentage of unknown words comes from a domain different than training data so the ways to handle out-of-domain words by adding in-domain data are often exploited. The second option is to use **factored models** (Koehn and Hoang, 2007) that are trained on a corpus with linguistic annotation. E.g. (Turchi and Ehrmann, 2011), (Bojar and Tamchyna, 2011) address the problem of how to reduce the OOV rate by introducing morphological information or using additional dictionary resources.

[11] The detailed linguistic evaluation of Moses output will be given in the next chapter

In order to include syntactic information, researchers exploit syntax-based approaches such as treelet translation (Quirk et al., 2005), dependency-base translation (Bojar and Hajič, 2008), decoding-as-parsing (Yamada and Knight, 2002) approaches and some others, but neither of them has any 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 – such as 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 and 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 needs to take into consideration in MT. Generally, factors include information on the part of speech and the morphological tag; syntactic or semantic components may be also used. We exploit factors as they were initially suggested in (Koehn and Hoang, 2007). In addition to baseline settings, we add models that exploit standard techniques – stemming[12] and lemmatization. Those models are introduced as backoff: if a word or a phrase is not found in a main phrase table, then it is searched in a backoff one. 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[13] and examine it within several models.

- The **Baseline** setup is based on a simple translation model from a word form to a word form. Following is an example entry from a phrase table,

[12] As we will explain further, our stemming is not linguistically motivated, it is just stripping off the last characters of a word.

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

the first n-gram before the delimiter ||| is in Czech, the second respective n-gram is in Russian. The n-gram includes a female surname Bhutto that receives an ending ‘ová’ in Czech and is declined according to a feminine adjective paradigm, whereas female surnames are indeclinable in Russian.

(3.3) dohodě s bhutto**vou**[14] ||| 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:

  (3.4) 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 the source side relies on the information coming from a MorphoDiTa tagger (described in Section 2.2.3). The back-off phrase table thus contains mapping from foreign lemmas into target form+lemma phrases:

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

  Note, that ’bhuttovou’ was not in the dictionary of lemmas, so it was left as the form.

- **Stemmed model – backoff**. We define stemming as stripping off a word ending – which often bears some morphological feature. As we have mentioned, the notion ‘stem’ here is not used in its linguistic sense, it is rather a technical term. We did not exploit any existent stemmer in our work.

  The problem of how much a word should be cut off a word was not that trivial. Several experiments have shown that leaving six characters in a stem brings better results in terms of BLEU. So, the back-off table provides a mapping between Czech words with up to six characters on the source side (note, that only 6 characters were left from the word form ‘bhuttovou’) and a respective Russian form+tag sequences:

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

[14] As we have mentioned, words in phrase tables are lowercased.

> Though very simple at the first sight and not as sophisticated as lemmatization, stemming brought a slightly bigger improvement in terms of the BLEU score, but the number of out-of-vocabulary words decreased three times in comparison with the baseline setup (see Table 3.2). The lead of stemming experiments over lemmatization can be possibly explained by the fact that the morphological dictionary of TreeTagger does not include lemmas for unknown words (which are rather infrequent) whereas the stemming model guesses 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 word form *bhuttové* (Genitive case) in our training data,[15] but there are words *bhuttová, bhuttovou.* Neither Simple, Factored-main and not even Lemmatized model can provide a translation because the word form *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.[16]

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 3.2: 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 4.

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

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

3.2.3 Data issues: genre

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

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

Table 3.3: BLEU score for simple model trained on different genres

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 Czech and Russian. Combination of the two corpora resulted in the increase of the BLEU score by almost two points. We have not trained a model on subtitles only, as the data are very specific and unreliable, but the data were exploited 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 Table 3.4, we compare three experiments with several data and setups. In the first baseline system, we just used the Russian side of the respective parallel corpus that served as training data for the translation model - UMC. The second row presents the experiment with factored setup using the same corpus. The third one is an overall experiment involving all the parallel and monolingual data under the best setup with a back-off model.

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

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 3.4: The best experiment: factored setup + more data

3.2.4 OOV: Named Entities

It is not straightforward to introduce linguistic information into the SMT systems as it is for the RBMT systems. Unlike in the rule-based MT systems, we cannot 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 (NE). This choice was motivated by the fact that there were quite a few unrecognized named entities, including multiword NEs.[17] Our approach below was quite standard, it was very similar to the one employed by (Tan and Pal, 2014).

We used a list of names and phrases extracted from Wikipedia headlines[18] as this was the only parallel Czech-Russian resource of NEs we managed to obtain.

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

In addition to BLEU, we calculated the number of OOV words with the same method as described in 3.2.2 – searching for non-Latin characters. In the first experiment, the BLEU score was 17.23% with 1216 OOV words. The BLEU score in the second experiment was slightly better – 17.90% with 1011 OOV words.

[17] For a detailed linguistic evaluation refer to Sections 4.2.16 and 4.2.1.

[18] The extraction of the headlines is described in (Bílek, 2014). In the latter work, these data were used as additional parallel data and it was not measured how adding this corpus affected the translation. We thought it might be an interesting experiment because the headlines present different type of data than the text.

We examined the list of OOV words in the output from the two experiments. Among those 205 words/MWEs that were not translated in the first setup and were translated correctly in the second experiment, there were NEs from the added resource, such as *Higgsův boson* – 'Higgs boson', *praní špinavých peněz* – 'money laundering', *paliativní péče* – 'palliative care' etc.

3.2.5 Impact of language relatedness on SMT

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

Intuitively, we assume that translation between related languages should bring better results than translation between those non-related. In our case (translation between Czech and Russian) this did note 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 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 3.5 demonstrates that the 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 language pair.

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 3.5: MT between related and unrelated languages

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 word forms leading to data sparseness. The second reason has also been mentioned with respect to other problem: the parallel corpus is better parallel for the English-Russian pair than for the Russian-Czech pair, because both Czech and Russian texts are translations from English.

However, we should also note that introducing morphological elements into SMT decreases the number of unrecognized words especially for the 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 value because some of the "improved" out-of-vocabulary words may be guessed incorrectly.

3.2.6 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 and the various experiments with data. While comparing the output of several MT systems in the next chapter, we will use the experiment with the best factored model trained on all data we managed to obtain, that reached 17.90% BLEU.

3.3 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 Translate[19] 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 extensive human resources (PC Translator in the case of the RBMT) or all the virtually available data (Google Translate in the case of the SMT).

3.3.1 Google Translate

Google Translate is currently one of the most popular online MT applications in the world. The system, based on statistical models, is developed and extended by new language pairs since 2002. As for all SMT systems, the main principle "The

[19] We will sometimes use just 'Google' for short.

more data the better" applies, and that is why the system is one of today's best, as it exploits tremendous amount of data indexed by the Google search engine.

Parallel data for some under-resourced language pairs are not that easy to obtain, so it is possible that some sort of pivoting is used sometimes. We should say that the system is being constantly developed and improved in a 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).[20] In this work, we use the sentences from our test set that were translated using Google Translate in May 2014.

3.3.2 PC Translator

PC Translator is a commercial MT system with a rule-based architecture. We are not aware of methods and modules that 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 its dictionary. The English-Czech dictionary contains almost million entries,[21] and the Czech-Russian pair 650,000. Dictionary entries are not only single words, they can also represent multiword expressions, covering above all many idioms.

The PC Translator output for the Czech-English pair took part in the WMT-2012 competition, and although the BLEU score was relatively low (10% at the 15th position, compared to 16.8% of the winner), the manual evaluation showed that the system achieved the third position (following Google and a Moses implementation exploiting Depfix).[22] We can only assume that for the English-Czech pair many manually written rules were implemented. It is obvious that as in the case of TectoMT, the less perspective Czech-Russian pair received much less attention and the relative score was lower. To translate our test set, we use PC Translator version 2010.

[20] The examples demonstrating this fact will be given in the next chapter, the most evident one is Example 4.61.

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

[22] `https://ufal.mff.cuni.cz/depfix`

3.4 Conclusions and discussion

In this chapter we have described basic properties of several MT systems, and showed the development procedure and preliminary evaluation. Table 3.6 summarizes different information on the systems involved in our comparison. The BLEU scores are computed on 3000 test sentences from WMT2013 test set.[23]

	System	BLEU
rbmt	TectoMT	9.38 %
	PC Translator	4.73 %
smt	Moses	18.57 %
	Google	14.44 %

Table 3.6: MT systems and the BLEU score

Table 3.6 indicates that SMT systems, in general, received considerably higher score in terms of BLEU than RBMT. As we have mentioned in other sections, we do not consider the BLEU score to be the absolute indicator of quality, nor is it suitable for the RBMT systems involved. We take it as a relative measure that can indicate the progress within some system. However, it is still the most popular metric of automatic evaluation and the WMT competitions showed the score correlates with human judgments.

In (Bojar, 2011), a very similar comparison is made for English-Czech machine translation. The author described commercial and research systems that are the same as we use. The conclusions are slightly different, because the language pair is very frequent, and all the English-Czech systems have been developed for years, involving many people. The main conclusion of the work was that commercial systems generally outperform research ones.[24] It should be noted that there is not such a big gap in quality between RBMT and SMT for Czech-English because this language pair received a lot of attention within PC Translator. On the contrary, if RBMT systems for Czech and Russian would have been tuned for years, we can expect that 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 sys-

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

[24] The situation has changed since the time this article was written, and research systems beat the commercial systems in the WMT competition, e.g. the Hybrid English-Czech system developed at ÚFAL (Bojar et al., 2013).

tem developed twenty years ago on old machines, based on a formalism not used nowadays with state-of-the-art systems that involve new methods, tremendous amount of data and highly modern and efficient machines.

In this Chapter, we discussed mainly the 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 **linguistic analysis** of the MT output of the presented systems.

4

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 the perspective of a linguist, focusing on specific language issues that pose challenge to MT. The chapter aims at answering the central questions of this thesis:

- What kinds of mistakes Czech-Russian MTs of various architectures tend to make?
- Which of these mistakes 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?

4.1 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 improvements and showing some gain (or sometimes loss) in the quality of translation. 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 excessive human annotation effort.

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

- Evaluation of the output of MT systems;
- Evaluation of the system architecture, 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).

Our setting of MT evaluation is therefore quite challenging because we compare several MT systems of different origin and of completely 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** (or **flags**) to mistranslated words. This annotation scheme allows to reveal specific weak points of the system and yield 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, e.g. (Bojar, 2011) for Czech-English MT. Automatic error detecting tools were proposed recently, e.g. Hjerson (Popovic and Burchardt, 2011) or Addicter (Zeman et al., 2011), exploiting the reference translation in order to spot troublesome places of MT outputs. (Zhou et al., 2008) proposed a system of pre-defined linguistic checkpoints to test if systems translate those linguistic phenomena correctly.

4.1.1 Error Taxonomy

Labels specify which type of error is involved, ex. **lexical**, **unknown word**, **missing word**, **word order**, **word form**. Those errors will be explained in more detail further. Following is an example of annotated sentences translated by the four MT systems:[1]

(4.1)

[1] 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 most 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.

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

(tmt) **form::**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:[2]

- 'surface' word issues: missing word, extra word, unknown word;
- morphosyntactic issues: agreement, surface valency, Genitive of negation constructions, incorrect part-of-speech, other errors in endings, word order, negation, reflexives, other syntactic constructions;
- lexical issues: wrong lexical choice, wrong disambiguation, totally bad word sense, wrong usage of multi-word expressions.

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 our own more fine-grained classification inspired partially by (Bojar, 2011). The difference from the classification in the latter work is in that we sort errors more according to language layers. We set the category 'surface' word issues for errors that we can spot relatively easy in a sentence. Morphological errors in the latter work were marked as **form::** mistake, and there ere no subtypes of those errors; syntactic errors were not defined at all, there was only a category 'word order' mistakes. Because the languages that we deal with are both morphologically rich, we decided that a more fine-grained classification was needed.

The whole scheme is shown in Figure 4.1.1.

During the annotation process we added a few additional specifications, indicating more linguistic information, e.g. when a wrong form of a word is used we include its part of speech (*verbform::ввести*); when another POS should be

[2] We admit that the classification can be arguable and misleading as some errors can be related to more layers.

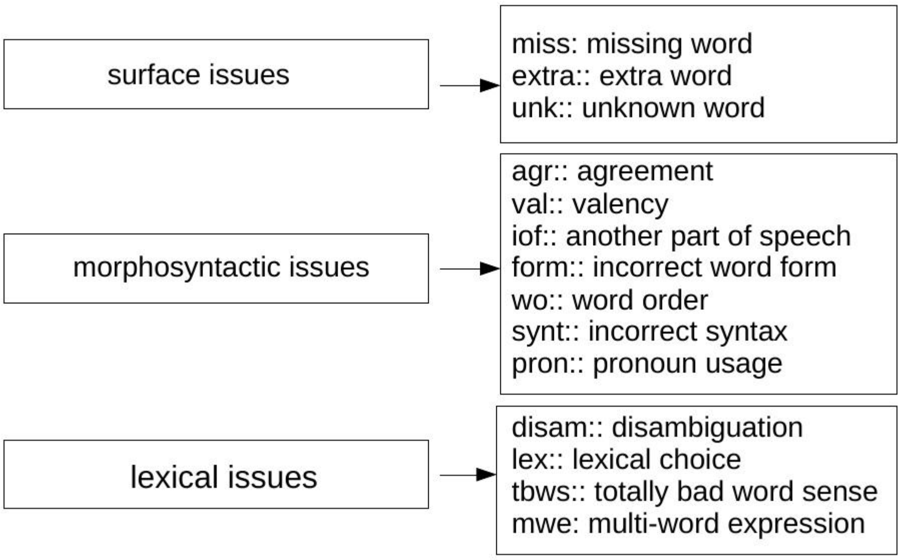


Figure 4.1: Error taxonomy adopted for Czech-Russian language pair

used, we specified both the wrong and the potentially correct versions (*n-iof-v::поддержке* - a noun instead of a verb), or in the cases when a word is missing, we specify the POS as well, as in the above example: *missAux-byt::*. Those tags will therefore help to identify various reasons that might have caused an error.

4.1.2 Some challenges of the annotation scheme

Before showing the resulting figures, we note some weak points and uncertainties of the presented approach and list some challenges we faced during the annotation process.

Annotation 'one tag per one word'

Firstly, we believe that in some cases it is more appropriate to use a single error tag for a string of words. 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 mistranslated multiword expression. 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:

(4.2) (src)	*volby*	*v*	*předstihu*
	elections.Pl	in.Prep	advance.Loc
	‘elections in advance’		
(mos)	*mwe::выборов*		
	elections.Pl.Gen		
	‘elections’		
(tmt)	*выбор*	*mwe1::в*	*mwe2::упреждение*
	election.Sg	in.Prep	advance.Acc
	‘election in advance’		

We have decided to mark each word with the respective tag even if this tag can get attached to more words to preserve annotation consistency.

Linearity of the approach

Secondly, we believe that the approach is somewhat “linear” as it is focused on a single 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 an example when a translation misses a verb. As the verb determines morphological properties of the dependent words, we could not exactly answer the question whether the ending of a noun (depending on the missing verb) is correct or not. It is not relevant for English, because if the dependent words are translated properly, there is no need to think about their morphological forms.

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 using the following example:

(4.3) (src)	*vystavený*	*svou*	*školou*
	issued.Participle.Masc.Sg	its.Pron.Ins	school.Ins.
	‘issued by its school’		
(mos)	*tbws::под*	*свою*	*школу*
	under.Prep	its.Acc	school.Acc
	‘under its school’		

In the Moses translation, the participle is mistranslated as a preposition, but there are some questions about the two dependencies – ‘its school’. With respect to the translated preposition, the form of the NP (the preposition *pod* – ‘under’ takes a noun in the Accusative case) is correct, but if the word "issued" was translated into Russian properly, it would need another case (Instrumental, as in the source).

Subjectivity of the evaluation

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

4.2 Error types in Czech-Russian MT output

As mentioned above, we have first classified 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 layer at the same time.[3] 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 surface language level to the deeper levels.

First we describe a group of ‘surface’ errors that are quite easy and unambiguous to annotate and generally do not have a deeper linguistic motivation.[4] 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: word form, 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

[3] 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.

[4] In spite of being surface, those mistakes can be considered as ‘serious’ because they affect dramatically the general perception of a sentence.

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 specific 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 4.1, the statistics of errors found in 100 sentences for all the four systems is presented.

Error/System	MOS	GOO	PCT	TMT
untranslated word	3	0	118	40
missing word	13	13	3	3
extra word	20	15	21	20
'surface errors' in total	36	28	142	63
word form	10	26	33	44
incorrect pos	8	10	14	0
valency	9	19	21	16
agreement	7	6	27	21
word order	7	13	10	7
pronoun usage	1	2	3	6
'to have'	1	3	3	0
'to be'	3	3	8	5
other syntactic errors	7	7	4	3
morphosyntactic errors in total	53	89	123	102
lexical choice	20	14	60	39
disambiguation	15	13	37	18
multi-word expression	1	2	7	11
totally bad word sense	7	6	41	8
lexical errors in total	43	35	145	76
errors in total	132	152	410	241

Table 4.1: Error types in Czech-Russian machine translation systems

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

Therefore we demonstrate examples of errors and discuss possible reasons why this very mistake may have occurred. We suggest if an error is a result of some discrepancy between the two languages, whether it is a merely a technical issue of the MT system or a mixture of both.

We should also note that all the conclusions and assumption made in this chapter are tightly linked with specific data samples. 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. The Moses output is even trickier to evaluate, because the output may be different each time the system is trained and tuned.

4.2.1 Unknown words – OOV

The most evident MT errors that impede understanding of a text are untranslated words (unknown words, out-of-vocabulary – OOV). Some other mistakes – such as wrong form of a word or syntactic 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 SMT, we have already described this problem in Section 3.2.2 where we explained in detail why the OOV words occur in SMT Moses and introduced some obvious improvements to the baseline experiment that helped us to significantly reduce the number of OOV. Another experiment in Section 3.2.4 concerned the lexical part of OOV – true OOV 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 the BLEU score, but whenever a word was present 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, although it could have occurred in some other morphemic form. 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 the news domain which are often translated into more languages)[5] and there were only 3 errors in the Moses output. We also calculated a number of OOV words in the whole WMT test set just by searching in the

[5] It is possible, that our test set is included into the Google's training data as well.

translation for words including a Latin character.[6] Google has 315 untranslated words (which is only 0,006% of all words) and Moses 250 (0,004%). So the OOV errors, when addressed properly (and in-domain), are not really a big challenge for the statistical systems.

As it is evident from Table 4.1, the RBMT systems produced many 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 the PC Translator pages) contains quite a large amount of entries (around 650,000 for the Czech-Russian pair).

TectoMT dictionary is based on the automatically extracted lexicon (see Section 2.2.2) 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 the system architecture – most probably in the analysis of the source sentence, and we can not really say which because we can not look inside the system. Words unknown to TectoMT were mainly named entities, and some of them were reflexive verbs that were not properly handled within the system.

4.2.2 Missing word

Another serious 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 the center of a sentence as it determines 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 (e.g., whether they have the proper morphological ending required by a verb).

The 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 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**, a verb is omitted quite often which affects the perception of the whole sentence and sometimes makes it impossible to further evaluate the

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

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:

(4.4)

(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 omitted words in the RBMT than in SMT systems. In the RBMT systems, those mistakes are all missing prepositions which are not that serious and do not affect the perception of the sentence. Still, these are grammatical errors and they can be justified by the difference between Czech and Russian surface valency.[7] 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 the Accusative case, and the system translated a respective noun with the 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:

(4.5)

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

‘to influence the outcome’

(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’

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

(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.

4.2.3 Extra words

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

For 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 the SMT systems because often the whole phrase is mistranslated and more errors are involved. For instance, the preposition was redundant in the following example:

(4.6)

(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 the 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 cannot, so the decision was to mark the preposition *при* – 'by' as an extra token and the noun *поддержке* – 'support' as a 'noun instead of verb'.[8]

[8] This very error might have also occurred because of an embedded time adverbial 'in year 2011' which interfered with the argument structure. We translated the phrase without the

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 the sentence. The 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 the pro-drop phenomena[9] in the Czech language. In short, whereas Russian (and English) uses a pronoun in the subject position, it is in most cases left out in 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:

(4.7)

(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 are rather a lexical issue and it is virtually impossible to make an exception to the pro-drop rules for all such troublesome cases.[10]

4.2.4 Agreement

A substantial number of morphological errors can be related to **agreement**. Agreement reflects the obligatory presence of a morphological feature in the form of a word, agreeing, e.g. in gender or number, determined by another word. Czech and Russian have the same types of agreement, but this fact does not mean a trouble-free translation even for the rule-based systems. Although there are many types of agreement, we will describe only the most typical cases that are

adverbial by Google Translate (*Legislators supported laws*), and it was translated properly – with a verb. We will discuss the issue of embedded clauses in more detail later.

[9] More on the pro-drop phenomenon can be found in Section 4.2.10.

[10] 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 other cases.

relevant for the machine translation - predicate-subject agreement and attribute-noun agreement.

Subject-verb agreement

The relation between a subject and a predicate is dual: on the one hand, the noun determines the suffix of the verb (agreement), and on the other hand, the predicate governs the specific case of the noun (surface 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 (in case of pronouns), but not in gender. The following example demonstrates verb agreement with a subject in person:

(4.8)	(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 do not stand far from each other, the number/gender morpheme has a better chance to be chosen properly because the respective n-gram is more likely to occur in the 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:

(4.9) (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, RBMT systems have better chances to cope with agreement issues as it is determined by rules. However, even if those rules are present in the system, it can fail to determine which words should agree with each other (due to errors in parsing). The second case is when a word with which a verb agrees is not translated properly or is not translated at all. It is therefore not evident how to mark a mistake (see Example 4.10). There was nothing to agree with in the PC Translator output and the wrong usage of infinitive form made it impossible to express an agreement in number for TectoMT.

(4.10)	(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) word 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-attribute 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) the governing noun,[11] so the SMT systems generally produce the correct hypothesis. From Table 4.1 it is evident that noun-adjective agreement errors in SMT are really rare 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 nouns and pronouns, which results in an error when the gender of a Czech noun is different from the Russian one. In the example below, 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.

[11] In this work, we will not touch non-projective constructions (informally called crossing dependencies). In the two languages, non-projectivity is allowed all the same, e.g. in Czech *Vánoční nadešel čas* – ‘lit. Christmas came time - Christmas time has come’; in Russian *Тяжелые настали времена* – ‘lit. Difficult came times - Difficult time has come’.

(4.11)

(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 4.12) it is evident that PC Translator does not properly handle 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. However, if the system had taken into account the relation between the two attributes and a noun, this morphological ambiguity would not have occurred.

(4.12)	(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’		

4.2.5 Incorrect part of speech

We propose a special category for mistakes when the word sense is chosen appropriately, but the part of speech is wrong, e.g. 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:

(4.13)

(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 participle or a transgressive that is confused with some other part of speech:

(4.14)

(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-**participle**	171	from	270	votes

‘states representing 171 from 270 votes’

4.2.6 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 including negation. The problem of Genitive of negation is a well-studied, e.g. (Mustajoki and 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 four Slavic languages – Czech, Russian, Polish and Slovenian.

Here, we will not go into a detailed description of this construction, but only point the errors that are made in sentences containing genitive of negation. Object marked as Accusative instead of Genitive case in the context of negation (Example 4.15) can be considered as an error in a word form of a noun, but we will treat this issue separately from other morphological mistakes.

(4.15)

[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

(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 the source sentence, the negated verb has dependents in Accusative case, and in the output all the systems produced objects in Accusative (or other improper cases), whereas the proper variant was Genitive (see ref example).

Though the most frequent morphological case used with negation is Genitive, there are cases when Accusative is used in this context. The choice Genitive vs. Accusative is determined by several properties of the object, e.g. the proper names of animated objects are used mostly in Accusative case.

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

(4.16)

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

[13] It was used in Old Czech, remains of this construction are some phrases like *nemám ani potuchy* – 'I don't have the slightest idea', see (Hausenblas, 1958) for a broader discussion.

(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 4.2.11.

Let us now demonstrate the difference between the two constructions in Russian, the one with Genitive of negation (Example 4.17(1ru)) and the one where a Nominative case is used:

(4.17)

(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] (Partee et al., 2011) and (Babby, 1980) showed that the two constructions can be distinguished with respect to their semantic properties: the Genitive of negation case (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 specific place (the viewers may be standing near the cinema).

Genitive of negation is an example of an obvious discrepancy between Czech and Russian, and sometimes it can cause mistakes in nominal 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 a transfer rule (substituting Genitive in the context of negation), but to ensure that this rule works a good analysis module (a syntactic parser) is needed.

4.2.7 Surface valency

In this thesis, we examine in detail one particular issue that we believe is crucial from the point of view of machine translation and language comparison as well.

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

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 a certain semantic type. Here, we will refer to valency with respect to its surface realization – morphemic endings of nouns or preposition required by a verb[15].

It was challenging to set distinct rules to distinguish surface valency errors from mere errors in word forms. In short, we attach the label 'valency' when dependencies of a predicate are used in a wrong form. On the surface, these errors look like morphological, but they result in syntactic and semantic unacceptability as well. So, in the following, we use the word 'valency' in the 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 SMTs and RBMTs. 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".

(4.18)

(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'

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

'detach voting rights from citizens'

In Czech the verb *odejmout* – 'to take away' requires an object in Dative case, whereas in Russian the verb *отнимать* – 'to take' with the preposition *у* – 'from' plus Genitive case should be used instead. The systems often use the Czech valency structure which sometimes results in a mistake like in the example above. This very case – verbs with the semantics of 'taking something from

[15] More sophisticated and broad definition of valency, theoretical and practical aspects of this linguistic phenomena will be given in Chapter 5. Here we will just present examples of errors not going deep into the comparative analysis of valency in the two languages.

somebody' is very tricky in Czech. We used Google Translate to translate from Czech into English, and the same error was made: the sentence 4.18(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 systems.[16] 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.'

(4.19)

(src)	*ovlivnit*	*výsledky*	*voleb*
	influence	results-Acc.Pl	elections-Gen

'To influence results of the elections'

(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 the translation was done directly without applying a rule on this discrepancy. Both statistical 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 Google Translate 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':

(4.20)

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

'Those restrictions concern mainly young people'

[16] We will present an experiment with adding surface valency rules in Section 5.2.3.

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

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

(4.21)

(src) *proti schválení zákonů, jež ..., se postavili demokratičtí*
against approval laws-Gen, which ..., refl stand (against)-*Verb*
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 the sentence has adopted a morphological marker of an object dependent on the preposition – the 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. A 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 make errors in cases when the valency structure of a verb in Czech and Russian was similar. Those errors might be the result of an improper analysis of the source phrase or some error in the transfer or generation phases.

In Chapter 5, we investigate the definition of valency more deeply. We also present some observations of discrepancies between Czech and Russian surface verb frames.

4.2.8 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.[17] The tag 'word order mistakes' does not indicate that if we change the order of translated words, the error will disappear. 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[18] 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 various types, see Example 4.22(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 4.22(tmt), except for one error that is related to the Genitive of negation, Section 4.2.6):

(4.22)

(src)	*nemají*	*američtí*	*občané*	*průkaz*
	not-have-3Plneg	american-adj	citizens	id

[17] 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.

[18] Here, we will use the notion 'marked', also non-standard word order for every word order other than SVO.

‘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**. In the example below, 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:[19]

(4.23)

(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

In the example above, the 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

[19] 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.

'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 4.24(src) or when a clause separates the two main arguments from one another.

(4.24)

(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, 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.[20]

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:

(4.25)

(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...'

(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'

[20] Just to mention, all the systems failed to properly disambiguate the word 'stát' – state which is ambiguous in Russian – for more information refer to Example 4.50.

(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 4.26, 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.

(4.26)

(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 adversative 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[21] position in a sentence: Czech obeys ‘a law of second position’, or Wackernagel’s law which does not apply in Russian language. In Czech,

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

clitics are 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 (pronouns 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,[22] 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 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 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:

(4.27)

(cs)	*Podařilo*	*by*	*se*	*mu*	*odejít*
	succeeded	would	refl	him	leave
	'He would succeed to leave'				

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

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

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:

(4.28)

(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)	*То*	*ему*	*вымстит*
	This	him	*unknown-word

*'This to him вымстит'

Google Translate was the closest to the right variant, whereas Moses, PC Translator and TectoMT generated completely inaccurate translations.

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

(4.29)

(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 meaning 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 (4.29), PC Translator confused the reflexive pronoun *se* with a preposition *se* – ‘with’[23]. 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 4.30. 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.

(4.30)

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

[23] This mistake occurs always!

‘This measures will partially undermine the US democratic system’

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

*‘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, e.g. *до 2004 года* – ‘lit. up to 2004 year’, in comparison to the corresponding constructions in Czech or English (*do roku 2004* – ‘up to year 2004’). This mismatch entails the following errors in RBMT systems:

(4.31)

(src) *do roku 2004*
to 2004 year
‘up to year 2004’

(pct/tmt) *до года 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 contrastive constructions. We have fixed it in TectoMT system with a block FixDateTime.pm.[24] The order of the two words – ’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%.

4.2.9 Constructions with the verb ‘to be’

The verb ‘to be’ has many meanings (copulative, existential, auxiliary, modal etc.) and has many translation equivalents in the languages. It is not always easy to distinguish between those meaning of the verb, so rule-based systems often mistranslate the constructions because the type of ‘to be’ verb is not recognized during the analysis. Here, we will discuss those functions of the verb ‘to be’ which, according to our data, impact the MT output most.

[24] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/FixDateTime.pm`

Copulative meaning - zero copula in Russian

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 (Česílko and TectoMT). Naturally, we want to test how MT systems handle this construction. The so-called copula drop, or zero copula, is a phenomenon when the verb 'to be' in its copulative meaning[25] is omitted. This phenomenon 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 Example 4.32 (cz-1) and (ru-1), but there are more translation variants. The form with Instrumental case (cz-2) can 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 copula in Present Tense, in the Past Tense the verb *byl(a,o)* – 'to be' is present in both languages and is expressed in the same way with the same cases.

(4.32)

(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)

We will now make some diachronic remarks. 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

[25] '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.

can be the reason that Russian has too many grammatical discrepancies with other Slavic languages.

Next, we will show several cases where MT systems made some mistakes while translating a sentence with the copula 'to be'. Statistical systems sometimes translate copula constructions properly and sometimes not. This can be presumably attributed to the fact that there is no unified translation pattern and different variants (see examples from 4.32) can occur in various combinations in the training data. In the following sentence, one mistake in the case (Instrumental instead of Nominative) occurred when the verb was not present:

(4.33)

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

(4.34)

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

'that they are American citizens'

(pct)	*что*	*являются*	*американскими*	**обцаны*
	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:

(4.35)

(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 Example 4.35 all the systems made different mistakes: Moses evidently used the Instrumental case from the source sentence with the copula which will be otherwise correct with the verb 'являться'.

Google has the opposite of Moses: the verb *являться* was used, but the Instrumental case of the dependent that is required with the verb was not. Moreover, the ordering of the verb *являться* and the negation particle is reversed (negation particle should come before a verb). This example is another proof of a fact that Google uses English as the pivot language to translate between Czech and Russian. The mistakes where a noun phrase is in its base form, the negation particle stands after the 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.

The TectoMT output included 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 recognize and identify the subject of the 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 new, more complex ones should be implemented. The simplest decision would be to stick to one 'basic' copula-drop style in the target Russian and to translate every copular construction, such as the sentence 4.32 (ru-1) while always changing the case to the Nominative. Again, a parser should recognize 'to be' as a copular verb and not mix it up with the existential 'to be' which will be described further. So, to handle this mismatch properly, we we need to recognize the copula during the analysis phase and write a proper rule of transfer and synthesis.

After this error analysis, we included the respective rule into the TectoMT system, to the transfer stage. In the baseline system, the copula verb was dropped, but the case of a predicative noun stayed the same as in the source – Instrumental (Example 4.34(tmt)). First, we tried to fix this error according to the sentence from Example 4.32(ru-2) – substituting the Czech copula with the verb *являться* without changing the case.

The BLEU score was not changed, and the translation was grammatical, but sometimes not fitting the language style (too official). So we decided to fix it according to the pattern 4.32(ru-1) – without a verb, just changing the case from Instrumental to Nominative. Again, the BLEU score changed only a little bit - by 0.001%.[26] However, the translation of the Czech sentence 4.34(src) into Russian is now correct (*что они американские граждане* – 'that they are American citizens'). As in the example with the year fix, and probably for the majority of such minor linguistic issues, the BLEU score is not really an appropriate measure.

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 specific problem can be the introduction of a post-editing rule on the same principle as for the RBMT.

[26] See the block `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2A/RU/DropCopula.pm`

Auxiliary ‘to be’

A similar problem concerns the Past Tense which is formed by the means of the auxiliary ‘to be’ plus the past participle of the main verb in Czech, but only for the first and the second person. The copula for the third person subjects is omitted. In Russian, it is omitted for all the persons (again, it was used in Old Russian and Old Church Slavonic).

(4.36)

(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:

(4.37)

(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 team’

From the example, we can see that statistical systems can handle the Past Tense properly (Moses), but not always (Google). Rule-based systems translated the verb in a correct tense, but either the preposition or the 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 often present in Russian sentences as well as in Czech. (Partee and Borschev, 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 4.38 (*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, and 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, see Example 4.38(tmt). Statistical systems produced the same (almost proper) translation as TectoMT.

(4.38)

(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, the verb *есть* – 'to be' should be used (*В москве сегодня* ***есть*** *хорошие консультанты* – 'Today in Moscow **there are** good consultants').

4.2.10 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,[27] but the languages chose different means to encode it. Russian is inclined to use a pronoun, like in the example 4.32(1-ru), whereas in Czech this information is encoded in the auxiliary verb 'to be', see ex. 4.32(1-cz)[28].

To illustrate this discrepancy we have calculated statistics of pronoun usage in the parallel Czech-Russian corpus (Klyueva and Bojar, 2008). In Table 4.2.10 we show that for the same sentences the usage of personal pronouns in Russian is far more frequent than in Czech.

pronoun	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 4.2: Pronoun usage in Czech and Russian

Due to such a disproportion in 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 4.39 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. PC Translator's output does not include the pronoun at all.

(4.39)

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

'They constituted only 11% of all the electors'

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

[28] 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 the sentence is not parsed correctly or the subject is not recognized at all. In the following example, obviously due to the parsing error, the subject was confused with the object, and the subject pronoun was wrongly inserted into the sentence.[29] This error can also belong to the word order/valency or parsing issues.

(4.40)

(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 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 4.39 (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.

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

4.2.11 Constructions with the verb ‘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 possessive construction *mít* – ‘to have’ is translated into Russian as *У кого-л. есть* – ‘Smb. has’ (see Example 4.41). 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’).

(4.41)

(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 4.42). Again, as with the verb ‘to be’, the variety of translation equivalents leads to more hypotheses from which the decoder has to choose. SMT systems often tend to translate Czech possessive constructions with ‘mít’ as ‘иметь’ which is sometimes acceptable in formal contexts. This can be attributed to the fact that the majority of training data come from the news domain, 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 4.42), 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 output of SMT and is relatively acceptable. PC Translator, on the contrary, has some rule for handling ‘to have’ constructions, but in many cases it confuses the possessive (have smth./smb.) with a modal (have to do smth.) functions of the verb ‘to have’, translating the possessive verb in its modal meaning:

(4.42)

(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 4.43) by Moses, though Google made a minor mistake leaving out the subject-possessor, so the sentence became impersonal:

(4.43)

(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 кого-л. есть*, but in most cases it is not applied properly, see Example 4.44. 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’.

(4.44)

(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 the auxiliary, modal usage or polarity and some other word order aspects.

4.2.12 Some other syntactic constructions

In the few points above, we have demonstrated the most frequent, obvious 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* cannot be related to any category as they are very unique to a specific language and are generally related to some lexical issue. Let us show a few constructions that are often translated wrongly.

The Czech construction from the Example 4.45 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.

(4.45)

(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 a phrase such as *то же самое происходило с* – ‘the same happened to’.

(4.46)

(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’

Transgressives

Transgressive is a non-finite form of a verb 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:

(4.47)

(src)	*zákony*	*vyžadující*	*předložení*
	laws	demanding-participle	presentation-**noun**

‘Laws demanding demonstration’

(mos)	*законы,*	*требующие*	**продемонстрировав*
	laws,	demanding-participle	*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.

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 three forms for a Czech transgressive 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 transgressive declension (a/ouc/ouce vs. e/-íc/-íce).[30]

As we have shown in (Klyueva, 2013), transgressives are used 40 times more in Russian than in Czech. The most frequent equivalent constructions to Russian

[30] This complexity might be the reason why they are disappearing from the language – native speakers just stopped using them.

gerunds in Czech are dependent clauses (Example 4.48) or coordination clauses (Example 4.49):

(4.48)

(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’

(4.49)

(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-**transgressive**	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.

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. *takřikajíc* – ‘so speaking’).

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

Above, we have shown only a few 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: the syntactic structure might be correct, but in a slightly different context it can be totally wrong.

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).

4.2.13 Disambiguation

When a word in a sentence is used in a wrong sense, we treat 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 the source (Czech) lexeme is polysemous and the target (Russian) lexical equivalent does not cover some sense that the Czech one has and those senses are expressed by different lexemes (see Example 4.50, word *stát*).
- There are more sense components in the target (Russian) lexeme than in the source Czech. E.g., the word *diplomat* can indicate only a human of a specific 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. E.g. the Czech word *sourozenci* – 'siblings' – cannot be translated into Russian by a single word (*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 Table 4.1 proves this fact. The choice of a proper sense depends on the context of the 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 the form of statistical post-editing, as, ex. presented in (Galuščáková et al., 2013).

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

The 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 cannot be sure about the architecture of this commercial system. On the basis of the analyzed data we have only concluded that PC Translator made far too many mistakes in the lexical semantic issues in comparison with the 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: in addition to other meanings, it can refer to a state as a country and a state as a province (esp. in texts about USA). In Russian, 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 the meaning of a state as a part of the USA, but all the systems chose the most frequent variant (country):

(4.50)

(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:

(4.51)

(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 can occur, see Example 4.52. A Czech word *let* can be translated either as a *рейс* – 'flight', or it can also be the suppletive genitive form of the word *years* in the context *10 years*:

(4.52)

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

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

This error evidently comes from the morphological analysis, during which the word *let* was not recognized properly – as the Genitive plural form of the lemma *rok* – 'years'. In TectoMT, the construction with age receives a special treatment, so it was translated correctly. Statistical systems also cope with this 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 Table 5 in (Bojar, 2011)) languages. The only significant difference in the table of Czech-Russian errors in comparison with that for Czech-English is that PC Translator scored significantly worse with respect to all lexical-semantic issues than other systems. Just for comparison, for the Czech-English pair, PC Translator had 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,[31] so it copes better with word sense issues for this language pair.

4.2.14 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. Errors tagged as **disam::** specify completely different sense of a word. Words marked with **lex::** bear mostly a proper sense, but some very close synonymous word is chosen instead.

Wrong lexical choice is a problem that is really hard to detect automatically because the difference between the two nearly 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:[32]

(4.53)

(src) *voliči* *předkládali* *průkazy*
voters demonstrated id's
'The voters showed their ID cards'

(mos) *избиратели* *выдвигали* *удостоверения*
voters proposed id's
'Voters put forward their ID cards'

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.*[33]

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 –

[31] The Czech-Russian pair is not at all a popular direction, which reflects the translation quality.

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

[33] This may seem as a disambiguation error from the point of view of English, but for Russian the two senses are closer synonyms.

'people', and the latter sense was used in the sentence 4.54 (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 led to a mistake.

(4.54)

(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 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%.

4.2.15 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 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:[34]

[34] According to our disclaimer about Google Translate, we used the text translated in May, 2014. Now, this mistake does not occur anymore and this phrase is translated correctly. We can only speculate about the reasons, but this will be a mere guess-work, because we do not know anything about the development process of Google Translate.

(4.55)

(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 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, 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 4.56. Some other words – mainly named entities were translated by very strange equivalents (like Indiana as ‘Hindi-Diana’) as well.

(4.56)

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

‘gaps in the model’

(pct)	*пробелы*	*зажечься*	*модели*
	gaps	light-inf	model

*‘Gaps to light a model’

4.2.16 Multi-word units

In this thesis, we will understand a multiword expression (MWE) as a sequence of words with non-compositional meaning - where the meaning of a phrase cannot 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, and how they are processed and incorporated into MT systems, e.g. (Anastasiou, 2010), (Bouamor et al., 2012). MWEs have been annotated within the Prague Dependency Treebank (Bejček and Straňák, 2010).

MWEs differ from language to language and are highly idiosyncratic. Even for the related Czech and Russian we can not be sure if the structure of MWE is similar. Both approaches – rule-based and statistical – experience difficulties when processing those units. Because meaning of multiword expressions is not compositional, the RBMT-based systems without appropriate information will translate the units word for word, which can lead to mistakes. 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 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.

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

(4.57)

(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 all three systems made an error in this case. Russian 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 Czech words correspond to one word in Russian are not so infrequent, e.g. *теракт* – 'act of terrorism' vs. *teroristický útok* – 'terrorist attack'.

Auxiliary multiword expressions

Auxiliary MWEs are 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 with handling them properly because their co-occurrence in 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, see Example 4.2. 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 the verb looses its initial meaning and the whole construction takes the semantics of the noun.[35] Following are several examples of Czech LVC in contrast to Russian. They can be equal for both (using the same verb):

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

[35] Researchers do not have a single standard definition of light verbs, sometimes it is quite hard to distinguish LVCs from idioms.

Or the languages can use different light verb:

(4.59) *brát zřetel* – 'lit. take consideration' vs. *принимать во внимание* – 'lit. accept into attention' - In English *take into consideration*

The examples above illustrate that some multiword verbs have identical component words in the two languages, and some do not. Generally, multiword expressions are translated properly in SMT when an LVC presents a bigram, but when the verb is separated from the noun, this LVC is generally mistranslated, see Example 4.60. The LVC from this example has the same structure in Czech and Russian, so the error cannot be explained by discrepancies between the languages.

(4.60)

(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'

(pct)	*прукопници*	*лазер*	*зажечься*	*белловэцћ*	*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 correct 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 correct structure using a good 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 the 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. As our data belong to the news domain, we have not found any idioms in the test set.

Czech and Russian are related languages, and the idiom expression might be equal in both languages, but that is not always the case.

Just for the sake of the experiment, we translated one idiom that has a different structure in Czech and Russian within Google Translate. The results were quite interesting. The translation from Czech into Russian copied the English variant of the Czech idiom (*He makes something out of nothing.*). As for the translation from Russian into Czech, it copied the structure of an English idiom *He makes mountains from molehills.* This can be another proof that in some cases Google Translate uses English as a pivot language.

(4.61)

(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)	*To*	*dělá*	*hory*	*z*	*molehills.*
	It	makes	mountains	from	molehills-**English**

*‘He makes mountains out of molehills’

All in all we can conclude that multiword expressions present a problem mostly for RBMT systems that need to have a bilingual lexicon of MWEs. Statistical systems cope with them as soon as a multiword unit fits into a respective n-gram which is relatively frequent in the training data, see the experiment in Section 3.2.4.

4.3 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. The error taxonomy and annotation schema are among the most used ones for the task of manual error analysis, but we have made a more fine-grained classification of errors.

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 all 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, the company LangSoft has been developing PC Translator Czech<->English pair for years, and there are still many errors in it, moreover, it did not even receive the best BLEU 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 the BLEU score, or the system performance for some phenomena.

The quality of translation also highly depends on the source: if the 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 to what other researchers have written about rule-based and statistical MT.

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 constitute the specific phenomenon do not stand far from one another. If dependent words stay far from each other in the source sentence, the respective n-gram will most probably not be found in the training data and thus the translation can easily be incorrect.

RBMT : Discrepancies between the languages have a much higher impact on the RBMT systems than on SMT. The 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 correctly. Another reason for a mistranslation is often an error in the analysis (parsing) or synthesis modules. We have fixed several errors by including new blocks into TectoMT. Though this system handles the specific phenomena correctly after the fixes, those changes are not reflected in the BLEU scores.

To sum up, what helps RBMT are years of hard manual work on lexicons, rules and the language processing modules. What helps SMT is mainly data (like adding larger translation and language models, domain adaptation etc.).

5

Valency in Czech and Russian

In this chapter, we discuss theoretical aspects of valency and focus on the comparative analysis of Czech and Russian surface valency using various linguistic resources. As we mentioned in the previous chapter, we will use the term 'valency' in the sense of 'surface valency' and take into consideration only surface realizations of valency frames.

The results of manual evaluation of MT output revealed that valency errors occurred in the output in all the systems. The amount of this type of errors (see Table 4.1) is not so big in comparison, for example, with untranslated words or lexical choice, but we found those errors to be interesting from the linguistic point of view as they concern several language layers: morphological, syntactic and semantic. Verb and its complements form a core of a sentence, so the mistakes in the surface form of the complements can lower the quality of a sentence.

Our initial assumptions 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 the Czech and Russian verbs had the same frame with the same morphological cases, see examples in Section 4.2.7. 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 generally 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.

- Phrase-based systems are hard to evaluate in linguistic terms. 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 an improper case, again independently of valency discrepancies/equality.

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 the Czech and Russian valency, aiming at building a Czech-Russian valency lexicon and integrating the data into the MT system. The only resource containing the data with valency information for the two languages was Ruslan lexicon (Section 2.3.1). This resource is quite outdated and not very reliable, so we decided to make an experiment on automatic extraction of Czech and Russian surface valency frames from parallel data. 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 a different surface valency in the two languages, semantically related verbs are likely to have this discrepancy as well.

Our experiments share similar ideas with a number of other research projects on valency within machine translation. In the following, we will name those that work with either Czech or Russian. In (Bojar and Šindlerová, 2010), the authors collect valency translation equivalents for Czech and English verbs exploiting a parallel treebank. (Rosa, 2013) built a simple probabilistic valency model for Czech and English and used 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.

This chapter is structured as follows. First, we define what we understand under the term ‘valency’, show the existing valency resources (Section 5.1). Then we describe the extraction of the surface valency frames from Ruslan dictionary, examine differences in valency and implement the extracted list of verbs + frames into the TectoMT system (Section 5.2). As Ruslan is a rather limited source of information, we also make an attempt to automatically derive a lexicon with surface frames using the valency resource Vallex, a bilingual dictionary and a parallel corpus (Section 5.3). Finally, in Section 5.4 we explore 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 dictionary.

5.1 Notion of valency

5.1.1 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.[1]

We have already made a disclaimer about the term 'valency' in the previous chapter. The way how we use 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 brevity, we will call this information 'valency', as it was done in other computationally-oriented works.

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 the 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 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.

In Prague, a valency theory was developed within the already mentioned FGD framework (see Section 3.1 for a detailed description of the language layers).

[1] 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.

Within FGD, valency can be spotted on the tectogrammatical layer in a form of valency frames. The valency frame of a verb is represented by a sequence of verbal[2] complements written in a form of **functors**. A functor has a number of morphemic realizations (e.g. case, preposition+case, relative clause etc.). A variety of examples will be presented further in the text.

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 complements based on a dialog test (Panevová, 1974). A wh-question about each complement is presented to the speaker and if the answer ''I don't know" in a coherent dialogue is possible, the complement is 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 in the semantic and the syntactic sense. Developed in roughly the same years, MTT and FGD theories share many common features: division into language 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). An extensive lexico-semantic resource based on the MTT theory – Explanatory Combinatorial Dictionary – was developed, which will be described in the next section on valency resources.

Prepositional vs. non-prepositional complements

In our work, we 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 the preposition governs its object. Actually, it does determine the case of the following noun. According to other theories (Kuryłowicz, 1960), a 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. However, on the tectogrammatical (more semantic) layer, the preposition becomes only an attribute to the respective noun, and that means that its function in a sentence is really more close to morphological. In the most recent studies (Universal Dependencies format),[3] a preposition is represented as a child of the noun.

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

[3] `http://universaldependencies.github.io/docs/`

As we study mainly the surface valency, we will often use phrases like 'a preposition entails the 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.

5.1.2 Valency Resources

Next, we will name the most reputable resources or those resources relevant to our work.[4]

FrameNet

FrameNet (Baker et al., 1998) presents a freely available lexicon of words organized into a semantic hierarchy. Words in FrameNet are assigned with a semantic frame reflecting roles of the main actants of the word. Each lexical unit in a sentence is assigned a semantic role – or frame element; frame elements, in turn, form a semantic frame of a lexical unit. Following is an example[5] of a frame element of a verb *to fry*:

(5.1) [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 and Palmer, 2002) – an abbreviation from 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

[4] The dictionary from Ruslan (Oliva, 1989) that we base our work on is described in Section 5.2.

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

roles are not that deep and granulated in comparison with the FrameNet, see the following example:[6]

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

The arguments are assigned with the specific numbers. Also, the modifiers are tagged with more semantic specifications like Manner, Time or Locative.

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 – sets of synonyms sharing the same subcategorizational pattern – or surface valency frame. Verbalex is organized more like an hierarchy of verb classes whereas Vallex semantic classification is just an additional feature.

Vallex

Vallex[7] 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 lexemes). The original valency entry of Vallex contains complex linguistic information:

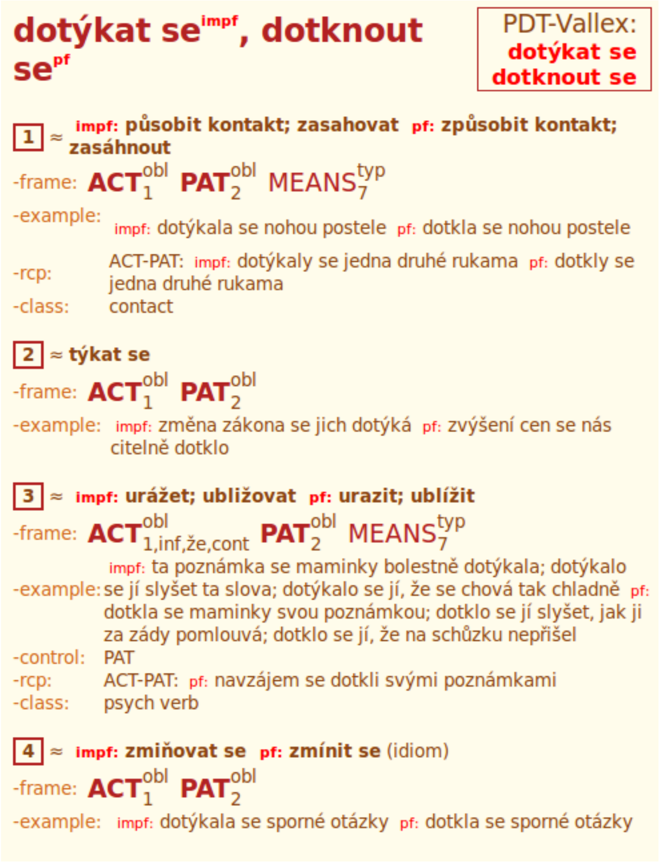
- a lemma – the basic form of a verb;
- a frame:
  - a functor - a rough analog of a 'deep role' (Actor, Patient, Addressee etc.);
  - a surface realization of the functor;

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

[7] http://ufal.mff.cuni.cz/vallex/2.5/

- a semantic class of the verb;
- examples of using the verb in a real context;
- information on reflexivity, aspect, idioms and some other.

Let us take as an example the verb *dotknout se* – ‘touch’, following is the Vallex entry for this lexeme:



dotýkat se^{impf}**, dotknout se**^{pf}

PDT-Vallex:
dotýkat se
dotknout se

1 ≈ impf: **působit kontakt; zasahovat** pf: **způsobit kontakt; zasáhnout**

-frame: $\mathbf{ACT}_{1}^{obl}$ $\mathbf{PAT}_{2}^{obl}$ MEANS_{7}^{typ}

-example: impf: dotýkala se nohou postele pf: dotkla se nohou postele

-rcp: ACT-PAT: impf: dotýkaly se jedna druhé rukama pf: dotkly se jedna druhé rukama

-class: contact

2 ≈ **týkat se**

-frame: $\mathbf{ACT}_{1}^{obl}$ $\mathbf{PAT}_{2}^{obl}$

-example: impf: změna zákona se jich dotýká pf: zvýšení cen se nás citelně dotklo

3 ≈ impf: **urážet; ubližovat** pf: **urazit; ublížit**

-frame: $\mathbf{ACT}_{1,inf,že,cont}^{obl}$ $\mathbf{PAT}_{2}^{obl}$ MEANS_{7}^{typ}

-example: impf: ta poznámka se maminky bolestně dotýkala; dotýkalo se jí slyšet ta slova; dotýkalo se jí, že se chová tak chladně pf: dotkla se maminky svou poznámkou; dotklo se jí slyšet, jak ji za zády pomlouvá; dotklo se jí, že na schůzku nepřišel

-control: PAT

-rcp: ACT-PAT: pf: navzájem se dotkli svými poznámkami

-class: psych verb

4 ≈ impf: **zmiňovat se** pf: **zmínit se** (idiom)

-frame: $\mathbf{ACT}_{1}^{obl}$ $\mathbf{PAT}_{2}^{obl}$

-example: impf: dotýkala se sporné otázky pf: dotkla se sporné otázky

Figure 5.1: Example of a Vallex entry

The entry for a lexeme *dotknout se* consists of 4 lexical units. Let us examine the first lexical unit (the first sense of a lexeme). This verb takes three complements: Actor, Patient and Means. Actor is expressed in the Nominative case (1), Patient is expressed as a direct object in the Genitive case (2), Means is an indirect object in the Instrumental case (7).

As we have mentioned, we are primarily interested in a surface valency, so we 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 5.3), 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.

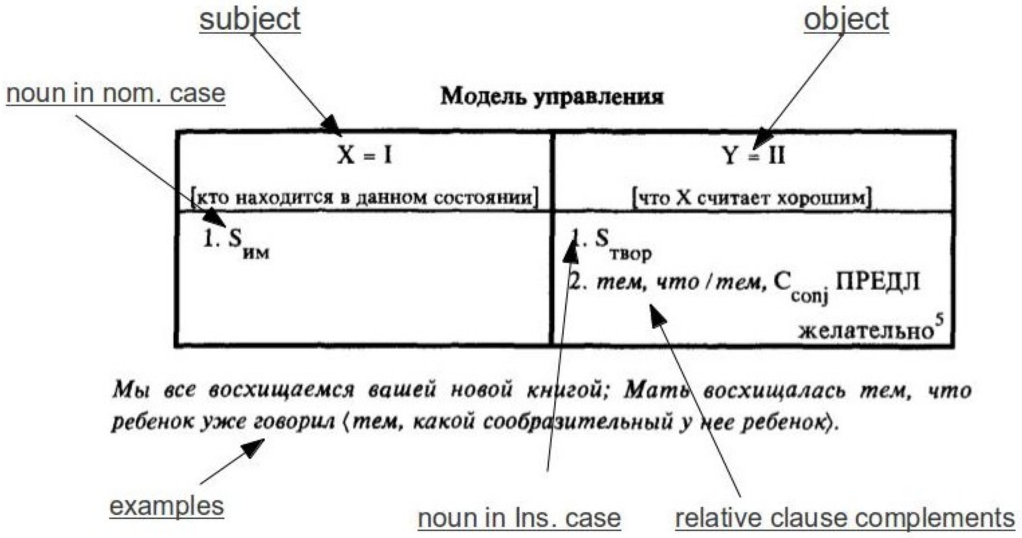
As in the case of Vallex, we will use only morphemic forms of the complements.

Explanatory combinatorial dictionary

The information on valency for Russian verbs is included into the Explanatory combinatorial dictionary (further, TKS - Tolkovo-Kombinatornyj Slovar') (Melčuk and 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 information on valency we are interested in is represented in the form of a **government pattern**.[8] Unlike in Vallex, the information on deep semantic roles of complements is not included, only a shallow syntactic function – **X** – subject and **Y** – object. Figure 5.2 depicts a government pattern of the verb *восхищаться* – 'to admire'.

A modern successor of the TKS dictionary, the *Active dictionary of Russian language* (Apresjan, 2011), which also includes the information on the government pattern of a verb, is still under development. We could have used both of the resources in our comparative analysis and tried to combine them into a bilingual valency resource, but, unfortunately, neither TKS, nor the Active dictionary are freely available online, only in a form of a book. It would have been an interesting idea to combine TKS and Vallex lexicons, but they are quite different resources. Vallex contains information on functors whereas the TKS includes only the syntactic functions of complements - subject and object.

[8] In Russian, *модель управления.*



Модель управления

X = I [кто находится в данном состоянии]	Y = II [что X считает хорошим]
1. $S_{им}$	1. $S_{твор}$ 2. *тем, что* / *тем,* C_{conj} ПРЕДЛ желательно[5]

Мы все восхищаемся вашей новой книгой; Мать восхищалась тем, что ребенок уже говорил ⟨*тем, какой сообразительный у нее ребенок*⟩.

Figure 5.2: Government pattern of a verb *восхищаться* – 'admire' in the TKS dictionary

5.2 Valency information extracted from the Ruslan lexicon

In the first stage of comparison of Czech and Russian valency we exploit the MT dictionary Ruslan, preliminarily described in Section 2.3.1.[9] Within the Ruslan MT system, verbs in the lexicon were assigned 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 5.3 demonstrates an entry from Ruslan dictionary for the verb *vystačit* – 'to be enough', the explanatory notes are given further:

(5.3) VYSTAC3==R(5,PRP,?(N(D),S(I,G)),39,CHVATIT6):

- *VYSTAC3* presents a a stem of the verb *vystačit* – 'be enough',

[9] This section contains selections from the paper (Klyueva and 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.

- *R* denotes a root of a tree,
- *5* is a symbol for a verb and PRP is a conjugation pattern of the Czech verb,
- *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 the Ruslan entry was written for Q-systems and is not easily comprehensible, we transformed the Ruslan entries into a more user-friendly format. First, we lowercased the entries and transfered Ruslan encoding of letters with diacritics (coded in numbers) into common letters and transformed the Cyrillic letters for a Russian translation. Then we selected the verbs and substituted the verb stem and the morphological information coded in special symbols with an appropriate verb ending. We did not use the semantic feature[10] 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 the transformed entry from the Example 5.3:

(5.4) 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 a different way – with letters in Ruslan dictionary or numbers in Vallex. In order to make the examples more comprehensible, we rewrite each example into a form with a contracted name of the case,[11] Example 5.2 will be therefore depicted as:

(5.5) Nom + vystačit + s + Ins -> Dat хватить + Gen- *to be enough*

[10] Examples containing semantic features in Ruslan dictionary were given in Section 2.3.1.

[11] Technically, the data that we use will have the original format.

5.2.1 The comparison of valency frames

Out of the 2080 verbal dictionary entries from Ruslan we have analyzed 1856 unique verbs.[12] We examined 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 brevity, we will call 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 the 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 a frame complement without a preposition both in Czech and Russian, ex.:

(5.6) 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[13] makes the comparison easier and it apparently also influences the number of identical frames. As we have already mentioned, because for the majority of verbs the Actor is in 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 5.3) in comparison with prepositional cases. Some examples:

[12] 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.

[13] Vocative case is not used in modern Russian unlike in Czech, and it is not relevant for our study of verb complements.

(5.7)

(1) povšimnout si + Gen -> заметить + Acc – *to notice*

(2) vyhýbat se + Dat -> избегать + Gen – *to avoid*

Table 5.1 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 analyzed verbs:						1856 (100%)

Table 5.1: Co-occurrence of the same cases in Czech and Russian based on Ruslan dictionary

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 the case when prepositions are translated straightforwardly or typically from Czech into Russian according to the dictionary default translation[14]. For example, the surface form with prepositions

[14] 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) has the same identical 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. However, there are cases when translation of the default forms does not match the surface forms, e.g. the Czech preposition *do* – 'to' corresponds to Russian *в* – 'to'.

Following is an example of a verb with identical prepositional complements:

(5.8) 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'.

(5.9) narazit na + Acc -> столкнуться с + Ins – *to rush into*

We sorted the list of 'preposition plus case' pairs from the Ruslan dictionary. Table 5.2 represents the top of the list with the frequencies of how often this frame occurred in Ruslan dictionary,[15] different prepositional cases are in bold.

5.2.2 Lexicon and a list of differences

The main result 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 Examples 4.18 or 4.19.

Table 5.3 shows statistics of those verbs with regard to our classification of simple and prepositional case frames.

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 is rather limited. In further experiments, we will provide a surface valency analysis for the two languages exploiting more large-scale language resources.

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))

[15] 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 5.2: Prepositional case correspondence – Ruslan dictionary

5.2.3 Exploiting valency information from Ruslan dictionary in machine translation

We have also exploited the entries from the Ruslan lexicon within the TectoMT (Section 3.1) system 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[16]:

(5.10) **narazit n:na+4 => столкнуться n:с+7** – *to run into smb*

[16] The notion 'formeme' was introduced in the Section 3.1.2.

Type of difference	Number of verbs	Percentage
Simple case	68	3.6%
Prepositional case	104	5.6%
Totally verbs with differences	172	9.2%
Total number of verbs	1856	100%

Table 5.3: Types of mismatches in surface valency frames

The list was incorporated into a system in the form of a block – FixValency.pm.[17] We evaluated the performance on the same test set that was used for linguistic evaluation - WMT 2013 test set (3000 sentences). We measured the BLEU score and manually checked the differences in the two outputs - before and after the new block was introduced. After implementing this block, some sentences with troublemaking verbs (verbs with different surface valency) were translated with a 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)[18] in the improved system (2TMT).

(5.11)

(SRC)	*využívali*	*obrovských*	*amerických*	*zakázek*
	used-3Pl	huge-**Gen**	american-**Gen**	contracts-**Gen**

' they made use of huge American contracts'

(1TMT)	*они*	*использовали*	*огромных*	*американских*	*заказов*
	*they	used	huge-**Gen**	american-**Gen**	contracts-**Gen**

' they made use of huge American contracts'

(2TMT)	*они*	*использовали*	*огромные*	*американские*	*заказы*
	they	used	huge-**Acc**	american-**Acc**	contracts-**Acc**

[17] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/FixValency.pm`.

[18] využívat n:2 => использовать n:4 in the block FixValency.pm

‘ they made use of huge American contracts’

However, there were cases when this rule worsened the translation. In Example 5.12, the prepositional complement was translated properly by (1TMT) because a rule for the preposition transfer from another module[19] 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 the Ruslan lexicon.[20]

(5.12)

(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 5.13, 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.

(5.13)

(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’

[19] `https://github.com/ufal/treex/blob/master/lib/Treex/Block/T2T/CS2RU/RuleBasedFormemes.pm`

[20] 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.

(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.

Manual evaluation

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 5.11
- worsened the translation like in Example 5.12
- did not have any effect as both variants were incorrect – Example 5.13

Effect	number of differences	Percentage
improved	28	58.3 %
worsened	3	6.2%
no effect	17	35.4%
Total	48	100%

Table 5.4: Manual evaluation of changes after adding FixValency.pm

From the table we can see that in the majority of cases the verbal valency is improved, or it has 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 the automatic evaluation metric BLEU.

5.2.4 Discussion on Ruslan dictionary

In this section we worked with and extracted surface valency information from the linguistic resource that was created more than 20 years ago. This information contributed to identification of discrepancies in the 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.

5.3 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) for the task of automatic extraction of surface valency frames.[21]

Some experiments on automatic valency extraction related to our languages can be found for example in (Bojar and Šindlerová, 2010), (Zeman and Sarkar, 2000), (Pala and Ševeček, 1997). The authors rely upon different methods, formats and language resources, and 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 at extracting surface valency 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, especially, the morphologically rich languages with free word order like Czech and Russian. In free word order languages it is impossible to rely on the order of individual complements of a verb and thus their identification constitutes a complex problem.

5.3.1 Setup of the Experiment

We aim at using the simplest possible means in our experiment. It is 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.

[21] Some passages in the section come from the related paper (Klyueva and Kuboň, 2014), where the author of the thesis conducted and described all the experiments.

Our experiment consists of the following stages:

- adapting valency frames
- extracting information on a Czech 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.[22]

Processing Vallex Frames

We are exploiting only surface realization of verb complements, typically having either 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[23] format: a verb plus a case of an argument without a functor. Following is an example of a Vallex frame slot in the original format representing the 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" />
  <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:[24] **vyžadovat+2**, **vyžadovat+4** (to demand + Genitive, to demand + Accusative), the information on subordinate and infinitive clauses was ignored.

[22] The script implementing the algorithm can be found here: `https://github.com/natalink/CzeRuValency/blob/master/valency.pl`.

[23] See Section 3.1.2 for the definition of the notion 'formeme'.

[24] Here we present cases as numbers because of the format of data involved in the experiment.

For prepositional valency, the format is **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[25] 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 2.2.1 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 Section 2.2.3. 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 find a verb with a surface valency frame matching the one from 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.

The bilingual dictionary then provides translations of the lemmas 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)[26] 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:

[25] `http://www.langsoft.cz/`

[26] We account always only for one complement per cycle.

(5.14) 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:

(5.15) 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 case. The dependent noun should be either in the Genitive or in the Accusative case, so we search[27] for a noun or a pronoun in the Genitive or the Accusative case. The 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 of 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 to this clause therefore provides a frame hypothesis:

(5.16) (cz)*vyžadovat+Acc* => (ru)*требовать+Gen*

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, the only necessary thing is 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 5.16) established in the preceding phases for a particular Russian verb and choose **the most frequent** Russian valency frame from this set.

[27] Here we will suppress some details such as optimization of search range (5 words around the verb) and a restriction within a clause.

The main statistics concerning the total number of patterns identified in the corpus and the number of the extracted patterns are presented in Table 5.5. 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 5.5: Statistics of the experiment

The last line in Table 5.5 shows only the identified patterns, it does not reflect whether these patterns are correct or not. The errors we have discovered are discussed below.

The second interesting observation was made while we compared the obtained results with those from the manually created Ruslan dictionary. 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 5.6. 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 5.1). A frequent co-occurrence cz:Acc vs. ru:Gen (196 times) reflects the most error-prone frame of our algorithm which we will discuss later in more detail.

Prepositional frames

The results for prepositional valency are presented in Table 5.7, Russian prepositions are transliterated. Due to a large number of very rare (and thus unreliable) correspondences we included only those which occurred more than 10 times into the table. The top of the table is quite similar to the one from manually created Ruslan, see Table 5.3 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 the most frequent pairs from the table just for the sake of completeness. As in the case with Ruslan, we consider a preposition in Czech an equivalent of the Russian one if it is a typical translation of each other (see Section 5.2.1 for examples) and the cases of the complements are the same.

		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 5.6: Co-occurrence of the same simple case in Czech and Russian

Out of the total of 841 prepositional pairs from Table 5.7 there were 154 different unique pairs. This represents 18.3% of the total. However, the 'comparison' from Tables 5.6 and 5.7 should be taken with caution because it is made on automatically extracted and highly erroneous data.

5.3.2 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 of the set of 4,286 extracted frames (Table 5.5) 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 the following pattern:

  Czech: Verb+Acc => Russian: Verb+Gen.

  This error pattern has its roots in the tagger inaccuracy due to the morphological ambiguity. In Russian, a masculine animate noun has the same form

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 5.7: Prepositional case correspondence

in the Genitive and the Accusative cases[28], 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:

(5.17) 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 animate and is often wrongly

[28] *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 the frame comes from the manually written Vallex.

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. 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:

  (5.18) *odebrat*+Acc (take smth.) => *отобратъ*+y+Gen (take from smb.)

  In this example, the Czech morphemic case Accusative is the surface realization of the functor Patient (PAT), whereas the Russian surface form *y+Gen* is the realization of the functor Experiencer (EXP). On 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 the lack of syntactic or semantic parsers, those two roles were confused.

  Let us look at this case more closely and examine the sentence from our corpus containing this example:[29]

  (5.19)

(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'			

  Although our algorithm 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:

  (5.20) *odebrat*+Dat => *отобратъ*+y+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.

[29] In order to simplify the text we leave only the relevant morphological tags.

Comparison with Ruslan frames

It would be a natural thing to compare the generated pairs [verb+frame] with the 'golden' data from Ruslan. We cannot call it *evaluation against 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 5.8: Examples of comparing 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 large 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 a 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.

5.4 Surface frame discrepancies and verb classes in Vallex

This section includes theoretical observations on discrepancies in surface valency frames. According to the Ruslan dictionary, about 9% of verbs in Czech and

Russian exhibit surface valency discrepancies. The question arises which verbs tend to have different valency frames 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 a Czech verb and its translation counterpart in Russian are different, there is a high probability 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 the semantic class of a verb. Vallex distinguishes 22 verb classes, such as verbs of communication, exchange, motion, perception, transport etc. Naturally, words that belong to the same semantic field or share the same component of meaning tend to have similar valency frames. Unlike the previous two experiments, here we will also make use of functors (semantic roles in the FGD nomenclature) since the surface discrepancies are often connected to certain functors.

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. This can be a consequence of a language contact (influence of German on Czech surface valency is mentioned, e.g., in (Berger, 2008)). Further, we will indicate when Czech and German frame slots match while Czech and Russian do not.

5.4.1 Frame comparison

We made a comparison of Czech and Russian frames with respect to verb classes in the following way: given a Czech verb and its semantic class, we check if its surface valency frame (for nominal complements only) fits the frame[30] of a Russian meaning-equivalent verb. We only examined such pairs that have a single "typical" translation equivalent in Russian whose valency includes the same semantic roles (functors).

In the following, we consider the Russian surface frame to be different from the Czech one if there is a frame slot for which the Czech and Russian surface realizations are different. If a surface form is represented by a preposition requiring a certain case, we consider the default translation of prepositions (see the discussion in the Section 5.2) as the equivalent realization. For example, the

[30] As in the two previous sections, we will refer to the 'surface frame of nominal complements' just as **frame** for shortness.

Czech frame element *PAT(před+Ins)* – 'in front of' corresponds to the Russian *PAT(перед+Ins)* – 'in front of' and is equal with respect to surface valency. Another sense of the same Czech prepositional phrase *PAT(před+Ins)* – 'from' has a different surface form in Russian: *PAT(от+Gen)* – 'from'.

Due to time reasons, it was impossible to evaluate all verb frames (in total 2,903 lexical units have a verb class assigned), so we have only examined selected verb classes: motion, communication, change, exchange, and mental action.

Example 5.21 shows a translation pair of a Czech and a Russian verb with an identical valency frame:

(5.21)

(cz) *obhajovat ACT(Nom) PAT(Acc)* – 'to defend'

(ru) *защищать ACT(Nom) PAT(Acc)* – 'to defend' – **matches**

The verb pair in Example 5.22 shows two differences:

(5.22)

(cz) *blahopřát ACT(Nom) PAT(Dat) CAUS(k+3)* – 'to congratulate'[31]

(ru) *поздравлять ACT(Nom) PAT(Acc) CAUS(с+7)* – 'to congratulate' – **does not match**

Here, different cases and different prepositions are used to express the semantic roles of Patient and Cause, respectively.

During the analysis of the individual classes, it became evident that differences in valency frames can be either regular or occasional within the given class. Next, we will present a description of surface frame differences in the selected semantic classes. The overall statistics will be then give in Table 5.9.

5.4.2 Verbs of Change

Verbs of the Change class often have the Difference 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 typical Czech realization of the DIFF slot, *o+Acc* – 'by', generally corresponds to *na+Acc* – 'on/to' in Russian (other variants are possible), see Examples 5.23 and 5.24.

(5.23)

[31] The German frame: *gratulieren PAT(zu+Dat)* matches the Czech one.

(cz) *ceny klesly **o** 20%* ‘prices fell **by** 20% (DIFF)’[32]

(ru) *цены упали **на** 20%* ‘prices fell **on** 20% (DIFF)’

(5.24)

(cz-o) *Administrace zkrátila dovolenou **o** 2 dny*[33]
‘administration shortened the holiday **by** 2 days (DIFF)’

(cz-na) *Administrace zkrátila dovolenou **na** 2 dny*
‘administration cut off the holiday **to** 2 days (EFF)’

(ru) *Администрация сократила отпуск **на** 2 дня*
‘administration cut off the holiday **on**/**to** 2 days (DIFF/EFF)’

In Example 5.24, we can spot the ambiguity of Russian surface form *на+Acc* for verbs of change. The sentence 5.24(ru) can be interpreted both as (cz-o) and (cz-na): that the administration shortened holidays by two days or to two days.

The surface form *o+Acc* as an expression of a difference is typical in Czech while Russian uses the preposition *o* – ‘about’ mainly with mental predicates (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 especially problematic in rule-based MT systems as the preposition is translated into Russian as *о* – ‘about’ by default. For the TectoMT system, some of the cases were covered by verbs from the converted Ruslan lexicon (Section 5.2.3), 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 shallow semantic parser. If the verb is identified as belonging to this class and the complement realization of DIFF is *o+Acc*, the surface form of the complement in Russian can be set to *на+Acc*.

5.4.3 Verbs of Motion

We have not found many dissimilarities in Czech and Russian frames within the class of Motion verbs. The most apparent one is that verbs with the semantic component ‘flee from something’ in Czech have the surface realization of PAT as *před+Ins* – ‘before, from’, but are translated into Russian with the respective verb plus the prepositional phrase *от+Gen* – ‘from’. Just to name some of these

[32] The German surface frame: *sinken DIFF(um+Acc)* matches the Czech one.

[33] The German surface frame: *kürzen DIFF(um+Acc)* matches the Czech one.

verbs: *prchat* – 'be on the run', *ujíždět* – 'speed off', *unikat* – 'escape'.

(5.25)

(cz) *prchat před policií* – 'run before the police'[34]

(ru) *убегать от полиции* – 'run from the police'

Roughly speaking, Russian prefers the preposition *от* – 'from' whereas Czech uses *před* – 'before' in this context. Verbs of other semantic classes with a similar component of meaning, as, e.g. the Location class, share this rule as well (e.g., Czech *schovat před+Ins* – 'hide before' vs. Russian *спрятать от+Gen* – 'hide from').

Other frame differences in this verb class seem to be rather coincidental as illustrated in the following example:

(5.26)

(cz) *trefit PAT(**Acc**)* – 'hit something'[35]

(ru) *попасть PAT(**в+Acc**)* – 'hit into something'

5.4.4 Verbs of Exchange

One of the regular and rather evident mismatches in Czech and Russian verbs of exchange with respect to surface frames was discussed in (Lopatková and Panevová, 2004). Several Czech verbs of exchange with the meaning of removing something from someone, e.g., *sebrat* – 'take away', *krást* – 'steal', *brát* – 'take' etc. exhibit a regular difference in surface frames in contrast to the equivalent Russian verbs. The Addresse (ADDR) functor denotes here a person or an object from whom something is taken. In Czech, it is realized using a simple case (Dative) on the surface,[36] whereas Russian uses a preposition with the Genitive case (*у+Gen* – 'from'):

(5.27)

[34] Only prepositions in the Czech and German frames match, but not the morphological case: *fliehen PAT(vor + Dat).* Just to note, there is no Instrumental case in German.

[35] The Czech frame of the verb *trefit+PAT(Acc)* – 'hit' corresponds to the German one *treffen PAT(Acc).*

[36] The same holds in German, e.g., *nehmen ADDR(Dat)* – 'take from smb'.

(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'

The Czech sentence in Example 5.27 is ambiguous since the Dative noun can either be interpreted as the Addressee or as the Benefactor (BEN), as in *take a toy for a baby.* BEN is the only possible interpretation of Dative with the verb *брать* in Russian, which can lead to translation errors if the Czech surface frame is left unchanged in Russian.

The same is true for a metaphorical usage of verbs of Exchange, e.g. the verb *zabírat ADDR(**Dat**)* – 'take(time)+Dat':

(5.28)

(cz) *studium mi zabírá hodně času*
study me.**Dat** takes lot time
'Study takes me a lot of time.'

(ru) *учеба отнимает у меня много времени*
study takes **from** me.**Gen** lot time
'Study takes me a lot of time.'

5.4.5 Verbs of Communication

There are many differences regarding surface frames between Czech verbs of this class and their Russian counterparts. Here we could not observe a single leading difference as in the previous classes. The mismatches 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. The Addressee (ADDR) with the surface form *na+Acc* – 'on' in Czech is presented differently in Russian, using a different preposition.

   (5.29)

   (cz) *mluvil na bratra*
   spoke **on** brother.**Acc**
   'He spoke to his brother'

(ru) *он говорил с братом*
He spoke with brother.**Ins**
'He spoke to his brother'

While the Russian preposition can also be used in Czech, albeit with a very slight change of meaning – *Mluvil s bratrem.* – 'He spoke with his brother.', it is not possible to transfer the Czech preposition *na+Acc* directly into Russian.

A similar case is the Czech verb *zavolat na + Acc* – 'to call on someone'. Russian allows only the surface realization of the Addresse as a plain Accusative: *позвать+Acc* – 'to call someone', while in Czech, both the Accusative and *na+Acc* are possible, with a minor difference in meaning.

2. The Patient (PAT) with the surface form *na+Loc* – 'on, upon' or *na+Acc* – 'on' in Czech may have another surface realization in Russian.

   Generally the morphemic form is *o+Loc* – 'about' for verbs of asking, such as *ptát se* – 'ask', *tázat se* – 'ask' with *na+Acc*:

   (5.30)

   (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**'

   The same surface frame is used in Russian in counterparts of other Czech verbs of speaking with the frame slot *na+Loc*, e.g., Czech *domlouvat se na + Loc* – 'agree on' vs. Russian *договориться о + Loc* – 'agree on/about'. In Czech, the surface realization *o+Loc* – 'about' is also possible.

3. Addressee (ADDR) in the Dative case for the following Czech verbs corresponds to Accusative in Russian:

   (5.31)

   (cz) *poblahopřát* ADDR(**Dat**) 'congratulate'[37]

   (ru) *поздравить* ADDR(**Acc**) 'congratulate'

   (5.32)

   (cz) *děkovat* ADDR(**Dat**) 'thank'[38]

[37] Compare with German *gratulieren ADDR(Dat)*

[38] German frame slot: *danken ADDR(Dat)*

(ru) *благодарить* ADDR(**Acc**) 'thank'

4. Patient (PAT) with the surface form *o+Acc* – 'about': Similar to the Change class (Example 5.23), some complements of the Czech communication verbs with this surface form have another surface realization in Russian due to the fact that the corresponding Russian preposition *o* does not combine with Accusative at all:

   (5.33)

   (cz) *hlásí se* *o* *slovo*
   asks **about** word.**Acc**
   'She asks for the word'

   (ru) *она* *просит* *слова*
   She asks word.**Gen**
   'She asks for the word'

5. There are several coincidental differences occurring only once or twice that do not fit any scheme, e.g.:

   (5.34)

   (cz) *doznávat se* PAT(**k+Dat**) 'confess to smth'[39]

   (ru) *признаваться* PAT(**в+Loc**) 'to confess in smth'

5.4.6 Class of Mental Action

The equivalent verbs of this class often show differences in surface frames, but these are rather coincidental; we were not able to identify any regular patterns.

Discrepancy occurs as a rule between Czech verbs requiring the Patient (PAT) in the surface form *na+Acc* – 'on'. This is similar to the verbs of the Communication class (see Example 5.30), where the Patient is regularly translated as *o+Loc* – 'about'. However, there is no common translation equivalent in the Mental Action class, and *na+Acc* corresponds to several surface forms in Russian:

(5.35)

(cz) *pamatovat* PAT(**na+Acc**) 'remember'

(ru) *помнить* PAT(**про+Acc**) 'remember'

(5.36)

[39] The German frame is equivalent to the Czech one: *sich bekennen PAT(zu+Dat)*.

(cz) *myslet* PAT(**na+Acc**) ‘think about’[40]

(ru) *думать* PAT(**о+Loc/про+Acc**) ‘think about’

(5.37)

(cz) *zvykat si* PAT(**na+Acc**) ‘get used to’[41]

(ru) *привыкать* PAT(**k+Dat**) ‘get used to’

The situation of the following verb pair is very similar to the one from Examples 5.31 and 5.32, though the functor here is PAT, not ADDR:

(5.38)

(cz) *rozumět* PAT(**Dat**) ‘understand’

(ru) *понимать* PAT(**Acc**) ‘understand’

There are many further differences that seem rather coincidental, as in the following example:

(5.39)

(cz) *pohrdat* PAT(**Ins**) ‘despise’

(ru) *презирать* PAT(**Acc**) ‘despise’

5.4.7 Overall results on verb class differences

In this experiment, we have compared Czech and Russian surface frames of verbs from 5 semantic classes (1,473 lexical entries in total) and examined the connection between surface valency discrepancies and verb classes. Some of the identified discrepancies between surface verb frames in the two languages are regular (semantically related words exhibit the same discrepancy in valency), others are coincidental. We made the following observations:

- Most mismatches occur in prepositional phrases.
- Within a verb class, we can often find surface frame patterns of Czech verbs which regularly correspond to certain Russian patterns.

[40] The German equivalent *denken PAT(an + Acc)* corresponds to the Czech surface frame.

[41] The German equivalent *sich gewöhnen PAT(an + Acc)* corresponds to the Czech frame.

- Quite a few Czech surface frames different from Russian follow the same pattern as German. This may be attributed to language contact between Czech and German, though we did not study this issue in detail.[42]

Table 5.9 presents the distribution of verbs with different frames according to their semantic classes.

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 5.9: Frame differences according to the verb classes

From this table we can see that verbs of physical activity (change, motion, exchange) are less likely to show mismatches in surface frames than verbs of mental activity (communication, mental action).

5.5 Summary

In this chapter, we have first discussed the practical aspects of valency in translation and focused on the discrepancies in surface valency of Czech verbs and their Russian translation counterparts.

The main contribution of this research is the lexicon of Czech and Russian verbs with their surface frames[43] that was extracted and converted from the dictionary of the Ruslan MT system. We have then identified the discrepancies in Czech and Russian surface frames (Section 5.2.1).

We have incorporated this lexicon into the TectoMT Czech-Russian rule-based MT system (Section 5.2.3). While the addition of the lexicon caused a slight decrease of the BLEU automatic evaluation score, a manual evaluation of the changes brought by the lexicon showed that in almost 60% of sentences with errors in surface valency, the wrong form was corrected .

[42] It is mostly the vocabulary, not grammar, that is generally borrowed from another language, but here we can see that surface frames can be calques from another language.

[43] https://github.com/natalink/CzeRuValency/blob/master/python/lexicon

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 (Section 5.3). The manual evaluation of a small sample revealed that the results are mostly correct (80% of generated frames were correct). To increase the precision of our approach, more detailed information would be needed, such as syntactic analysis of the sentences or fine-grained word alignment.

Finally, we presented a detailed linguistic observation on which semantic classes of verbs tend to show discrepancies in the surface form of their arguments (Section 5.4). We have explored five selected semantic classes of Czech verbs and their Russian translation counterparts, and we confirmed our initial hypothesis that in certain classes, Czech-Russian pair of verbs with a similar meaning tend to have the same 'discrepancy pattern' in surface valency. In addition, the total number of verb pairs where discordant surface frames were found (8%) is approximately the same as the number of such calculated on the Ruslan lexicon.

6 Conclusion

The results achieved in our research are theoretical and practical. One of the main outputs of our research is data that we collected/created for the purpose of MT between Czech and Russia, which can be exploited in other experiments. The research contributes to comparative linguistics by analyzing error types in Czech-Russian machine translation output and linking these errors to specific differences between Czech and Russian.

Data

The following list presents the data that we gathered for our machine translation experiments:

- **Czech-Russian parallel corpus** We automatically downloaded a parallel corpus UMC that was subsequently tagged. It served mainly as a part of training data for SMT. The corpus was used for extracting the Czech-Russian dictionary and surface valency frames. Also, some linguistic phenomena like usage of pronouns, transgressives were explored using this parallel corpus.
- **The Czech-Russian dictionary** was automatically extracted from the parallel corpus. To the best of our knowledge, there is no other freely available Czech-Russian dictionary. The dictionary was used for implementing the experimental Czech-to-Russian machine translation within the RBMT systems TectoMT and Česílko.
- **A list of surface valency frames** was extracted from the dictionary Ruslan. This list was therefore used in the comparative analysis of Czech and Russian surface valency frames.

MT systems

The main objective of this work was to build 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 system Moses for Czech-Russian. Also, we compared their performance with 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 better than RBMT (TectoMT and PC Translator).

The manual evaluation of errors in the output of the four systems suggested that BLEU correlates with the manual evaluation as there were many more words marked as errors in the RBMT output than in those of SMT.

One of the initial hypotheses – that MT between related languages under the similar settings could bring better results than when constructing an MT between unrelated languages – turned out to be false. The morphological complexity of the two languages is detrimental and their relatedness does not help.

Differences between Czech and Russian

Our work contributes to comparative studies of the two languages by exploring the MT output and relating the errors to specific language discrepancies. We proposed a classification of errors relevant for the language pair under consideration. It was concluded that discrepancies between the two languages have more impact on RBMT systems than on SMT ones.

We focused especially on the problem of surface valency of verbs in Czech and Russian. Using several valency resources, we calculated that the number of mismatches in surface valency between the two languages is about 10%. We examined the verbs that show this difference and found out that there are more discrepancies in the class of verbs of mental activity than in the class of physical activity. Another observation was that quite often a Czech surface valency frame different from the respective one in Russian was similar to the one in German.

RBMT and SMT strategies 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 resemble learners who started to learn a language exploiting traditional methods 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 considerable amount of efforts and time, which is true both for RBMT and for a typical second-language learner. The learners and the 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 or add a new word into the dictionary.

SMT systems can thus be 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 (like n-grams), but the statistical ‘approach’ to language learning seems to be more efficient than the rule-based. It allows us to acquire a language in a short amount of time without implicit knowledge of abstract rules.

Statistical and rule-based approaches often benefit from borrowing some features from each other (Hybrid systems), and this may be true for the language learners as well.

Our future work may be pursued in the direction of finding common features between language acquisition and machine translation.

List of Figures

List of Tables

A

Appendix – DVD content

The enclosed DVD contains the following files:

- **Czech-Russian-factored** Parallel Czech-Russian Corpus UMC with morphological annotation: `http://hdl.handle.net/11858/00-097C-0000-0001-4909-7`;
- **cz-ru-dictionary** Czech-Russian dictionary, 17,122 entries, cleaned;
- **lexicon-ruslan** Czech-Russian surface frames extracted from Ruslan;
- **lexicon-autogen** automatically extracted Czech-Russian surface frames.

Bibliography

Anastasiou, D. (2010). *Idiom Treatment Experiments in Machine Translation.* Cambridge Scholars Publishing.

Apresjan, V. (2011). Active Dictionary of the Russian Language: Theory and Practice. In *Proceedings of the 5th International Conference on Meaning-Text Theory. Barcelona.*

Babby, L. H. (1980). *Existential Sentences and Negation in Russian.* Ann Arbor:Caroma Publishers.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baobao, C., Danielsson, P., and Teubert, W. (2002). Extraction of translation unit from Chinese-English parallel corpora. In *Proceedings of the first SIGHAN workshop on Chinese language processing.*

Bejček, E. and Straňák, P. (2010). Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Benešová, V., Lopatková, M., and Hrstková, K. (2008). 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*, pages 18–25, Hong Kong, China. City University of Hong Kong.

Berger, A. L., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Giuett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., and Urei, L. (1994). The Candide system for machine translation. In *Proceedings of the ARPA Conference on Human Language Technology*, pages 157–162.

Berger, T. (2008). Deutsche Einflüsse auf das grammatische System des Tschechischen. In *Studien zur historischen Grammatik des Tschechischen*, Bohemistische Beiträge zur Kontaktlinguistik, pages 57–69. München.

Boguslavsky, I. (1995). A bi-directional Russian-to-English machine translation system (ETAP-3). In *Proceedings of the Machine Translation Summit V. Luxembourg.*

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., and Frid, N. (2000). Dependency treebank for Russian: Concept, tools, types of information. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 987–991. Association for Computational Linguistics Morristown, NJ, USA.

Bojar, O. (2011). Analyzing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 95:63–76.

Bojar, O. (2012). *Čeština a strojový překlad*, volume 11 of *Studies in Computational and Theoretical Linguistics.* ÚFAL, Praha, Czechia.

Bojar, O. and Hajič, J. (2008). Phrase-based and deep syntactic English-to-Czech statistical machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 143–146, Columbus, Ohio. Association for Computational Linguistics.

Bojar, O., Homola, P., and Kuboň, V. (2005). An MT System Recycled. In *Proceedings of MT Summit X*, pages 380–387.

Bojar, O. and Prokopová, M. (2006). Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 1236–1239.

Bojar, O., Rosa, R., and Tamchyna, A. (2013). Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, pages 92–98, Sofija, Bulgaria. Bălgarska akademija na naukite, Association for Computational Linguistics.

Bojar, O. and Šindlerová, J. (2010). Building a bilingual vallex using treebank token alignment: First observations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 304–309, Valletta, Malta. ELRA, European Language Resources Association.

Bojar, O. and Tamchyna, A. (2011). Forms Wanted: Training SMT on Monolingual Data. In *Proceedings Research Workshop of the Israel Science Foundation University of Haifa, Israel.*

Bojar, O. and Tamchyna, A. (2013). The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Bílek, K. (2014). A Comparison of Methods of Czech-to-Russian Machine Translation. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Bílek, K., Klyueva, N., and Kuboň, V. (2013). Exploiting Maching Learning for Automatic Semantic Feature Assignment. In Boonthum-Denecke, C. and Youngblood, M., editors, *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2013*, pages 297–302, Palo Alto, California. FLAIRS, AAAI Press.

Čermák, F. and Rosen, A. (2012). The case of intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

Chandioux, J. (1988). Meteo: An operational translation system. In *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications), March 21-25, 1988.*, pages 829–839.

Clancy, S. J. (2010). The Chain of Being and Having in Slavic. *Studies in Language Companion Series 122.*

Čmejrek, M., Cuřín, J., and Havelka, J. (2003). Czech-english dependency-based machine translation. In Copestake, A. and Hajič, J., editors, *EACL 2003 Proceedings of the Conference*, pages 83–90, Budapest, Hungary. Association for Computational Linguistics.

Colmerauer, A. (1970). *Les Systemes-Q: Ou Un Formalisme Pour Analyser & Synthesiser Des Phrases Sur Ordinateur.* Univ.Montreal.

Doddington, G. (2002). 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, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dušek, O., Žabokrtský, Z., Popel, M., Majliš, M., Novák, M., and Mareček, D. (2012). Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada. Association for Computational Linguistics.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Felipe Sánchez-Martínez, G. R.-S., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144. Special Issue: Free/Open-Source Machine Translation.

Galuščáková, P., Popel, M., and Bojar, O. (2013). Phrasefix: Statistical post-editing of tectoMT. In *Proceedings of the Eight Workshop on Statistical Machine Translation*, pages 141–147, Sofija, Bulgaria. Bălgarska akademija na naukite, Association for Computational Linguistics.

Galuščáková, P. and Bojar, O. (2012). 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*, volume 247 of *Frontiers in AI and Applications*, pages 58–65, Amsterdam, Netherlands. IOS Press.

Gispert, A., Mariño, J. B., and Crego, J. M. (2005). Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Eurospeech 2005, Lisbon, Portugal*, pages 3185–3188.

Hajic, J. (1987). 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, pages 113–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hajič, J. (2001). *Disambiguation of Rich Inflection - Computational Morphology of Czech*, volume I. Prague Karolinum, Charles University Press. 334 pp.

Hajič, J., Homola, P., and Kuboň, V. (2003). A Simple Multilingual Machine Translation System. In Hovy, E. and Macklovitch, E., editors, *Proceedings of Machine Translation Summit IX*, pages 157–164, New Orleans, USA.

Hajič, J., Hric, J., and Kuboň, V. (2000a). Machine Translation of Very Close Languages. In *Proceedings of the 6th Applied Natural Language Processing Conference.*

Hajič, J., Kuboň, V., and Hric, J. (2000b). Česílko - an MT system for closely related languages. In *ACL2000, Tutorial Abstracts and Demonstration Notes*, pages 7–8. ACL, ISBN 1-55860-730-7.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková Razímová, M. (2006). Prague Dependency Treebank 2.0. LDC2006T01, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, Jul 2006.

Hana, J. (2004). Czech clitics in Higher Order Grammar. In *Working Papers in Slavic Studies*. Department of Slavic and East European Languages and Literatures, Columbus, Ohio.

Hartley, A., Babych, B., and Sharoff, S. (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI, pp. 412–418, Copenhagen.*

Hausenblas, K. (1958). *Vývoj předmětového genitivu v češtině*. Studie a práce lingvistické III, Praha.

Hladná, V. (2012). Valence sloves v českých a ruských publicistických textech. Diplomová práce, Univerzita Palackého v Olomouci, Filozofická fakulta.

Hlaváčková, D. (2005). Verbalex – new comprehensive lexicon of verb valencies for czech. In *Proceedings of the Slovko Conference*.

Homola, P. (2009). *Syntactic Analysis in Machine Translation*, volume 6 of *Studies in Computational and Theoretical Linguistics*. Institute of Formal and Applied Linguistics, Praha, Czechia.

Hutchins, W. J. (1986). *Machine translation : past, present, future*. Ellis Horwood series in computers and their applications. E. Horwood New York Chichester Brisbane, Chichester.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *Language Resources and Evaluation*.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2007). A large-scale classification of english verbs. *Language Resources and Evaluation*.

Kirschner, Z. and Rosen, A. (1989). APAC - An experiment in machine translation. *Machine Translation*, 4(3):177–193.

Klyueva, N. (2013). Usage of some non-finite constructions in czech and russian. In *6th Annual International Conference on Languages and Linguistics*. Athens Institute for Education and Research, Atiner.

Klyueva, N. and Bojar, O. (2008). UMC 0.1: Czech-russian-english multilingual corpus. In *Proceedings of the Conference Korpusnaja lingvistika - 2008*, pages 188–195.

Klyueva, N. and Kuboň, V. (2014). Automatic valency derivation for related languages. In Boonthum-Denecke, C. and Eberle, W., editors, *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014*, pages 437–442, Palo Alto, California.

Klyueva, N. and Kuboň, V. (2010). 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.

Koehn, P. (2010a). An Experimental Management System. *Prague Bulletin of Mathematical Linguistics*, 94:87–96.

Koehn, P. (2010b). *Statistical Machine Translation*. Statistical Machine Translation. Cambridge University Press.

Koehn, P. (2011). What is a better translation? Reflections on six years of running evaluation campaigns. *Tralogy 2011*.

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *EMNLP-CoNLL*, pages 868–876. ACL.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions., Prague, Czech Republic*, pages 177–180.

Kolovratník, D., Klyueva, N., and Bojar, O. (2009). UMC003: Czech-english-russian tri-parallel test set for MT. Institute of Formal and Applied Linguistics.

Kuryłowicz, J. (1960). *Esquisses linguistiques*. Wrocław – Kraków, Polska Akademia Nauk.

Lavie, A. and Denkowski, M. J. (2009). The Meteor Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, 23(2-3):105–115.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago Press, University.

Lopatková, M. and Panevová, J. (2004). Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi). 5:348–356.

Mareček, D., Popel, M., and Žabokrtský, Z. (2010). Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–201, Uppsala, Sweden. Uppsala Universitet, Association for Computational Linguistics.

McDonald, R., Lerman, K., and Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *CONLL*, pages 216–220.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Melčuk, I. and Zholkovsky, A. (1984). Explanatory Combinatorial Dictionary of Modern Russian. Vienna: Wiener Slawistischer Almanach.

Mustajoki, A. and Heino, H. (1991). *Case selection for the direct object in Russian negative clauses*. University of Helsinki.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics 29*, pages 19–51.

Oliva, K. (1989). A Parser for Czech Implemented in Systems Q. Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague.

Oliveira, F., Wong, F., Li, Y., and Zheng, J. (2005). Unsupervised word sense disambiguation and rules extraction using non-aligned bilingual corpus. *Natural Language Processing and Knowledge Engineering.*

Pala, K. and Ševeček, P. (1997). Valence českých sloves. In *Sborník prací FFUB.*

Panevová, J. (1974). On verbal frames in Functional generative description. *Prague Bulletin of Mathematical Linguistics*, (22):3–40.

Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Partee, B. H. and Borschev, V. (2002). Existential Sentences, BE, and the Genitive of Negation in Russian. *Conference on Existence: Semantics and Syntax.*

Partee, B. H., Borschev, V., Paducheva, E., Testelets, Y., and Yanovich, I. (2011). Russian Genitive of Negation Alternations: The Role of Verb Semantics. *Scando-Slavica*, 57(2):135–159.

Popel, M. (2010). English-Czech Machine Translation Using TectoMT. In Šafránková, J. and Pavlů, J., editors, *WDS 2010 Proceedings of Contributed Papers*, pages 88–93, Praha, Czechia. Univerzita Karlova v Praze, Matfyzpress, Charles University.

Popel, M. and Žabokrtský, Z. (2009). Improving English-Czech Tectogrammatical MT. *The Prague Bulletin of Mathematical Linguistics*, (92):1–20.

Popovic, M. and Burchardt, A. (2011). From human to automatic error classification for machine translation output. In *15th International Conference of the European Association for Machine Translation (EAMT 11)*. European Association for Machine Translation.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 271–279, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raab, J. (2007). *Morče - Czech morphological tagger.* ÚFAL MFF UK, Prague, Czech Rep.

Rosa, R. (2013). Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

Sgall, P., Hajicová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Springer.

Skwarska, K. (2002). Záporový genitiv v současné češtině, ruštině, polštině a slovinštině. In *IX. Zborník materiálov z IX. kolokvia mladých jazykovedcov.*

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Spoustová, D., Hajič, J., Raab, J., and Spousta, M. (2009). 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*, pages 763–771, Athina, Greece. Association for Computational Linguistics.

Tan, L. and Pal, S. (2014). Manawi: Using multi-word expressions and named entities to improve machine translation. *ACL 2014*, page 201.

Tesnière, L. (1959). *Eléments de syntaxe structurale.* Klincksieck Paris.

Thurmair, G. (2004). Comparing rule-based and statistical MT output. In *Proceedings of Workshop on the amazing utility of parallel and comparable corpora.*

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *European Conf. on Speech Communication and Technology*, pages 2667–2670.

Trask, R. L. (1944). *Language and linguistics : the key concepts.* Abingdon England ; New York: Routledge, 2nd ed. edition.

Tufis, D. (2002). A cheap and fast way to build useful translation lexicons. In *In Proceedings of the 19th international Conference on Computational Linguistics - Volume 1*, pages 1236–1239.

Turchi, M. and Ehrmann, M. (2011). Knowledge Expansion of a Statistical Machine Translation System using Morphological Resources. In *Polibits 43.*

Urešová, Z. (2012). Building the PDT-VALLEX valency lexicon. In *Proceedings of the fifth Corpus Linguistics Conference*, pages 1–18, Liverpool, UK. University of Liverpool, University of Liverpool.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress (2)*, pages 1114–1122.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy.

Yamada, K. and Knight, K. (2002). A decoder for syntax-based statistical mt. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 303–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

Žabokrtský, Z. (2005). Resemblances between meaning-text theory and functional generative description. In *Proceedings of the 2nd International Conference of Meaning-Text Theory*, pages 549–557, Moskva, Russia. Slavic Culture Languages Publishers House.

Zeman, D., Fishel, M., Berka, J., and Bojar, O. (2011). Addicter: What is wrong with my translations? *Prague Bulletin of Mathematical Linguistics*, 96:79–88.

Zeman, D. and Sarkar, A. (2000). Learning verb subcategorization from corpora: Counting frame subsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 227–233, Athens, Greece. European Language Resources Association.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao, T. (2008). 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, pages 1121–1128.