РАЗЛИЧИЯ МЕЖДУ ЧЕШСКИМ И РУССКИМ ЯЗЫКАМИ НА МАТЕРИАЛЕ ПАРАЛЛЕЛЬНОГО КОРПУСА

Н. М. Клюева (netsie@yandex.ru)
Карлов Университет в Праге. Чехия.

В данной статье мы приводим примеры некоторых различий в конструкции предложений чешского и русского языков, которые были выявлены на материале параллельного чешско-русского корпуса. Статья не претендует на описание всех различий между языками, ограничиваясь лишь самыми частотными отличиями.

Ключевые слова: чешский язык, русский язык, параллельный корпус, языковые отличия

SOME DIFFERENCES BETWEEN CZECH AND RUSSIAN: A PARALLEL CORPUS STUDY

Klyueva N. M. (netsie@yandex.ru)
Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

We present a comparative study of some constructions in Czech and Russian. Though Czech and Russian are closely related Slavic languages, they have a few differences at the level of syntax, morphology and their semantics. We discuss incongruities that we found in a parallel Czech-Russian corpus, mainly reflecting differences in the sentence structure. The linguistic evidence presented in the paper will be used while constructing the transfer module of a rule-based machine translation system between Czech and Russian.

Key words: Czech, Russian, parallel corpus, language differences

1 The research was supported by the grant P406/2010/0875 GAČR and GAUK 639012.
1. Introduction

In this paper we make a corpus-based research in order to find the most frequent differences between Czech and Russian with the respect to the sentence structure. The languages share a lot of common features on every language level, still they have differences that might present a challenge for example to the Machine Translation system as well as to the language teaching.

The contrastive study of Slavic languages and our concrete pair of languages is a well studied topic. Just to name some of them, the paper [6] presents structure similarity measuring mainly on the material of Slavic languages. The authors in [2] compare Czech and Russian aspect and negation.

Here we neither suggest some global algorithm for finding differences nor focus in detail on one concrete linguistic problem, but we rather list the most frequent correspondences between Czech and Russian.

Our work was initially motivated by the task of creating transfer rules for the Machine Translation system between Czech and Russian [4] that can capture the main differences between the languages, and the results presented here are the first step towards the set of such rules.

The paper is divided into two main parts — in the Section 2 we provide a short description of a parallel corpus and a method how we searched for the examples. Section 3 lists the most frequent differences between the languages.

2. Parallel Corpus Study

One of the most popular technique of linguistic investigation now is the study of some language fact on an annotated corpus. We have made our research on a parallel segmented and tokenized Czech-Russian corpus\(^2\) that contains about 100.000 sentences on each side. For our task we have chosen 88.000 sentences with a sentence alignment one-to-one, where one Czech sentence is aligned to one Russian.

The corpus contains news texts mainly with political, social or economic themes downloaded from the site www.project.syndycate.org.

The corpus is tagged with a morphological tagger on both sides. For Russian we have used the TreeTagger [10] and for Czech language the Positional Tagger [3], the result annotation looks like follows:

\[(1cz) \text{Chápu|chápat|VB-S---1P-AA---jejich|jeho|PSXXXXP3-------}
\text{postoj|postoj|NNIS4------A----}
\]

\[(1ru) \text{Я|я|P-1-sn понимаю|понимать|Vmip1s-a-p их|они|P-3-pa}
\text{позицию|позиция|Ncfsan}
\]

'I understand their position'\n
As the two taggers exploit different annotation schemes, we did not make a full table of tag correspondences for Czech and Russian. We took only the first letter from

\(^2\) https://ufal.mff.cuni.cz/umc/cer/
the tag which reflects the word class and unified it for Czech and Russian according the schema in [3], so we will not go into the detail of both annotation schemes as the rest information will not be used for this research. We assigned each sentence with a clue encoding order of sentence constituents (first letters of part of speech), ex (2).

(2cz) Chápu jejich postoj VPN: (VerbPronounNoun)
Understand.1Sg.Pres their position
(2ru) Я понимаю их позицию PVPN: (PronVerbPronNoun)
I understand their position
'I understand their position'

Therefore we have calculated a Levenshtein’s distance between those sequences of part of speech tags that measures how different two strings (in our case the order of the word classes) are. Levenshtein’s distance reflects the minimal edit distance — a number that shows how many edit steps need to be introduced into the Czech string to transform it to the Russian. So the sentences in the example (1) have the edit distance 1, which illustrates that to transform a Czech sequence of tags into the Russian only one letter (P-pronoun) should be added (which is actually the personal pronoun “Я” — I). So the more is the Levenshtein’s distance, the more are two sentences different from each other. Those sentences are therefore the main focus of our research. The statistics on the distribution of sentences with the respect to the distance is presented in Table 1.

<table>
<thead>
<tr>
<th>Levenshtein’s distance</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>&gt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td># of sentences</td>
<td>296</td>
<td>721</td>
<td>1503</td>
<td>2423</td>
<td>74372</td>
</tr>
</tbody>
</table>

It appeared that only a small amount of sentences have the same structure in Czech and Russian (edit distance 0), ex:

(3cz) Oficiálně Čína zůstává komunistickou zemí.
AdverbNounVerbAdjNoun
(3ru) Официально Китай остается коммунистической страной.
AdverbNounVerbAdjNoun
both: 'Officially China remains Communist country'

This fact was really astonishing because we expected much more correlations in the sentence structure.

---

3 Abbreviations for POS in [3] are: N=noun, V=verb, A=adjective,P=pronoun, R=preposition, D=adverb etc.

4 Levenshtein distance was calculated using Perl module https://metacpan.org/module/Text::Levenshtein
One of the reasons might be that Czech and Russian sentences in this corpus are translated from English original in different, often novel way, ex. (4).

(4cz) Evropa je krátkozraká
Europe is myopic

(4ru) Европа ведет себя недальновидно
Europe behaves itself without foresight

'Europe is being myopic'

Even it is not always the fact that sentences with the distance is 0 will have similar structure. It is often the difference in phraseology that becomes the source of the incorrespondence. In the example (5) Czech and Russian use totally different constructions — fixed phrase units 'v sázce je' vs. 'на карту поставлена', though the sequences of parts of speech are the same:

(5cz) V sázce je bezpečnost lidstva
At stake is security mankind.gen

(5ru) На карту поставлена безопасность человечества
On map put.passive security mankind.gen

'At stake is the security of mankind'

Moreover, the news sentences are generally long and have complicated structure which will increase the amount of incorrespondences.

The third factor is that we have studied only the order of part of speech sequences. Probably if we have a more deep annotation the percentage of sentences with corresponding structure will be higher. More deep syntactic analysis would have help to detect more non-trivial differences and to handle such incorrespondences as for example ellipsis. To provide such an analysis, the high-quality syntactically annotated parallel treebank would be needed, which we do not have for both Czech and Russian. From the experience of the Prague Czech-English Dependency Treebank[1], a lot of manual annotation work should be used. Moreover, the parallel texts chosen for the treebank should be translated manually with the instructions for annotators to translate as close as possible to the original, so that it would be easier to align the annotated trees. For this goal it is not sufficient to take parallel texts already translated as the sentences can be translated from one language to another in a novel way.

As we do not have so far human resources for such a long-lasting goal, we made our study on what we have — a parallel corpus with simple morphological annotation — exploiting the corpus only in a linear word-for-word manner.

2.1. Differences in a Sentence Structure between Czech and Russian

To illustrate the cases where Czech and Russian use the different construction we have taken those sentences that have the Levenshtein's distance 1, 2 or 3. They
reflect some of the relevant differences in the sentence structure and at the same time
do not overload the sentence with too much incorrespondences. Below we describe
the most frequent differences according to the parallel corpus. If there exist several
options to translate a construction from Czech into Russian, they are given with an ap-
propriate statistics from the corpus.

2.2. Structural differences

One of the most big challenges while translating from Czech into Russian is the
omission of a subject pronoun in Czech and a verb 'to be' which is dropped in Rus-
sian (6ru) but is present in Czech⁵. According to the corpus, in 1102 cases the copula
construction in Czech corresponds to a zero-copula construction with a dash symbol
in Russian, in 1673 cases there is no dash and no verb like in (6ru). The copula verb
was translated as a verb 'являться' — 'appear to be' in 1331 cases(7ru, first clause).
This verb occurs generally in the written texts and sounds officially.

(6cz) Vlády jsou zkorumpované
Governments are corrupt

(6ru) Правительства Ø коррумпированы
Governments corrupt
'Governments are corrupt'

(7cz) První strategie je krátkozraká a druhá je ošklivá.
First strategy is shortsighted and second Ø nasty.

(7ru) Первая стратегия является недальновидной, а вторая — отвратительной.
First strategy is shortsighted and second — nasty.

Analytical past in Czech is formed by the appropriate form of the verb “to be”
and the past participle whereas in Russian the copula is omitted as in the example (8).

(8cz) Přišla jsem pozdě
come.Past.Fem to_be.1Sg.aux late

(8ru) я пришла поздно
I.1Sg come.Past.Fem late
'I came late'

Reflexive particle in Russian is incorporated into a verb, and in Czech — though
considered to be a part of a lemma [5] — is written separately from the verb:

⁵ In the deep syntactic analysis within the treebanks the missing sentence elements are gener-
ally annotated with zero mark Ø.
Some differences between Czech and Russian: a parallel corpus study

(9cz) Proč se Shiller mýlil? PronPartNounVerb
Why refl.part Shiller mistake.3Sg.Past
(9ru) Почему Шиллер ошибся? PronNounVerb
Why Shiller mistake.3Sg.Past

'It is sometimes misleading when the particle stands far away from the verb and it can not be easily identified to which verb it belongs or even in case if the particle 'se' if it is really a particle or a vocalized preposition 'se' (with):

(10cz) Popsali nové odrůdy rýže s, o něž jsou připraveni se svými africkými protějšky podělit.
Described new sorts of rice, about which are.ready se.refl se.prep their African colleagues share.inf

(10ru) Они рассказали о новых сортах риса, которыми они готовы поделиться со своими африканскими коллегами.
'They described new sorts of rice, with which they are ready to share with their African colleagues'

Often things get complicated when some of the listed phenomena (ex. pro-drop, past tense, reflexive verbs) got gathered in one sentence and the clitics go in Wacker-nagel's position which mixes the sentence structure even more:

(11cz) Dlouho jsem se smál. AdvAuxReflVerb
Long to_be.aux refl.particle laugh.Past
(11ru) Я долго смеялся. PronAdvVerb
I long laugh.Past

'It would have been more appropriate to make a deep syntactic analysis of the evidence in the above examples (6)–(11) within either Prague Dependency Treebank[5] or SynTagRus treebank of Russian[9], where those 'trivial' differences would be eliminated on the deepest level of annotation. However, we should have a big amount of annotated parallel data, which is difficult to built and it is not realistic so far.

On the clause level the obvious difference is the usage of some coordinating conjunctions with contrastive meaning, namely the order of elements in such clauses.

(12cz) Trest však mohl být tvrdý NounConjVerbVerbAdj
Punishment but might be hard
(12ru) Но наказание могло быть суровым ConjNounVerbVerbAdj
But punishment might be hard

'But the punishment might be hard'
The Czech contrastive conjunction *však* usually takes the second position in the sentence (12cz), which causes dissimilarity with the Russian translation equivalent (12ru). The conjunction *ale* may also take the second position (13cz), so the order of elements may be similar in Russian (13ru). There is also a special meaning of the word *ale* that expresses amazement or surprise (14) that does not exist in Russian:

(14cz) *To byla ale cesta!*

That was but trip

(14ru) *Ну и дорога была!*

Well and trip was

'What a trip was it!'"}

According to the corpus statistics Russian tends to use a sentence with the interrogative particle *ли* more often than Czech with the respective -li. This particle occurs in 1873 sentences in Russian and only in 208 for Czech both in interrogative sentences and relative clauses. Instead, in Czech sentences other particles with similar meaning are used: zda — 454 translations, ex. (15), jestli — 37, or there is no particle at all in 993 cases, ex. (16).

(15cz) *Otázka tedy nezní, zda Evropa existuje, ale zda jsme spokojeni s tím, jak funguje.*

(15ru) *Вопрос заключается не в том, существует ли Европа, а в том, удовлетворены ли мы тем, как она функционирует.*

'& The question is not whether Europe exists, but if we are content with the way it function'"

(16cz) *Praskne další bublina?*

Burst next bubble?

(16ru) *Лопнет ли очередной пузырь?*

Burst if(int. part.) next bubble

'Will the next bubble burst?'

2.3. Differences in connection with lexicology and idioms

A multilingual expression or a certain fixed lexical phrase in one language is translated into the other language in the other way or just can be translated descriptively. Here we suggest three examples of differences that belong to the field of phraseology and idiomatics. The differences in lexicology are not easy
Some differences between Czech and Russian: a parallel corpus study

to detect automatically, so here we provide several created examples not from the corpus.

The sentence (17cz) reflects that the woman has not done the action herself, but someone else did it. In Russian (17ru) it cannot be identified from a sentence whether the hair was cut by woman herself or by someone else, but native speakers know, that generally it is the second option. The Czech causative construction (17cz) can sound strange to the native speaker of Russian, and vice versa a Russian variant may seem awkward and even funny for the Czech speaker (a lady cutting her hair herself at a hairdresser).

(17cz) Nechala si ostříhat vlasy v kadeřnictví
Let.Past.Fem.Sg refl.part cut hair at hairdresser
(17ru) Она подстригла волосы в парикмахерской
She cut.Past.Fem.Sg hair
'She had her hair cut at hairdresser'

In (18) it is not so obvious which Russian equivalent should be used to the Czech construction 'slyšet na' — 'hear at' as this meaning of the verb 'hear' is specific to the Czech language only.

(18cz) Rusové slyší na české lázně
Russians hear at Czech spa
(18ru) Русские интересуются чешскими курортами
Russians interest.3Pl Czech spa
'Russians like Czech spa' or 'Russians are interested in Czech spa'

**Idiomatic expressions** present the big challenge for our task because their word-for-word translation will result in a construction that may sound awkward or even ungrammatical in another language:

(19cz) agenda svobody nezmění přístup lidí přes noc
agenda freedom.gen change.neg.fut attitude people.gen over night
(19ru) план свободы не изменит отношение людей в один миг
agenda freedom.gen not change.fut attitude people.gen in one second
'a freedom agenda will not change people 's attitudes overnight'

3. Conclusion

In this paper we have presented a comparative study of Czech and Russian with the respect to the differences mainly in the sentence constructions. The examples illustrating differences are based on the parallel corpus research. The method of finding differences proposed here was not sufficient enough as it showed a very low percentage of correspondences in sentence structure which intuitively should be much higher. Still, we have selected some frequent differences in Czech and Russian that
may serve as a basis for the detailed research in the future involving more deep syntactic analysis.

The evidence described here is also planned be exploited in the Rule-Based Machine Translation from Czech into Russian.

References