SOME DIFFERENCES BETWEEN CZECH AND RUSSIAN: A PARALLEL CORPUS STUDY

Natalia Klyueva
Institute of Formal and Applied Linguistics, Charles University in Prague

1. Introduction

Motivation

Though Czech and Russian are closely related languages, they have a few differences on the level of syntax, morphology and semantics. Here we will discuss those incorrespondences that we have found in a parallel Czech-Russian corpus mainly in the sentence structure.
Our main aim is to create a set of transfer rules for the Rule-Based Machine Translation system, so here we show some constructions in two languages that need to be properly handled.

2. Parallel Czech-Russian Corpus

Parallel corpus
We have made our research on a parallel segmented and tokenized Czech-Russian corpus 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. Full statistics of the corpus is given in the table below.
The corpus contains news texts mainly with political, social or economic thematics downloaded from the site www.project.syndycate.org

	Czech	Russian
Words	1,747,997	1,815,550
Tokens	2,022,990	2,152,326
Sentences	96,335	101,528

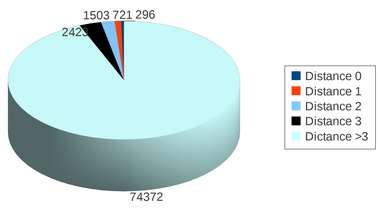
Morphological annotation
The corpus is tagged with a morphological tagger on both sides, the format of annotation is **form|lemma|tag,** the first letter of the tag is a part of speech of a word which we will exploit in our work. For Russian we have used the TreeTagger and for Czech language the Positional Tagger, the result annotation looks like follows:
(1cz)Chápu|chápat|VB-S---1P-AA--- jejich|jeho|PSXXXXP3------- postoj|postoj|NNIS4-----A----
(1ru)Я|я|P-1-sn понимаю|понимать|Vmip1s-a-p их|они|P-3-pa позицию|позиция|Ncfsan
'I understand their position'

Sentences with the same structure:
Oficiálně Čína zůstává komunistickou zemí.
Официально Китай остается коммунистической страной.
Sentences with different structure:
V sázce je bezpečnost lidstva
На карту поставлена безопасность человечества

3. Measuring distance between sentences

Levenshtein's distance between sequences of word classes
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 the tag which reflects the word class and unified it for Czech and Russian as follows:
N=noun, V=verb, A=adjective,P=pronoun, R=preposition, D=adverb etc.

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. Example of an annotated pair of sentences:
Chápu jejich postoj(cz) vs. *Я понимаю их позицию(ru)*
VPN:(VerbPronNoun) vs. PVPN:(PronVerbPronNoun) — edit distance 1.



Why are there so many incorrespondences?
- sentences are generally long and have complicated structure
- Czech and Russian sentences are translated not directly, but from English original in different way
- Our method is very superficial. We have studied only the order of part of speech sequences, more deep annotation is needed

Set of sentences to evaluate:
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. We have analyzed this set of sentences and detected several groups of linguistic dissimilarities(see second column).

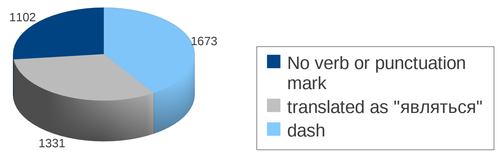
This project was supported by grants P406/2010/0875 GAČR and GAUK 639012.

4. Differences

Ellipsis in Czech – pronoun drop
- Czech tends to incorporate a person morpheme into a verb and leaves out (almost always, see the table) a personal pronoun
- Jsem student vs. Я – студент
- Usage of personal pronouns according to the corpus in Czech and Russian:

	Ja/я	Ty/ты	On(a,o)/ On(a,o)	My/мы	Vy/вы	Oни/oni(y,a)
Czech	143	8	264	462	18	167
Russian	5433	24	5102	2361	334	4131

Ellipsis in Russian – copula drop
- verb 'to be' is dropped in Russian but is present in Czech
- Several variants of translation of Czech copula into Russian, statistics of frequency of translation is in the chart after the examples.
  - *Vlády **jsou** zkorumpované Правительства коррумпированы* (no verb or punctuation mark)
  - *První strategie **je** krátkozraká Первая стратегия **является** недальновидной (more official variant)*
  - *A druhá **je** ošklivá → А вторая **-** отвратительна (the dash symbol is used)*



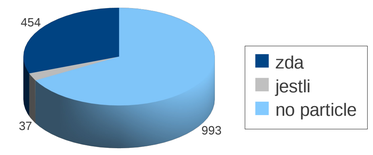
Analytical past
Analytical past in Czech is formed by the appropriate form of the verb "to be" and the past participle whereas in Russian the copula is omitted :
(cz) *Přišla jsem pozdě*
(ru)я пришла поздно

Reflexives
Reflexive particle in Russian is incorporated into a verb, and in Czech – though considered to be a part of a lemma – is written separately from the verb :
*Proč **se** Shiller mýlil?*
*Почему Шиллер ошиб**ся**?*

Contrastive conjunctions
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 :
*Trest **však** mohl být tvrdý*
***Но** наказание могло быть суровым*

Particle -li
The languages are related, so the closed-class (function) words generally have very similar surface realization, but this might be tricky: Czech particle **li** has totally another usage than Russian **ли**. This particle occurs in 1873 sentences in Russian and only in 208 for Czech both in interrogative sentences and relative clauses. In Czech sentences other particles with similar meaning are used: zda(ex a), jestli, or there is no particle at all(ex. b)Translation variants for Russian **ли** are shown on the chart.
(a)O*tázka tedy nezní , **zda** Evropa existuje , ale **zda** jsme spokojeni s tím , jak funguje .*
*Вопрос заключается не в том , существует **ли** Европа , а в том , удовлетворены **ли** мы тем , как она функционирует .*

(b)Praskne další bublina?
*Лопнет **ли** очередной пузырь?*



5. Differences in lexicology and idiomatics

Those differences are:
- not easy to detect automatically in a corpus
- not easy to cope within the Rule-Based MT system – specific lexicon needed
- addressed properly In the Statistical Machine Translation if seen in the training data

Causative Construction:
(cz)Nechala si ostříhat vlasy v kadeřnictví
(lit. Она дала отстричь волосы в парикмахерской*)*
(ru)Она подстригла волосы в парикмахерской
(like the one in English) She had her hair cut

*Czech construction "**slyšet na**"(слышать на):*
Rusové slyší na české lázně (lit. Русские слышат на чешские курорты)
Русские интересуются чешскими курортами

Idiomatic expressions:
nosit dříví do lesa (lit. носить дрова в лес)
ездить в Тулу со своим самоваром