Czech-Russian Corpus via a Simple Web Interface

Natalia Klyueva, Radovan Garabík, Ondřej Bojar

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Slavicorp 2012, Mainz

Motivation

- Czech-Russian corpus was created and used:
  - for the purpose of Machine Translation,
  - in a linguistic research – comparing Czech and Russian languages
- The corpus has been so far available to download only in a machine-readible format as one file
- Radovan Garabík has put it into a user-friendly interface

Parallel Czech-English-Russian UMC Corpus

- Intercorp has a Czech-Russian section, but...
- Texts downloaded from the single source, Project Syndicate, news, politics, economics (2.186 texts) http://www.project-syndicate.org/
- Texts in Czech are tagged by the Positional Tag system, English and Russian ones by the TreeTagger
- Annotation: each word form is assigned by a lemma and a morphological tag: Cz: mnohé|mnohý|AAFP1----1A----, En: happens|happen|V|VVZ, Ru: указывают|указывать|V|Vmip3p-a-p

Statistics of the corpus

	Czech	Russian	English
Words	1,747,997	1,815,550	1,920,164
Tokens	2,022,990	2,152,326	2,255,901
Sentences	96,335	101,528	97,250

Corpus view

```
16953 1-1     0.709711        European|European|J|JJ attitudes|attitude|N|NNS demonstrate|demonstrate|V|VVP the|the|D|DT
      consequences|consequence|N|NNS .|.|S|SENT       Такие|такой|P|P---pn европейские|европейский|A|Afp-pnf подходы|подход|
      N|Ncmpnn имеют|иметь|V|Vmip3p-a-p свои|свой|P|P---pa последствия|последствие|N|Ncnpan .|.|S|SENT
16954 1-1     0.655   Americans|Americans|N|NPS are|be|V|VBP right|right|J|JJ to|to|T|TO point|point|V|VV out|out|R|RP that|
      that|D|DT Europeans|Europeans|N|NPS spend|spend|V|VVP an|an|D|DT undue|undue|J|JJ amount|amount|N|NN of|of|I|IN time|
      time|N|NN consulting|consulting|N|NN among|among|I|IN themselves|themselves|P|PP ,|,|,|, and|and|C|CC then|then|R|RB
      come|come|V|VV up|up|R|RP with|with|I|IN very|very|R|RB little|little|J|JJ .|.|S|SENT   Американцы|американец|N|Ncmpny
      правы|правый|A|Afp-p-s ,|,|,|, когда|когда|C|C указывают|указывать|V|Vmip3p-a-p на|на|S|Sp-a то|тот|P|P--nsa ,|,|,|,
      что|что|C|C европейцы|европеец|N|Ncmpny тратят|тратить|V|Vmip3p-a-p непозволительно|непозволительный|A|Afpns-s много|
      много|R|R времени|время|N|Ncnsgn на|на|S|Sp-a консультации|консультация|N|Ncfpan друг|друг|N|Ncmsny c|c|S|Sp-i другом|
      друг|N|Ncmsiy ,|,|,|, а|а|C|C затем|затем|R|R предлагают|предлагать|V|Vmip3p-a-p слишком|слишком|R|R мало|мало|R|R .|.|
      S|SENT
16955 1-1     0.233333        Most|most|J|JJS of|of|I|IN the|the|D|DT time|time|N|NN ,|,|,|, they|they|P|PP agree|agree|V|
      VVP only|only|R|RB to|to|T|TO try|try|V|VV to|to|T|TO put|put|V|VV the|the|D|DT brakes|brake|N|NNS on|on|I|IN firm|
      firm|N|NN action|action|N|NN .|.|S|SENT Чаще|частый|A|Afc всего|все|P|P--nsg они|они|P|P-3-pn соглашаются|соглашаться|
      V|Vmip3p-m-p лишь|лишь|Q|Q на|на|S|Sp-a то|тот|P|P--nsa ,|,|,|, чтобы|чтобы|C|C попытаться|попытаться|V|Vmn----m-e
      сдержать|сдерживать|V|Vmn----a-e решительные|решительный|A|Afp-pnf действия|действие|N|Ncnpan .|.|S|SENT
16956 1-1     0.404425        They|they|P|PP are|be|V|VBP reluctant|reluctant|J|JJ .|.|S|SENT Они|они|P|P-3-pn на|на|S|Sp-a
      все|весь|P|P---pa идут|идти|V|Vmip3p-a-p c|c|S|Sp-i большой|большой|A|Afpfsif неохотой|неохота|N|Ncfsin .|.|S|SENT
16957 1-2     0.183871        "|"|'|'' Don|Don|N|NP '|'|P|POS t|t|N|NN go|go|V|VVP too|too|R|RB fast|fast|R|RB in|in|I|IN
      welcoming|welcome|V|VVG Russia|Russia|N|NP to|to|T|TO NATO|NATO|N|NP !|!|S|SENT "|"|'|'' "|"|`|`` Don|Don|N|NP '|'|P|
      POS t|t|N|NN go|go|V|VVP too|too|R|RB far|far|R|RB in|in|I|IN supporting|support|V|VVG Israel|Israel|N|NP and|and|C|CC
      neglecting|neglect|V|VVG the|the|D|DT Palestinians|Palestinians|N|NPS !|!|S|SENT "|"|'|'' "|"|`|`` Don|Don|N|NP '|'|P|
      POS t|t|N|NN extend|extend|V|VV the|the|D|DT fight|fight|N|NN against|against|I|IN terrorism|terrorism|N|NN to|to|T|TO
      the|the|D|DT producers|producer|N|NNS of|of|I|IN weapons|weapon|N|NNS of|of|I|IN mass|mass|J|JJ destruction|
      destruction|N|NN !|!|S|SENT "|"|`|`` What|What|W|WP should|should|M|MD be|be|V|VB done|do|V|VVN instead|instead|R|
      RB ?|?|S|SENT   "|"|-|- Не|не|Q|Q надо|надо|R|R слишком|слишком|R|R спешить|спешивать|V|Vmn----a-e c|c|S|Sp-i
      привлечением|привлечение|N|Ncnsin России|россия|N|Ncfsgn в|в|S|Sp-l НАТО|НАТО|N|Ncnsln !|!|S|SENT "|"|-|- .|.|S|SENT
      "|"|-|- Не|не|Q|Q заходите|заходить|V|Vmi-2p-a-p слишком|слишком|R|R далеко|далеко|R|R c|c|S|Sp-i поддержкой|поддержка|
      N|Ncfsin Израиля|израиль|N|Ncmsgn и|и|C|C пренебрежением|пренебрежение|N|Ncnsin палестинцами|палестинцами|N|Ncmpin
      "|"|-|- !|!|S|SENT
16958 1-2     -0.3    "|"|'|'' Political|political|J|JJ solutions|solution|N|NNS "|"|`|`` should|should|M|MD be|be|V|VB
```

A pair of sentences from the Czech-Russian Corpus

Dobře|dobře|Dg-------1A---- zapadají|zapadat_:T|VB-P---3P-AA--- běloši|běloch|NNMP1-----A---- ,|,|Z:---------- Asiaté|Asiat_;E|NNMP1-----A---- i|i-1|J^---------- lidé|člověk|NNMP1-----A---1 ze|z-1|RV--2---------- Středního|střední|AAIS2----1A---- východu|východ|NNIS2-----A---- .|.|Z:----------

Здесь|здесь|R прекрасно|прекрасно|R уживаются|уживаться|Vmip3p-m-p Белые|белый|Afp-pn ,|,|, Азиаты|азиат|Ncmpny и|и|C представители|представитель|Ncmpny Среднего|средний|Afpmsg Востока|восток|Ncmsgn .|.|SENT

The corpus via the web interface

http://korpus.sk:8095/

Query

Прага Hľadať v korpuse cer-csru-ru

áäčdéěíĺľňóôöŕřšťúüůýžÁÄČĎÉĚÍĹĽŇÓÔÖŔŘŠŤÚÜŮÝŽ

абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ

[]^*„“"\=

O korpuse •

#38

79676	Для них перемена выглядит чем - то иррациональным , или даже катастрофичным . Как мы видели на примере протестов против МФВ , прошедших недавно в **Праге** , и прошлогодних протестов против Мировой Торговой Организации в Сиэттле , этот страх распространился далеко и широко . Люди похоже все больше рассматривают перемены не как что - то , что укрепляет их свободу и достоинство , а как силу , способствующую распространению несправедливости и жадности .	Zdá se jim proto , že jsou to změny absurdní , nesmyslné a dokonce až katastrofické . Jak jsme mohli pozorovat na nedávných pražských protestech proti MMF , ale stejně tak i na pouličních protestech proti Světové obchodní organizaci (WTO – World Trade Organization) v americkém Seattle , šíří se tyto obavy dál a dál . Lidé stále více vidí ve změnách nikoliv cosi , co zvyšuje jejich vlastní svobodu a důstojnost , ale sílu , která povzbuzuje lakotu , hrabivost a nespravedlnost .
105108	И все же демократическое будущее Украины пока еще не гарантировано . Украина претерпевает истинную либеральную революцию , родственную великим европейским либеральным революциям 1848 года и напоминает Бархатную революцию в **Праге** в 1989 году	To však ještě nezaručuje demokratickou budoucnost Ukrajiny . Ta zažívá skutečnou liberální revoluci , která je blízkou příbuznou velkých evropských liberálních revolucí z roku 1848 a připomíná i pražskou

Done

Usage of the corpus

- Theoretical research
- Machine Translation

A playground for experiments

- Measuring phonetic differences
- Comparing valency in Czech and Russian (10% of verbs have different valency frame, ex. doufat v neco – надеяться на что-либо)
- Prepositions in Czech and Russian
- Ellipsis in Czech and Russian
- Word order issues

Sample search – machine readible

- Copula translation from Czech into Russian?
- cat Czech-Russian | grep являться | egrep "být\|VB-.---..-AA---"

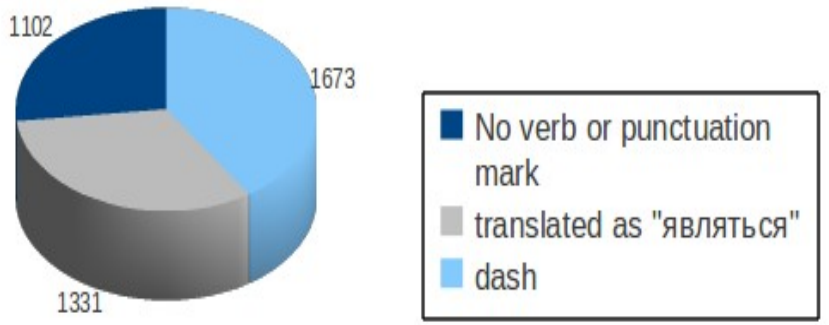
```
|,|, по|по|Sp-d которому|который|P--msd разворачиваются|разворачиваться|Vmip3p-m-p дебаты|дебаты|Nc-pn как|как|C в|в|Sp-l сири
йском|сирийский|Afpmslf правительстве|правительство|Ncnsln ,|,|, так|так|R и|и|C среди|среди|Sp-g подданных|подданный|Ncmpgy .
|.|SENT
Ammar|Ammar|NNIS1-----A---- Abdulhamid|Abdulhamid|NNXXX-----A---- je|být|VB-S---3P-AA--- syrským|syrský|AAMS7----1A---- spisov
atelem|spisovatel|NNMS7-----A---- a|a-1|J^------------ sociologem|sociolog|NNMS7-----A---- působícím|působící_^(*3it)|AGIS7--
---A---- v|v-1|RR--6---------- Damašku|damašek|NNIS6-----A---- ,|,|Z:------------ kde|kde|Db------------ je|být|VB-S---3P-AA
--- spolumajitelem|spolumajitel|NNMS7-----A---- malého|malý|AANS2----1A---- vydavatelství|vydavatelství|NNNS2-----A---- Etana|
Etana_;S|NNMS1-----A---- Press|Press-1_;K_,t_^(souč._názvů,_např._tisk._agentur)|NNIXX-----A---- .|.|Z:------------ Он|он
|P-3msn живет|жить|Vmip3s-a-p в|в|Sp-l Дамаске|дамаск|Ncmsln ,|,|, где|где|R является|являться|Vmip3s-m-p одним|один|P--msi из
|из|Sp-g руководителей|руководитель|Ncmpgy небольшого|небольшой|Afpmsgf издания|издание|Ncnsgn Этана|Этана|Ncmsgn Пресс|пресс|
Ncmsnn .|.|SENT
Toto|tento|PDNS1---------- dění|dění_^(*3it)|NNNS1-----A---- je|být|VB-S---3P-AA--- samozřejmě|samozřejmě_^(*1ý)|Dg-------1A--
-- důsledkem|důsledek|NNIS7-----A---- invaze|invaze|NNFS2-----A---- do|do-1|RR--2---------- Iráku|Irák_;G|NNIS2-----A---- pod|
pod-1|RR--7---------- vedením|vedení|NNNS7-----A---- USA|USA_;G|NNIPX-----A---- .|.|Z:------------ Ta|ten|PDFS1---------- pro
měnila|proměnit_:W|VpQW---XR-AA--- nezřetelnou|zřetelný|AAFS4----1N---- a|a-1|J^------------ vzdálenou|vzdálený_^(*3it)|AAFS4
----1A---- hrozbu|hrozba|NNFS4-----A---- v|v-1|RR--4---------- impozantního|impozantní|AAMS4----1A---- souseda|soused|NNMS4---
--A---- ,|,|Z:------------ jehož|jenž_^(který_[ve_vedl.větě])|P1XXXZS3------- úmysly|úmysl|NNIP1-----A---- nejsou|být|VB-P---
3P-NA--- ve|v-1|RV--6---------- vztahu|vztah|NNIS6-----A---- k|k-1|RR--3---------- syrskému|syrský|AAIS3----1A---- baasistické
mu|baasistický|AAIS3----1A---- režimu|režim|NNIS3-----A---- rozhodně|rozhodně_^(*1ý)|Dg-------1A---- přátelské|přátelský|AANS1
----1A---- .|.|Z:------------ Эти|этот|P---pn изменения|изменение|Ncnpnn ,|,|, конечно|конечно|Q ,|,|, являются|являться|Vm
ip3p-m-p следствием|следствие|Ncnsin оккупации|оккупация|Ncfsgn Ирака|ирак|Ncmsgn американцами|американец|Ncmpiy и|и|C их|они|
P-3-pa союзниками|союзник|Ncmpiy ,|,|, превратившей|превращать|Vmps-sfafei туманную|туманный|Afpfsaf и|и|C отдалённую|отдалённ
```

..and more user friendly

	сирийского общества , замалчивается подавление основных свобод .	a náboženské fráze .
2586	Амар Абдулхамид - сирийский писатель и общественный аналитик . Он живет в Дамаске , где **является** одним из руководителей небольшого издания Этана Пресс . Его последний роман " Менструация " , написанный и изданный на английском лондонским издательством Саги в 2001 , переведен на несколько языков .	⊘ Ammar Abdulhamid je syrským spisovatelem a sociologem působícím v Damašku , kde je spolumajitelem malého vydavatelství Etana Press . V roce 2001 vyšel u londýnského nakladatelství Saqi Books jeho zatím poslední , anglicky psaný román Menstruace , který již byl přeložen do řady dalších jazyků .
2813	Каждая группа сосредоточена на утверждении своих собственных привилегий , в то время как сохранение территориальной целостности и национального единства страны становится всё более неуправляемым и скользким делом . Эти изменения , конечно , **являются** следствием оккупации Ирака американцами и их союзниками , превратившей туманную и отдалённую угрозу во внушительного соседа , намерения которого в отношении сирийского баасистского режима никак не назовёшь дружественными . В результате до правящих кругов , ранее лишённых чувства реальности , наконец начала доходить необходимость радикальных изменений в структуре и стиле режима .	Všechny skupiny se zaměřují na vymezení svých konkrétních výsad , ovšem při zachování územní celistvosti a národní jednoty země , která je čím dál podrážděnější a křehčí . Toto dění je samozřejmě důsledkem invaze do Iráku pod vedením USA . Ta proměnila nezřetelnou a vzdálenou hrozbu v impozantního souseda , jehož úmysly nejsou ve vztahu k syrskému baasistickému režimu rozhodně přátelské . Proto se konečně začalo usazovat povědomí o potřebě drastických změn v uspořádání a stylu režimu , jenž dříve býval k realitě slepý .
3084	Даже несмотря на это , такая перемена всё равно значительна по сирийским стандартам , поскольку на независимые инициативы по традиции смотрят неодобрительно . Действительно значимым событием **является** создание сирийским инженером Айманом Абдул Нуром пресс - службы All 4 Syria (www . all 4 syria . org) . В службу входит электронный информационный бюллетень , содержащий относящиеся к Сирии доклады и статьи из разнообразных источников , часто включающие комментарии деятелей оппозиции внутри страны и за её пределами .	Navzdory tomu představuje tento vývoj na syrské poměry stále pokrok , neboť nezávislé iniciativy bývají tradičně v nemilosti . Skutečně zásadní je tiskový servis All 4 Syria (www . all4syria . org) , založený syrským inženýrem Ajmanem Abdulem Nourem . Servis zahrnuje elektronický zpravodaj , jenž přináší zprávy a články týkající se Sýrie , shromažďované z různých zdrojů , často včetně vyjádření představitelů domácí i zahraniční opozice .
	Я участвовал в реализации другой инициативы - проекта Thawra (www	Já jsem se podílel na zahájení Projektu Thawra (www

Sample search - copula

- *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)*



Valency differences

- Valency in Czech and Russian, prepositional valency
  - (cz)utíkat před +Ins vs. (ru)убегать от + Gen
  - (cz)pro + Acc vs. для + Gen

Searching for valency differences

(ru)отказывать в + Acc vs. (cz)odepírat +Acc

Query

отказывать [tag="N.*"] Hľadať v korpuse cer-csru-ru

áäčďéěíĺľňóôöŕřšťúüůýžÁÄČĎÉĚÍĹĽŇÓÔÖŔŘŠŤÚÜŮÝŽ

абвгдеёжзийклмнопрстуфхцчшщъыьэюяАБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ

[]^*„“"\=

O korpuse •

#13

840804	Должно быть " Аль – Каеда " надеется заманить американских солдат вместе со всем снаряжением на узкие высокогорные тропы , где до этого уже нашли свою смерть российские солдаты . Со стороны американцев мудро было бы **отказать Усаме** Бен Ладену в таком удовольствии . Таким образом , принципал использует уполномоченных для того , чтобы самому не увязнуть в трясине .	Al - Kajdá určitě doufá , že do vysokých a úzkých horských průsmyků , kam si Sověti přišli pro jistou smrt , naláká dlouhé konvoje amerických vojáků a jejich výzbroje . Rozumná americká strategie tuto radost bin Ládinovi odepře . Hlavní aktér tedy prostředníka používá proto , aby sám neskončil v bažině .
878036	Спустя одиннадцать лет после своей первой связи с Всемирной Паутиной (WWW) , доступ Китая к Интернету все еще охраняется защитными системами , встроенными в его серверы - посредники , которые доказали , что они практичнее и неприступнее Берлинской Стены . Более того , увеличение спроса на широкополосную связь привело к запуску " Проекта Jin Dun (Золотой Щит) " стоимостью $ 800 миллионов , автоматической цифровой системы общественного порядка , которая поможет продлить коммунистическое правление , **отказывая китайцам** в праве на информации . Принцип , лежащий в основе Золотого Щита , состоит в том , что " когда	Jedenáct let po úvodním připojení se k celosvětové síti World Wide Web (WWW) hlídají čínský přístup k internetu stále firewally zabudované v proxy serverech , které se ukázaly jako praktičtější a neprostupnější než berlínská zeď . Růst poptávky po širokopásmovém připojení navíc přinesl spuštění „ Projektu Ťin Tun “ (Zlatý štít) , automatického digitálního systému veřejného dozorování v ceně 800 milionů dolarů , který pomůže prodloužit komunistickou vládu tím , že odepře čínskému lidu právo na informace . „ Zlatý štít “ je založen na principu , že , zatímco ctnost se zvýší o jednu stopu

Some more verbs

- (cz)Ceny klesly **o** 20%
  (ru)ceny upali **na** 20%
- (cz)Prchat, ujíždět,unikat **před** + Ins
  (ru)скрываться, уезжать, убегать **от** + Gen
- (cz)brát + Dat - (ru)брать **у** +Gen
- (cz)ptát se **na**+ Acc - (ru)спросить **о** +Loc

Phrase table from Moses – translation of prepositions

pro Nájdi Stratégia: Presne v: CS-BG KO KO-DE slovake BG-CS cercsru

Výsledky hľadania pre: **pro**

CS-RU frázy Založené na paralelnom česko-ruskom korpuse

a nadace pro reformu ≈ и китайского фонда реформ
a ne pro jednoho ≈ а не за одного
a nikoliv pro bohaté ≈ а не для богатых
a to nejen pro ≈ и не только для
a tudíž nepodstatné pro ≈ конструкции , не играющие для
a tudíž nepodstatné pro ≈ не играющие для человечества
a udělala hodně pro ≈ и многое сделала для
a urážení lidí pro ≈ и оскорбление людей на
...

Vstup: UTF-8 Výstup: UTF-8 In English · Po slovensky

Naposledy hľadané slová:
pro,

Machine Translation – testing the corpus quality

- A number of experiments with MT systems were done using the corpus as training data:
- Statistical MT Moses between related and non-related languages (BLEU score in brackets):
  - ru->cs (11%, with morph. 13%)
  - en->cs (14% with morph. 15%)
  - cs->ru (9%)
- Rule-Based MT Cesilko cs->ru(3%)
- Translation quality is low, we need more data

Work in progress and plans for future

- Collecting ebooks:
  - We have a parallel Czech-English corpus
  - Search for the respective Russian texts on lib.ru
  - Making a tri-parallel corpus of ebooks
- Collecting film titles

Thank you!

This work was supported by grants P406/10/0875 and GAUK 639012