

ST PETERSBURG STATE UNIVERSITY
INSTITUTE FOR LINGUISTIC STUDIES (RAS)
HERZEN STATE PEDAGOGICAL UNIVERSITY OF RUSSIA

PROCEEDINGS
OF THE INTERNATIONAL CONFERENCE
«CORPUS LINGUISTICS–2017»

June 27–30, 2017, St. Petersburg

SAINT PETERSBURG
2017

TEACHING CORPUS LINGUISTICS WITH ARANEA WEB CORPORA¹

Abstract. Our paper describes our experience in introducing the new subject *Introduction to Corpus Linguistics* for the students of language-related programmes at our University. We describe both the technical infrastructure, and the pedagogical aspects related to the subject.

1. Introduction

The first stage of the Aranea Project [Benko 2014; 2016; Benko & Zakharov 2016] has been targeted to creation of a family of dozen+ Giga-word web corpora for languages spoken in Slovakia and its neighbouring countries, as well as for the main foreign languages taught at Slovak universities. This stage is next to completed and the Aranea family currently contains corpora for 18 languages in (usually) two sizes, with some languages having also region-specific variants.

In parallel to building the corpora, works have been done to introduce this resource into teaching within the programmes of foreign language and translation studies at our University. After first four semesters of teaching, we would like to summarize some experiences with the newly introduced subject *Introduction to Corpus Linguistics*.

2. The *Aranea* Corpus Portal

While building the Aranea corpora needed a considerable hardware infrastructure (servers with a lot of RAM and free disk space), the corpus portal itself could be maintained with a moderate hardware configuration. In our case, a new hardware has been recently assigned to our Project — a quad-core virtual machine with 4 GB of main memory and 2 TB of disk space within the University sever cloud. The portal runs the *NoSketch Engine*² corpus manager under the *Ubuntu Linux* operating system. The decision for this corpus manager has been mainly motivated by its user-friendliness, rich set of features and ability to cope with very large (i. e., larger than 2 Gigaword) corpora. It is, however, worth noticing that the competing *CQPweb*³ system would be more user-friendly for the system administrator.

¹ This work has been partially financed by the Slovak KEGA Grant Agency, Project No. K-16-022-00.

² <https://nlp.fi.muni.cz/trac/noske>

³ <http://cwb.sourceforge.net/cqpweb.php>

The need to migrate to the virtual system has been accelerated by the crash of the data disk array at our (8-years old) own server. Our initial worries concerning the performance on the virtual machine were not approved and the overall speed of query operations seems to be even higher than those on the “real” machine.

The *Appendix I* shows the home page of our Portal.⁴

Our Portal offers two modes of operation. The *Guest Access* mode (without a password) allows users to work with the smaller (100 Megaword) editions of all corpora, while the *Full Access* mode requires a (free) registration by name and e-mail address. Besides having more corpora at hand, registered users can also profit from some extra corpus manager features, such as saving the default display parameters, creation of subcorpora, etc.

The Guest mode can be conveniently used during the first lessons of a course, as no previous setup is required to start querying the corpus. Though smaller corpora are only available, this is usually not an issue for corpus linguistics beginners. Moreover, 100 Megaword corpora are not really small, are they?

3. The Computer Lab

The minimal configuration for teaching a practically oriented corpus subject is a classroom with a good wireless connection where students can connect their own laptops or tablets. In optimal case, however, a computer lab with workstations having large screens is preferred. It is also important that the projector conveying the contents of the teacher’s monitor is able to do it in full resolution and project it at a sufficiently large screen.

Our computer lab contains 20+1 *MS-Windows* workstations with 21” screens. We have decided to use machines in “all-in-one” configuration requiring less table space than traditional desktop computers. As the corpus manager is fully accessible via a web browser, the only special arrangement was installation of different keyboard layouts for the respective languages. If needed, however, a virtual keyboard, such as that accessible on various web sites⁵, can be used.

4. Syllabus of the Course

To make maximal use of the corpora at hand, our new subject has been designed as a series of “hands-on” workshops, with most of the lessons con-

⁴ <http://unesco.uniba.sk/guest/index.html>

⁵ <http://translit.net/>

sisting of small research tasks to be performed by the students themselves. It has been also decided that the syllabus should be created as “language-independent” as possible, which would enable mixed-language groups.

The overall course syllabus is divided into three parts.

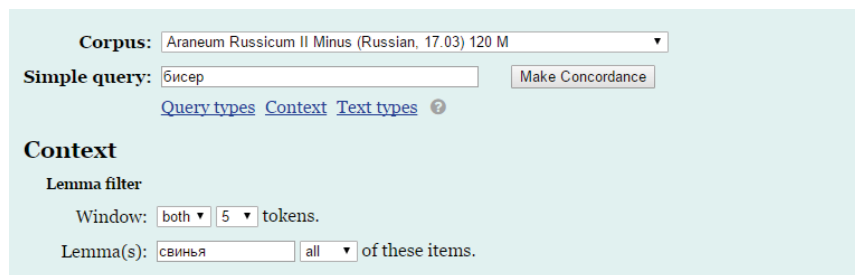
4.1 The first part (3–4 lessons) is a “theory-less” introduction into corpus query procedures in an “annotation-agnostic” way. During this stage, the students are shown the main differences between “linguistic querying” provided by corpus managers, and “information querying” provided via search engines. The main idea conveyed here is that “*Google is a very bad concordancer*” [Sharoff, 2006] and the ability to cope with morphological forms is really crucial for linguistic querying (not only) for languages with rich morphology.

The topics covered in the first part are:

1. Typing characters with foreign diacritics
2. Aranea portal in Guest mode
3. Basic queries: word form, lemma, phrase, “Simple query”
4. Frequency distributions and Context search

After the first part of the course, students are able to perform queries without having to know the details of more sophisticated search tools, such as Corpus Query Language.

For example, the context search can be conveniently used for looking for idioms. A query at *Fig. 1* will look for an expression containing keywords “*бисер*” (“pearl”) and “*свинья*” (“swine”) in any morphological form within a window of 5 words left/right.



The screenshot shows the Aranea corpus query interface. At the top, the 'Corpus' is set to 'Araneum Russicum II Minus (Russian, 17.03) 120 M'. Below this, the 'Simple query' field contains 'бисер'. To the right of the query field is a 'Make Concordance' button. Below the query field are links for 'Query types', 'Context', and 'Text types', along with a help icon. The 'Context' section is expanded, showing a 'Lemma filter' section. Under 'Lemma filter', the 'Window' is set to 'both' and '5' tokens. The 'Lemma(s)' field contains 'свинья' and is followed by a dropdown menu set to 'all' and the text 'of these items.'

Fig 1.

Fig 2 shows the result of this query operation.

Query **бисер** 797 > Positive filter (excluding KWIC) **свинья** 10 (0.08 per million) ⓘ

oshoworld....	Послушайте его: То, что свято, киньте собакам перед свиньями мечите бисер .
oshoworld....	Вы слышали противоположное высказывание: не бросайте собакам, не мечите бисер перед свиньями , потому что они не поймут.
lib.pushki...	Так этого ты от нас не добьешься, ибо мы, по завету Господа, воздержимся от того, чтобы давать святыню псам и метать чистый и светоносный, богоукрашенный бисер перед свиньями .
forum.deti...	Так зачем перед свиньями бисер метать?
minus5.ru ...	Не метать бисер перед свиньями , перестать быть хорошим для всех, прекратить лгать, лстыть, возводить напраслину, осуждать.
dal.by	Яркая иллюстрация к поговорке: "Метать бисер перед Свиньями ". шустер решил задницу свою прикрыть?Мол я вот правду пытался найти,а мне самого дезинформировали...Это на случай,что ополчение дойдёт до киева,хотя такой киевский запаведник фашизма лучше оставить,как зоопарк. ¶
realschool...	А.С.Пушкин в письме П.Вяземскому вовсе отказывает Чацкому в уме, ибо признак умного человека – знать кому и зачем ты это говоришь, а не метать бисера перед свиньями .
astrosyste...	Ведь новые сорта бисера тем же самым свиньям кидать не хочется, пока вы продолжаете видеть людей как свиней.
new.krasfa...	Перед началом мероприятия Василий Борисов - гриль-мен гастробаба " Свинья и бисер " показал мастерское приготовление на гриле говядины, телятины и свинины.
new.krasfa...	Жюри чемпионата стейков состояло из профессионалов лучших ресторанных заведений Красноярска: Лавренович Сергей - председатель жюри, шеф-повар ресторана «Суриков» Борисов Василий – гриль-мен гастробаба « Свинья и бисер » и победитель краевого Чемпионата стейков 2011 года, который проводила Ассоциация гостеприимства в апреле в МВДЦ «Сибирь», Лаевский Александр – гриль – мен Бара «Харлей».

Fig. 2.

4.2. The intermediate lesson is a lecture introducing the basic concepts of corpus linguistics, Web as Corpus (WaC) technology [Kilgarrieff, 2001; Kilgarrieff and Grefenstette, 2003] and corpus annotation, both external and internal (linguistic).

4.3. The remaining lessons progressively cover topics as follows: Corus Query Language (CQL)

1. Morphosyntactic annotation, Slovak *SNK tagset*⁶
2. *Araneum Universal Tagset (AUT)*⁷
3. “Native” tagsets for other (foreign) languages
4. Regular expressions and their use in corpus queries
5. Collocations and statistical association measures
6. Parallel corpora: *InterCorp*⁸ and *Treq*⁹

⁶ http://korpus.sk/morpho_en.html

⁷ http://aranea.juls.savba.sk/aranea_about/aut.html

⁸ <https://ucnk.ff.cuni.cz/intercorp/>

⁹ <http://treq.korpus.cz/>

4.4. The course is completed by a final assignment having a form of a “crowdsourcing” project. Each student is given a spreadsheet containing 1,000 tokens derived from a frequency list of word forms from the latest version of *Araneum Slovacum* that have not been recognized by the morphological analyzer. Their task is to lemmatize and/or correct the PoS tag for the respective items in the table. As each data file is being processed by two independent annotators, it can be later evaluated and used to amend the morphological lexicon during the next round of tagging.

5. The Textbook

While the previous text describes the already materialized results, the creation of a textbook is currently “work in progress”.

The need for a new textbook is dictated not only by absence of any Slovak educational resource on the topic, but also lack of suitable textbook in (say) English, that would cover:

- Introduction into corpus linguistics in a compact form
- Problems of morphosyntactic annotation of morphologically-rich languages, such as Slovak
- Problems of creation and using web corpora
- Direction to use the NoSketch Engine corpus manager

We would like to take inspiration from the unique book of James Thomas [Thomas, 2016]. The planned publication is to appear both in paper and electronic form.

References

1. Benko V. (2014), Aranea: Yet another Family of (Comparable) Web Corpora. In: Text, Speech, and Dialogue. 17th International Conference, TSD 2014 Brno, Czech Republic, September 8–12, 2014, Proceedings. Ed. P. Sojka et al. Cham; Heidelberg; New York; Dordrecht; London: Springer, 2014, pp. 247–256. ISBN 978-3-319-10816-2.
2. Benko V. (2016), Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016, pp. 4245–4248. ISBN 978-2-9517408-9-1.
3. Benko V., Zakharov V. P. (2016), Very large Russian corpora: new opportunities and new challenges. Rec. Alexej Vladimirovič Bajtin, Igor Michajlovič Boguslavskij. In: Kompjuternaja lingvistika i intellektualnyje tehnologii: po materialam meždunarodnoj konferencii “Dialog” (2016), vypusk 15 (22). Otv. red. A. A. Belkina. Moskva: Rossijskij gosudarstvennyj humanitarnyj universitet, 2016, pp. 79–93. ISSN 2221-7932.

4. Kilgarriff A. (2001), Web as corpus. In: P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Kioja (eds.) *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster (29 March — 2 April 2001). Lancaster: UCREL, pp. 342–344.
5. Kilgarriff A., Grefenstette G. (2003), Introduction to the Special Issue on the Web as Corpus. In: *Computational Linguistics*. E-ISSN 1530-9312, 2003, vol. 29, no. 3, pp. 333–347.
6. Sharoff, S. (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In: *WaCky! Working Papers on the Web as Corpus*. ISBN 88-6027-004-9, Bologna: Gedit Edizioni, 2006. pp. 63–98.
7. Thomas J. (2016), *Discovering English with Sketch Engine*, 2nd. edition. Versatile, 2016.

Vladimír Benko

Comenius University in Bratislava (Slovakia)

Slovak Academy of Sciences

E-mail: vladimir.benko@uniba.sk

Anna Butašová

Comenius University in Bratislava (Slovakia)









































































































E-mail: anna.butasova@uniba.sk

Comenius University in Bratislava

UNESCO Chair in Plurilingual and Multicultural Communication

Aranea Project Main NoSketch Engine Site (Guest Access) 

Free registration is required for work with the *Maius* and *Maximum* class of corpora.
To register, please fill in and submit [this form](#).

Language	Aranea Corpora	Minus 120 M	Maius 1,20 G	Maximum
Arabic (not tagged yet)	Araneum Arabicum	  Q	  Q *	
Bulgarian	Araneum Bulgaricum	  Q	  Q	
Chinese (simplified script)	Araneum Sinicum	  Q	  Q	
Czech	Araneum Bohemicum	  Q	  Q	5,17 G   Q
Dutch	Araneum Nederlandicum	  Q	  Q	
English	Araneum Anglicum	  Q	  Q	
English (<i>African TLDs</i>)	Araneum Anglicum Africanum	  Q	  Q	
English (<i>Asian TLDs</i>)	Araneum Anglicum Asiaticum	  Q	  Q	
Finnish	Araneum Finnicum	  Q	  Q	
French	Araneum Francogallicum	  Q	  Q	
French (<i>African TLDs</i>)	Araneum Francogallicum Africanum	  Q	  Q *	
Georgian (not tagged yet)	Araneum Georgianum	  Q		
German	Araneum Germanicum	  Q	  Q	
Hungarian	Araneum Hungaricum	  Q	  Q	
Italian	Araneum Italicum	  Q	  Q	
Polish	Araneum Polonicum	  Q	  Q	
Portuguese	Araneum Portugalicum	  Q	  Q	
Russian	Araneum Russicum	  Q	  Q	13,7 G   Q
Russian (<i>Russia-only TLDs</i>)	Araneum Russicum Russicum	  Q	  Q	
Russian (<i>non-Russia TLDs</i>)	Araneum Russicum Externum	  Q	  Q	
Slovak	Araneum Slovaccum	  Q	  Q	2,68 G   Q
Spanish	Araneum Hispanicum	  Q	  Q	
Swedish	Araneum Suedicum	  Q	  Q	
Language	Other Corpora	Minus 120 M	Maius 1,20 G	Maximum
Arabic (not tagged yet)	Ajdir Arabicum	  Q *		
Croatian	Zagrabia Croatica (hrWaC)	  Q	  Q	
Slovene	Aemona Slovena (ccGigafida)	  Q		

* Parvum (< 120 M) and Medium (< 1,2 G) class corpora are only available for some languages.

Appendix I.