

## Aranea Web-Crawled Corpora: A Source of Diverse and Unified Language Data for NLP

Vladimír Benko

Comenius University Science Park, UNESCO Chair in Plurilingual and Multicultural Communication  
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics

Relevant Working Groups: **WG1, WG3, WG4**

In the framework of the Aranea Project (Benko, 2014), a family of web-crawled corpora for languages taught at Slovak universities has been built since 2013. As of the beginning of 2023, more than two dozen different languages have been processed already, with several more being in preparation.

All corpora are processed by (almost) identical pipeline, using the tools as follows:

- SpiderLing<sup>1</sup> (Suchomel and Pomikálek, 2012) web crawler optimized for downloading textual data in a specified language that also incorporates utility for on-the-fly language identification and removal of 100% duplicate documents.
- Unitok<sup>2</sup> (Michefeit et al., 2014) universal tokenization utility.
- Onion<sup>3</sup> (Pomikálek, 2011) utility for detection and removal of semi-duplicate contents at the document or paragraph level.
- Set of language-specific filters (Benko, 2016) for (secondary) language identification (to remove documents that escaped the SpiderLing language detection procedure), deletion of texts with errors in character encoding, “non-discursive” texts (e.g., containing tables with too many numerical values), web spam, etc.
- Tools for lemmatization and PoS tagging – TreeTagger<sup>4</sup> (Schmid, 1994), UDPipe<sup>5</sup> (McDonald et al., 2012; Straka et al., 2016) and/or CSTlemma<sup>6</sup> (Jongejan and Dalianis, 2009) are used when available, complemented by other tools for some languages, such as MorphoDiTa<sup>7</sup> (Straková et al., 2014), Apertium<sup>8</sup> (Khana et al., 2021), Hunpos<sup>9</sup> (Halácsy et al., 2007), and even Hunspell<sup>10</sup> (Németh et al., 2004). Since recently, new corpora are tagged and lemmatized by an “ensemble” approach, i.e. independently using all tools available and aggregating their outputs. We expect to be able to reprocess all older corpora utilizing this approach in the near future as well.
- NoSketch Engine<sup>11</sup> (Rychlý, 2007) corpus manager.

In our efforts to make the corpora as “comparable” as possible, most design decisions were applied in a project-wide manner, as follows:

- The tokenization policy is always “compatible”, even in cases when this may be suboptimal for the respective tagger(s)/lemmatizer(s), which means:
  - Period-final abbreviations (such as “Mr.”, “approx.”) are treated as single tokens, though sequences of such abbreviations (“U.S.A.”, “Ph.D.”) are split.
  - Hyphenated words (“multi-lingual”, “Austro-Hungarian”), even if they contain digits (“64-bit”, “K-12”), are treated as single tokens.
  - E-mail addresses (“foo@bar.com”), URLs (“https://google.com”), hashtags (“#WeLoveYou”), etc., are treated as single tokens.
  - Multiple occurrences of the same punctuation (“...”, “????”) or special graphic characters “☺☺☺☺” are treated as single tokens. Sequences of different ones, however, are split.

---

<sup>1</sup> <https://corpus.tools/wiki/SpiderLing>

<sup>2</sup> <https://corpus.tools/wiki/Unitok>

<sup>3</sup> <https://corpus.tools/wiki/Onion>

<sup>4</sup> <https://www.cis.lmu.de/~schmid/tools/TreeTagger/>

<sup>5</sup> <https://ufal.mff.cuni.cz/udpipe/1>

<sup>6</sup> <https://cst.dk/online/lemmatiser/uk/>

<sup>7</sup> <https://ufal.mff.cuni.cz/morphodita>

<sup>8</sup> [https://wiki.apertium.org/wiki/Main\\_Page](https://wiki.apertium.org/wiki/Main_Page)

<sup>9</sup> <https://github.com/mivog/hunpos>

<sup>10</sup> <http://hunspell.github.io/>

<sup>11</sup> <https://nlp.fi.muni.cz/trac/noske>

- “Native” tagsets are mapped to an “universal” Araneum tagset<sup>12</sup> (denoting the main word classes only).
- Special attribute indicating the success status of the morphological lexicon lookup is added to each token (if the respective tagger provides for such information).
- The resulting corpora are sampled to get two basic “compatible” sizes (125 million and 1.25 billion tokens, respectively, which approximately gives 100 million and 1 billion words). For some “large” languages even larger corpora are created, reaching usually approx. 10 billion tokens.
- Corpora bear “language-neutral” (Latin) names denoting language, variety, and size.
- All corpora are accessible for online access via the NoSketch Engine corpus manager at the Aranea Corpus Portal<sup>13</sup>
- Source format of the corpora, both in raw text format, and in fully tagged or lemmatized, have often been provided for non-commercial purposes to other research groups.

We would like present to the *UniDive* community the opportunity of using (in multiple WGs) our multi-lingual data in several unified formats for various purposes.

The Appendix shows the home page of the Aranea Corpus Portal, indicating the list of languages already published for online use. In preparation, there are languages as follows: Danish, Belarusian, Kazakh, Tatar, Korean, Slovene, Croatian, Serbian, and Maltese. Some of them will be already available before the Paris *UniDive* event.

## References

- 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. Eds. P. Sojka et al. Cham – Heidelberg – New York – Dordrecht – London : Springer, 21–29. ISBN 978-3-319-10816-2.
- Benko, V. 2016. Feeding the “Brno Pipeline”: The Case of Araneum Slovacum. In: A. Horák, P. Rychlý, A. Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 19–27, 2016. © Tribun EU 2016
- Halácsy, P., Kornai, A. and Oravecz, Cs. 2007. HunPos: an open source trigram tagger. Proceedings of the ACL 2007 Demo and Poster Sessions, pages 209–212, Prague, June 2007.
- Jongejan, B. and Dalianis, H. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore : Association for Computational Linguistics, pp. 145-153.
- Khanna, T., Jonathan N. Washington, J. N., Tyers, F. M., Bayatli, S., Swanson, D. G., Pirinen, T. A., Tang I. and Alòs i Font, H. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. Machine Translation <https://doi.org/10.1007/s10590-021-09260-6>
- McDonald, R., Nivre, J., Quirnbach-Brundage, Y., Goldberg, Z., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckstrom, O., Bedini, C., Bertomeu Castelló, N., Lee, J. 2013. Universal Dependency Annotation for Multilingual Parsing. In Proceedings of ACL.
- Michelfeit, J., Pomikálek, J., Suchomel, V. 2014. Text Tokenisation Using unitok. In 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU, pp. 71-75.
- Németh, L., Viktor Trón, V., Halácsy, P., Kornai, A., Rung, A. and Szakadát, I. 2014. Leveraging the open source ispell codebase for minority language analysis. SALT MIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages.
- Pomikálek, J. 2011. Removing boilerplate and duplicate content from web corpora. PhD thesis, Masaryk University, Faculty of informatics, Brno, Czech Republic.

<sup>12</sup> [http://unesco.uniba.sk/aranea\\_about/aut.html](http://unesco.uniba.sk/aranea_about/aut.html)

<sup>13</sup> <http://unesco.uniba.sk/guest/>, <http://aranea.iuls.savba.sk/guest/>

Rychlý, P. 2007. Manatee/Bonito – A Modular Corpus Manager. 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. p. 65–70.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester.







































































































































































Straka, M., Hajič J., Straková J. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia.

Straková J., Straka M. and Hajič J. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13-18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Suchomel, V., Pomikálek, J. 2012. Efficient Web Crawling for Large Text Corpora. In Adam Kilgarriff, Serge Sharoff. Proceedings of the seventh Web as Corpus Workshop (WAC7). Lyon, pp. 39-43.

## Appendix

### Aranea Project Mirror Site powered by NoSketch Engine

Language	Aranea Corpora	Minus 125 M	Maius 1.25 G	Maximum
Arabic (not tagged yet)	Araneum Arabicum	 		  978 M *
Bulgarian	Araneum Bulgaricum	 	 	
Chinese (simplified script)	Araneum Sinicum	 	 	
Czech	Araneum Bohemicum IV	 	 	  7.10 G
Dutch	Araneum Nederlandicum	 	 	
English	Araneum Anglicum II	 	 	  11.4 G
English ( <i>Africa</i> )	Araneum Anglicum Africanum	 	 	
English ( <i>Asia</i> )	Araneum Anglicum Asiaticum	 	 	
Estonian	Araneum Estonicum II	 	 	
Finnish	Araneum Finnicum	 	 	
French	Araneum Francogallicum III	 	 	  10.9 G
French ( <i>France</i> )	Araneum Francogallicum Gallicum	 	 	  3.29 G
French ( <i>Belgium</i> )	Araneum Francogallicum Belgicum	 		  365 M *
French ( <i>Canada</i> )	Araneum Francogallicum Canadiense II	 		  406 M *
French ( <i>Switzerland</i> )	Araneum Francogallicum Helveticum	 		  229 M *
French ( <i>Africa</i> )	Araneum Francogallicum Africanum II	 		  310 M *
Georgian	Araneum Georgianum	 		  254 M *
German	Araneum Germanicum III	 	 	  8.91 G
German (Germany)	Araneum Germanicum Germanicum	 	 	  5.59 G
German (Austria)	Araneum Germanicum Austriacum	 		  441 M *
German (Switzerland)	Araneum Germanicum Helveticum	 		  381 M *
Hungarian	Araneum Hungaricum	 	 	
Italian	Araneum Italicum	 	 	
Latin	Araneum Latinum			  109 M *
Latvian	Araneum Lettonicum	 		  671 M *
Norwegian	Araneum Norvegicum II Beta	 	 	  3.53 G
Persian	Araneum Persicum Beta	 	 	  3.09 G
Polish	Araneum Polonicum	 	 	
Portuguese	Araneum Portugalicum	 	 	
Romanian	Araneum Dacoromanicum	 	 	
Russian	Araneum Russicum III	 	 	  19.8 G
Russian ( <i>Russia</i> )	Araneum Russicum Russicum	 	 	
Russian ( <i>non-Russia</i> )	Araneum Russicum Externum	 	 	
Slovak	Araneum Slovaccum VI Beta	 	 	  4.34 G
Spanish	Araneum Hispanicum	 	 	
Swedish	Araneum Suedicum	 	 	
Ukrainian	Araneum Ucrainicum Beta	 	 	
Uzbek	Araneum Uzbecicum	