Two Years of *Aranea*: Increasing Counts and Tuning the Pipeline

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, SK-81101 Bratislava, Slovakia
E-mail: vladob@juls.savba.sk

Abstract

The *Aranea* Project is targeted at creation of a family of Gigaword web-corpora for a dozen of languages that could be used for teaching language- and linguistics-related subjects at Slovak universities, as well as for research purposes in various areas of linguistics. All corpora are being built according to a standard methodology and using the same set of tools for processing and annotation, which – together with their standard size and– makes them also a valuable resource for translators and contrastive studies. All our corpora are freely available either via a web interface or in a source form in an annotated vertical format.

Keywords: web corpora, open-source and free tools

1. Introduction

During the last decade, creation of web-derived corpora has been recognized as an effective way of obtaining language data in situations where building traditional corpora would be too costly or too slow (Baroni et al., 2004, 2009; Saharoff, 2006; Jakubíček et al., 2013; Schäfer & Bildhauer, 2013). The recently released open-source tools for this purpose made it possible (even for low-funded educational and research institutions in Central Europe) to undertake projects of creation large-scale web corpora.

Within the framework of our *Aranea* project, a family of Gigaword web corpora for some two dozens of languages is being created. In our previous paper (Benko, 2014a), we presented the state of the Project after its first year, introduced the methodology and tools used, and demonstrated some results for the first eight languages processed (i.e., *Dutch*, *English*, *French*, *German*, *Polish*, *Russian*, *Slovak* and *Spanish*). In this paper, we would like to show the new developments within the *Aranea* family of corpora during the past year; summarize the results and discuss the possible future of the Project.

2. The *Aranea* Project

The main reason why we decided to start our Project was the lack of suitable corpora that could be used for teaching purposes at our Universities. When starting our Project in Spring 2013, the main design decisions of were set as follows:

- The corpora family will be "Slovak-centric" (i.e., it will consist of languages used and/or taught in Slovakia and its neighbouring countries;
- All corpora will bear "language-neutral" (Latin) names ;
- They will be designed as "comparable" (i.e., the same size(s), methodology of processing and annotation);
- Will be (freely) available via a web interface;
- A compatible sketch grammar for the Sketch Engine will be written for each language.

The second year of the *Aranea* Project was dedicated to completing the language list of the "inner circle" (i.e., Czech, Hungarian and Ukrainian – the latter, however, without any annotation yet), and adding four foreign languages taught at our University (Chinese, Finnish, Italian, and Portuguese). For English and Russian, territory-specific corpora have also been created by restricting web crawling to national top-level domains of countries where English/Russian have an official or semi-official status.

3. The Processing Pipeline

The procedures:

(1) Crawling web data by *SpiderLing* (Suchomel & Pomikálek, 2012); seed URLs obtained by *BootCaT* (Baroni & Bernardini, 2004) using list of 100 medium-frequency seed words from the respective language. Language identification, boilerplate removal, conversion HTML to text and basic deduplication (removing exact duplicates) is performed by *SpiderLing* on the fly.

(2) Filtration: fixing irregularities in document headings, removing surviving HTML markup, and documents with misdetected language and misinterpreted character encoding encoding. Filtration tools have been developed in a step-by-step manner based on analysis of already processed corpus data.

(3) Tokenization by *Unitok* universal tokenizer (Michelfeit, Pomikáek & Suchomel, 2014) using the respective language definition. For Chinese, the word segmentation procedure was also used (Sharoff, 2015).

(4) Segmentation on sentences using a rudimentary algorithm based on regular expressions.

(5) Document-level deduplication by *Onion* (Pomikálek, 2011) using 5-grams with similarity threshold set to 0.9. Detected partial duplicate documents removed. Deduplication parameters were chosen after an experiment with different settings had been performed (Benko, 2013).

(6) Conversion of utf-8 punctuation to ASCII. This step has been introduced in attempt to make the source text more compatible with the respective language model used for tagging. As the data used for training

taggers mostly originated in pre-utf-8 times, the language models produced by training are not aware of the punctuation characters introduced by

(7) Tagging: Most of the corpora have been tagged by *Tree Tagger* (Schmid, 1994). The Hungarian data has been tagged by the Research Institute for Linguistics at the Hungarian Academy of Sciences.
(8) Recovering the original punctuation.
(9) Marking out-of-vocabulary tokens and ambiguous lemmas.
(10) Mapping native tagset to *Araneum Universal Tagset* (*AUT*). This provides an alterantiv
(11) Paragraph-level deduplication, again by Onion (5-grams, similarity threshold 0.9), near-duplicate paragraphs marked.
(12) Sampling to receive two standard sizes: 1.2 billion and 120 million tokens.
(13) Processing ("compiling") by *NoSketch Engine*, as well as by *Sketch Engine* using the compatible sketch grammar (Benko, 2014b).

Out of the 13 steps listed, only 5 involve language specific processing, most notably the tagging and tagset mapping. To speed up the most compute-intensive operations (tokenization and tagging), they can performed in several parallel processes to take the advantage of the multiple-core processor of our server.

The results of the processing steps (1) to (5) are summarized in Table 1

Corpus name	*Language code*	*Crawling time (days)*	*Downloaded*		*Deduplicated*		*Dedup ratio (%)*
			*Docs (*10^6)*	*Tokens (*10^9)*	*Docs (*10^6)*	*Tokens (*10^9)*	
Bohemicum	*cs*	24	5.803	n/a[1]	3.593	3.199	n/a
Germanicum	*de*	4	4.936	2.527	3.888	1.998	79.07
Anglicum	*en*	4	3.535	3.321	2.808	2.694	81.12
Anglicum Africanum	*en.af*	4	2.884	2.459	1.927	1.412	57.42
Anglicum Asiaticum	*en.as*	4	4.753	3.172	3.318	2.005	63.21
Hispanicum	*es*	4	3.415	2.440	2.520	1.864	76.39
Finnicum	*fi*	16	4.531	2.813	3.018	1.628	57.87
Francogallicum	*fr*	7	5.945	3.915	4.367	2.962	75.66
Hungaricum	*hu*	8	1.894	1.629	1.802	1.113	68.32
Italicum	*it*	26	6.363	3.142	4.944	2.379	75.72
Naderlandicum	*nl*	2	3.912	2.151	2.881	1.592	74.01
Polonicum	*pl*	3	4.349	2.718	2.664	1.667	61.33
Portugallicum	*pt*	7	3.184	1.807	2.417	1.381	76.43
Russicum	*ru*	2	1.490	1.726	1.313	1.519	88.01
Russicum Externum	*ru.ex*	4	2.835	1.790	2.317	1.419	79.27
Russicum Russicum	*ru.ru*	2	3.860	2.571	3.382	2.664	86.85
Slovacum	*sk*	263	7.221	4.412	4.023	2.417	50.57
Ucrainicum	*uk*	12	2.721	2.025	1.907	1.313	69.88
Sinicum	*zh*	4	1.869	1.624	1.528	2.317	77.22

Table 1: Crawling, Tokenization and Deduplication

In the terminology of *SpiderLing*, a "document" is the contents of a single web page converted from HTML to "pure text" format.

The crawling was usually performed in several 48-hour sessions (with longer sessions for "small" languages) to get about two Gigawords of raw data, that could be further subject to filtration and deduplication. As it can be seen, for "large" languages this has been achieved in about four days crawling, with the first Gigaword usually obtained after 24 hours. The situation with the "smaller" languages, however, was much less favourable. The large number of crawling days for Slovak was the result of an attempt to produce a multi-Gigaword corpus of that language (i.e. using the strategy "as much as can get") – here we probably cope with the problem of limits of Slovak texts available from the web by the technology used.

The deduplication scores (showing the proportion of the tokens that remained after deleting the partially duplicate documents) indicate that for "smaller" languages more duplicate contents has been downloaded, which is again an obstacle in building large-scale web corpora for those languages. Nevertheless, a Gigaword corpus could be obtained for all participating languages.

4. Morphosyntactic Annotation

With the exception of Czech and Hungarian (and Ukrainian, where no tagger was available), all other corpora have been tagged by Tree Tagger. The speed of tagging by Tree Tagger mostly depended on the complexity of the respective tagset, ranging from 8 hours per Gigaword (for English) to more than 36 hours per Gigaword for Russian. Tree Tagger has proved to be a reliable, robust and fast tool. The quality of tagging, however, varied from one language to another, and depended not only on the number of tags in the respective tagset, but also on the coverage of the lexicon and the size of the

manually tagged training data – an information that was not available for many language models supplied with Tree Tagger.

For languages with complex inflectional morphology, such as Slovak and Russian, the basic word classes were typically assigned correctly, while the precision of assignment of subcategories was negatively influenced by the tagging strategy used by Tree Tagger, i.e, that only previous tags are being considered in calculations. This often fails in situations like agreement of *adjective + noun* combinations in gender, number and case, where the adjective subcategories can only be disambiguated according to those of the following noun. The process of tagging is summarized in Table 2.

Language code	*Tagger*	*Language model timestamp (YY/MM)*	*Tagset*
cs	MorphoDiTa	13/11	Jan Hajič
de	Tree Tagger	12/12	STTS
en	Tree Tagger	15/02	Penn
es	Tree Tagger	14/05	simplified CRATER
fi	Tree Tagger	14/07	simplified Finnish Treebank
fr	Tree Tagger	10/01	Achim Stein French
hu	Hunpos	n/a	HNC
it	Tree Tagger	14/10	Achim Stein Italian
nl	Tree Tagger	14/09	?
pl	Tree Tagger & Morfeusz	14/06 & 08/02	NKJP
pt	Tree Tagger	12/03	EAGLES Portuguese
ru	Tree Tagger	12/06	MULTEXT-East Russian
sk	Tree Tagger	15/01	SNK
uk	n/a		
zh	Tree Tagger	07/07	LCMC

Table 2: Morphosyntactic annotation

Language models available for Tree Tagger do not provide for version numbers, though they often exist in several versions. We decided to identify them by the timestamp of the respective parameter files

MorphoDiTa in a new tagger developed at the Charles University in Prague (Straková et al, 2014). The language models available include Czech (and English) at present, though it can be expected that new languages will appear soon. Besides the much better quality of tagging for morphologically rich languages, another advantage of *MorphoDiTa* is the speed of tagging – the three-Gigaword Czech data was tagged in some 9 hours.

The most problematic language in our set was Hungarian. As the Hunpos tagger is not utf-8 compliant, special pre- and post-processing was necessary so that the full contents of our data be retained – this has been performed for our Project at the Research Institute for Linguistics in Budapest.

5. The Universal Tagset

All native tagsets have been subsequently mapped into the *Araneum Universal Tagset* (*AUT*) to be used within the compatible sketch grammar (Benko, 2014b). This tagset can be described as "PoS-only" and contains tags for 11 traditional word classes plus additional tags that accommodate information from the respective "native" tagsets. The list of *AUT* tags is summarized in Table 3.

Tag	*PoS*
Dt	determiner/article
Nn	noun
Aj	adjective
Pn	pronoun
Nm	numeral
Vb	verb
Av	adverb
Pp	preposition
Cj	conjunction
Ij	interjection
Pt	particle
Ab	abbreviation/acronym
Sy	symbol
Nb	number
Xx	other (content word)
Xy	other (function word)
Yy	unknown/alien/foreign
Zz	punctuation

Table 3: Araneum Universal Tagset

6. Corpus Access

For on-line use all corpora are sampled to receive the uniform 1.2-billion-token *Maius* size (i.e., approximately 1 billion of words), as well as the 120-million-token *Minus* size (for educational purposes).

The annotated data of all corpora are further being indexed ("compiled") by the *NoSketch Engine*[1] (Rychlý, 2007) corpus management system at the Web Corpora Portal of our Department[2]. The smaller *Minus* series corpora are accessible in without any password, accounts for full access are available after a free registration.

The *Aranea* family of corpora is also being hosted at the Institute of the Czech National Corpus web site[3].Users having an account at the *Sketch Engine*[4] site can also find our corpora there.

The source versions of the corpus data are available for download (for research and educational purposes). Note, however, that the copyright status of the data is not clear and users from countries where this might cause legal problems will have to solve this issue themselves.

7. Further Developments

We would like to develop the *Aranea* Project in several directions:

(1) Complete the language list with languages needed in foreign language teaching at our University (Bulgarian, Romanian, Modern Greek, Swedish, Japanese, and Korean) and provide the region-specific variants for "large" languages (American vs. Iberian Spanish, Canadian and African French, non-Germany German);

(2) Improve the processing pipeline by incorporating user feedback (abbreviation lists for better tokenization, language-specific filtration, additional lexicon entries based on analysis of most frequent out-of- vocabulary tokens);

(3) Provide additional layers of alternate morphosyntactic annotation for languages where more than one tagger and/or language model is available.

8. Conclusion

After the second year of our Project, we can conclude that creation of Gigaword corpora by open-source and free tools is feasible, and requires (almost) no additional programming. The processing pipeline has been standardized, the language-independent parts have been identified, with greatly increased the productivity of the whole process. It has been also shown, that increasing the sizes of the corpora, especially for small languages, will require much crawling time and processing resources.

9. Acknowledgement

This research has been, in part, funded by the VEGA Grant Agency (Grant Number 2/0015/14).

10. Bibliographical References

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In: Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisbon

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43(3), pp. 209–226.

Benko, V. (2013). Data Deduplication in Slovak Corpora. In Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning, pp. 27–39. Lüdenscheid: RAM-Verlag.

Benko, V. (2014a). Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. pp. 257–264.

Benko, V. (2014b). Compatible Sketch Grammars for Comparable Corpora. In Andrea Abel, Chiara Vettori, Natascia Ralli (Eds.): Proceedings of the XVI EURALEX International Congress: The User In Focus. 15–19 July 2014. Bolzano/Bozen: Eurac Research, 2014. pp. 417–430.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. (2013). The TenTen Corpus Family. In: Proc. Int. Conf. on Corpus Linguistics, Lancaster.

Michelfeit, J., Pomikálek, J., and Suchomel, V. (2014). Text Tokenisation Using unitok. In Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 71–75, 2014. Brno: NLP Consulting 2014.

Oravecz, Cs., Váradi, T., Sass, B. (2014). The Hungarian Gigaword Corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik: ELRA, pp. 1719–1723.

Pomikálek, J. (2011). Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno.

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

Schäfer, R., Bildhauer, F. (2013). Web Corpus Construction. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Sharoff, S. (2006). Creating General-Purpose Corpora Using Automated Search Engine Queries. In: Broni, M. and S. Bernardini (Eds.): WaCky! Working Papers on the Web as Corpus. pp. 63—98. Gedit Edizioni: Bologna.

Sharoff, S. (2015). Parsonal Communication.

Straková J., Straka M., 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.

Suchomel V., Pomikálek J. (2012). Efficient Web Crawling for Large Text Corpora. In: 7th Web as Corpus Workshop (WAC-7), Lyon, France.

[1] http://nlp.fi.muni.cz/trac/noske
[2] http://unesco.uniba.sk/guest/index.html
[3] https://kontext.korpus.cz/
[4] http://www.sketchengine.co.uk/