Araneum Nederlandicum Maius
A New Family Member

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Comenius University in Bratislava, UNESCO Chair in Translation Studies
vladob@juls.savba.sk

Aranea* — A family of (comparable) web corpora

- **Slovak-centric** (languages spoken and/or taught in Slovakia and the neighbouring countries)
- Crawled and pre-processed by ***SpiderLing*** at (approximately) the same time
- Language-independent filtration by the same tools
- Language-dependent filtration by the same methodology
- Compatible tokenization strategy
- Morpho-syntactically annotated by free tools (***Tree Tagger***, etc.)
- Sentence-segmented & sentence-level deduplicated, duplicate sentences marked
- Word sketches with compatible sketch grammars
- "Language-neutral" (Latin) names denoting the language and size

---

* ***Araneum*** (pl. *Aranea*, n.) is the Latin expression for (cob)web

Four sizes for each corpus planned

- **Maius** (greater) … basic version, approx. 1.2 billion tokens
- **Minus** (smaller) … 10 % sample of Maius (for teaching purposes)
- **Minumum** (minimal) … 1 % sample of Maius (for toolchain and sketch grammar experiments)
- **Maximum** (maximal) … as much as we can get

- Eight Aranea family members available by now
  - **Araneum Russicum** (Russian)
  - **Araneum Francogallicum** (French)
  - **Araneum Germanicum** (German)
  - **Araneum Hispanicum** (Spanish)
  - **Araneum Polonicum** (Polish)
  - **Araneum Nederlandicum** (Dutch)
  - **Araneum Anglicum** (English)
  - **Araneum Slovacum** (Slovak)
- In preparation
  - **Araneum Bohemicum** (Czech)
  - **Araneum Hungaricum** (Hungarian)

- The "Minus" versions accessible at `http://sketchengine.co.uk`

Compatible sketch grammars

- Common set of rules for all languages
- Fixed order of tables in word sketches
- Rule names represent collocational relationships (i.e. not syntactic)
- Word class (PoS) of keyword is not indicated (each rule works for any PoS)

Araneum Nederlandicum

- Crawled 16–18 November 2013 (48 hours) by ***SpiderLing*** (includes encoding and language detection module and boilerplate removal tool)

TLD	%
nl	73,90
be	10,06
com	8,79
net	1,64
org	1,47

TLD	%
eu	1,14
nu	1,02
info	0,67
other	1,31
total	**100,00**

- **14 GB** of document-level deduplicated text
- **3.9 M** documents, 48.1 M paragraphs
- **2.15 G** tokens
- tagged by ***Tree Tagger*** (OOV rate **8.21 %**)
- punctuation & special chars normalized and retagged
- Dutch tagset mapped into ***Araneum universal tagset***

tokens	share	tag	PoS
236320119	10.98%	**Dt**	determiner
511447021	23.77%	**Nn**	noun
174547408	8.11%	**Aj**	adjective
142108486	6.60%	**Pn**	pronoun
57497923	2.67%	**Nm**	numeral
283587979	13.18%	**Vb**	verb
128404111	5.97%	**Av**	adverb
219821945	10.21%	**Pp**	preposition
118408377	5.50%	**Cj**	conjunction
2706822	0.13%	**Ij**	interjection
22261630	1.03%	**Pt**	particle
254942011	11.85%	**Zz**	punctuation

Araneum Nederlandicum Maius

- downsized to **1.2 G** tokens
- segmented on sentences: **71.6 M** sentences
- sentence-level deduplicated by fingerprint method (punctuation, special graphics chars and numbers ignored in hash generation)

sentences	**count**	**%**
non-duplicate	44844413	62.62
duplicate	**26770074**	**37.38**
total	**71614487**	**100.00**

tokens	**count**	**%**
non-duplicate	826247099	68.85
duplicate	**373753738**	**31.15**
total	**1200000837**	**100.00**

- tokens in duplicate sentences marked
- compatible Dutch sketch grammar written
- processed by **Sketch Engine**

References:

Benko, V. (2013): *Data Deduplication in Slovak Corpora*. In: SLOVKO 2013. Proceedings of the Seventh International Conference, Bratislava, Slovakia, 13–15 November 2013. (Forthcoming.)

Benko, V. (2013): *Compatible Sketch Grammar Experiment*. Proceedings of the International Conference Corpus Linguistics – 2013, June 25–27, 2013, St. Petersburg. pp. 21–29.

Kilgarriff, A. et al. 2004. *The Sketch Engine*. In: G. Williams and S.Vessier (eds.), Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6-10, 2004. Lorient : Université de Bretagne-Sud, pp. 105–116.

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

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

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

Handouts available at
`http://milo.juls.savba.sk/~vladob/20140116_leiden.zip`