

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

Vladimír Benko | [vladimir.benko@uniba.sk](mailto:vladimir.benko@uniba.sk)











































































































































































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

- A family of comparable corpora for 30+ languages
- Crawled and pre-processed by “Brno pipeline”:  
SpideLing for crawling (includes justText and chared modules)  
Onion for deduplication & Unitok for tokenization
- Compatible tokenization policy for all languages
- Custom tools for language-dependent filtration
- FLOSS tools for lemmatization and PoS Tagging
- Ensemble approach adopted for newer corpora
- “Language-neutral” (Latin) names
- 2 basic sizes (125 M and 1.25 G tokens), “as much as can get” size for some languages
- Free online access via a NoSketch Engine corpus manager
- Source data available for non-commercial purposes
- <http://aranea.juls.savba.sk> | <http://unesco.uniba.sk>

## Comenius University in Bratislava UNESCO Chair in Plurilingual and Multicultural Communication

Aranea Project Main Site powered by NoSketch Engine (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 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
Danish	Araneum Danicum Beta	 	 	
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 Portugallicum	 	 	
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 Slovacum VI Beta	 	 	  4.34 G
Spanish	Araneum Hispanicum	 	 	
Swedish	Araneum Suedicum	 	 	
Ukrainian	Araneum Ucrainicum Beta	 	 	
Uzbek	Araneum Uzbekicum	