Feeding the "Brno Pipeline"
The Case of *Araneum Slovacum*

Vladimír Benko

vladob@juls.savba.sk

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

RASLAN 2016
Karlova Studánka, 2 December 2016

Aranea

"Brno Pipeline" (term coined by Nikola Ljubešić)

- ***SpiderLing***
  chared *... web page encoding identification*
  trigrams *... language identification*
  jusText *... boilerplate removal*
- ***Unitok*** *... tokenization*
- ***Onion*** *... document & paragraph level deduplication*
- ***NoSketch Engine*** *... corpus manager*

Aranea

Other tools needed

- ***Tagger***
  - *TreeTagger*
  - *Hunpos* (Hu)
  - *MorphoDiTa* (Cs)

Optional tools

- ***Sentence splitter***
- ***Filter(s)***

Aranea

A family of web corpora intended (primarily) for teaching purposes

Motivation (negative):

Available corpora

- Do not cover all languages needed (**WaC**, **COW**, …)
- Have different user interface and are (mostly) not available for download
- **Sketch Engine**: **sketch grammars are too different**

Aranea

A family of web corpora intended (primarily) for teaching purposes

Motivation (positive):

- New (open-source) tools became available in 2013
- Good experience with two Slovak web corpora
  - **Wanda** (2010 … 200 M tokens, iterative downloading by WebBootCat)
  - **Bruna Slovaca** (*skTenTen*, 2011 … 800 M tokens, created at the Masaryk University in Brno)

Aranea

- **Slovak-Centric** (languages spoken and/or taught in **Slovakia** and its neighbouring countries)
- Crawled and pre-processed by at (approximately) **the same time**
- Language-independent processing by **the same tools**
- Language-dependent processing using **the same methodology**
- **The same size** (2 “comparable” versions),
- “Language-neutral” (Latin) names

Aranea

- **Name Araneum** (*pl.* **Aranea**, *n.*)

araneum, aranei

noun

declension: 2nd declension

gender: neuter

Definitions:

1. mass of threads resembling a spider web
2. spider web, cobweb

Age: In use throughout the ages/unknown

Area: Agriculture, Flora, Fauna, Land, Equipment, Rural

Geography: All or none

Frequency: For Dictionary, in top 20,000 words

Source: "Oxford Latin Dictionary", 1982 (OLD)

Aranea

	masculine	*feminine*	*neuter*
positive	**parvus**	**parva**	**parvum**
comparative	**minor**	**minor**	**minus**
superlative	**minumus**	**minima**	**minimum**

positive	**magnus**	**magna**	**magnum**
comparative	**māior**	**māior**	**māius**
superlative	**maximus**	**maxima**	**maximum**

Aranea

- **Language**
  Anglicum, **Hispanicum**, **Francogallium**, **Russicum**, **Slovacum**...
- **Variants**
  Duplex (parallel alternate annotation), ...

Aranea Comparable Web Corpora

Four sizes

- **Maius** (greater) ... basic version, **1.2 billion tokens**, approx. **1 billion words**
- **Minus** (smaller) ... **10 %** sample of Maius (for **teaching purposes**)
- **Minimum** (minimal) ... approx. **1 %** sample of Maius (not accessible by users, used for **toolchain and sketch grammar experiments**)
- **Maximum** (maximal) ... **as much as we can get**

Aranea

- Compatible **tokenization**
- **Sentence-segmented**
- Document, paragraph & sentence-level **deduplicated**
- **PoS-tagged** by (possibly) **free tools**
- native tagsets mapped to ***Araneum Universal Tagset***
- Word sketches with ***compatible sketch grammars***

Aranea

<s>
Такого понятия в общем-то и нет,
хотя мы-то с вами знаем, что есть.
</s>

Aranea

<s>
Такого
понятия
в
общем-то
и
нет
,
хотя
мы-то
с
вами
знаем
,
что
есть
.
</s>

Aranea

<s>

Такого	такой	P--nsga
понятия	понятие	Ncnsgn
в	в	Sp-a
общем-то	общем-то	P--nsan
и	и	C
нет	нет	R
,	,	,
хотя	хотя	C
мы-то	мы-то	P--nsnn
с	с	Sp-i
вами	вы	P-2-pin
знаем	знать	Vmip1p-a-e
,	,	,
что	что	C
есть	быть	Vmip3s-a-e
.	.	SENT

</s>

Aranea

<s>

Такого	такой	P--nsga	1
понятия	понятие	Ncnsgn	1
в	в	Sp-a	1
общем-то	общем-то	P--nsan	0
и	и	C	1
нет	нет	R	1
,	,	,	1
хотя	хотя	C	1
мы-то	мы-то	P--nsnn	0
с	с	Sp-i	1
вами	вы	P-2-pin	1
знаем	знать	Vmip1p-a-e	1
,	,	,	1
что	что	C	1
есть	быть	Vmip3s-a-e	1
.	.	SENT	1

</s>

Aranea

<s>

Такого	такой	Pn	P--nsga	1
понятия	понятие	Nn	Ncnsgn	1
в	в	Pp	Sp-a	1
общем-то	общем-то	Pn	P--nsan	0
и	и	Cj	C	1
нет	нет	Av	R	1
,	,	Zz	,	1
хотя	хотя	Cj	C	1
мы-то	мы-то	Pn	P--nsnn	0
с	с	Pp	Sp-i	1
вами	вы	Pn	P-2-pin	1
знаем	знать	Vb	Vmip1p-a-e	1
,	,	Zz	,	1
что	что	Cj	C	1
есть	быть	Vb	Vmip3s-a-e	1
.	.	Zz	SENT	1

</s>

Aranea

Available corpora (November 2016)

Araneum Anglicum (15.04)
Araneum Anglicum Africanum (15.04)
Araneum Anglicum Asiaticum (15.04)
Araneum Bohemicum (15.04, **Maximum** 16.10)
Araneum Bulgaricum (16.05)
Araneum Finnicum (15.04)
Araneum Francogallicum (15.03)
Araneum Germanicum (15.02)
Araneum Hispanicum (15.04)
Araneum Hungaricum (14.12)

Aranea

Available corpora (December 2016)

Araneum Italicum (14.12)
Araneum Nederlandicum (15.02)
Araneum Polonicum (15.02)
Araneum Portugallicum (15.05)
Araneum Russicum (15.02, **Maximum** 16.04)
Araneum Russicum Russicum (15.03)
Araneum Russicum Externum (15.03)
Araneum Sinicum (15.03)
Araneum Slovacum (15.04, **Maximum** 16.01)

Aranea

Own tools

Sentence segmentation

- Rule-based (9 + 2 flex rules)
- (more or less) language independent
- Recall & precession not evaluated yet

Sampling and splitting utility

- Head (first *n* tokens alligned to doc)
- Sample (per mille in docs)
- Split to *n* parts (4 by default)

Aranea

Own tools

Normalization

- Unicode spaces to ASCII spaces
- Multiple spaces to single space
- Composite diacritics
- Soft hyphens
- Ligatures

Filtration with correction

- Survived HTML markup
- Empty paragraphs

Aranea

Own tools

Filtration with removal of affected docs

- Incorrectly detected language
- Encoding issues
- Spam

"Lemmatization" of punctuation and special graphic chars