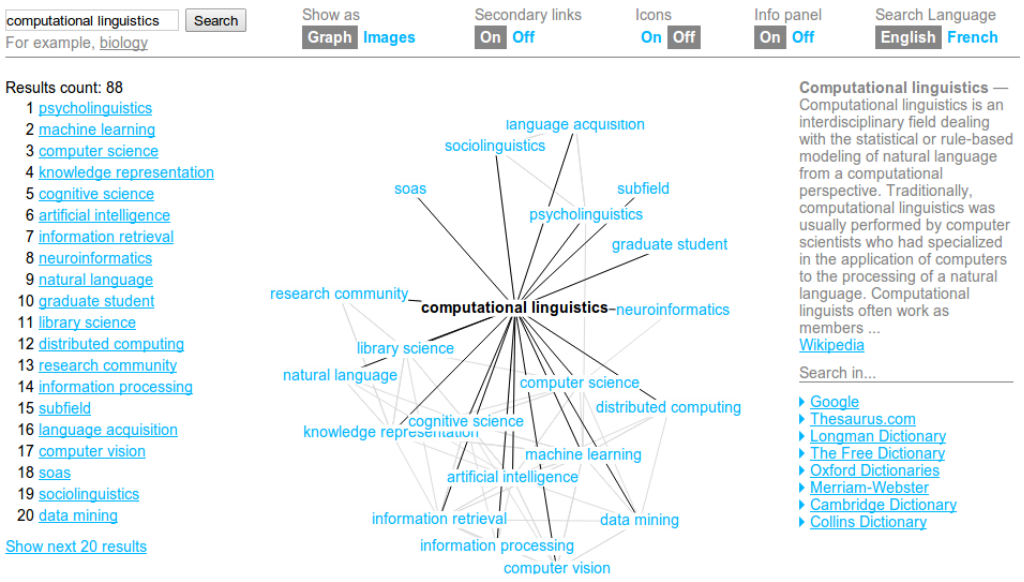
Extraction of Russian Hypernyms and Co-Hyponyms with Lexico-Syntactic Patterns[1]

Kristina Sabirova (bezadresa.net@gmail.com), South Ural State University
Artem Lukanin (ice_lc@mail.ru), South Ural State University
Alexander Panchenko (alexander.panchenko@uclouvain.be), DigSoLab & Université catholique de Louvain

SERELEX

- The system provides a list of related words, given a query: **http://serelex.cental.be/**
- It uses patterns to extract hypernyms, co-hyponyms and synonyms from English and French texts [1, 2].



RUSSIAN PATTERNS

- **Method:** We translated existing English patterns and added several new ones. We encoded 10 lexico-syntactic patterns, in the form of FSTs in Unitex 3.1beta [3].
- **Notation:** X – hypernym, Y – hyponym.

1. Такие/таких/таким X, как Y[, Y] и/или Y

```
В Индии зародились такие {[религии]=HYPER} как
{[индуизм]=HYPO}, {[буддизм]=HYPO}, {[сикхизм]=H
YPO} и {[джайнизм]=HYPO}
```

2. X, такие/таких/таким как Y[, Y] и/или Y

```
…{систем [верований]=HYPER}, таких как
{[шаманизм]=HYPO},{[политеизм]=HYPO},
{[пантеизм]=HYPO}, {[анимизм]=HYPO}
```

3. X: Y[, Y] и/или Y

```
…мир, передаваемый человеку через {его [ощущения
]=HYPER}: {[зрение]=HYPO}, {[слух]=HYPO}, {[обон
яние]=HYPO}, {[осязание]=HYPO} и другие
```

4. Y[, Y][, а также/также как [и]/и/или] другие/другим/других/о других X

```
…использованием зараженных вирусом
{[шприцев]=HYPO}, {[игл]=HYPO} и {других
{медицинских и парамедицинских
[инструментов]=HYPER}}
```

5. Y [и по сей день/в <A> время] это/есть/является/был [самый] <A> X

```
{Рентгеноструктурный [анализ]=HYPO} и по сей
день является самым {распространенным
[методом]=HYPER} определения структуры вещества…
```

6. X, включая/в том числе [и]/включительно, Y[, Y] и/или Y

```
…сосудистую систему {[позвоночных]=HYPER}
животных, в том числе {[человека]=HYPO} и
некоторых беспозвоночных
```

7. X, [а] особенно/в особенности/особо, Y[, Y] и/или Y

```
L-Аргинин входит в {состав пептидов и
[белков]=HYPER}, особенно {высоко содержание
[аргинина]=HYPO}…
```

8. X, как например/в частности/например/к примеру, Y[, Y] и/или Y

```
Это является одной из основных проблем при
проектировании устройств {цифровой
[электроники]=HYPER}, в частности, {цифровых
[фотоаппаратов]=HYPO}…
```

9. Виды/типы/формы/разновидности/сорта X, как Y[, Y] и/или Y

```
Такие виды {[оружия]=HYPER} как {[шпага]=HYPO} и
{[рапира]=HYPO} тоже причисляют к мечам, что не
совсем верно
```

10. Y - вид/тип/форма/разновидность/сорт X

```
{[Хобби]=HYPO} — вид {человеческой
[деятельности]}, некое занятие, увлечение…
```

THE FIRST RESULTS

- **Corpus:** Russian DBPedia (extended abstracts) [4].
- **Size:** 1,325,858 sentences, ~47,000,000 tokens.

Pattern	Extracted sentences	Extracted hypernyms	Extracted hyponyms
1	1572	1573	3907
2	1175	1182	2627
3	1825	1844	6149
4	1166	1166	3364
5	454	454	454
6	2342	2348	3185
7	232	232	232
8	544	544	648
9	31	31	36
10	1104	1104	1104
Total:	10445	10478	21706

REFERENCES

1. Panchenko A., Morozova O., Naets H. A Semantic Similarity Measure Based on Lexico-Syntactic Patterns. // In Proceedings of KONVENS 2012, — Vienna (Austria), pp.174--178, 2012.
2. Panchenko A., Romanov P., Morozova O., Naets H., Romanov A., Philippovich A., Fairon C. "Serelex: Search and Visualization of Semantically Similar Words" . In Proceedings of ECIR 2013. LNCS 7814.
3. Unitex corpus processor. Available under LGPL license: http://www-igm.univ-mlv.fr/~unitex/
4. Russian Dbpedia Corpus without accents: http://cental.fltr.ucl.ac.be/team/~panchenko/data/serelex/corpus-ru-dbpedia-short-dea.csv

1 Исследование выполнено при финансовой поддержке РГНФ (проект № 13-04-12020 «Новый открытый электронный тезаурус русского языка»).