Towards named entity annotation of Latvian National Library corpus

Pēteris Paikens, Ilze Auziņa, Ginta Garkāje, Madara Paegle
Institute of Mathematics and Computer Science, University of Latvia

The LNB digital corpus

- Books, journals and newspapers
- 4.5 billion words, 240 000 documents
- 18th-20th century, 1920-1930'ies focus
- A lot of OCR problems
- Significant orthography changes

Training corpus

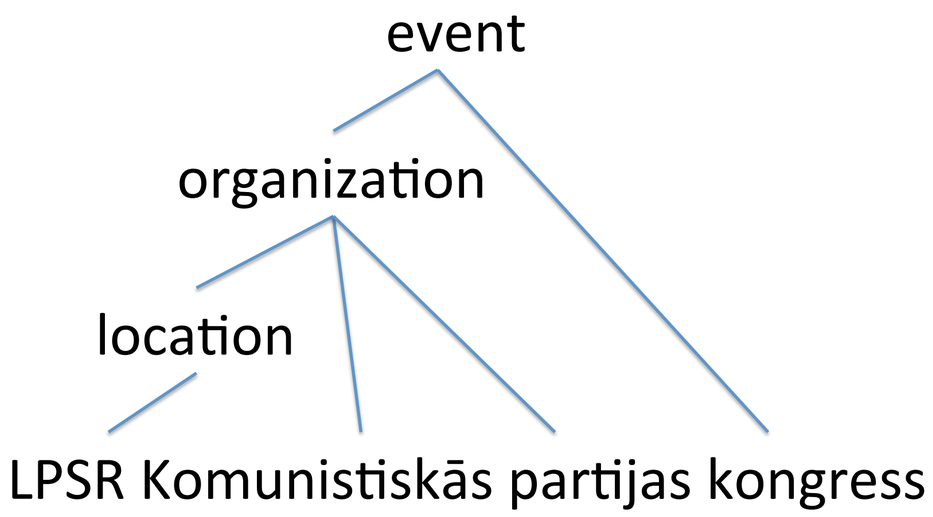
150 000 words, 9000 entities

Year	Type	Title	Size (words)
1861	Newspaper	Latviešu Avīzes	5 224
1863	Book	Tahiti salas ļaudis	5 612
1882	Newspaper	Arājs	10 346
1918	Newspaper	Baltijas Ziņas	19 152
1928	Magazine	Atpūta	12 354
1934	Book	Madonas vadonis tūristiem	2 471
1935	Newspaper	Mūzikas Apskats	13 129
1942	Newspaper	Sendergruppe Ostland	8 234
1957	Newspaper	Cīņa	15 568
1966	Book	Krusttēvs Oskars : atmiņas	11 505
1988	Newspaper	Padomju jaunatne	15 181
1988	Book	Kārlis Ulmanis	5 724
1999	Magazine	Zīlīte	2 460
2005	Book	Ugāles baznīca	5 206
2007	Magazine	Dadzis	18 791

Index of NE mentions in corpus

- Authoritative database of entities
- Pseudonyms and multiple names
- Spelling variations
- Renamed objects
- Automatical identification of people frequently mentioned in press

Hierarchical annotation of named entities



Annotation tool

Rau. mūsu redakcijai iesūtītā vēstule, kuru šāgada 3. jūlijā Vljpā. «Gaudeamus X» svētku laikā parakstījuši 1132 Latvijas Padomju Sociālistiskās Republikas delegācijas dalībnieki: Latvijas PSR Augstākajai Padomei. Izvērtējot vēsturi atbilstoši Jaunajam domāšanas veidam, nedrīkst atstāt neievērotu Jautājumu par latviešu tautas nacionālo simboliku. Ziņas par tās rašanos sarkanbaltsarkano krāsu veidā sniedzas līdz XII gadsimtam, tā saglabājusies ari turpmāk, izmantota pirmajos latviešu Dziesmu svētkos (1873. g.), to lietojuši latviešu studenti Tērbatā, kā ari latviešu sar kanie strēlnieki. Sai simbolikai ir būtiska nozīme visā latviešu tautas vēsturē, un to nedrīkst saistīt tikai ar buržuāziskās republikas periodu, kura pastāvēja nedaudz vairāk kā divdesmit gadus. Tādēļ mēs. Baltijas republiku

studentu Dziesmu un deju svētku «Gaudeamus-X» Latvijas PSR delegācijas dalībnieki, lerosinām Augstāko Padomi izskatīt Jautājumu par latviešu tautas nacionālas simbolikas atjaunošanu. So vēstuli LPSR Augstākajai Padomei delegācijas vārdā lesniedza LĻKJS CX studējošās jaunatnes nodaļas vadītājs A. Amerlks. kora «Juventus» prezidents A. Telk manis un Latvijas ĻKJS prēmijas laureāta, studentu pūtēju orķestra diriģents J. Pūriņš. J. Stradiņa publikācija laik rakstā «Ciņa» un I. Latkovska raksts «Padomju Jaunat nē», kā ari «Ciņas» apaļais galds lielā mērā Izgaismo «apstrīdamo krāsu» vēsturi. Vēstules autori gan kļūdījušies, ziņas par sarkanbaltsar kanā karoga parādīšanos pārceļot vēl par gadsimtu senākā pagātnē, taču tas nav galvenais. Nav būtiski, manu prāt, lauzt šķēpus par to, vai

Named entity recognition

- Based on Stanford NER
- 80-85% precision
- 70-80% recall

Mākslīgā intelekta laboratorija LU MII
L N B DIGITĀLĀ BIBLIOTĒKA
ERAF EIROPAS REĢIONĀLĀS ATTĪSTĪBAS FONDS
Projects LNB/2011/33/ERAF and 2DP/2.1.1.2.0/10/APIA/VIAA/011