Advance Search in CLARIN Text Corpora

Pavel Rychlý
Natural Language Processing Centre, Masaryk University
Botanická 68a, Brno, Czech Republic
pary@fi.muni.cz

Keywords: Corpus, Federated Content Search, Corpus Search

1. Introduction

CLARIN is a large-scale pan-European initiative to create, coordinate and make language resources and technology available and readily usable. Many of CLARIN resources are text corpora and one of services provided by CLARIN centres is *Federated Content Search* (CLARIN-FCS). It is a service for searching content of individual corpora in a standard way. Using a formal protocol the results of corpus queries could be processed by different systems or applications to provide a user interface for end users or to further compute or combine results into more sophisticated information.

One of corpus search engines which provides the mentioned FCS service is NoSketch Engine. It is a powerful and complex system for corpus querying and overall handling which is easy to install and customise. We have added the FCS support to the system and this paper shows how the system was prepared for Czech CLARIN endpoint.

The structure of this paper is the following: the next two sections describes CLARIN-FCS and NoSketch Engine in more details, the following section presents installation of NoSketch Engine as an CLARIN-FCS endpoint service in Czech CLARIN (LINDAT) project.

2. Federated Content Search

The CLARIN-FCS Interface Specification defines a set of capabilities, an extensible result format and a set of required operations. CLARIN-FCS is based on open standards like XML and SRU.[1] The original specification requires to search content of one source (language resource), but CLARIN-FCS exploits SRU's extension mechanism to search in several different sources from one URL. NoSketch Engine uses this extension.

The CLARIN project not only defines the format of input and respective output of the service there are also reference implementations and automatic endpoint testing services. NoSketch Engine fulfils all the required formats.

3. NoSketch Engine

NoSketch Engine is an open-source project combining Manatee and Bonito into a powerful and free corpus management system (Rychly, 2007). Manatee is a corpus management tool including corpus building and indexing, fast querying and providing basic statistical measures. It utilise a fast indexing library called Finlib. Bonito is a graphical user interface to corpora maintained by Manatee. It is a web interface written in Python which can be run under any web-server supporting the CGI.

NoSketch Engine is a limited version of the software empowering the famous Sketch Engine service (Kilgarriff et al., 2004), a commercial variant offering word sketches, thesaurus, keyword computation, user-friendly corpus creation and many other excellent features.

Manatee is able to deal with extremely large corpora, it is successfully used for a 70+ billion word corpus (English part of the ClueWeb collection (Pomikálek et al., 2012)) and is able to provide a platform for computing a wide range of lexical statistics. The system provides the following modules:

- Concordancer: show words in context
  The powerful query language can be used to find interesting words or phrases in a corpus, any annotation could be used in the queries, there are many filtering and sampling options. The resulting concordance could be sorted in various ways and collocations and frequency distributions of keywords, context or text types could be computed.

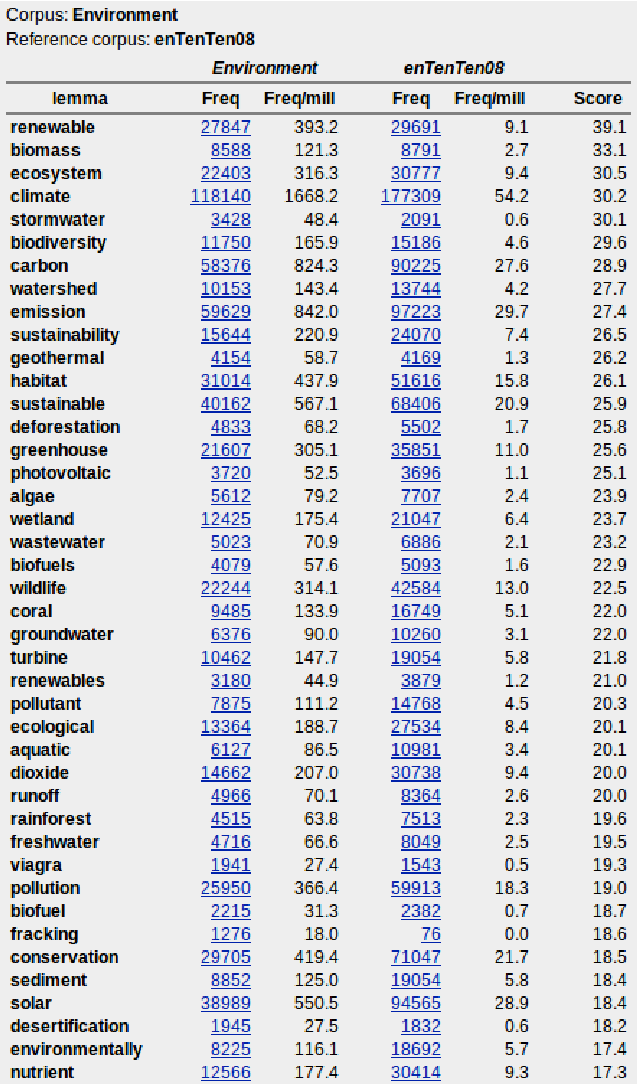
- Word-lists: searching and browsing of words or any annotation (lemma, part of speech, morphology tags, text types).

- Keywords: comparing corpora or sub-corpora to a reference corpus. It finds those words that are most frequent in the corpus, in comparison with their frequency in a reference corpus. Keywords also let us discover the distinctive vocabulary of a domain. For example, keywords extracted from corpus Environment using reference corpus enTenTen08 are listed in Figure 1.

- Parallel corpora: search many parallel corpora and display respective translations.

NoSketch Engine is an open source system, it is distributed under the GNU GPL license. The system serves as a base for several other corpus systems or applications. NoSketch Engine and derived systems are used by thousands of users worldwide. The lead users have been dictionary publishers and it is in day-to-day use for lexicography at Oxford University Press, Cambridge University Press, Harper Collins, Macmillan, Cornelsen and the Instituut voor Nederlandse

[1] Search/Retrieval via URL, for specification and more information see http://www.loc.gov/standards/sru/



Corpus: **Environment**
Reference corpus: **enTenTen08**

	Environment		*enTenTen08*		
lemma	**Freq**	**Freq/mill**	**Freq**	**Freq/mill**	**Score**
renewable	27847	393.2	29691	9.1	39.1
biomass	8588	121.3	8791	2.7	33.1
ecosystem	22403	316.3	30777	9.4	30.5
climate	118140	1668.2	177309	54.2	30.2
stormwater	3428	48.4	2091	0.6	30.1
biodiversity	11750	165.9	15186	4.6	29.6
carbon	58376	824.3	90225	27.6	28.9
watershed	10153	143.4	13744	4.2	27.7
emission	59629	842.0	97223	29.7	27.4
sustainability	15644	220.9	24070	7.4	26.5
geothermal	4154	58.7	4169	1.3	26.2
habitat	31014	437.9	51616	15.8	26.1
sustainable	40162	567.1	68406	20.9	25.9
deforestation	4833	68.2	5502	1.7	25.8
greenhouse	21607	305.1	35851	11.0	25.6
photovoltaic	3720	52.5	3696	1.1	25.1
algae	5612	79.2	7707	2.4	23.9
wetland	12425	175.4	21047	6.4	23.7
wastewater	5023	70.9	6886	2.1	23.2
biofuels	4079	57.6	5093	1.6	22.9
wildlife	22244	314.1	42584	13.0	22.5
coral	9485	133.9	16749	5.1	22.0
groundwater	6376	90.0	10260	3.1	22.0
turbine	10462	147.7	19054	5.8	21.8
renewables	3180	44.9	3879	1.2	21.0
pollutant	7875	111.2	14768	4.5	20.3
ecological	13364	188.7	27534	8.4	20.1
aquatic	6127	86.5	10981	3.4	20.1
dioxide	14662	207.0	30738	9.4	20.0
runoff	4966	70.1	8364	2.6	20.0
rainforest	4515	63.8	7513	2.3	19.6
freshwater	4716	66.6	8049	2.5	19.5
viagra	1941	27.4	1543	0.5	19.3
pollution	25950	366.4	59913	18.3	19.0
biofuel	2215	31.3	2382	0.7	18.7
fracking	1276	18.0	76	0.0	18.6
conservation	29705	419.4	71047	21.7	18.5
sediment	8852	125.0	19054	5.8	18.4
solar	38989	550.5	94565	28.9	18.4
desertification	1945	27.5	1832	0.6	18.2
environmentally	8225	116.1	18692	5.7	17.4
nutrient	12566	177.4	30414	9.3	17.3

Figure 1: Example of keywords extracted from corpus Environment

Lexicologie (INL, Institute of Dutch Lexicology) among others. It is used for big national corpora in the Netherlands, Slovenia, Croatian, Hungary, Russia, and many others.

4. LINDAT FCS Service

The *LINDAT/CLARIN Centre for Language Research Infrastructure* is a Czech part of CLARIN. The project is funded by the Ministry of Education, Youth and Sports of the Czech Republic. It is a *type B* Centre and provides CLARIN-FCS service for selected corpora.

There are about 60 language resources of *corpus* type in the LINDAT repository. About half of that are corpora not suitable for text search – these include for example speech corpora, annotation-only data (additional annotation to a different language resource without the content itself).

All suitable corpora[2] was transformed into vertical format and encoded for NoSketch Engine.

The CLARIN-FCS support was added into NoSketch Engine. There are two non-trivial oprations (*scan* and *searchRetrieve*) which are translated into two NoSketch Engine methods (*wordlist* and *concordance*). The translation of an input is quite straightforward and translation of the output into FCS XML format is done via an Bonito's templating framework. The biggist problem of the implementation was finding the right format (or set of features) of many corner cases which are not defined by the SRU specification but required by conformance tests.

All available corpora could be accessed not only by FCS queries which are suitable for further processing, but also by Bonito user interface. Using Bonito, we can use much more advances queries: use all available annotation (POS tags, lemma, etc.), find phrases or syntactic structures, or use the full Manatee's query language. An example of a concordance is at Figure 2.

The system is installed on one server of the Brno part of the LINDAT project. It is easy to modify graphical appearance of the system, all LINDAT specific graphical styles was incorporated in the installation. An example of the system page is at Figure 3.

5. Conclusion

This paper presents an off-the-shelf system NoSketch Engine which is ready to provide Federated Content Search endpoint service for text corpora. It is easy to install and easy to prepare any corpus for indexing and searching. The system scales well in many ways, it can be used for corpora of size up to 100 billion tokens, it is language and tag-set independent, the number of positional and structural attributes are not limited. It provides both CLARIN-FCS services and advance corpus search including word list and concordance creation, filtering and sorting of a concordance.

6. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIN project LM2010013.

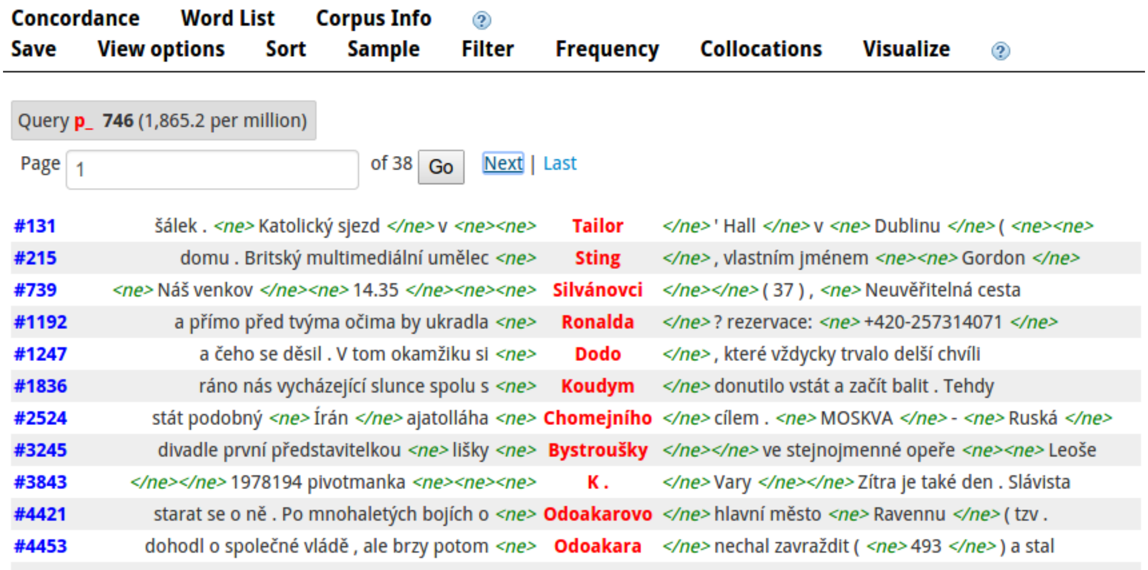
7. References

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. In *Proceeding of Euralex*, pages 105–116, lorient, France.

Pomikálek, J., Jakubíček, M., and Rychlý, P. (2012). Building a 70 billion word corpus of english from clueweb. In *LREC*, pages 502–506.

Rychly, P. (2007). Manatee/bonito - a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masaryk University.

[2] With the exception of *W2C – Web to Corpus* – collection of 120 automatically generated corpora of very low quality

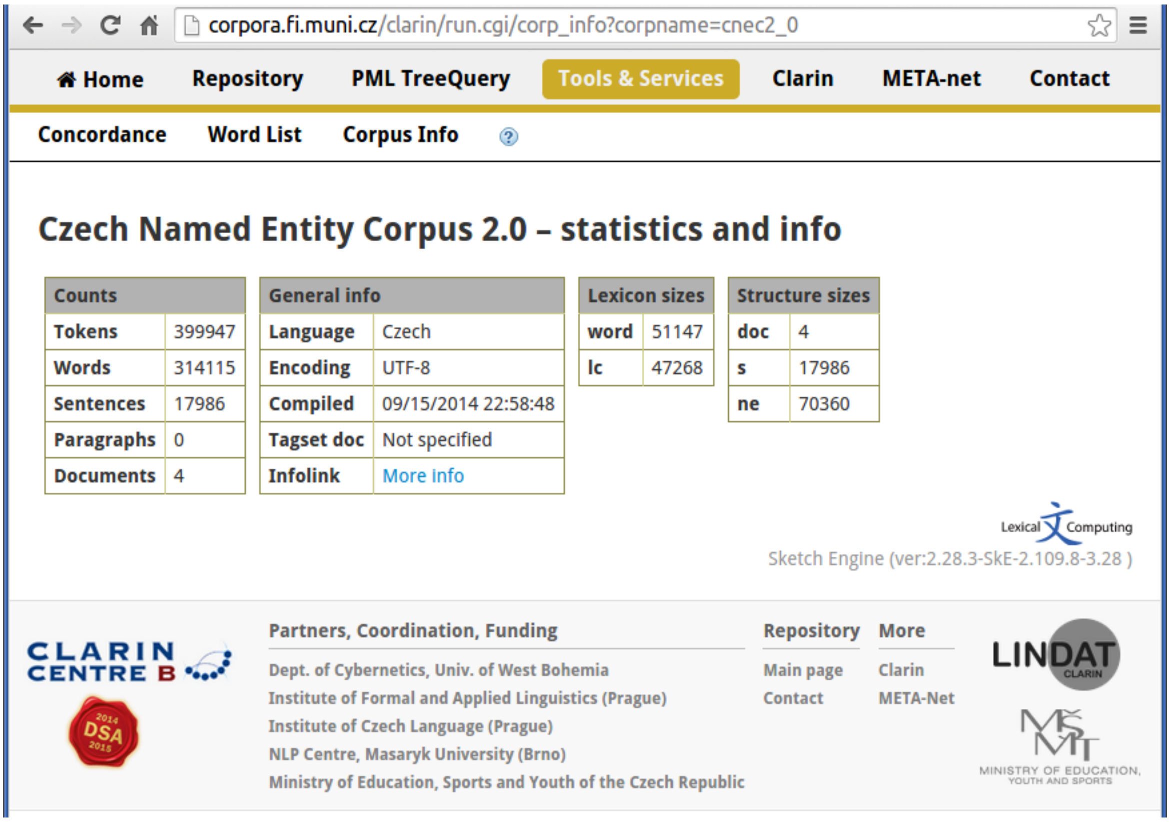


Query p_ 746 (1,865.2 per million)

Page 1 of 38

#131	šálek . <ne> Katolický sjezd </ne> v <ne><ne>	Tailor	</ne> ' Hall </ne> v <ne> Dublinu </ne> (<ne><ne>
#215	domu . Britský multimediální umělec <ne>	Sting	</ne> , vlastním jménem <ne><ne> Gordon </ne>
#739	<ne> Náš venkov </ne><ne> 14.35 </ne><ne><ne>	Silvánovci	</ne></ne> (37) , <ne> Neuvěřitelná cesta
#1192	a přímo před tvýma očima by ukradla <ne>	Ronalda	</ne> ? rezervace: <ne> +420-257314071 </ne>
#1247	a čeho se děsil . V tom okamžiku si <ne>	Dodo	</ne> , které vždycky trvalo delší chvíli
#1836	ráno nás vycházející slunce spolu s <ne>	Koudym	</ne> donutilo vstát a začít balit . Tehdy
#2524	stát podobný <ne> Írán </ne> ajatolláha <ne>	Chomejního	</ne> cílem . <ne> MOSKVA </ne> - <ne> Ruská </ne>
#3245	divadle první představitelkou <ne> lišky <ne>	Bystroušky	</ne></ne> ve stejnojmenné opeře <ne><ne> Leoše
#3843	</ne></ne> 1978194 pivotmanka <ne><ne><ne>	K .	</ne> Vary </ne></ne> Zítra je také den . Slávista
#4421	starat se o ně . Po mnohaletých bojích o <ne>	Odoakarovo	</ne> hlavní město <ne> Ravennu </ne> (tzv .
#4453	dohodl o společné vládě , ale brzy potom <ne>	Odoakara	</ne> nechal zavraždit (<ne> 493 </ne>) a stal

Figure 2: Example of a concordance (Czech Named Entity Corpus 2.0, query `<ne type="p-"/>`)



Czech Named Entity Corpus 2.0 – statistics and info

Counts	
Tokens	399947
Words	314115
Sentences	17986
Paragraphs	0
Documents	4

General info	
Language	Czech
Encoding	UTF-8
Compiled	09/15/2014 22:58:48
Tagset doc	Not specified
Infolink	More info

Lexicon sizes	
word	51147
lc	47268

Structure sizes	
doc	4
s	17986
ne	70360

Sketch Engine (ver:2.28.3-SkE-2.109.8-3.28)

Partners, Coordination, Funding

Dept. of Cybernetics, Univ. of West Bohemia
Institute of Formal and Applied Linguistics (Prague)
Institute of Czech Language (Prague)
NLP Centre, Masaryk University (Brno)
Ministry of Education, Sports and Youth of the Czech Republic

Repository

Main page
Contact

More

Clarin
META-Net

Figure 3: Example of and corpus description in LINDAT style. (Czech Named Entity Corpus 2.0)