Bilingual terminology extraction

Vít Baisa

Sketch Engine

vit.baisa@sketchengine.co.uk

6th Sketch Engine Workshop
Herstmonceux, August 10, 2015

Terminology extraction: recap

- combination of rules & statistics
- languages: Czech, Dutch, English, French, German, Chinese Simplified, Chinese Traditional, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish
- you can help us to add your language
- currently in progress: Turkish, Hungarian

Terminology extraction: what is a term?

Terminology extraction: what is a term?

unithood

how it is grammatically defined? (e.g. noun phrases)

Terminology extraction: what is a term?

unithood

how it is grammatically defined? (e.g. noun phrases)

termhood

does it belong to a domain?

Unithood

- Sketch Grammar formalism
- CQL (corpus query language) rules

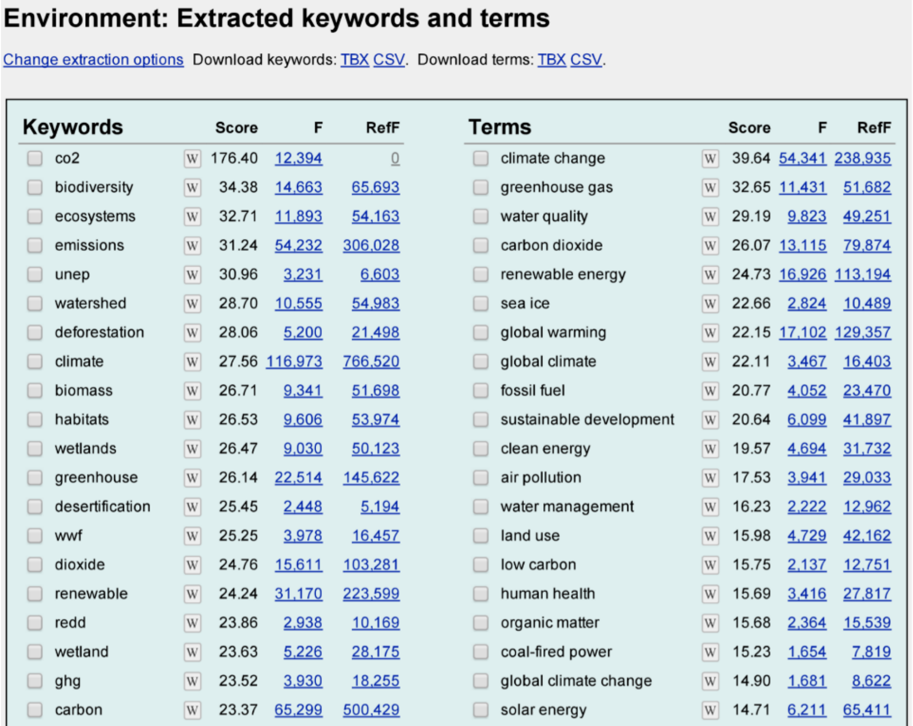
```
# English, computer mouse
*COLLOC "%(2.lc)_%(1.lc)"
    2:[tag=="NN" | tag=="JJ" | tag=="VVG"] 1:[tag=="NN"]

# German, kleines Haus
*COLLOC "%(2.adj_stem)%(1.gender_ending)_%(1.lemma_cap)"
    2:[kind="ADJA"] 1:[kind="N"] & 1.case = 2.case

# Czech, Ústav národního zdraví
*COLLOC "%(1.gender_lemma)_%(2.lc)_%(3.lc)"
    1:noun 2:adj_genitive 3:noun_genitive & agree(2,3)
```

Termhood

- simple math parameter N

$$\frac{f_{focus} + N}{f_{ref} + N}$$

- f is relative (per million) frequency of a term
- the formula is used also for keyword extraction
- N influences whether rare or frequent words are preferred
- a reference corpus in the same language is needed



Environment: Extracted keywords and terms

Change extraction options Download keywords: TBX CSV. Download terms: TBX CSV.

	Keywords		Score	F	RefF
□	co2	W	176.40	12,394	0
□	biodiversity	W	34.38	14,663	65,693
□	ecosystems	W	32.71	11,893	54,163
□	emissions	W	31.24	54,232	306,028
□	unep	W	30.96	3,231	6,603
□	watershed	W	28.70	10,555	54,983
□	deforestation	W	28.06	5,200	21,498
□	climate	W	27.56	116,973	766,520
□	biomass	W	26.71	9,341	51,698
□	habitats	W	26.53	9,606	53,974
□	wetlands	W	26.47	9,030	50,123
□	greenhouse	W	26.14	22,514	145,622
□	desertification	W	25.45	2,448	5,194
□	wwf	W	25.25	3,978	16,457
□	dioxide	W	24.76	15,611	103,281
□	renewable	W	24.24	31,170	223,599
□	redd	W	23.86	2,938	10,169
□	wetland	W	23.63	5,226	28,175
□	ghg	W	23.52	3,930	18,255
□	carbon	W	23.37	65,299	500,429

	Terms		Score	F	RefF
□	climate change	W	39.64	54,341	238,935
□	greenhouse gas	W	32.65	11,431	51,682
□	water quality	W	29.19	9,823	49,251
□	carbon dioxide	W	26.07	13,115	79,874
□	renewable energy	W	24.73	16,926	113,194
□	sea ice	W	22.66	2,824	10,489
□	global warming	W	22.15	17,102	129,357
□	global climate	W	22.11	3,467	16,403
□	fossil fuel	W	20.77	4,052	23,470
□	sustainable development	W	20.64	6,099	41,897
□	clean energy	W	19.57	4,694	31,732
□	air pollution	W	17.53	3,941	29,033
□	water management	W	16.23	2,222	12,962
□	land use	W	15.98	4,729	42,162
□	low carbon	W	15.75	2,137	12,751
□	human health	W	15.69	3,416	27,817
□	organic matter	W	15.68	2,364	15,539
□	coal-fired power	W	15.23	1,654	7,819
□	global climate change	W	14.90	1,681	8,622
□	solar energy	W	14.71	6,211	65,411

Fine-tuning: options

- stoplists (blacklists)
- simple math parameter
- minimum frequency
- minimum score
- minimum character length
- only alphanumerical strings
- . . .

Bilingual (multilingual) terminology extraction

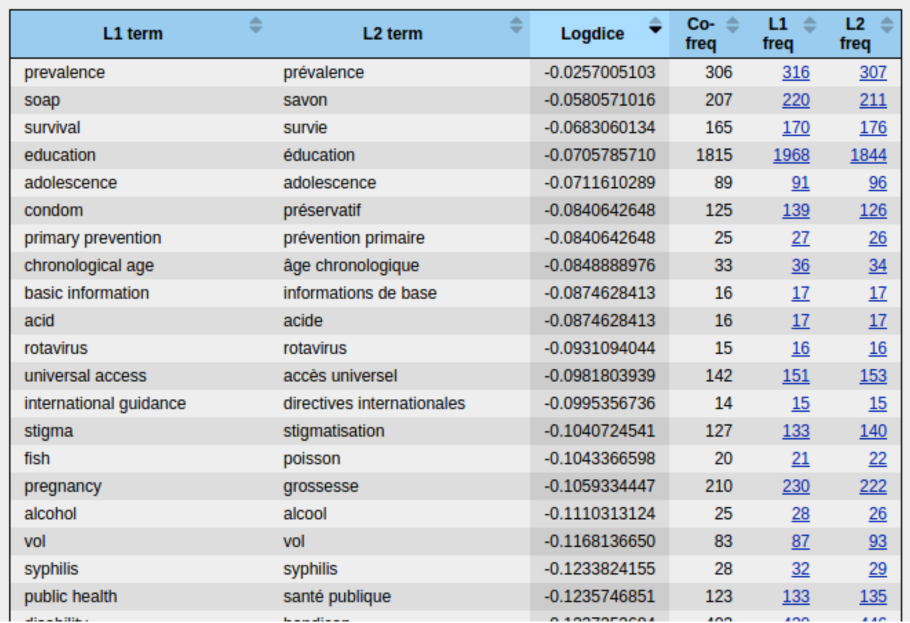
- recent development
- parallel corpora needed

Bilingual (multilingual) terminology extraction

- recent development
- parallel corpora needed

Two-step procedure

1. extraction of terms in source and target languages
2. counting co-occurrences of the terms



L1 term	L2 term	Logdice	Co-freq	L1 freq	L2 freq
prevalence	prévalence	-0.0257005103	306	316	307
soap	savon	-0.0580571016	207	220	211
survival	survie	-0.0683060134	165	170	176
education	éducation	-0.0705785710	1815	1968	1844
adolescence	adolescence	-0.0711610289	89	91	96
condom	préservatif	-0.0840642648	125	139	126
primary prevention	prévention primaire	-0.0840642648	25	27	26
chronological age	âge chronologique	-0.0848888976	33	36	34
basic information	informations de base	-0.0874628413	16	17	17
acid	acide	-0.0874628413	16	17	17
rotavirus	rotavirus	-0.0931094044	15	16	16
universal access	accès universel	-0.0981803939	142	151	153
international guidance	directives internationales	-0.0995356736	14	15	15
stigma	stigmatisation	-0.1040724541	127	133	140
fish	poisson	-0.1043366598	20	21	22
pregnancy	grossesse	-0.1059334447	210	230	222
alcohol	alcool	-0.1110313124	25	28	26
vol	vol	-0.1168136650	83	87	93
syphilis	syphilis	-0.1233824155	28	32	29
public health	santé publique	-0.1235746851	123	133	135
disability	handicap	-0.1237253684	403	420	446

Bilingual terminology extraction

- we need to evaluate the extraction properly
- data can be saved as TBX
- granularity affects quality

The (not so distant) future for BTE

- parallel vs. comparable corpora
- definition finding
- term hyper-, hyponyms finding
- term thesaurus
- the ultimate goal: one-click terminology :)
- terminology consistency checking
- multi- instead of bilingual extraction

The last slide

- API available
- IntelliWebSearch configurations
- plugins for SDL, Kilgray products planned
- one-off terminology extractions
- promissing results so far