NHH

Deling av norske språkdata: terminologi og flerspråklige ressurser

ELRC Workshop
8. juni 2016

Gisle Andersen, NHH

Innhold

- Språkressurser for maskinoversettelse
  - Enspråklige og flerspråklige korpus
  - Allmenne og fagspesifikke korpus
  - Allmenne og fagspesifikke ordbøker
  - → Terminologi (flerspråklig)
- Terminologi og terminologihåndteringssystemer
  - CLARINO/Termportalen
- Termekstraksjon
  - Enspråklige og flerspråklige metoder

Terminologihåndtering: status i Norge

Per i dag gode på

- tilgjengeliggjøring
  - termbaseteknologi
  - konverteringsalgoritmer
  - felles formater og standarder (f.eks. TBX)

- utvikling av nytt innhold
  - begrepsanalyse
  - definisjonsskriving

Betydelig potensial for bedre å utnytte

- teknologier for parallellspråklig og enspråklig maskinell analyse
- termekstraksjon
- kollokasjonsanalyse

CLARINO, CLARIN og CLARA

- CLARIN: felles europeisk forskningsinfrastruktur
  - legger føringer for formater, arbeidsmåter, tilgangskontroll og juridiske forhold
- CLARINO: nasjonalt prosjekt, finansiering fra NFR
  - etablerer en nasjonal infrastruktur for språkressurser
- CLARINO WP7 *Terminology integration*: NHHs arbeidspakke i CLARINO-prosjektet
  - etablerer en nasjonal portal for terminologiressurser

CLARIN ERIC
Common Language Resources and Technology Infrastructure

CLARINO

NHH Termportalen

Termportalen på terminologi.no
Enhetlig tilgang til distribuerte ressurser

rmportalen

Søkeside | Om Termportalen | Termbasene | Om CLARINO | Publikasjoner

Termbasene

baser er tilgjengelige?

Beskrivelse	Ansvarlig institusjon
Økonomisk-administrativ terminologi	NHH
Termbase for Mikroøkonomi utviklet ved NHH	NHH
Maritim ordbok	NHH
Tolketjenesten i Bergens termbase	Tolketjenesten i Bergen
Marine evertebrater	NHH
Norsk-tysk juridisk terminologi	NHH
Termbaser utviklet av Norsk termbank	Uni Research
Termbaser utviklet av Rådet for teknisk terminologi	Fagbokforlaget
English for business	NHH
Artsdatabanken	Havforskningsinstituttet
Maritim terminologi	Sjøfartsdirektoratet

NHH

Termportalen

Forsiden | Søkeside | Om Termportalen | Termbasene | Om CLARINO

Du er her: Forsiden > Termbasene > MRT > 89824

Domene: Marine terminology

hvd abyssopelagisk sone

def nivå i pelagisk sone fra 4000 m ned til dyphavsslettene

mrk

hvd abyssopelagic zone

def

mrk

Termekstraksjon

- Teknologi for maskinell termekstraksjon (ekserpering, utdragning) av termer
  - som alternativ til rådende metode: manuell termekstraksjon
- Korpusbasert tilnærming
  - Enspråklig metode: sammenlikning av fagspråklige tekster med allmennspråklige tekster
  - Flerspråklig metode: sammenlikning av termer som forekommer i parallelle fagspråklige tekstkilder
  - Oftest behov for betydelig manuelt arbeid
  - Behov for flerordsprosessering (n-gram, kollokasjoner, assosiasjonsmål)

Termekstraksjon

Enspråklig termekstraksjon – maskinelt finne frem til:

- Enkeltord og ordsekvenser
- … som bare forekommer i fagspråklig tekstkorpus (FTK) men ikke i allmennspråklig tekstkorpus (ATK)
- … som forekommer hyppigere i FTK enn ATK

  → **termkandidater**

Ekstraksjon av termkandidater ved kollokasjonsanalyse (flerordsuttrykk)

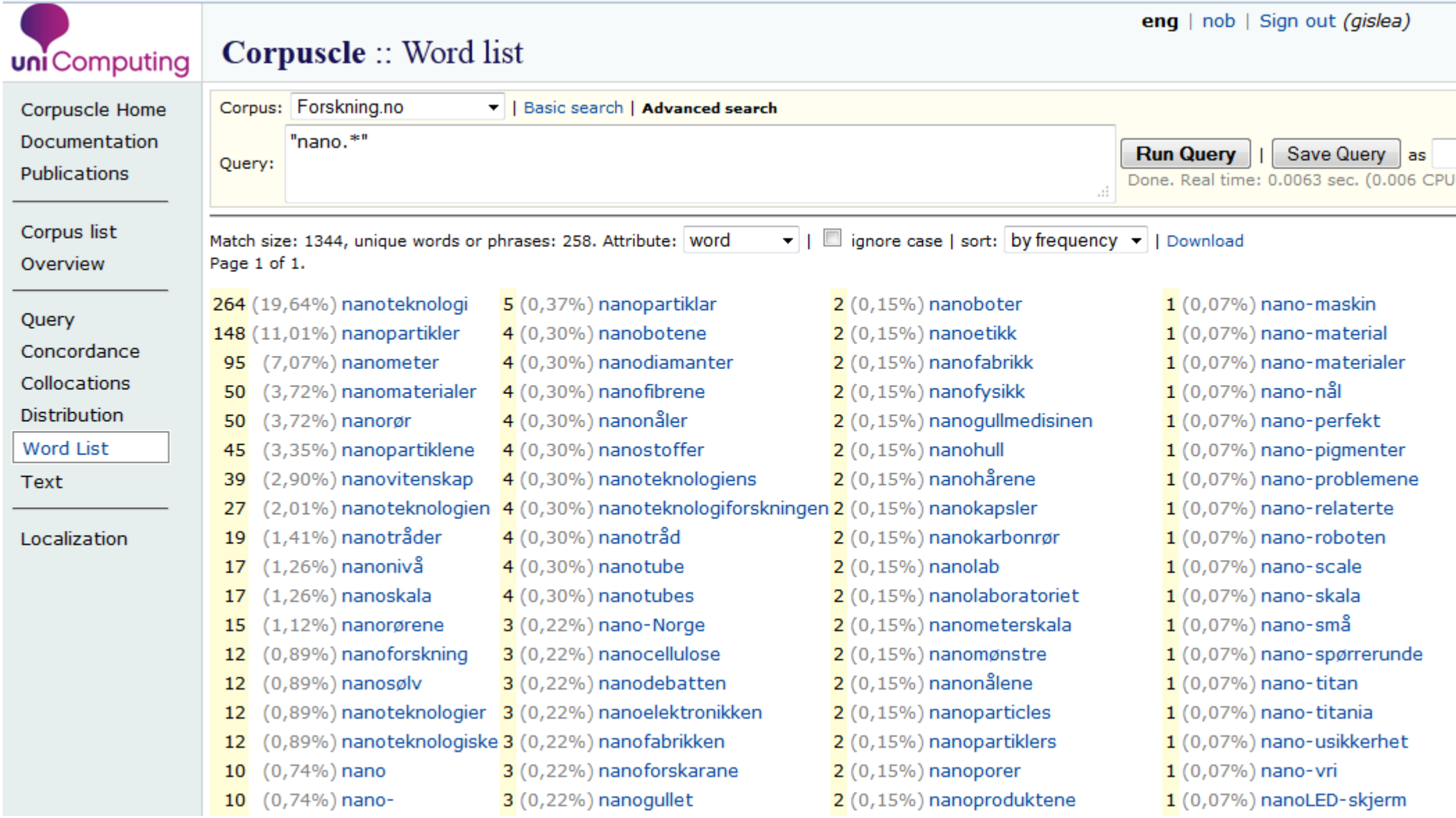
24.337948	tardive dyskinesier	term candidate
23.767403	patagonske tannfisken	term candidate
23.767403	kilhodet dvergkaiman	term candidate
23.636066	lipsum lorem	term candidate
23.60035	interpleural regionalanalgesi	term candidate
22.960312	retinitis pigmentosa	term candidate
22.888853	nucleus accumbens	term candidate
22.63496	solar plexus	term candidate
22.301065	respiratorisk syncytialt	term candidate
21.838442	perfektum partisipp	term candidate
21.754522	amyotrofisk lateralsklerose	term candidate
21.71328	sri lankiske	term candidate
21.538925	spinal muskelatrofi	term candidate
21.331665	erektil dysfunksjon	term candidate
21.202454	pankreas nekrose	term candidate
21.182045	vitro fertilisering	term candidate
21.035398	methyl isocyanat	term candidate
20.822964	kaffir limeblader	term candidate
20.813917	patagonsk tannfisk	term candidate
20.72288	elektrolytisk manganmetall	term candidate
20.643866	polyklorerte bifenyler	term candidate
20.548527	resultatbaserte omfordelingsmodellen	term candidate
20.471565	monokrystallinske silisiumskiver	term candidate
20.464855	kerrs pink	term candidate
20.330614	konjugert linolsyre	term candidate
20.31515	malignt melanom	term candidate
20.233053	residerande festivalmusikaren	term candidate

Kilde: (Lyse & Andersen 2012)

Forskning.no-korpuset

Korpusassistert terminologiarbeid: termkandidater

NHH



Korpusassistert terminologiarbeid: kollokasjoner

NHH

uniComputing

eng | nob | Sign out *(gislea)*

Corpuscle :: Collocations

- Corpuscle Home
- Documentation
- Publications
- Corpus list
- Overview
- Query
- Concordance
- Collocations
- Distribution
- Word List
- Text
- Localization

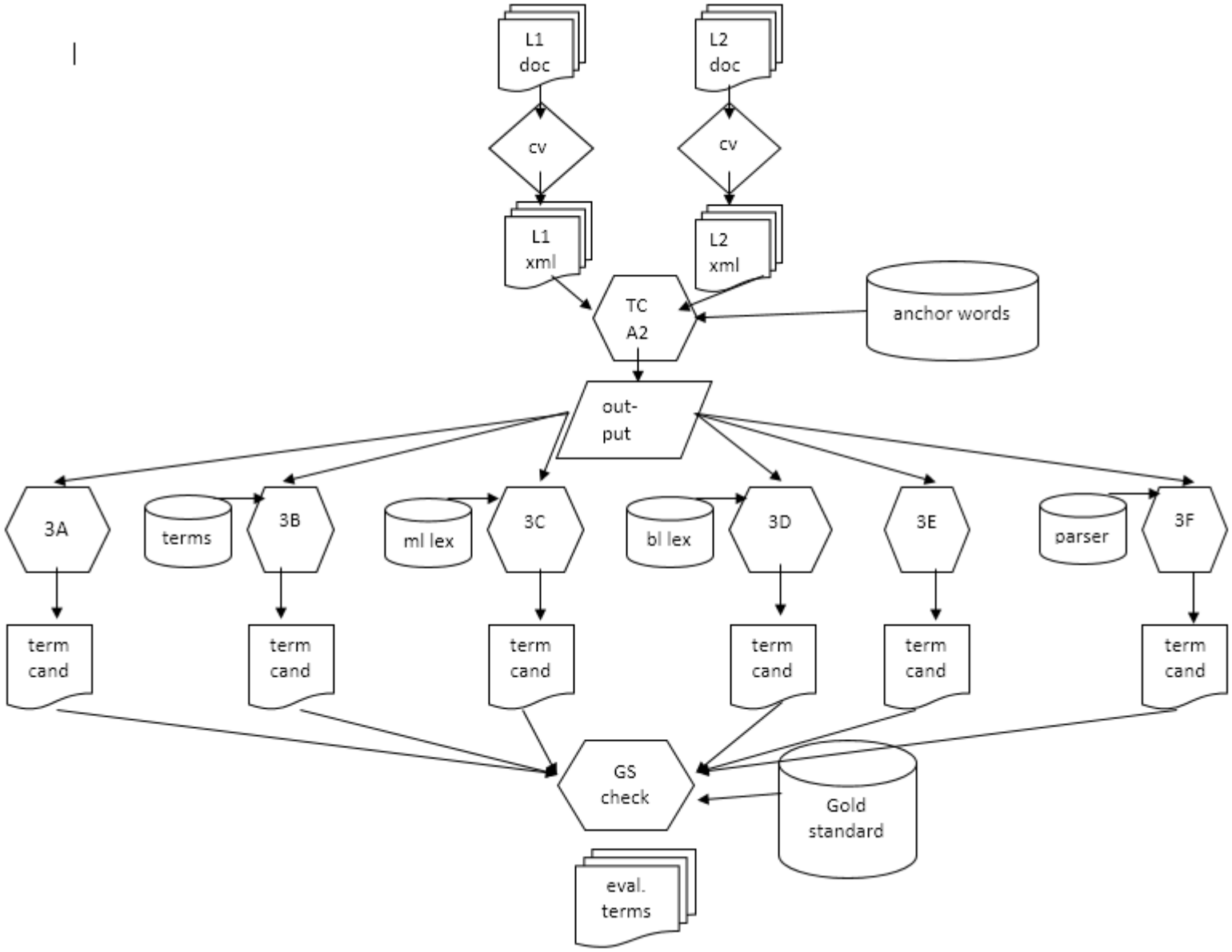
Query: "nano.*" = wordlist

Show collocations by word, left context: 1, right context: 1, sorted by MI * log(Freq) | Download

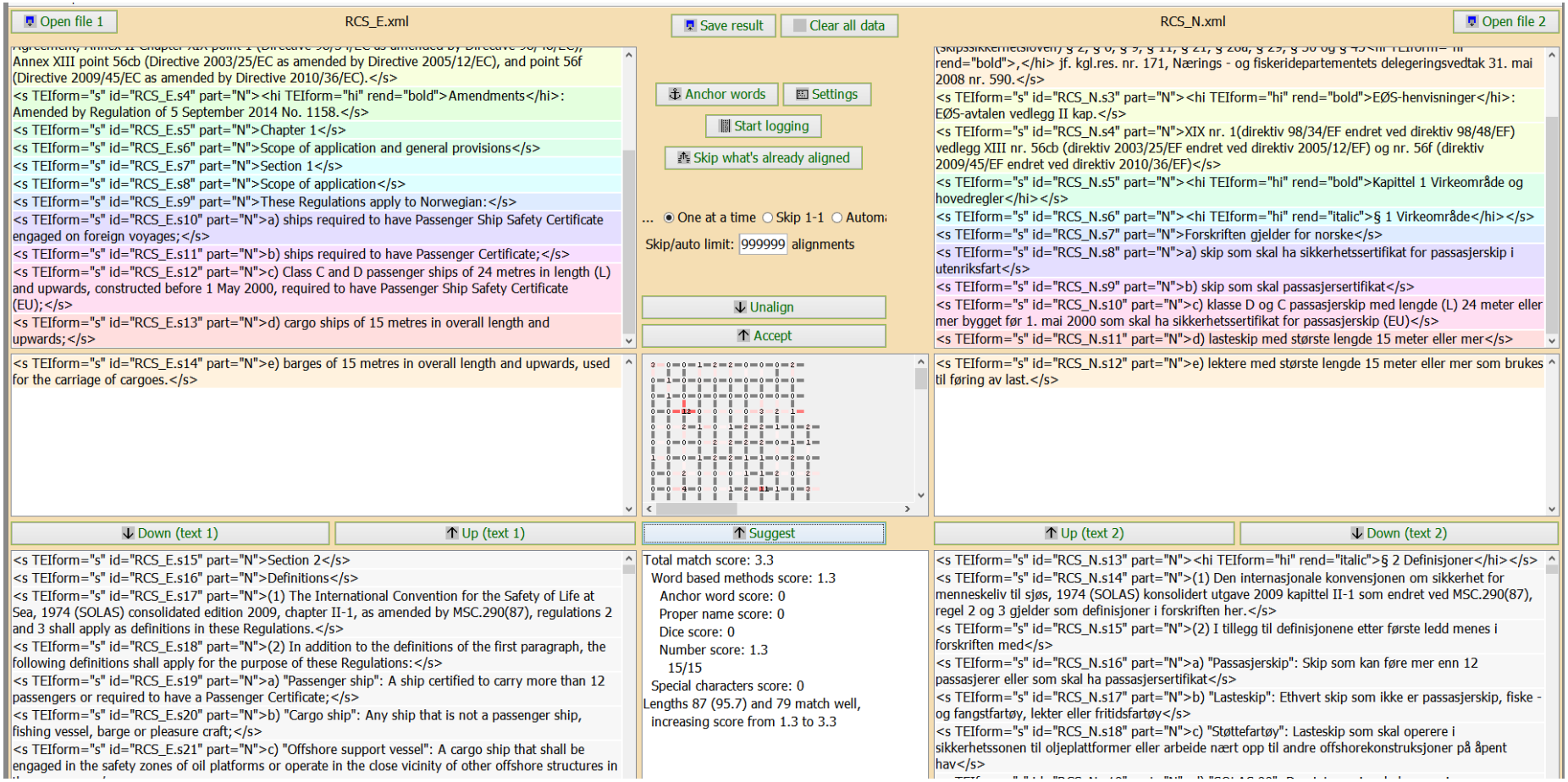
217 collocations calculated; page 1 of 5. Previous Next | Show concordance for selection

Freq.	Rel. freq.	MI	LL	Delta	Collocate
1	0.2500	19.1794	23.5233	-1	*omhandlar* nanopartiklar
1	0.0435	17.9778	126.9900	1	nanopartiklers *giftighet*
1	0.0238	17.1090	266.7489	-1	*Kartlegger* nanopartiklers
3	0.0067	13.6972	3854.6467	1	nanopartiklenes *virkning*
1	1.0000	16.2919	1492.6072	-1	*10^{20}* nanopartikler
1	1.0000	16.2919	1492.6072	-1	*fareklassifisere* nanopartikler
1	1.0000	16.2919	1492.6072	-1	*plateformede* nanopartikler
1	0.0125	16.1794	597.3529	1	nanopartiklers *mobilitet*
1	0.0294	16.0919	132.5654	-1	*milliardar* nanopartiklar
3	0.1000	12.9699	424.6375	-1	*framstilte* nanopartikler
5	0.0388	11.6026	77.3414	-1	*produserte* nanopartikler
1	0.5000	15.2919	1277.4412	-1	*Hoecke* nanopartikler
1	0.5000	15.2919	1277.4412	-1	*ensartete* nanopartikler
1	0.0030	15.1350	3837.6367	-1	*generell* nanopartikkel-effekt
6	0.0282	11.1421	127.7598	-1	*kunstige* nanopartikler
1	0.0027	14.9699	4385.7750	1	nanopartikkelens *overflate*
1	0.0027	14.9699	4385.7750	1	nanopartikkels *overflate*
2	0.0263	12.7615	62.8505	-1	*selvlysende* nanopartiklene
1	0.1250	13.2919	831.5433	-1	*ensartede* nanopartikler
1	0.0028	12.7077	3000.5625	1	nanopartiklar *kunstig*

System architecture for multilingual TE



Parallellstilling av tekster fra Sjøfartsdirektoratet



Termekstraksjon: metoder og teknologi

1. Kandidatekstraksjon ved regulære uttrykk
   - fra parallellstilling på setningsnivå til ordnivå
2. Sjekk av ordtilfang i termbaser:
   - kontroll av enkeltord og ordsekvenser mot termer i Termportalen
   - sjekk om samme oversettelsesrelasjon
3. Nyordsanalyse
   - sjekk ord og ordsekvenser (n-gram) mot ordtilfang i Norsk aviskorpus
4. Statistiske assosiasjonsmål
   - sterke kollokasjoner = termkandidater (*spooling device*; *universell utforming*)
5. Sjekk av ordtilfang i leksikalske databaser
   - engelsk-norsk/norsk-engelsk
6. Parsing av norsk og engelsk vha. dataverktøy fra INESS (jf. under)

Step 3A: Pattern matching

- Premise: recognisable patterns in sentence and paragraph structure, punctuation, etc. suggesting termhood
- Extraction based on regular expressions (perl)

<s>b) barges;</s>	<s>d) lektere</s>
<s>The spooling device shall:</s>	<s>Spoleapparatet skal:</s>
<s>a) initial certification upon changes in use;</s>	<s>a) førstegangssertifisering ved endret bruk</s>
<s>e) handrails, corridors and passageways, doorways, doors, lifts, vehicle decks, passenger lounges, accommodation and washrooms shall be …	<s>e) Håndlister, korridorer og ganger, døråpninger, dører, heiser, bildekk, passasjersalonger, innredning og toaletter skal være …
Wire/chain stoppers shall be dimensioned for a safe working load …	En wire- og kjettingstopper skal være dimensjonert for en sikker arbeidsbelastning …

Step 3B: Check of terminological inventory

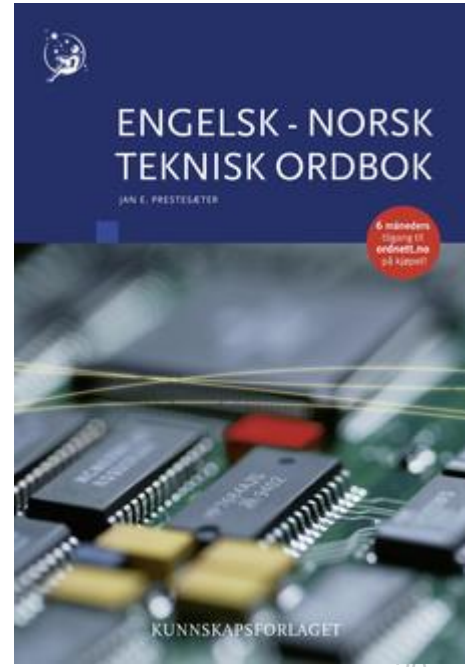
- Premise: if word/sequence of words is already registered as term in other component of *Termportalen*, it has high termhood (it is likely to constitute a term in current context also)
- Question 1: same or different translation relation
- Question 2: same or different domain
- Methodological issue: inflected forms in texts; base form in term base

Step 3C: Neology detection

- Premise: if word/sequence of words can be shown to be a neologism (domain-specific vocabulary), it has high termhood (is likely to be a term)
- Check against inventory of words in large general language corpus (GLC); *Norsk aviskorpus* (Norwegian Newspaper Corpus, NNC; cf. Andersen 2012; Andersen & Hofland 2012)
- Check among neologisms registered in NNC's neology database

Step 3D: Monolingual/bilingual lexicon lookup

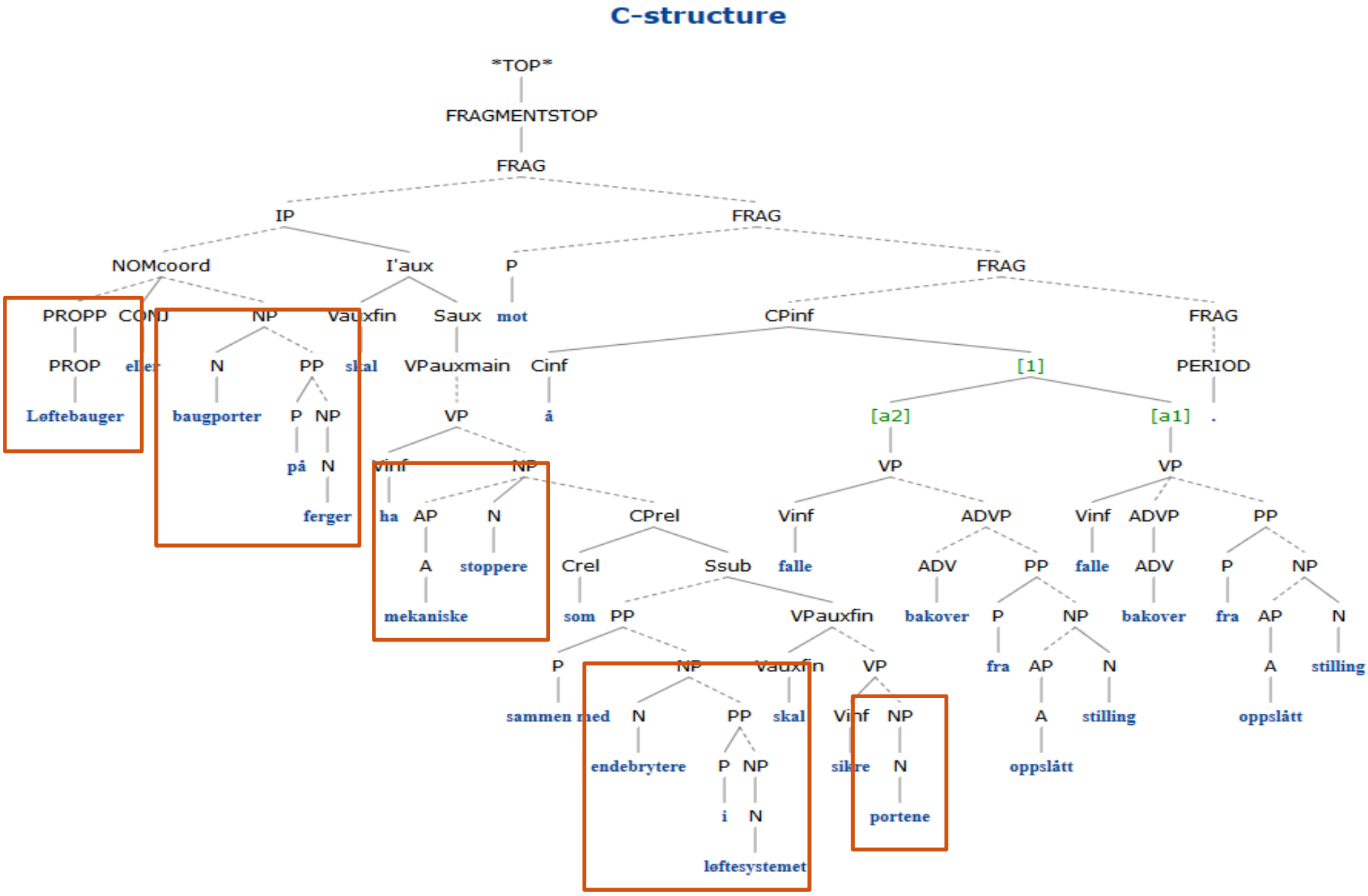
- Premise: if word/sequence of words is found among the lexical inventory in a mono/bilingual technical or specialised dictionary, it has high termhood
- Agreement with *Kunnskapsforlaget* to reuse some of their manuscripts



Step 3E: Association measures (AMs)

- Premise: terms are often constituted as collocations, i.e. words with a strong tendency to co-occur, so strong collocations may be seen as indicators of termhood
- Association measures, statistical measures of unithood/termhood (Heylen & De Hertog 2015)
- Important to select adequate AM for TE, e.g. Pointwise Mutual Information, Chi-square (cf. Lyse & Andersen 2012)
- Collocation patterns should be compared with GLC data (NNC)

Step 3F: Parsing techniques (INESS)



References

Andersen, Gisle, ed. 2012. *Exploring Newspaper Language - Using the web to create and investgate a large corpus of modern Norwegian*. Amsterdam: John Benjamins.

Andersen, Gisle, and Knut Hofland. 2012. Building a large monitor corpus based on newspapers on the web. In *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian*, edited by G. Andersen. Amsterdam: John Benjamins.

Heylen, Kris, and Dirk De Hertog. 2015. Automatic term extraction. In *Handbook of Terminology*, edited by H. J. Kockaert and F. Steurs. Amsterdam: John Benjamins.

Lyse, Gunn Inger, and Gisle Andersen. 2012. Collocations and statistical analysis of n-grams. In *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian*, edited by G. Andersen. Amsterdam: John Benjamins.

Rosén, Victoria. 2012. Exploring corpora through syntactic annotation. In *Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian*, edited by G. Andersen: John Benjamins.