

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**



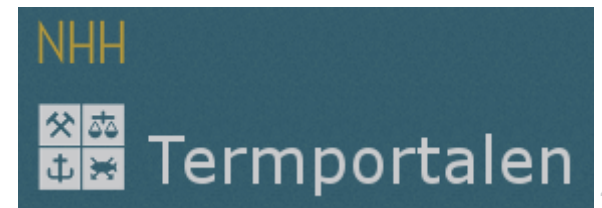
CLARINO, CLARIN og CLARA

- 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 terminologiresurser

CLARIN ERIC
Common Language Resources and Technology Infrastructure



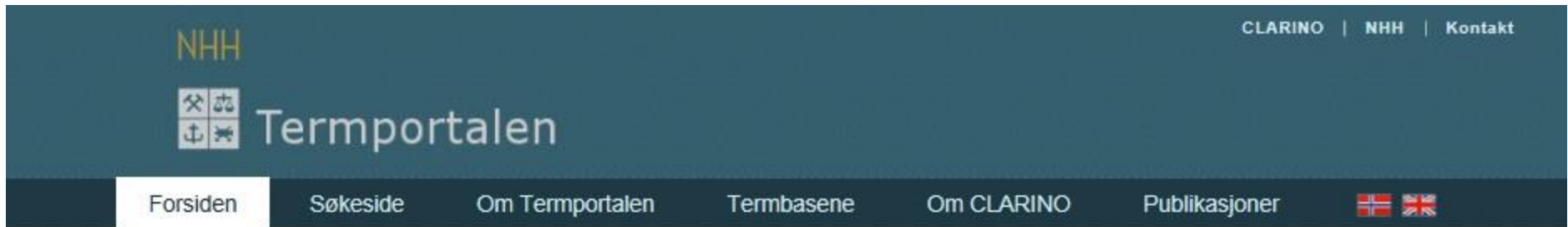
CLARINO



Termportalen på terminologi.no

Enhetlig tilgang til distribuerte ressurser

NHH



Du er her: [Forsiden](#)

Termportalen ved NHH

Velkommen til Terminologi.no – en nasjonal portal for terminologi!

I Norge har det i lang tid vært etterlyst et initiativ for å etablere en nasjonal termportal. Dette arbeidet er nå i gang i regi av CLARINO, et nasjonalt prosjekt for forskningsinfrastruktur finansiert av Norges Forskningsråd (NFR).

Termportalen vil gi brukere fri tilgang til norsk terminologi innenfor en lang rekke fagområder, og gjøre det mulig å søke på tvers av termbaser og fagområder.

Termportalen er under utvikling og er i ferd med å fylles med innhold. Arbeidet med Termportalen er organisatorisk tilknyttet Institutt for fagspråk og interkulturell kommunikasjon, Norges Handelshøyskole (NHH), og er et samarbeid mellom NHH, Universitetet i Oslo og Universitetet i Bergen.

NHH

Versjon: Alfa 1

Drift: Eining for digital dokumentasjon, UiO

Termbaser 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



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	erekttil 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



Corpuscle :: Concordance

eng | nob | Sign out (gi)

Corpuscle Home
Documentation
Publications

Corpus list
Overview

Query

Concordance

Collocations
Distribution
Word List
Text

Localization

Corpus: | [Basic search](#) | [Advanced search](#)

Query: "nanoteknolog.*" :: language = "nob" & author = "Thomas Evensen" & date = "2005-01.*"

Run Query

Save Query

Done. Real time: 0.0061 sec.

Hit 1 - 30 of 60 | | Go to: | Download (Excel mode) | Type: | Show line filter | Show: | Hide

count	cpos	match
1	29657	lanomat-programmet i Forskningsrådet, ønsker nå en offentlig debatt om rammebetingelser for nanoteknologi
2	29685	positivt med denne teknologien. </s> <s> På sikt kan både syn og hørsel forbedres ved hjelp av nanoteknologi
3	29771	eget prosjekt med en arbeidsgruppe som nå har utredet nasjonale forskningsbehov knyttet til nanoteknologi
4	30046	dningen prøver vi derfor å skape en balansert offentlig debatt om muligheter og risiko innenfor nanoteknologi
5	30160	g Høvik mener at næringslivet vil bli nødt til å være mer kunnskapsintensivt i framtiden ettersom nanoteknologi
6	30244	for konkurrere internasjonalt med kompetanse som hovedfortrinn. </s> <s> Da vil kunnskap om nanoteknologi
7	30261	nologi være viktig, sier Høvik. </s> </div> <div> <s> En av de største enkeltsatsningene innenfor nanoteknologi
8	30350	lle våpen. </s> </div> <div> <s> – Det kan være betenkelig at mye av forskningen i USA innenfor nanoteknologi
9	30447	erialer til mulige helse og miljøtrusler. </s> </div> <div> <s> – Effektene av hybridsystemer, hvor nanoteknologi
10	30714	b i menneskeorganismen? </s> </div> <div> <s> Hudkrem produsenten L'Orèals bruker allerede nanoteknologi
11	30752	disse utfordringene, og den konkluderer med at «føre-vår-prinsippet» må gjelde også innenfor nanoteknologi
12	30797	> Forskning og kompetansebygging på etiske, rettslige og samfunnsmessige aspekter innenfor nanoteknologi
13	4157184	lanomat-programmet i Forskningsrådet, ønsker nå en offentlig debatt om rammebetingelser for nanoteknologi
14	4157212	positivt med denne teknologien. </s> <s> På sikt kan både syn og hørsel forbedres ved hjelp av nanoteknologi
15	4157298	eget prosjekt med en arbeidsgruppe som nå har utredet nasjonale forskningsbehov knyttet til nanoteknologi
16	4157573	dningen prøver vi derfor å skape en balansert offentlig debatt om muligheter og risiko innenfor nanoteknologi
17	4157687	g Høvik mener at næringslivet vil bli nødt til å være mer kunnskapsintensivt i framtiden ettersom nanoteknologi
18	4157771	for konkurrere internasjonalt med kompetanse som hovedfortrinn. </s> <s> Da vil kunnskap om nanoteknologi

Korpusassistert terminologiarbeid: termkandidater



Corpuscle :: Word list

Corpus: Forskning.no | [Basic search](#) | [Advanced search](#)

Query: "nano.*"

[Run Query](#) | [Save Query](#) as
Done. Real time: 0.0063 sec. (0.006 CPU)

Match size: 1344, unique words or phrases: 258. Attribute: word | ignore case | sort: by frequency | [Download](#)
Page 1 of 1.

264 (19,64%) nanoteknologi	5 (0,37%) nanopartiklar	2 (0,15%) nanoboter	1 (0,07%) nano-maskin
148 (11,01%) nanopartikler	4 (0,30%) nanobotene	2 (0,15%) nanoetikk	1 (0,07%) nano-material
95 (7,07%) nanometer	4 (0,30%) nanodiamanter	2 (0,15%) nanofabrikk	1 (0,07%) nano-materialer
50 (3,72%) nanomaterialer	4 (0,30%) nanofibrene	2 (0,15%) nanofysikk	1 (0,07%) nano-nål
50 (3,72%) nanorør	4 (0,30%) nanonåler	2 (0,15%) nanogullmedisinen	1 (0,07%) nano-perfekt
45 (3,35%) nanopartiklene	4 (0,30%) nanostoffer	2 (0,15%) nanohull	1 (0,07%) nano-pigmenter
39 (2,90%) nanovitenskap	4 (0,30%) nanoteknologiens	2 (0,15%) nanohårene	1 (0,07%) nano-problemene
27 (2,01%) nanoteknologien	4 (0,30%) nanoteknologiforskningen	2 (0,15%) nanokapsler	1 (0,07%) nano-relaterte
19 (1,41%) nanotråder	4 (0,30%) nanotråd	2 (0,15%) nanokarbonrør	1 (0,07%) nano-roboten
17 (1,26%) nanonivå	4 (0,30%) nanotube	2 (0,15%) nanolab	1 (0,07%) nano-scale
17 (1,26%) nanoskala	4 (0,30%) nanotubes	2 (0,15%) nanolaboratoriet	1 (0,07%) nano-skala
15 (1,12%) nanorørene	3 (0,22%) nano-Norge	2 (0,15%) nanometerskala	1 (0,07%) nano-små
12 (0,89%) nanoforskning	3 (0,22%) nanocellulose	2 (0,15%) nanomønstre	1 (0,07%) nano-spørrerunde
12 (0,89%) nanosølv	3 (0,22%) nanodebatten	2 (0,15%) nanonålene	1 (0,07%) nano-titan
12 (0,89%) nanoteknologier	3 (0,22%) nanoelektronikken	2 (0,15%) nanoparticles	1 (0,07%) nano-titania
12 (0,89%) nanoteknologiske	3 (0,22%) nanofabrikken	2 (0,15%) nanopartiklers	1 (0,07%) nano-usikkerhet
10 (0,74%) nano	3 (0,22%) nanoforskarane	2 (0,15%) nanoporer	1 (0,07%) nano-vri
10 (0,74%) nano-	3 (0,22%) nanogullet	2 (0,15%) nanoproduktene	1 (0,07%) nanoLED-skjerm

Korpusassistert terminologiarbeid: kollokasjoner



Query: "nano.*" = wordlist

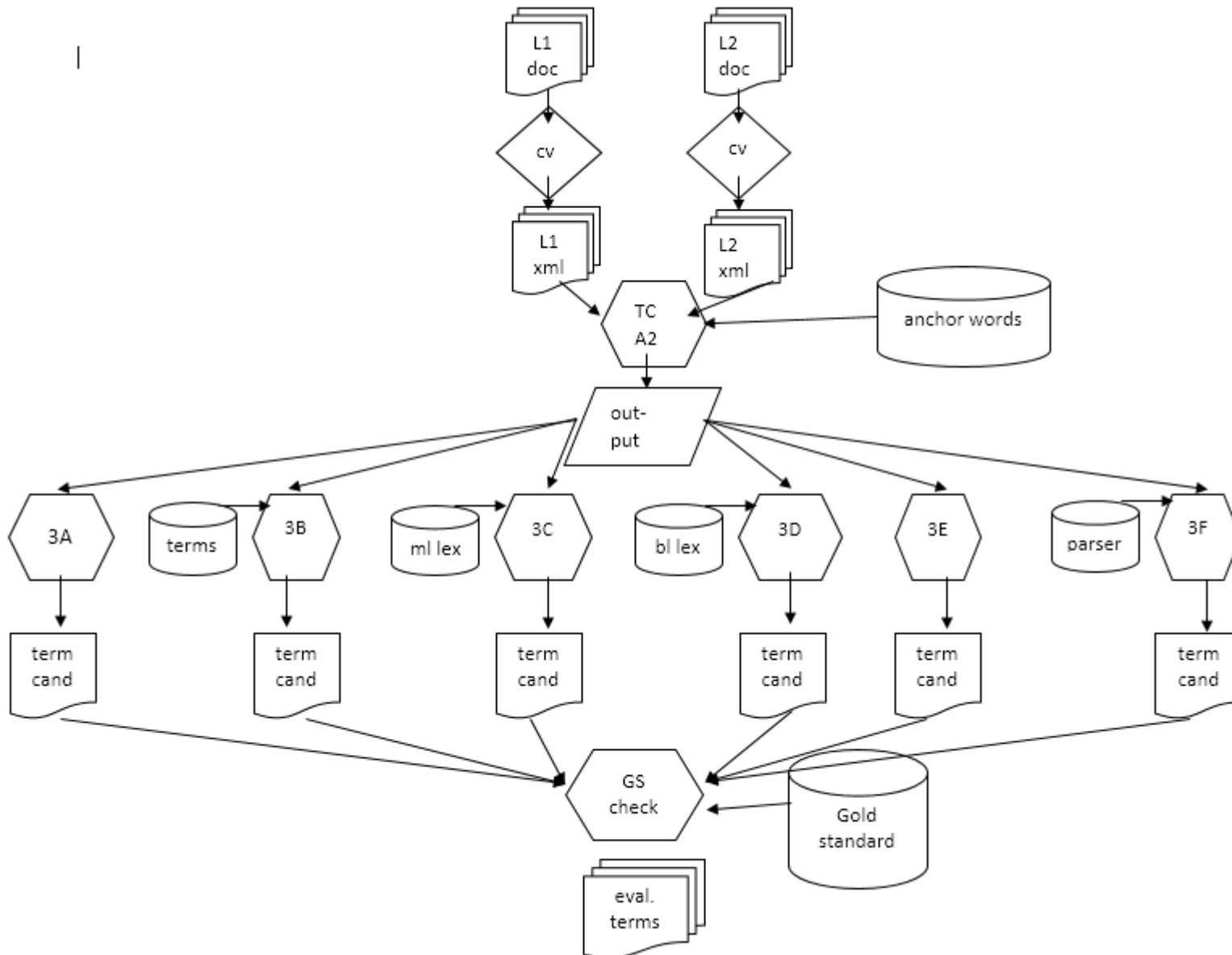
Show collocations by word, left context: 1, right context: 1, sorted by MI * log(Freq) | Download

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	<input type="checkbox"/> omhandlar nanopartiklar
1	0.0435	17.9778	126.9900	1	<input type="checkbox"/> nanopartiklers <i>giftighet</i>
1	0.0238	17.1090	266.7489	-1	<input type="checkbox"/> <i>Kartlegger</i> nanopartiklers
3	0.0067	13.6972	3854.6467	1	<input type="checkbox"/> nanopartiklens <i>virkning</i>
1	1.0000	16.2919	1492.6072	-1	<input type="checkbox"/> 10 ²⁰ nanopartikler
1	1.0000	16.2919	1492.6072	-1	<input type="checkbox"/> <i>fareklassifisere</i> nanopartikler
1	1.0000	16.2919	1492.6072	-1	<input type="checkbox"/> <i>plateformede</i> nanopartikler
1	0.0125	16.1794	597.3529	1	<input type="checkbox"/> nanopartiklers <i>mobilitet</i>
1	0.0294	16.0919	132.5654	-1	<input type="checkbox"/> <i>milliarder</i> nanopartiklar
3	0.1000	12.9699	424.6375	-1	<input type="checkbox"/> <i>framstilte</i> nanopartikler
5	0.0388	11.6026	77.3414	-1	<input type="checkbox"/> <i>produserte</i> nanopartikler
1	0.5000	15.2919	1277.4412	-1	<input type="checkbox"/> <i>Hoecke</i> nanopartikler
1	0.5000	15.2919	1277.4412	-1	<input type="checkbox"/> <i>ensartete</i> nanopartikler
1	0.0030	15.1350	3837.6367	-1	<input type="checkbox"/> <i>generell</i> nanopartikkel-effekt
6	0.0282	11.1421	127.7598	-1	<input type="checkbox"/> <i>kunstige</i> nanopartikler
1	0.0027	14.9699	4385.7750	1	<input type="checkbox"/> nanopartikkelens <i>overflate</i>
1	0.0027	14.9699	4385.7750	1	<input type="checkbox"/> nanopartikkels <i>overflate</i>
2	0.0263	12.7615	62.8505	-1	<input type="checkbox"/> <i>selvlysende</i> nanopartiklene
1	0.1250	13.2919	831.5433	-1	<input type="checkbox"/> <i>ensartede</i> nanopartikler
1	0.0028	12.7077	3000.5625	1	<input type="checkbox"/> nanopartiklar <i>kunstig</i>



System architecture for multilingual TE



Parallellstilling av tekster fra Sjøfartsdirektoratet



The screenshot shows a text alignment application with two panes: RCS_E.xml on the left and RCS_N.xml on the right. The central control panel includes buttons for 'Anchor words', 'Settings', 'Start logging', 'Skip what's already aligned', 'Unalign', and 'Accept'. A central window displays a grid of alignment scores. The left and right panes show XML text with various tags like <s>, <hi>, and <hi>rend='bold'</hi>.



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>The spooling device shall:</s>

<s>a) initial certification upon changes in use;</s>

<s>e) handrails, corridors and passageways, doorways, doors, lifts, vehicle decks, passenger lounges, accommodation and washrooms shall be ...

Wire/chain stoppers shall be dimensioned for a safe working load ...

<s>d) lektere</s>

<s>Spoleapparatet skal:</s>

<s>a) førstegangssertifisering ved endret bruk</s>

<s>e) Håndlister, korridorer og ganger, døråpninger, dører, heiser, bildekk, passasjersalonger, innredning og toaletter skal være ...

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

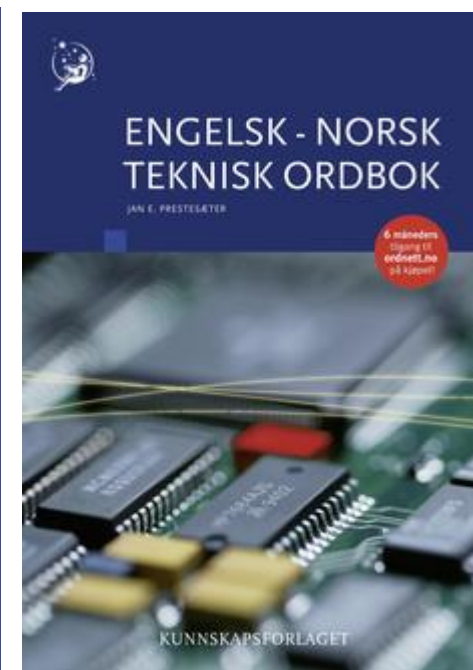


Step 3C: Neology detection

- 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





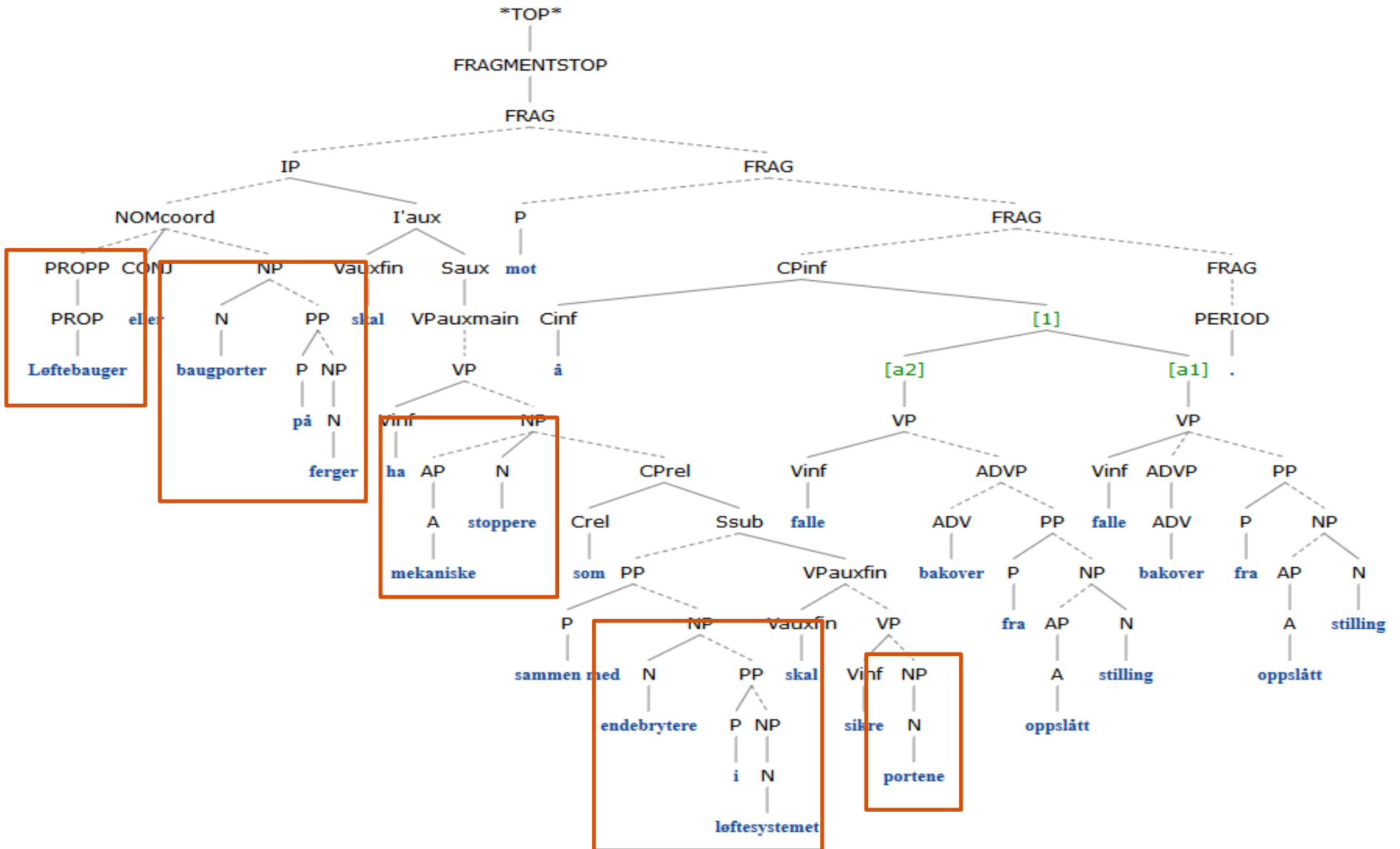
Step 3E: Association measures (AMs)

- 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)

C-structure





References

Andersen, Gisle, ed. 2012. *Exploring Newspaper Language - Using the web to create and investigate 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.