KB-N (KunnskapsBank for Norsk økonomisk-administrativt domene): presentasjon av et korpusbasert terminologiprosjekt

Kai Innselset & Magnar Brekke

NHH

Abstract

This paper reports early results of a 3-year project at the Department of Professional and Intercultural Communication, NHH, aiming to establish a knowledge-bank for economic-administrative domains. The underlying assumption is that domain-focal special knowledge is embedded in text produced typically by domain experts for documentary, argumentative, didactic or general communicative purposes. It is further assumed that the essential knowledge content is embedded in relatively language independent concepts and manifested through relatively language specific terminology (*in casu* English and Norwegian), and that such terminology is stratified with respect to domain specificity ranging from general shared terms down to a small set of domain-focal terms.

KB-N represents the culmination of efforts to refine and integrate computational strategies and tools in NLP for corpus design and analysis, automatic and semi-automatic extraction, representation, and retrieval of terminology, dynamic thesaurus building, dynamic display of authentic collocational and phraseological evidence, etc. The paper attempts to demonstrate (via screen-shots) a working version of the integrated KB-N software suite for handling corpus-based concordancing, term extraction and selection, thesaurus building on-the-fly etc., in the context of a discussion of general theoretical and methodological issues. A range of applications is envisaged for the knowledge-bank. Designed as a web-enabled resource it will be available for e.g. systematic terminology registration and look-up, textbook authoring, e-learning as well as machine translation.

1. Teoretisk perspektiv

KB-N (KunnskapsBank for Norsk økonomisk-administrativt domene) (http://www.nhh.no/fsk/sff/kbn/) er et treårig prosjekt innenfor NFRs KUNSTI-program (http://program.forskningsradet.no/kunsti/om/) i samfinansiering med NHH og i samarbeid med Aksis (Avdeling for kultur, språk og informasjonsteknologi) ved UiB (http://www.aksis.uib.no/).

KB-N-konseptet kombinerer et sett av noe ulike moduler og subsystemer, med røtter innenfor ulike vitenskapshistoriske tradisjoner og med ulike anvendelser innenfor moderne språkteknologi og kunnskapsrepresentasjon. KB-N utfordrer den dikotomien som lenge har dominert forskningstradisjonene i krysningsfeltet mellom språk og teknologi, som i stikkords form kan karakteriseres som motsetningen mellom teoretisk uinformert samleaktivitet og empirisk uinformert parametrisering. I et overordnet komplementært perspektiv søker KB-N å bidra til den tilnærming som i senere år har skjedd mellom empiriske og teoretiske innfallsvinkler nettopp gjennom kombinasjonen og integrasjonen av aktuelle metoder og resultater.

Datamaskinen er blitt for tekstbasert empiri (herunder korpuslingvistikken) det mikroskopet lenge har vært for f.eks. biologisk empiri, men tilgangen til nærmest uendelige datamengder har tydeliggjort behovet for videre teori- og metodeutvikling. “Knowledge management”, som

hadde en viss blomstringstid på 90-tallet, er nå igjen blitt et sentralt anliggende for mange virksomhetsområder, men overgangen fra "data" via "informasjon" til "kunnskap" har vist seg å være ganske enkelt uhåndterlig uten automatiserte prosedyrer.

Det teoretiske fundamentet for fagspråklingvistikken ligger i antagelsen av at et identifiserbart utsnitt av "virkeligheten" utgjør et definerbart fagkunnskapsdomene, og videre at den essensielle teoretiske kunnskapen innenfor fagdomenet lar seg representere i språklig/tekstlig form. Den språklige/tekstlige kunnskapsrepresentasjonen utkrystalliserer seg i særlig grad i det som av profesjonen regnes som fagterminologi, og her spesielt i form av komplekse nominalfraser.

En av våre utgangshypoteser var at det langs disse linjer skulle være mulig å prøve holdbarheten i eksisterende fagdomenekategorier med tilhørende kriterier og forsøke å etablere en matrise som i større grad fanger opp prototypiske språklige korrelater for det som ofte er hovedsakelig pragmatiske eller administrative faginndelinger. Vi øynet her muligheten for å avsvekke de Beaugrandes påstand om at "there are indefinitely many 'degrees of freedom' between a 'reality' and its discourse representations", men de foreløpige erfaringene i så måte kan snarere tjene til å bekrefte påstanden.

Økonomisk-administrative fagdomener er som kjent ikke i særlig grad "rene" fag, men består ofte av hybrider avledet fra fag som matematikk, statistikk, organisasjonsteori, atferds-psykologi osv. Mange av de aktuelle fagene beveger seg nettopp i dette skjæringsfeltet mellom "opprinnelige" begreper og termer som er blitt "overtatt" av andre fag og gitt et til dels avvikende begrepsinnhold. Ofte endrer domenekunnskapen seg i prosessen, særlig dersom den i tillegg oversettes til et annet språk der det "samme" fagdomenet har utviklet ganske andre språklige konvensjoner.

Enhver språklig representasjon av et kunnskapsunivers krever presis håndtering også av innholdssiden. Innenfor den klassiske Wüsterske "Allgemeine Terminologielehre" inntar **begrepet** en sentral stilling, men det har vist seg vanskeligere å gi det et presist innhold innenfor mindre eksakte fagdomener som f.eks. innen samfunnsvitenskapene og humaniora. Her har det foregått en interessant teoriutvikling både innenfor nordiske miljøer og den romanskspråklige verden. Dynamisk begrepssystematisering er allerede en integrert funksjon i vår språkteknologiske plattform. Dette vil utgjøre en helt grunnleggende onomasiologisk dimensjon innenfor kunnskapsbasen og bidra til å sikre validiteten, reliabiliteten såvel som den flerspråklige integriteten av de data som registreres, lagres og gjenfinnes.

2. Systemarkitektur

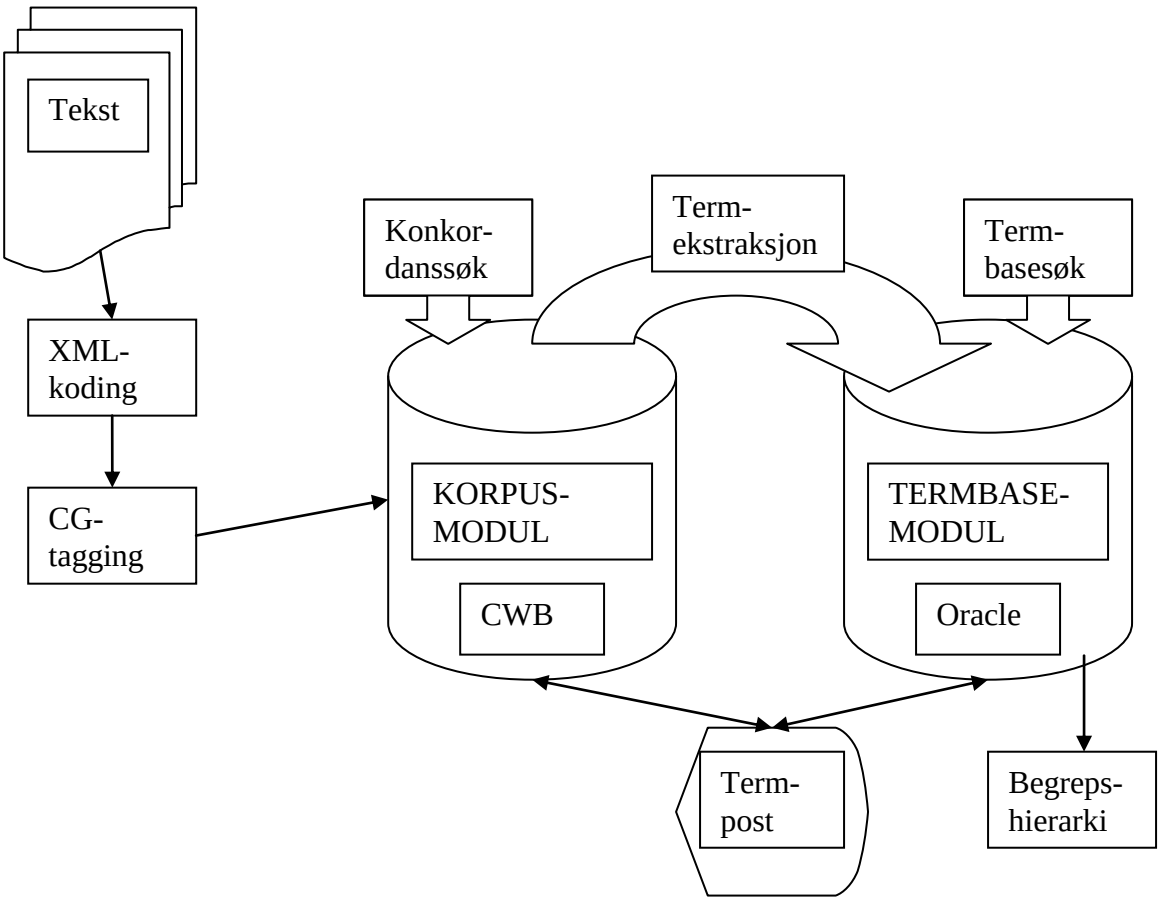
Kunnskapsbank for norsk økonomisk-administrativt domene (KB-N) har som ambisjon å etablere

a) *en språkteknologisk plattform* for å utvikle og utnytte teoretiske og metodiske innsikter i tekstbasert kunnskapsrepresentasjon, med særlig henblikk på å kunne utnyttes i maskinoversettelse (MT) for profesjonell kommunikasjon mellom norsk og engelsk.
b) *en språkressursbank*, bygget på den språkteknologiske plattformen, som videreutvikler, tar i bruk og kvalitetssikrer den fagkompetanse, de elektroniske verktøy og de språkressurser som skal til for at norsk fagspråk sikres fortsatt eksistens

og utvikling på fagdomener som har grunnleggende betydning for norsk samfunns- og næringsliv.

c) et grensesnitt mot *IKT-baserte applikasjoner* som utnytter språkressursbanken for effektiv håndtering av skriftbasert fagkommunikasjon mellom norsk og engelsk.

Den generelle systemarkitekturen framgår av figur 1 nedenfor, et prinsippdiagram for KB-Ns verktøysuite.



Figur 1: KB-N Verktøysuite, prinsippdiagram

Med utgangspunkt i programmoduler utviklet primært av Paul Meurer og Knut Hofland ved Aksis, og i NHHs spesialkompetanse på domenekunnskap, fagspråkforskning og terminologiske databaser (ved Marita Kristiansen, Kari Øvsthus, Magnar Brekke og Kai Innselset) har KB-N under utvikling en integrert programsuite med funksjoner for

- tekst/korpus-håndtering, ordklassemerking og generering av søkbar konkordans med lenking mot aktuell tekstforekomst og termpost
- XML-koding av tekstens makrostruktur
- parallellstilling (alignment)
- termekstraksjon: 1) automatisk 2) forhåndsekserpert
- hierarkisk begrepssystematisering:
  interaktiv bearbeiding av begrepssystem og termhierarkier
- termbankregistrering: tekstbasert flerspråklig kunnskapsrepresentasjon med egne felter for domene, kollokasjon, grammatisk informasjon o.l. i termposten
- termbanksøking

3. Metodologisk tilnærming

3.1 Innhenting av tekstmateriale

KB-Ns prosjektplan legger til grunn materiale fra tre ulike tekstlige kommunikasjonstyper som antas å representere ulike faglighetsnivå manifestert gjennom ekspositoriske, didaktiske, og populariserende tekstfunksjoner. Disse kan eksemplifiseres som henholdsvis forskings-artikkel, lærebok, og avis-/tidsskriftartikkel. Tilgang til domenespesifikk tekst fra økonomisk-administrative subdomener byr på svært ulike utfordringer når det gjelder engelsk og norsk, spesielt på høyere faglighetsnivåer. Gjennom NHHs biblioteksabonnementer har vi rikelig tilgang til aktuell engelsk tekst, men lite på norsk, i og med at de fleste fagfolk publiserer på engelsk. For didaktisk stoff stiller saken seg litt annerledes, og gjennom spesifikke avtaler med norske forlag tar vi sikte på å innhente de nødvendige kvanta norsk tekst. Spesielt Fagbokforlaget har stilt seg svært imøtekommende i så måte. Opphavsrettsproblemet er under drøfting med NFF, som også har signalisert interesse for å finne fram til smidige løsninger som ikke truer forfatternes rettigheter. Bruken av WWW som kyberkorpus byr på spesielle utfordringer, ikke minst når det gjelder kvalitetssikring av form og innhold.

3.2 Termekstraksjon

Tradisjonelt har dette skjedd gjennom s.k. ekserpering, ved at en terminolog leser gjennom et aktuelt dokument og markerer sekvenser som det er grunn til å tro representerer fagtermer. Kontroll av dette skjer normalt i konsultasjon med domeneekspert før innskriving i termbase. Nyere utvikling innen korpuslingvistikken har lagt det metodologiske grunnlaget for termekstraksjon direkte fra et maskinleselig tekstkorpus, som på forhånd er kompilert av utvalgte og (i en eller annen forstand) representative e-tekster. Tekstene gjennomgår ofte ordklassemerking og en grunnleggende statistisk analyse (frekvenslister, konkordanser), som danner basis for avdekking av signifikante forekomster og kompilering av liste med termkandidater. Kvaliteten/treffsikkerheten av termkandidatlisten avhenger av mange ting, men ikke minst graden av filtrering/sanering av ikke-termer og annen støy.

I praksis fungerer "ekserpering" og "ekstrahering" som komplementære tilnærmingsmåter. I KB-N velges fagtekster i nøye konsultasjon med fageksperter før innlegging i dokumentbasen. Et sett av data-algoritmer ekstraherer termkandidater fra korpus (se tabell 1, s. 5), og terminolog/domeneekspert velger termer fra kandidatlisten, etablerer en begrepsstruktur og identifiserer manglende termer, noe som selvfølgelig ikke lar seg automatisere.

Automatisk termutvinning fra elektronisk tekst foregår i KB-N langs to hovedlinjer. Den ene er ekvivalensfinning i parallellstilte flerspråklige tekstversjoner, som også anvendes med stort hell i det nordiske NorNa-prosjektet der gruppen deltar. Innen de fleste subdomener vil dette være en grunnleggende, men svært begrenset metode. Den andre er automatisert termfinning i frittstående tekster på norsk eller engelsk, en betydelig større metodologisk utfordring. Disse tekstene er sammenlignbare i den forstand at de representerer samme kunnskapsdomene og kommunikasjonsform, men er ikke oversatte tekster. Relativt enkle algoritmer for å identifisere lavfrekvente, men faglig signifikante forekomster, har gitt brukbare resultater for engelsk (System Quirks s.k. "Weirdness"-funksjon og "Ferret"-modul), men for norsk er dette tilnærmet jomfruelig mark. Automatisk ordklassemerking (tagging) åpner opp for å utnytte generaliserbare mønster av ordklassesekvenser. Planen er å raffinere metodene i betydelig

grad, og Paul Meurers prosjektinnsats har allerede nådd lovende resultater i å videreutvikle og utprøve kriteriesett og algoritmer. Nedenfor vil vi gi eksempler på bruk av denne metoden.

Som første ledd i arbeidet med å fange opp termkandidater blir det generert en liste over nominalfraser i materialet som etter bestemte kriterier kan representere mulige fagbegreper. Tabell 1 viser typer av kriterier som legges til grunn for filtreringen.

Termkandidatfilter
1. lingvistisk filter: (adj. i positiv form)* + subst (minus genitivform) adj +"og/eller" + adj + subst subst + "-" + "og/eller" +subst 2. Navnegjenkjenner: Denne overstyrer det lingvistiske filteret etter bestemte kriterier (utvikles i et eget prosjekt uavhengig av KB-N) 3. Signifikanskalkyle ("Weirdness") For hver ordform sammenlignes forekomsten i en relevant fagtekst med forekomsten i et stort allmennkorpus (Hoflands norske aviskorpus ved Aksis, ca. 350 mill. ord). "Sjeldne" ord i en fagtekst er ofte fagtermer.

Tabell 1: Kriterier for automatisk ekstraksjon av norske termkandidater

KB-N vil her kunne fungere som en prøvebenk for utvikling og raffinering av slike ekstraksjonsalgoritmer. En termkandidatliste vil også være en pekepinn for hva et dokument omhandler, og følgelig for hvilket subdomene dokumentet skal tilordnes. Her er angivelsen av antall forekomster i dokumentet pr. ekstrahert streng særlig nyttig, samt spredning på tvers av tekster og domener.

3.3 Begrepssystematisering

Begrepssystematisering uttrykker tradisjonelt hovedsakelig generiske/partitive begrepsrelasjoner innenfor et avgrenset fagdomene. Men tekster hentet fra økonomisk-administrative fagdomener viser seg som nevnt ofte å være konglomerater av ulike subdomener og genrer, slik at analysen kan ende med et helt sett av fragmenter av ulike begrepshierarkier, en lite fruktbar framstillingsform. Her har pilotstudier vist at det kan være mer sakssvarende å legge til grunn tekstens "virksomhetsområde" og søke å gi dette et tesaurus-lignende begrepshierarki. KB-N representerer her en mulighet til utprøving av en slik metode i stor skala, både hva angår strukturering og grafisk visning av underliggende elementer og relasjoner.

Med utgangspunkt i de termpostene som er registrert i basen kommer nå den innledende begrepssystematiseringen. Nødvendig fagkunnskap vil komme inn i konsultasjon mellom terminolog og domeneekspert etter at begrepene er grovsystematisert av prosjektmedarbeiderne. Hovedverktøyet i denne prosessen er et dynamisk begrepshierarki som avspeiler den aktuelle kunnskapsstrukturen, der nye begreper fortløpende føyes inn i hovedhierarkiet.

Domeneeksperten vil på dette punktet lett kunne fastslå hvorvidt viktige begreper og termer faktisk mangler, noe som en automatisk ekstraksjonsalgoritme i prinsippet ikke kan dersom disse ikke er representert i korpus-utvalget. Poenget er at språkteknologiske løsninger alltid må søke en optimal interaksjon mellom menneske og maskin.

4. KB-Ns programverktøy og brukergrensesnitt: en kort presentasjon

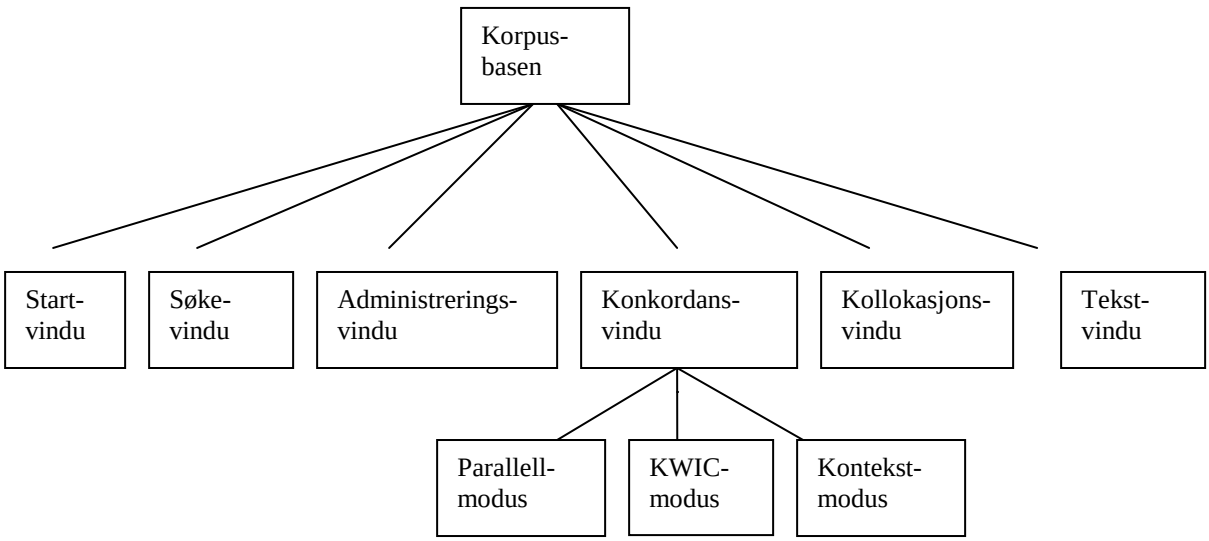
4.1 Innledning

Programvaren som skal presenteres nedenfor er utviklet av forsker Paul Meurer ved AKSIS, UiB i forbindelse med prosjektet "Norsk språkbank - et verktøy for korpussøk og -administrering". I samarbeid med KB-N-prosjektet har Meurer gjort en hel rekke spesialtilpasninger for KB-N. Programvaren er under stadig utvikling, og de "skjermbilder" som vises representerer et foreløpig stadium av det som forventes å bli det endelige. Den foreliggende presentasjonen kan anses som en skriftlig og noe oppdatert versjon av den som ble holdt ved NHHs Institutt for fagspråk og interkulturell kommunikasjon 14. mai 2004. Beskrivelsen av grensesnittet er knyttet opp mot konkrete eksempler i forbindelse med automatisk termekstraksjon, term- og ekvivalentregistrering, begrepssystematisering, og visning av konkordanser og kollokasjoner. Disse termene forklares underveis.

Programvaren består av to hovedmoduler; Korpus-basen og Term-basen.

4.2 Korpus-basen

I skrivende stund er arkitekturen for brukergrensesnittet som vist i figur 2.

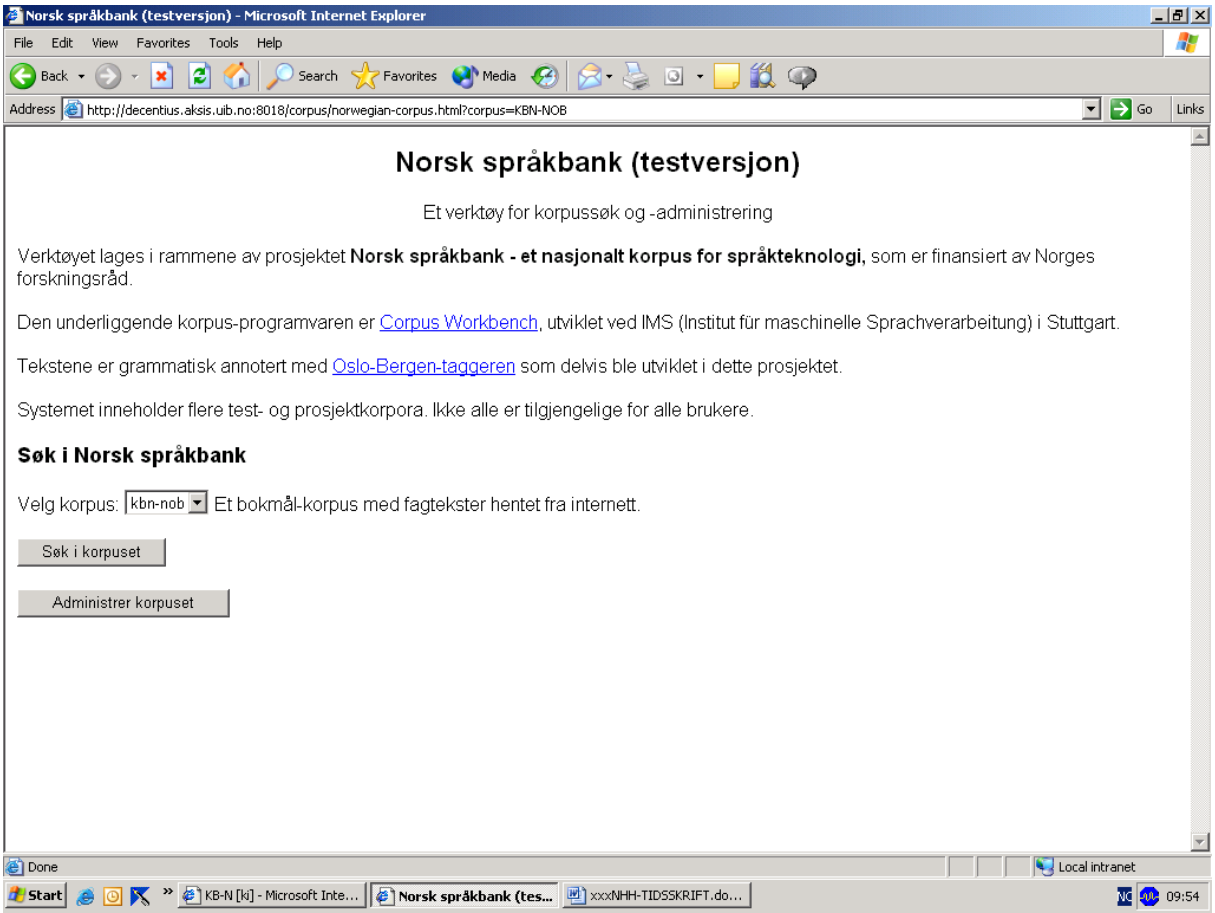


Figur 2: Korpus-basen

I dette avsnittet skal vi ta for oss Start-vinduet, Søke-vinduet og Administrerings-vinduet. De øvrige vinduene med tilhørende modi vil bli beskrevet i forbindelse med visning av konkordanser og kollokasjoner på slutten av demonstrasjonen.

Bilde 1 nedenfor viser Korpus-basens startvindu, som er et fellesvindu for alle korpusbaser som administreres av Paul Meurer ved Aksis, UiB. Her kan man velge blant de basene man har bruksrett til. I øyeblikket har KB-N-prosjektet liggende inne to baser, en norsk base og en engelsk. Disse basene er bygd opp av parallelltekster i strengeste forstand. Hver enkelt dokumentfil i den norske basen har sin motsvarighet i en fil i den engelske basen som representerer henholdsvis måltekst eller kildetekst i oversettingssammenheng. Base-suiten vil

senere bli utvidet med baser for norske og engelske separatfiler, dvs. sammenlignbare filer som ikke inngår i et oversettingsforhold.



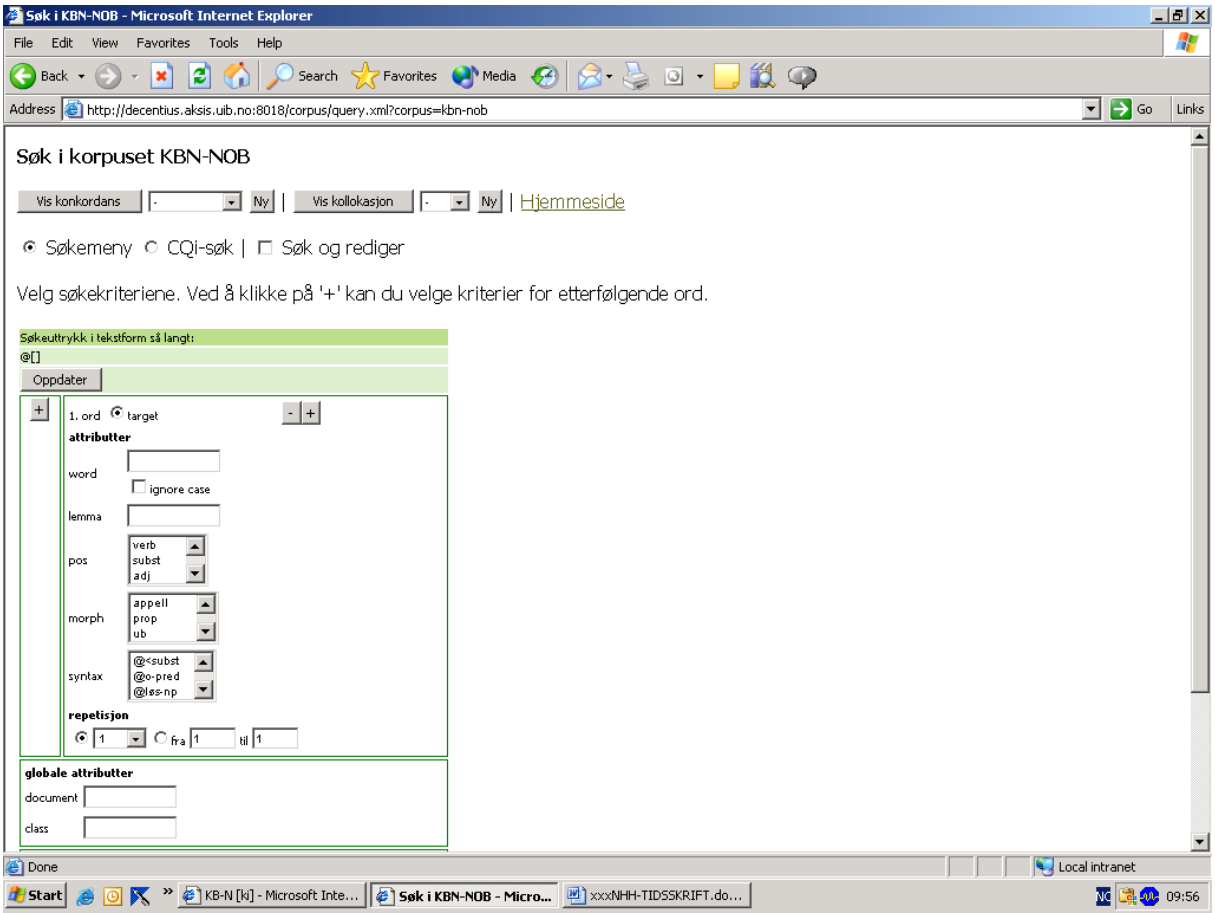
Bilde 1: Korpus-basens startvindu

Fra dette vinduet kan man gå videre til enten Søke-vinduet eller Administrerings-vinduet ved å trykke på henholdsvis “Søk i korpuset” og “Administrer korpuset”.

Bilde 2 (s. 8) viser Korpus-basens søkevindu. Vinduet inneholder foruten selve søkeruten to knapper for henholdsvis konkordans og kollokasjon. Fra rullegardinmenyen for konkordans kan man velge mellom tre visningsmodi: KWIC, kontekst og parallellstilling. I søkeruten kan man blant annet velge mellom søk på lemma (dvs. normalisert oppslagsform) eller ordformer, eller alle ord som tilhører en gitt ordklasse. Søkeruten er dessuten repeterbar. Ved å trykke på plusstegnet kan man få opp ekstra søkeruter, noe som gjør det mulig å foreta mer avanserte søk.

Avanserte brukere kan også velge å skrive inn søkeuttrykk direkte i CQi-spørrespråket i tekstform. Når et søk er lagt inn i søkeruten, vises dette søket i tekstuell form ved overgang til CQi-søk. På denne måten, ved først å bygge så mye av søkeuttrykket som mulig i søkeruten, og så arbeide videre i tekstmodus, kan man lage kompliserte søkeuttrykk som ikke lar seg formulere fullt ut via søkeruten. CQi-søk vil ikke bli demonstrert i denne korte presentasjonen.

Vi kommer tilbake til dette vinduet senere i forbindelse med omtalen av konkordans- og kollokasjonsvisning.



Bilde 2: Korpus-basens søkevindu

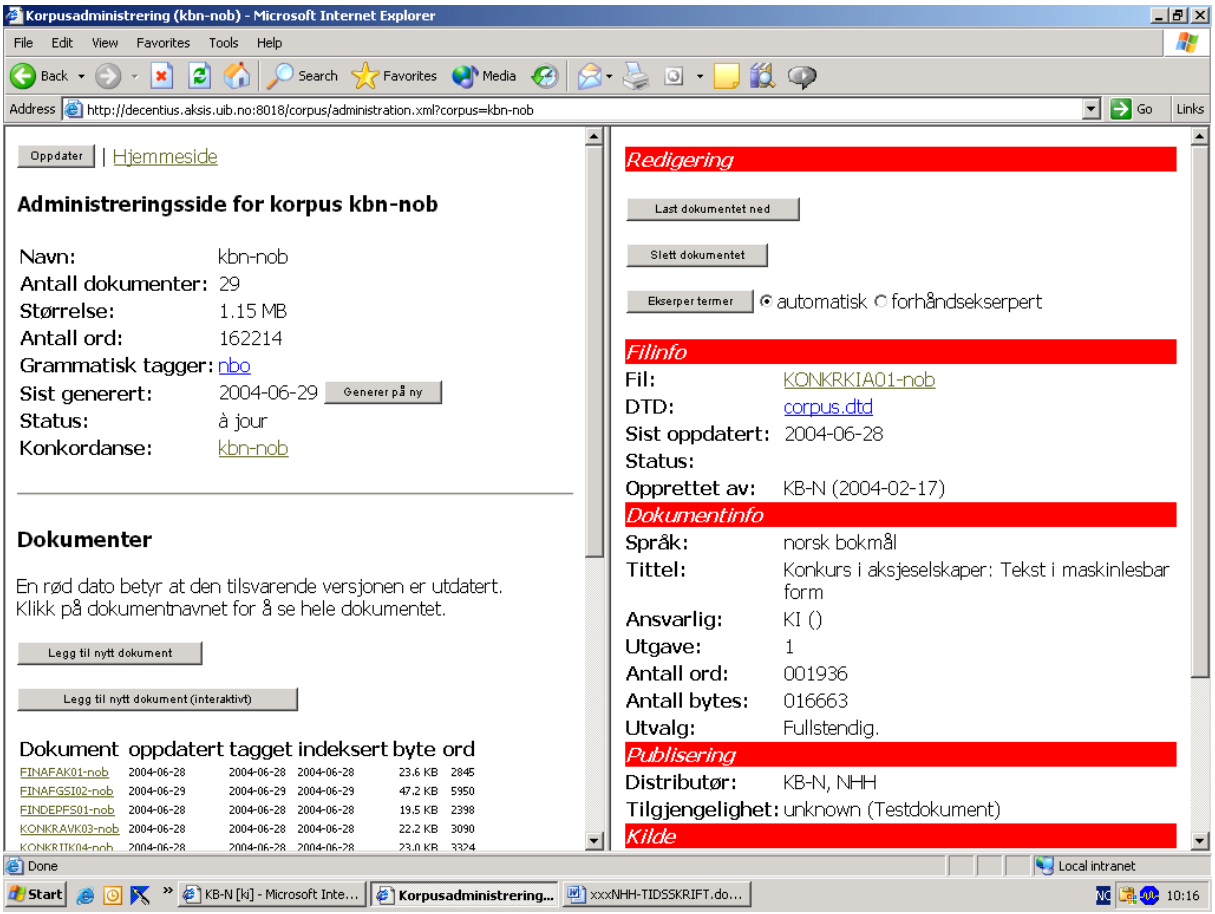
Bilde 3 (s. 9) viser Korpus-basens administreringsvindu. Vinduet består av to rammer. Rammen til venstre inneholder overordnede opplysninger om basen. Øverst finner vi blant annet opplysninger om antall dokumenter (filer) som er lagt inn og den totale størrelsen på dette materialet uttrykt i megabyte og antall ord. I den nederste delen av rammen ligger filkatalogen med tilsvarende størrelsesopplysninger for hver enkelt fil. Det er også fra denne rammen man legger inn nye filer i basen. Disse er forhåndskodet i XML i henhold til XCES-kodestandarden (Corpus Encoding Standard for XML). Ved innlegging i Korpus-basen blir materialet automatisk grammatisk annotert ved hjelp av Oslo-Bergen-taggeren.

Rammen til høyre er i utgangspunktet tom. Innholdet vi her ser, er kommet fram ved å dobbeltklikke på et filnavn i filkatalogen. Rammen inneholder metaopplysninger som forfatter, publikasjonsår o.l. for den valgte filen. De fleste av disse opplysningene gjenspeiler de opplysningene som er lagt inn i filens "header".

Øverst i rammen ser vi tre knapper. Den tredje knappen ovenfra aktiverer termekstraksjon fra den valgte filen. Her kan man velge mellom automatisk ekstraksjon og ekstraksjon basert på manuell forhåndskoding av termer. Vi skal i det følgende vise arbeidsgangen ved termeks-traksjon og påfølgende termregistrering i Term-basen. Dette arbeidet foregår i sin helhet i

Term-basens tredelte vindu, som beskrevet nedenfor, men valg av fil og iverksettingen av ekstraksjonen skjer altså fra Korpus-basens administreringsvindu.

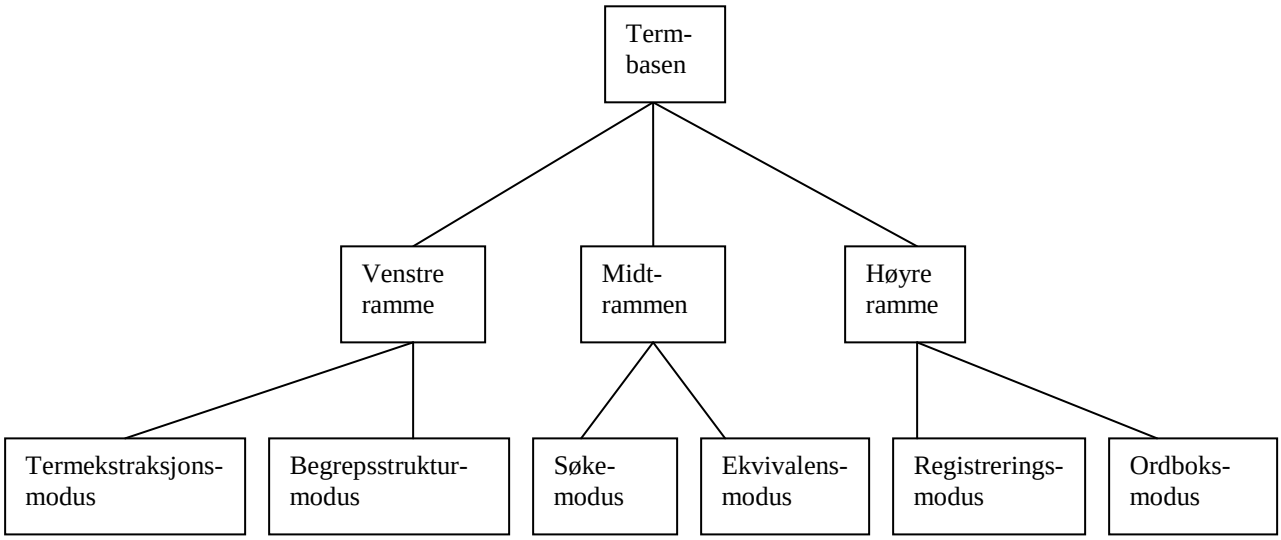
Den filen vi har valgt å arbeide med her, er ikke forhåndskodet med henblikk på termekstraksjon, og vi velger derfor automatisk termekstraksjon som vist i bilde 3 nedenfor.



Bilde 3: Korpus-basens administreringsvindu

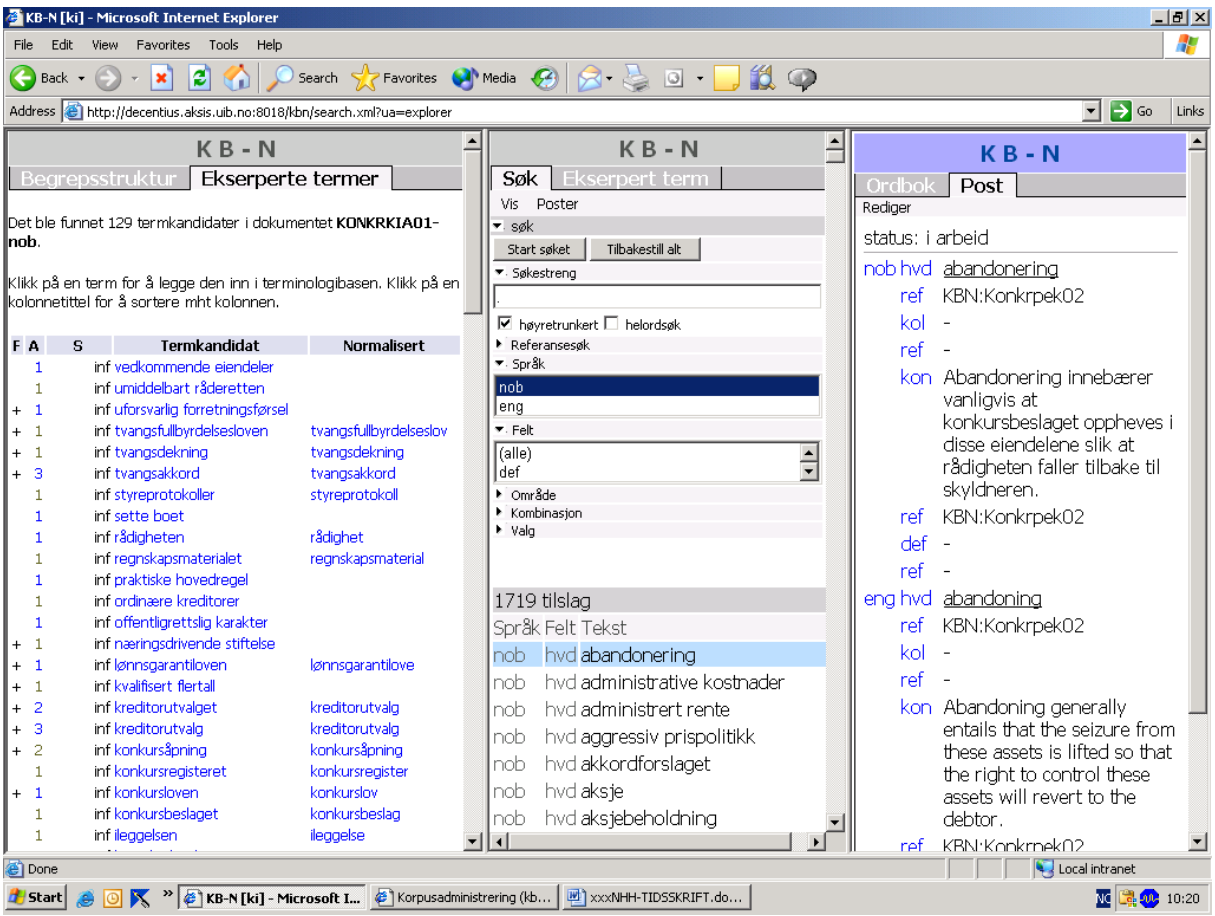
4.3 Term-basen

Fra Korpus-basens administreringsvindu ovenfor valgte vi automatisk termekstraksjon. Resultatet vises i venstre ramme i bilde 4 (s. 10). Vi befinner oss nå ikke lenger i Korpus-basen, men i Term-basen. I skrivende stund er arkitekturen for brukergrensesnittet som vist i figur 3 (s. 10).



Figur 3: Termbase-modulen

Før vi går videre med å demonstrere termregistrering med utgangspunkt i automatisk termekstraksjon, skal vi gi en beskrivelse av Term-basen. Denne modulen har kun ett vindu som til gjengjeld har tre rammer, hver med to modi. Vi skal ta for oss rammene enkeltvis, knyttet opp mot det konkrete eksempelet vi er i gang med.



Bilde 4: Termbase-vinduet med termkandidater i venstre ramme

4.3.1 Venstre ramme

Venstre ramme i bilde 4 har et Termekstraksjons-modus og et Begrepsstruktur-modus. Vi er nå i Termekstraksjons-modus, og rammen viser derfor de termkandidatene som er automatisk ekstrahert fra den valgte filen i korpuset i henhold til kriteriene beskrevet ovenfor i tabell 1. Vi må igjen understreke at det er termkandidater som blir ekstrahert; ikke termer.

Kandidatlisten inneholder følgelig strenger som Oslo-Bergen-taggeren tolker som forekomster av et sett med forhåndsdefinerte typer av nominalfraser (lingvistisk filter) der forekomstene ligger over en definert terskelverdi for frekvens i et allmennspråklig referansekorpus ("weirdness"-filter). Vi arbeider også med å skreddersy en navnegjenkjenner for prosjektet for å kunne ekstrahere navn på ulike institusjoner og organisasjoner med relevans for våre domener. Denne navnegjenkjenneren vil overstyre de ovennevnte filtrene. I tillegg vil det bli utarbeidet en stoppliste primært for å skrelle bort uønskede leksemer (særlig adjektiver) i flerords-termkandidater. F.eks. ønsker vi å få ut 'vedkommende' og 'umiddelbar' i bilde 4, men beholde 'uforsvarlig' og 'ordinær'.

Termkandidatlisten har fem kolonner. De tre kolonneoverskriftene til venstre angir henholdsvis termer som allerede er registrert i Term-basen (F=Finnes), antall forekomster av ekstrahert tegnstreng (A=antall), og "Weirdness"-resultat etter sammenligning av forekomst i aktuelt dokument med forekomst i et stort referansekorpus (S=signifikans). Kolonnen med overskriften "Normalisert" angir lemmatisert form av de ulike ordformene.

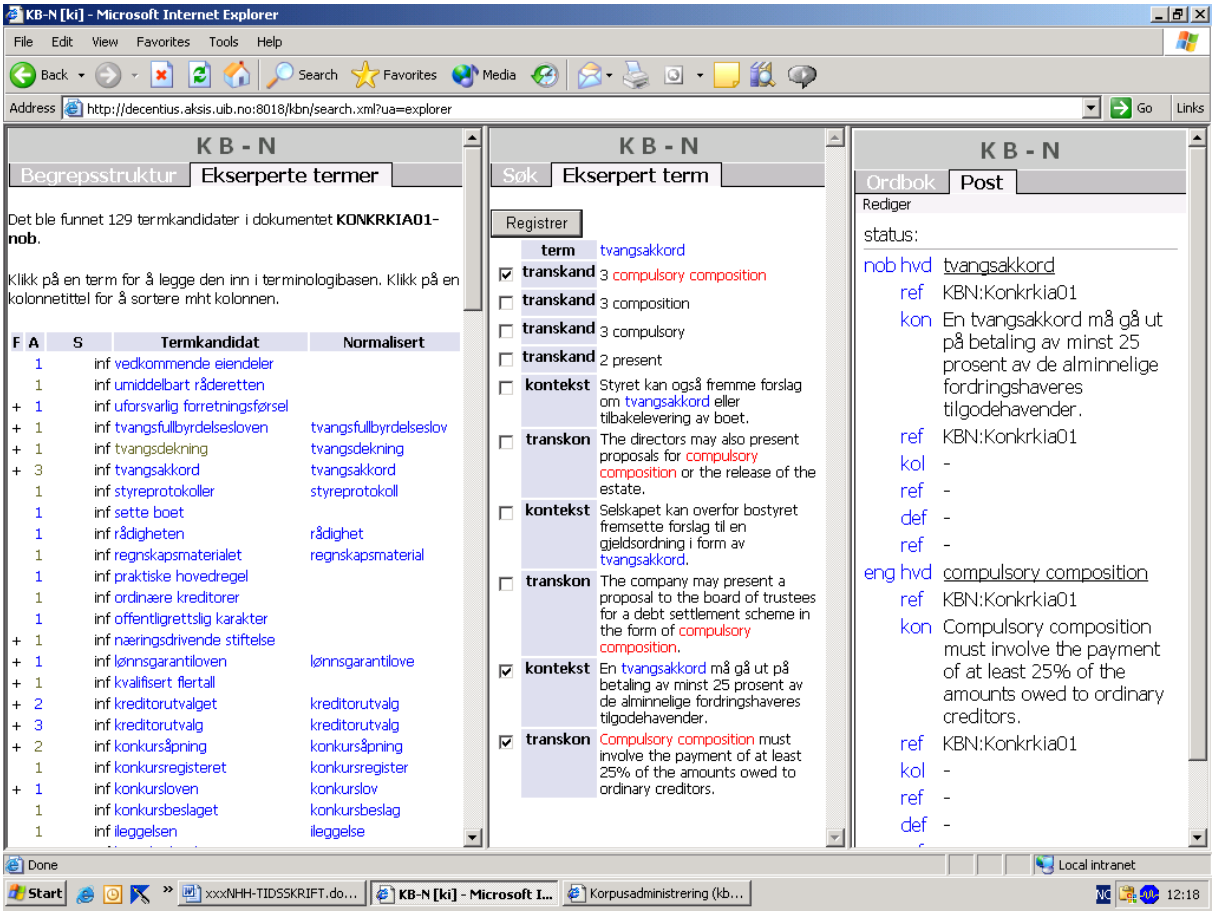
Venstre rammes alternative modus, Begrepsstruktur-modus, kommer vi tilbake til i forbindelse med bygging av hierarkier med utgangspunkt i registrerte termer.

4.3.2 Midtrammen

Midtrammen har modiene Søke-modus og Ekvivalens-modus. Bilde 4 viser rammen i Søke-modus. Her har vi søkt fram samtlige termer som allerede ligger inne i Term-basen. Vi ser begynnelsen av termlisten nederst i rammen. Ved å klikke på en term i denne liste kommer tilsvarende begrepspost opp i høyre ramme der den kan studeres og eventuelt redigeres. Ekvivalens-modus blir beskrevet nedenfor.

4.3.3 Høyre ramme

Denne rammen viser begrepspostene. Rammens to modi er henholdsvis Registrerings-modus og Ordboks-modus. I terminologiarbeidet er det kun Registrerings-modus som er relevant. Ordboks-modus er et visningsmodus for sluttbrukere. Bilde 4 viser rammen i Registrerings-modus ("Post").



Bilde 5: Termbase-vinduet med ekvivalentforslag og kontekster på begge språk i midtrammen

Alle relevante opplysninger om et gitt begrep samles i én begrepspost som er bygd opp av et selvvalgt sett av forhåndsdefinerte felttyper, de fleste repeterbare. Nye opplysninger legges til etter hvert som arbeidet med begrepet skrider fram. Feltene i posten er basert på en fast mal som brukes i forbindelse med automatisk termekstraksjon. I likhet med selve feltinnholdet kan felttypene redigeres etter behov (sletting, nye felt fra lukket liste, repetisjoner osv.).

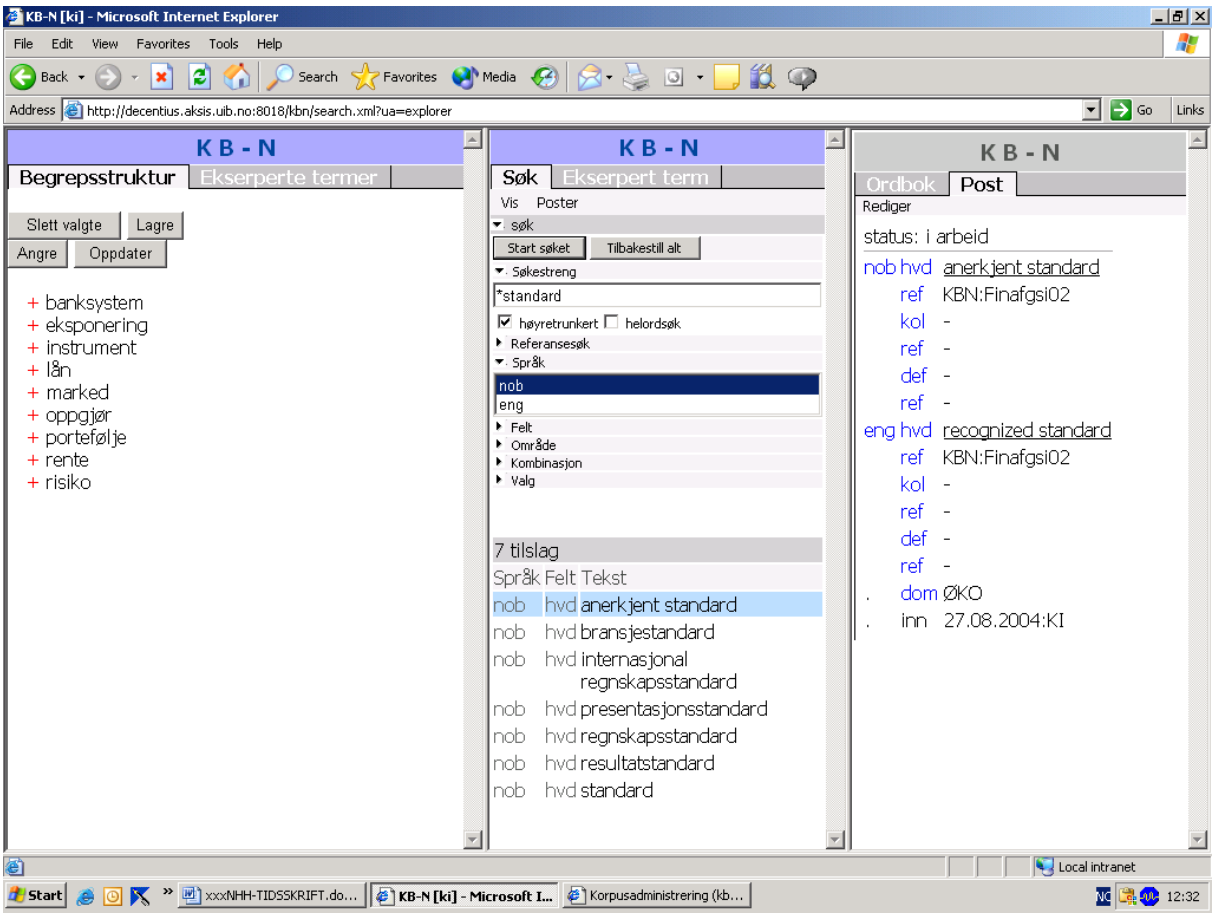
---------- *** ----------

Vi skal nå gå videre med å vise hvordan vi registrerer en norsk term, dens engelske ekvivalent og eventuelle interessante kontekster i en begrepspost. Med utgangspunkt i bilde 4 velger vi ut en termkandidat fra listen i venstre ramme. Vi konsentrerer oss om termkandidat nummer 6 ovenfra, 'tvangsakkord'.

Først dobbeltklikker vi på tallet som markerer antall forekomster. Midtrammen, som før var i Søke-modus, skifter nå til Ekvivalens-modus ("Ekserpert term") og viser den valgte termen, forslag til engelsk ekvivalent, alle kontekster termen forekommer i i filen, og den engelske oversettelsen av disse kontekstene (se bilde 5). Ved å stille markøren over et gitt ekvivalent-forslag vil dette bli uthevet i kontekstene slik at det blir lettere å parallell-lese norsk og engelsk kontekst for å verifisere ekvivalensen. Riktig ekvivalentforslag kan deretter velges

ved at man klikker på avkryssingsruten til venstre for forslaget. På samme måte kan man krysse av for kontekster man vil ha med.

Vi krysser av for ekvivalentforslaget 'compulsory composition' og den siste konteksten på begge språk. Så klikker vi på knappen "Registrer". Alt vi har krysset av for blir automatisk lest inn i en ny termpost som dukker opp i høyre ramme. Denne kan om nødvendig redigeres før lagring.



Bilde 6: Termbase-vinduet innstilt for hierarkibygging

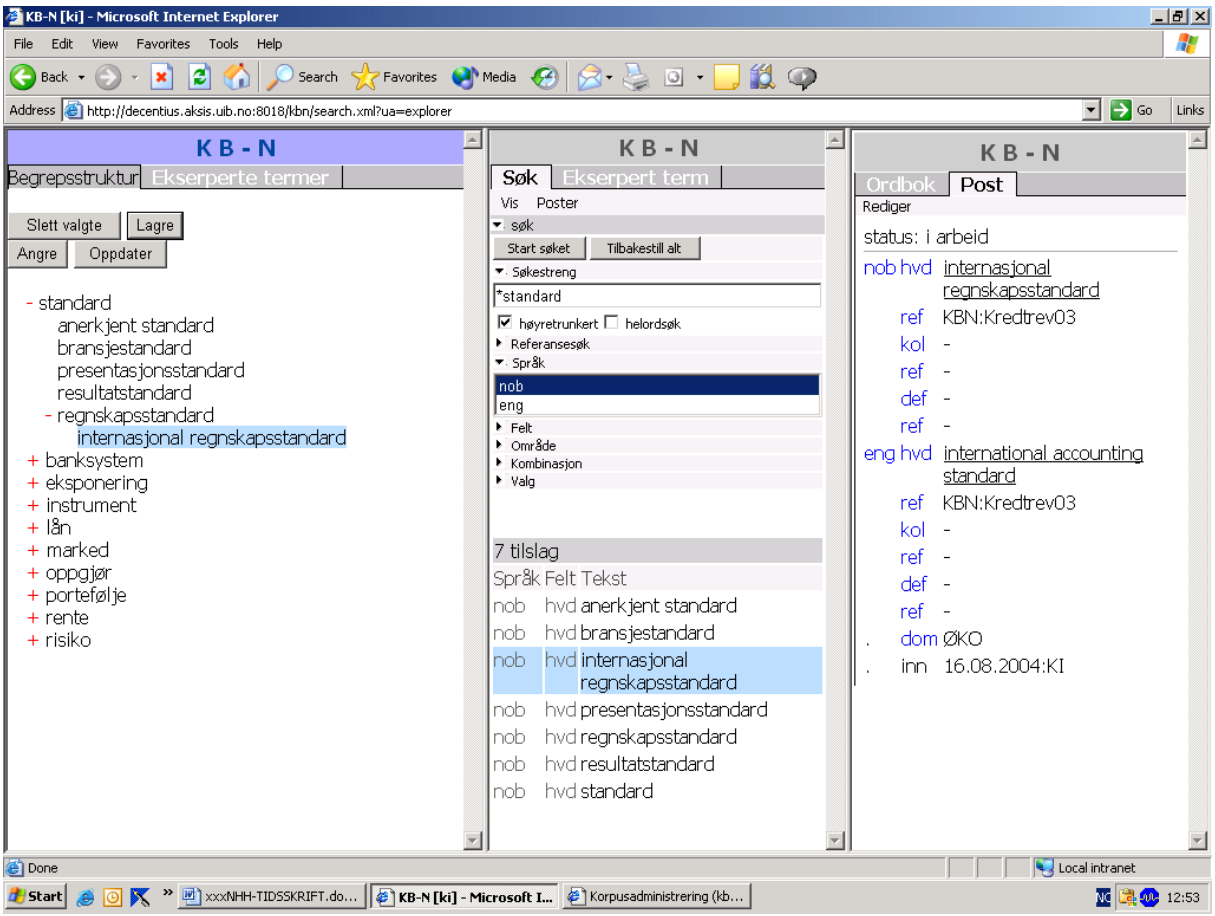
Vi skal nå vise med et konkret eksempel hvordan man kan systematisere termene i basen ved å bygge hierarkier. Vi bestemmer oss for å bygge et hierarki av alle termene i basen som betegner "standarder".

Vi befinner oss fortsatt i Termbase-vinduet. Vi stiller først inn venstre ramme på Begrepsstruktur-modus. Rammen vil da vise overtermene i de hierarkiene som allerede ligger inne. Hierarkiene kan aktiveres ved at man trykker på plusstegnet foran overtermene.

I midtrammen går vi inn i søke-modus og angir et venstre- og høyretrunkert søk på 'standard'. Resultatet er vist i bilde 6 ovenfor.

Nå kan vi begynne å bygge hierarkiet. Dette foregår etter dra-og-slipp-metoden. Vi drar termene en etter en fra trefflisten i midtrammen over til venstre ramme. Vi starter med det som skal være overtermen i hierarkiet, nemlig "standard". Vi plasserer musepekeren på

termen, trykker ned venstre musetast, drar termen til en posisjon like over den øverste termen (‘banksystem’) i venstre ramme, og slipper musetasten. Termen ”standard” er da kommet på plass. Så plasserer vi undertermene i hierarkiet ved å slippe dem på plass med musepekeren stående på overtermen. Resultatet vises i bilde 7 nedenfor. Dette gir et første leksikalsk fundament for videre hierarkisering i samsvar med subdomenets begrepsstruktur, som forutsetter spesifikk domenekunnskap.



Bilde 7: Termbase-vinduet med hierarki

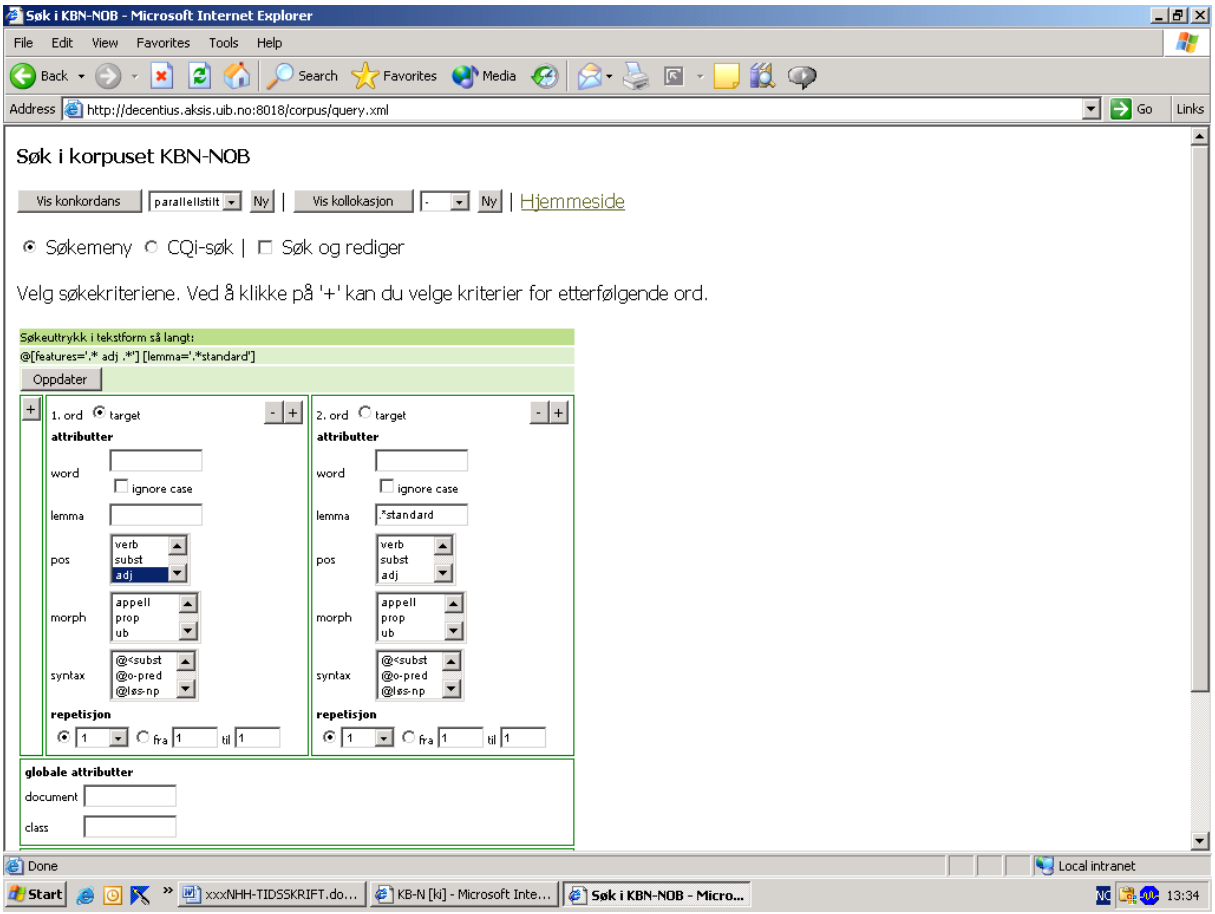
4.4 Visning av konkordanser og kollokasjoner

Ovenfor har vi først demonstrert hvordan vi registrerer termer i Term-basen med utgangspunkt i automatisk termekstraksjon fra filer i Korpus-basen, deretter hvordan vi kan bygge hierarkier av registrerte termer.

Vi forlater nå Term-basen og går tilbake til Korpus-basen for å vise noen av søke- og visningsmulighetene der. En **konkordans** er kort fortalt en sammenstilling av alle tekstlinjer eller setninger som inneholder en gitt søkestreng. Som eksempel kan vi fortsatt holde oss til søk på ulike typer av standarder. Vi velger et søk som skal fange opp alle forekomster av flerordstermer som inneholder ‘standard’ som kjerne, altså et adjektiv etterfulgt av ‘standard’ alene eller et sammensatt ord med ‘standard’ som kjerne.

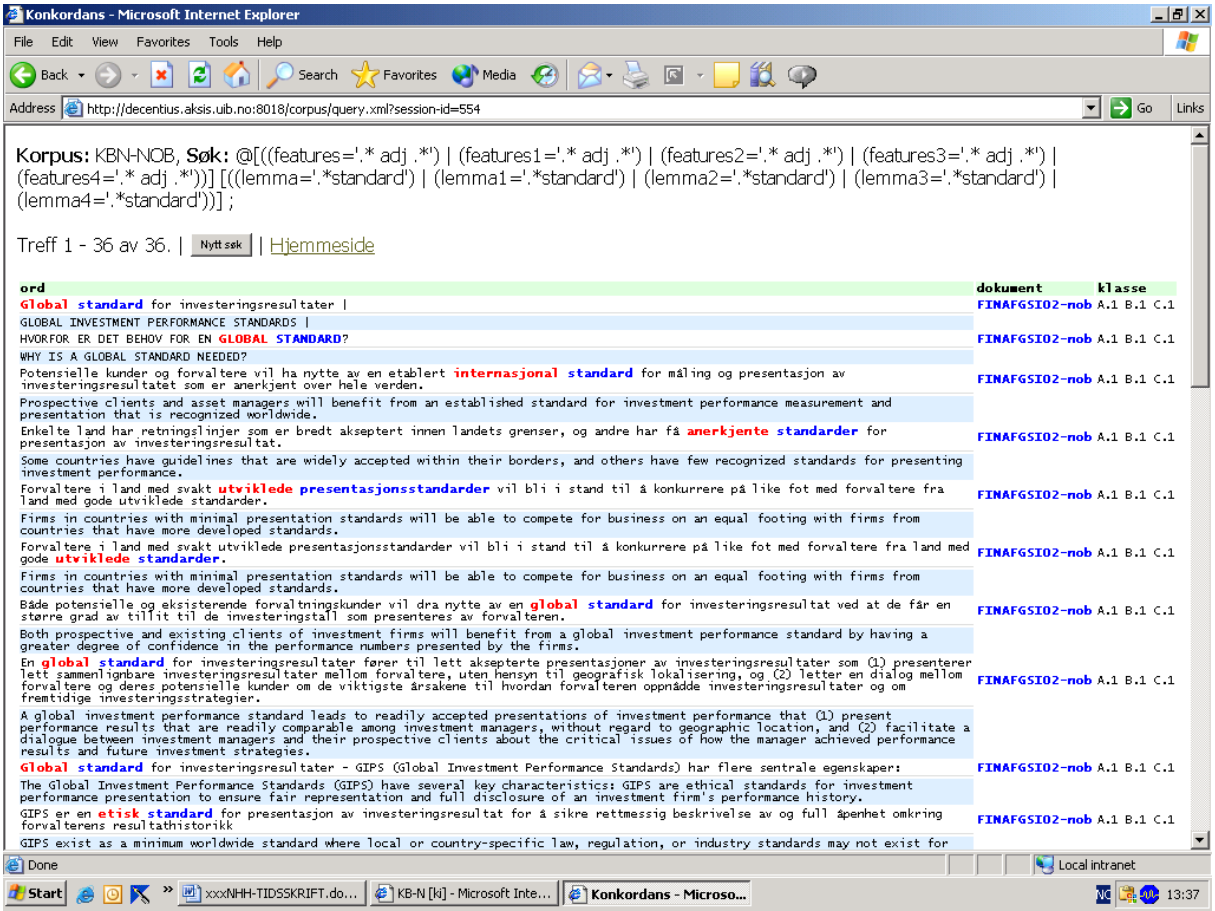
Vi går til Korpus-basens søkevindu og formulerer søket som vist i bilde 8 (s. 15). Vi ser at vi her har brukt to søkeruter. I den første har vi angitt søk på ordklassen adjektiv uten å spesifisere tegnstreng i lemma-feltet. I den andre ruten har vi angitt et venstretrunkert søk på

'standard' i lemma-feltet. Legg også merke til at fordi vi har tilgang til eksakt parallelle tekster på norsk og engelsk, har vi her valgt visningsmoduset "parallellstilt" fra rullegardinmenyen øverst til venstre.



Bilde 8: Korpus-basens søkevindu. Søk på uspesifisert adjektiv fulgt av venstretrunkert "standard" i lemma-feltet

Resultatet av søket ser vi på neste side som en konkordans i bilde 9. Programmet går gjennom hver enkelt fil i den norske Korpus-basen, plukker ut alle setninger som inneholder forekomster som tilfredsstiller søket, og sammenstiller disse setningene en etter en med den tilsvarende oversettelsen i den engelske Korpus-basen.



Bilde 9: Korpus-basens konkordansvindu i parallell-modus

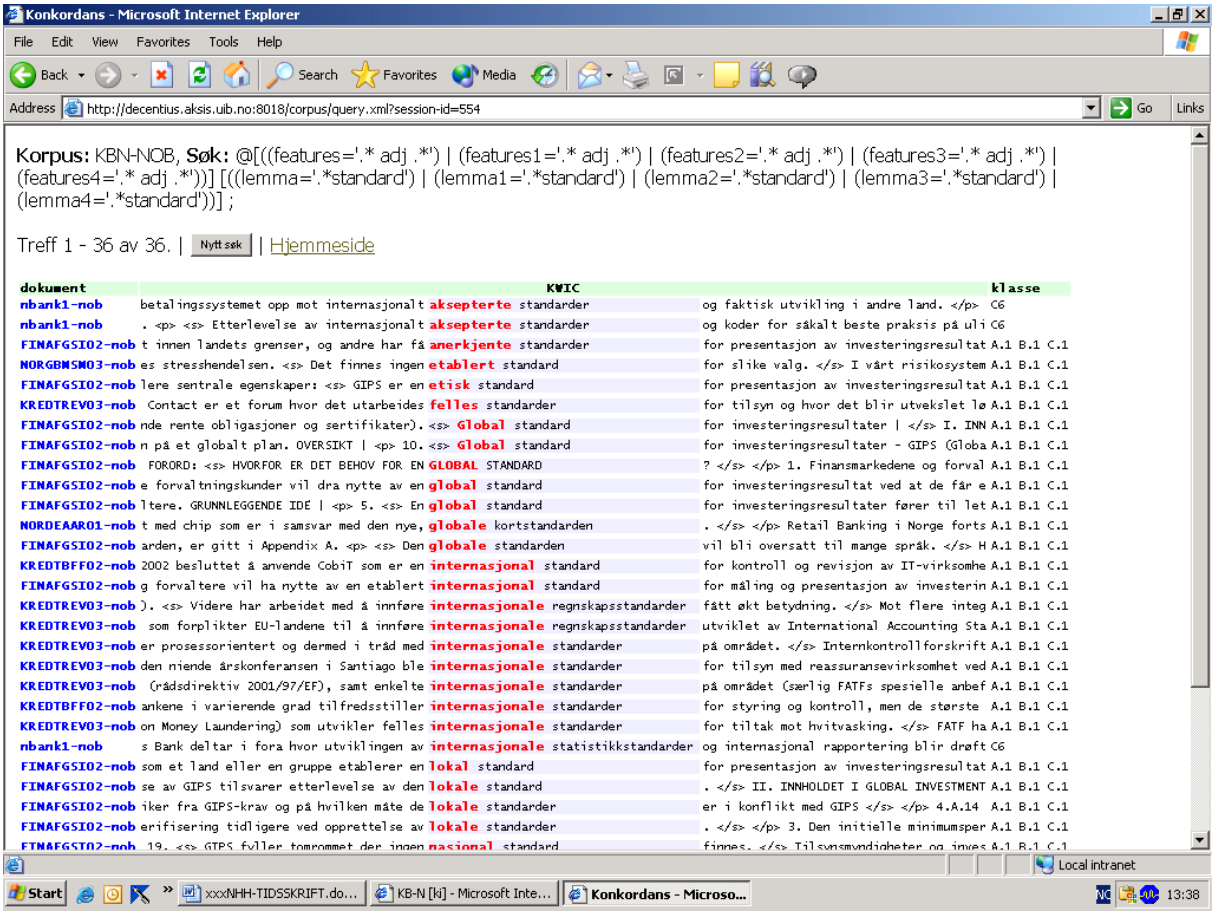
Foruten å kunne studere hvilke norske termer for ulike standarder som er brukt i dokumentene i Korpus-basen, har man her lett tilgang til de engelske ekvivalentene slik at man blant annet kan avdekke eventuelle synonymer. Dette skjermbildet er også et utmerket utgangspunkt for å studere fraseologien som knytter seg til termene på begge språk.

Som nevnt tidligere finnes det i tillegg til Parallell-modus også to andre modi, nemlig Kontekst-modus og KWIC-modus. Begge disse modiene viser konkordans kun for den basen man søker i (altså norsk eller engelsk). Skjermbildet for Kontekst-modus svarer til en enspråklig visning av Parallell-modus, og vil derfor ikke bli nærmere omtalt her.

I bilde 10 på neste side ser vi konkordansresultatet av samme søk i KWIC-modus. Konkordanslinjene er her listet opp alfabetisk etter søketreff. Videre er treffene plassert i et midtfelt med tilhørende kontekst til høyre og venstre på linjen så langt det er plass på skjermen.

I likhet med de andre modiene er treffene også her klikkbare slik at man kan aktivere et nytt vindu som viser en utvidet og ekspanderbar kontekst. Hvis vi klikker på treffet i konkordanslinje 5, 'etisk standard', får vi det resultatet som er vist i bilde 11 (s. 18), Korpus-basens tekstvindu.

Her kan man klikke på [Mer] for ekspandering av konteksten, [Tagget] for visning av grammatisk annotasjon og [XML] for visning av tekstkoding.

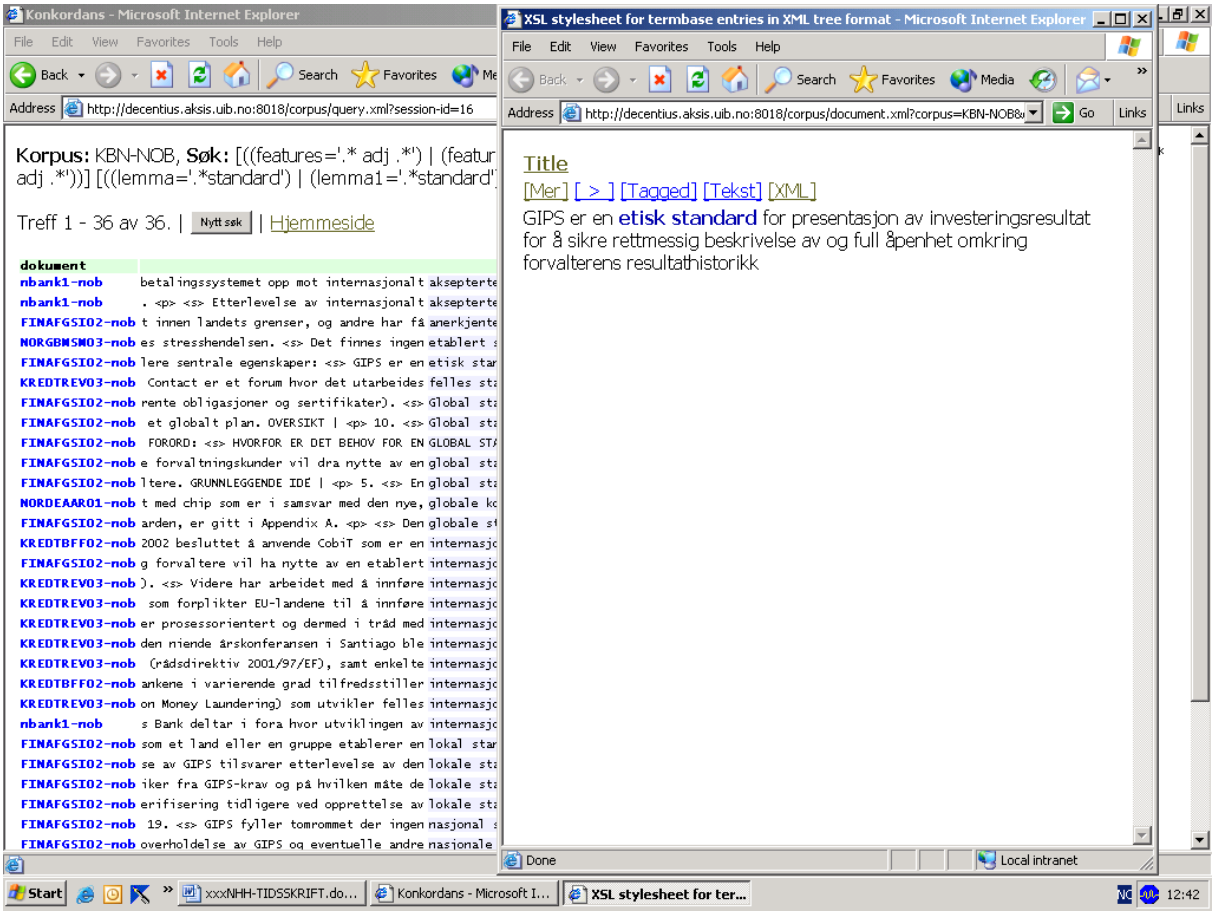


Bilde 10: Korpus-basens konkordansvindu i KWIC-modus

Fra Korpus-basens søkevindu (se bilde 8) kan vi også velge visning av **kollokasjoner**, dvs. leksemer som hyppig opptrer nær hverandre, vanligvis innenfor et "vindu" på 3-4 ord til høyre eller venstre. En klassisk kollokasjonsvisning er en tabellarisk oversikt over de ulike kontekstene et gitt ord forekommer i, sammen med statistisk informasjon om hyppigheten (absolutt og relativ frekvens) og relevans ("Mutual Infomation" (MI)) av hver kontekst. MI framhever leksemer som opptrer signifikant hyppigere sammen enn enkeltelementene gjør hver for seg.

Treffene kan sorteres enten etter frekvens, etter MI eller alfabetisk, og innstillingene kan lagres for senere bruk. I avkryssingsrutene til venstre kan man merke av de kollokasjonene man vil studere nøyere. Ved så å klikke på "Vis konkordans" vises alle konkordanslinjene for de avkryssede valgene. Bilde 12 (s. 19) viser resultatet av søk på uspesifisert adjektiv fulgt av venstretrunkert 'standard' i lemma-feltet. F.eks. utgjør kombinasjonen 'internasjonal standard' en viktig kollokasjon i et gitt korpus når den der opptrer vesentlig hyppigere enn leksemene 'internasjonal' eller 'standard' hver for seg.

Slike konkordans- og kollokasjonsstudier kan bidra til å identifisere viktige elementer av domenespesifikk fagspråklig flerordsterminologi samt synonymvalg og fraseologi og dermed gi viktige kriterier for "word sense disambiguation". Strukturen i den enkelte begrepspost tar høyde for at essensiell informasjon av denne type kan registreres og gjøres tilgjengelig for bl.a. analyse og oversettelse.



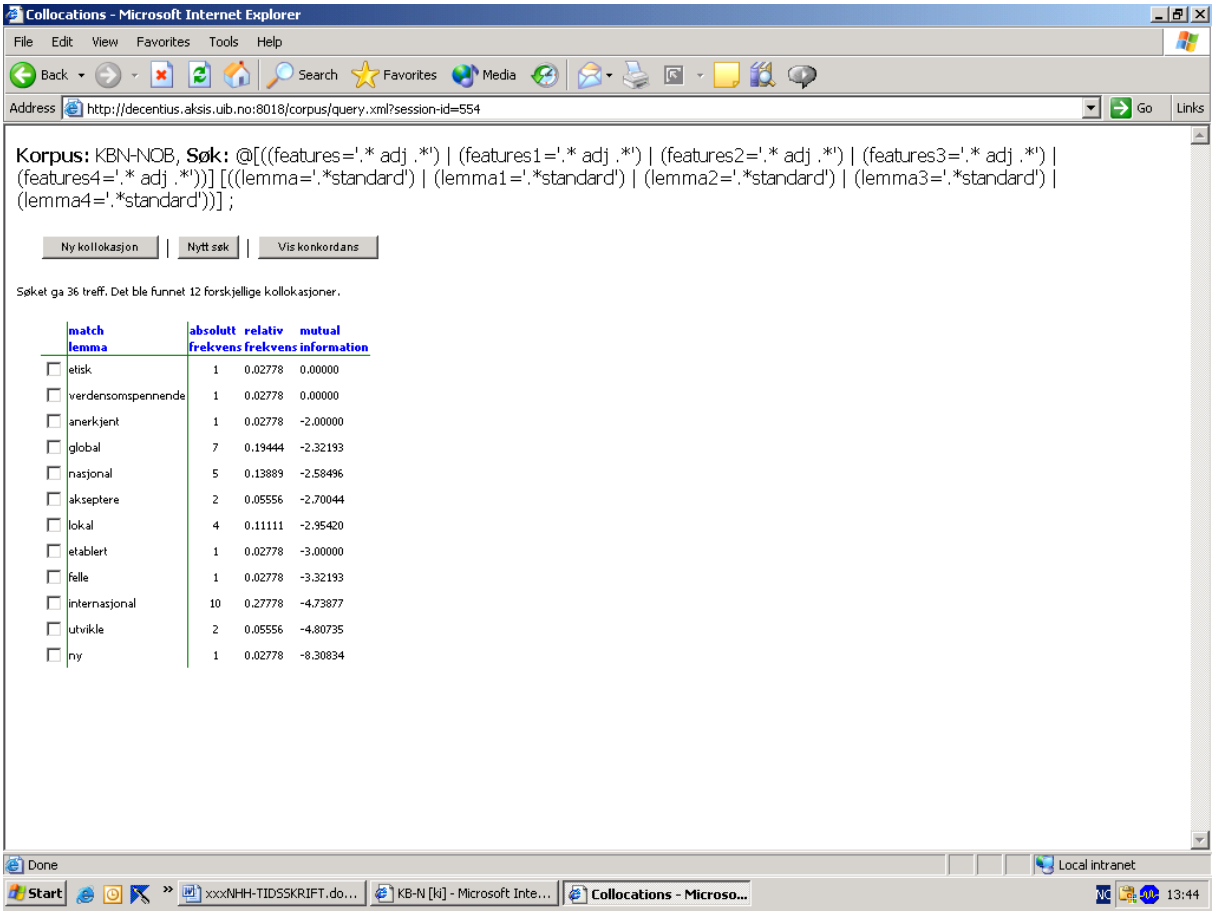
Bilde 11: Korpus-basens tekstvindu

5. "Kan det brukast til noko?"

Når innholdet i Korpus-basen og Term-basen, de to substansielle hovedmodulene i språkressursbanken, har nådd en kritisk masse og sammen med tilhørende funksjoner er integrert på den språkteknologiske plattformen, er KB-N-konseptet i prinsippet realisert. Vi vil da ha et begrepsorientert tekst- og termbasert kunnskapshåndteringssystem innrettet mot språkteknologiske applikasjoner, primært innen oversettelse, dokumentasjon, publisering, men med e-læring og undervisning samt formidling som viktige forlengelser. Norsk Språkbank vil dermed kunne være en realitet hva angår økonomisk-administrativt domene.

Innenfor rammen av KUNSTI-programmet er det én hovedbruksarena som peker seg ut som særlig aktuell: maskinoversettelse (MT). Gjennom LOGON-prosjektet (http://www.norskdok.uib.no/projects/?logon) foretas det her en storstilt satsing på utvikling av norsk-til-engelsk MT, og en kobling mellom KB-Ns kunnskapsbase (spesielt termdatabasen) og LOGONs framtidige MT-system vil kunne utgjøre en interessant og realistisk utprøvingsarena.

Men ved siden av og i forlengelse av dette perspektivet ser vi selvfølgelig for oss en bred vifte av distribusjonskanaler. Som tradisjonelle trykksaker kan det lett produseres domeneorienterte enspråklige termsamlinger så vel som to/flerspråklige fag-glossarer for studier eller oversettelse. I større format kan stoffet trykkes i form av en (fag)ordbok, med CD-ROM eller "USB-drive" som åpenbare alternative distribusjonsmedier. Men hovedkanalen vil uten tvil bli en WWW-tilgjengelig termbase tilgjengelig via standard nettlesere.



Bilde 12: Korpus-basens kollokasjonsvindu

Ut over den spesifikke MT-applikasjonen som er nevnt, hvilke behov er det vi har ambisjoner om å imøtekomme? Planen er å bli en økonomisk-administrativ kunnskapsbank eksempelvis for Ekspert som søker faguttrykk, for Lekmann som søker definisjon, for Språkstudent som søker språkbrukseksempel, for Forfatter som søker bruksomfang (scope), eller for Translatør som søker ekvivalent og kontekst. Et bruksområde som er spesielt aktuelt innenfor NHH-miljøet, er direkte kobling til e-læring.

Skal slike ambisjoner kunne oppfylles, stilles det store krav til kvalitetssikring av så vel Korpus-base som Term-base. Systemet må legges til rette for maskinoppslag så vel som menneskeoppslag. Kunnskapsinnholdet krever kontinuerlig validering, standardisering og normering i henhold til ISOs standarder og prosedyrer, med profesjonell ivaretaking av referanser, dokumentasjon, sitering, og ikke minst opphavsrettslige spørsmål.

Endelig kommer de økonomiske og praktiske spørsmål omkring drift, vedlikehold og videre utvikling av databasesystem og distribusjonsnett, rutinemessig oppdatering av innhold, representasjonsform og grensesnitt. Dette krever presis kontroll med termenes livssyklus for å unngå opphoping av foreldet materiale. Som en gjennomgående dimensjon vil vi legge til rette for aktiv brukermedvirkning.