Hvordan slippe inn i Platons hage[1]? – Om kartlegging og dokumentasjon av norsk akademisk vokabular

Ruth Vatvedt Fjeld & Arash Saidi

In the last few years several dictionaries of academic words have been developed, especially for English, but also for other languages, based on automatic analysis of academic corpora. Different methods have been used. For the Norwegian Academic Vocabulary list, we have experimented with different types of stop lists, and the article presents some results and a discussion of the lexical results of these experiments. We also recommend a phraseological analysis of academic corpora to map out some of the typical phrases in academic language.

1. Hvorfor er akademiske ordlister viktig?

Manglende kjennskap til akademisk vokabular blir ofte regnet som årsak til store forskjeller mellom studenters akademiske prestasjoner. Både første- og andrespråksbrukere i alle utdanningssituasjoner klarer seg bedre dersom de behersker et større og mer akademisk ordforråd:

> Control of academic vocabulary, or the lack thereof, may be the single most important discriminator in the 'gate-keeping' tests of education. (Gardner & Davies 2013:1)

Det gjelder også for norske studenter. Dessuten leser de mye faglitteratur på engelsk, og engelsk blir ofte krevd i akademisk skriving. Dermed mangler studentene gjerne norske ekvivalenter og får probmeler når de skal skrive på morsmålet. Pedagogiske lister over norsk akademisk vokabular kan bøte på dette. Målet er å utvikle norsk til et fullverdig moderne akademisk språk.

[1]Ordet akademi stammer fra Platons skole for filosofi og læring i Athen, grunnlagt ca. 385 f.Kr. Skolen lå i nærheten av et tempel i utkanten av byen, som etter tradisjonen skulle være oppkalt etter helten Akademos. (Kilde: Wikipedia)

Vi har få undersøkelser av ordforråd eller forståelse av ord i norsk. Mest kjent er en studie av 80 ord i nyhetsspråk (Vinje & Østby 1981). I undersøkelser av ordforståelse testes forståelse av vanskelige ord, sjelden hvordan de blir brukt i utforming av tekster. For akademisk framgang er produksjon av akademisk tekst like viktig som resepsjon. I tillegg til å kartlegge vanskelige ord vil finne ut hvilke ord og formuleringer som brukes for å organisere og strukturere tekst etter akademiske prinsipper. Det kan være alminnelige ord med spesiell funksjon i akademisk uttryksmåte.

I arbeidet med å kartlegge norsk akademisk vokabular følger vi opplegg fra liknende prosjekter, med systematiske analyser av ordforrådet i allmennspråk, i akademisk språk og i teknisk fagspråk. En sammenlikning av ordbruken etter teksttype kan vise hva som kjennetegner det leksikalske inventaret i typisk akademisk tekst. Som materiale anvendes et elektronisk korpus av akademiske tekster der frekvens og spredning av ord analyseres. Ord som regnes som typiske akademiske, blir tatt med i den akademiske ordlisten. Resultatet skal utgjøre den første listen over norsk akademisk ordforråd, kalt The Norwegian Academic Vocabulary List (NAV).

2. Hva er akademisk vokabular?

Leksikografiens hovedoppgave er å dokumentere og beskrive ords egenskaper. Å dokumentere typiske ord i akademisk språk for å lage hjelpelister for akademisk opplæring er dels en pedagogisk, dels en leksikografisk oppgave. Problemet er å skille akademiske ord fra andre typer ord. Det har vært lange diskusjoner om skillet mellom fagspråk og allmennspråk, uten at man har funnet en klar definisjon på forskjellen, f.eks. i Molde (1976:6) og Engberg (1992:94). Det er nok en glidende overgang mellom allment og faglig språk. De fleste forklaringer tar utgangspunkt i sender/mottaker-modeller, altså hvem som skriver, og for hvem, mer enn språktrekk i tekstene. Et unntak er Ralph (1981), som mener at forskjellen er at ordforrådet er annerledes i fagtekster enn i allmenntekster på grunn av fagtermene. Målet vårt er å finne generelt og emneuavhengig vokabular som som skiller den akademiske diskursens vokabular fra allmennspråkets.

3. Tidligere kompilerte akademiske ordlister

Den mest kjente akademiske ordlisten er *Academic Word List* (AWL) (Coxhead 2000), men det fins mange flere, jf. Gardner & Davies (2013) for utfyllende oversikt. Ordlistene er laget for engelsk i pedagogisk hensikt, både for andrespråksbrukere og utrente studenter. For nordmenn egner listene seg som hjelpemidler i arbeid med engelske akademiske tekster, men det er klart behov for hjelpemidler også til slik tekstproduksjon på norsk.

I Norden arbeides det også med akademiske ordlister. Hittil fins *Svensk Akademisk Ordlista* (SAO) (Ribeck et al. 2013).

4. Norsk akademisk vokabular (NAV)

4.1. Kartlegging av akademiske enkeltord

I Oslo er nylig første versjon av en liste over akademisk vokabular i norsk bokmål ferdig. Prosjektet ble satt i gang ved Institutt for lingvistiske og nordiske studier ved Universitetet i Oslo av Janne Bondi Johannessen, Ruth Vatvedt Fjeld, Kristin Hagen og Arash Saidi. Innen LUNAS-nettverket arbeides det med å lage tekstdatabaser med søkefunksjonalitet som gjør det mulig å søke på tvers av dansk, norsk og svensk, som grunnlag for pedagogiske akademiske ordlister. I første omgang var målet å kompilere en liste av enkeltord som er typisk akademiske.

4.2. Materiale

Materialet er et akademisk korpus av masteroppgaver, phd-avhandlinger, tidsskriftsartikler og liknende fra DUO (Digitale utgivelser ved Universitetet i Oslo). Korpuset består av 9689 dokumenter og utgjør ca. 310 millioner løpeord. Av disse er ca. 100 millioner løpeord lemmatisert med Oslo-Bergen-taggeren. Listen og de øvrige resultatene er derfor foreløpige, men vi regner med at de viser reelle tendenser i det norske akademiske bokmåls-vokabularet.

Tekstene kommer fra humaniora (44 mill. ord), pedagogikk (17 mill. ord), medisin (10 mill. ord), samfunnsfag (14 mill. ord), matematikk og

naturvitenskap (6 mill. ord), juss (4 mill. ord), teologi (3 mill. ord) og odontologi (4 mill. ord). Tekster fra alle fakulteter ved Universitetet i Oslo gjør at flest mulig akademiske stilarter ligger til grunn for analysene.

4.3. Metode for lemmaseleksjon

I tidligere akademiske ordlister har man unngått allmennspråklige ord, allmenne fremmedord og fagspesifikke, tekniske termer. Gardner & Davies (2013:8) framhever at typisk akademiske ord skal forekomme med en viss frekvens og spredning i flere typer akademiske tekster og i forskjellige akademiske disipliner, samt at de ikke er høyfrekvente i allmennspråklige tekster, for å bli tatt opp i Academic Vocabulary List (AVL).

Det anvendes bare negative kriterier, dvs. fravær av visse egenskaper ved ord eller uttrykk ved utvelgelsen av akademiske ord, det Gardner & Davies (2013:8) kaller akademiske kjerneord: "academic core words are those that appear in the vast majority of the various academic disciplines". Det stilles altså krav til statistisk spredning over de forskjellige disiplinene og teksttypene. Med negative kriterier og heuristiske metoder kan resultatene bli tilfeldige. Og når lemmaseleksjonen er basert på statistiske metoder, kan resultatene variere avhengig av type og størrelse på korpus. Kravet om spredning i akademiske tekster er et forbedrende tilleggskriterium, men like fullt er det uklart hva et akademisk ord egentlig er.

En tidlig studie av akademisk ordforråd er utført av Martin (1976:92). Hans definisjon er positiv: "Academic vocabulary has in common a focus on research, analyses and evalutation - those activities which characterize academic work". Martin regner dermed med at det ordforrådet som er best egnet til akademisk analyse og vurdering, kan regnes som akademisk. For å beskrive slik virksomhet kreves spesielle tekststrukturerende ord og uttrykk, ofte adverbialer, konjunksjoner og subjunksjoner i tillegg til de tekniske termene på fagfeltet.

Lemmautvalget i de presenterte akademiske ordlistene inneholder likevel mange alminnelige ord; AVLs ti viktigste er: *study, group, system, social, provide, however, research, level, result*, og SAOs ti første *dock, studie, beskriva, social, enligt, innebära, samt, form, betydelse, fall.* Det er vanskelig å se at de er typisk akademiske. Resultatene fra statistisk lemmaseleksjon må derfor diskuteres ut fra hva som er "adekvat ordforråd"

for akademisk virksomhet (Ribeck et al 2013:370). Siden fagtermer er silt vekk fra korpus, vil resultatlistene med nødvendighet inneholde allmenne ord. Fordi de er hyppige i akademiske tekster, kalles de akademiske. Seleksjonen er da fundert på klassisk stilistisk metode (jf. Engberg 1998:37). En mer pragmatisk fundert tilnærming, der man konsentrerer seg om å undersøke hvilke språklige midler som brukes til å utføre visse faglige funksjoner (Engberg, ibid.), kan gi et annet resultat. I vårt tilfelle vil det være de funksjoner som kreves i akademisk argumentasjon.

Vi anvender den samme metoden som ble brukt i SAO, dvs. å kartlegge ord som ikke faller innenfor det allmenne og som ikke er fagterminologi. Det er likevel en viss forskjell mellom de to listene, da DUO-korpuset har spredning på flere fagfelt enn SweAk-korpuset (jf. Ribeck et al. 2013:375), som bare har tekster fra humaniora og samfunnsvitenskap. Som i SAO er det i NAV tatt utgangspunkt i lemmaer og ikke ordfamilier.

Vår metode var tredelt. Første steg var å finne et ords ”keywordness”, som definerer om et ord har uvanlig frekvens i korpus. Dette målet er beskrevet av Scott (1997:233-245), som framhever at ethvert ord med lavere ”keywordness” enn 1.1 skal lukes ut av korpuset. Andre steg var å finne den reduserte frekvensen for hvert lemma i korpuset. Vi beregnet først antall forekomster av lemmaet i korpuset, delte korpuset i intervaller i forhold til frekvensen, og talte opp antall intervaller som inneholdt lemmaet. (Et forklarende eksempel kan være ordet *analyse* i et tenkt korpus på 1 million ord, der dette ordet har en frekvens på 1000. Da deles korpuset i 1000 like store tekstdeler og antall tekstdeler som inneholder ordet *analyse*, telles. Så telles den reduserte frekvensen per million ord i hele korpuset for å luke ut ord med mindre enn 15 i redusert frekvens per million ord.) Dette fungerte som kontroll for ord med høy redusert frekvens, men kun i en del av korpuset. Som tredje og siste steg brukte vi en stoppliste for å luke ut høyfrekvente ord som de to tidligere prosedyrene ikke fjernet.

4.4. Stopplister

Stopplister brukes for å unngå at resultatlisten ikke inneholder svært vanlige ord. Coxhead (2000) fjernet de 2000 mest frekvente ordene i The General Service List (West 1953) fra sine resultater, og Ribeck et al (2013) fjernet de

1000 mest frekvente ordene i lettlestkorpuset LäsBarT på 1 million ord. Vi antok at både størrelsen på det korpuset stopplisten ble hentet fra, og antall ord som ble tatt med på listen, ville ha betydning for resultatet, og testet det derfor ut. NoWac-korpuset, som er laget ved automatisk innsamling av tekst på .no-domenet på internett for bokmål i perioden 2009-2010 og er på 700 mill. løpeord (Guevara 2010), ble valgt som utgangspunkt. Vi brukte tre stopplister, en med de 1000 mest frekvente ordene i NoWac-korpuset, en med de 2000 vanligste, og en med de 5000 vanligste ordene i dette store internettkorpuset. Det ga tre forskjellige resultatlister fra samme korpus.

De 500 mest frekvente ordene på alle tre resultatlistene ble gjennomgått manuelt for å undersøke om de allmenne ordene og tekniske termene var fjernet og de akademiske var med. Uaktuelle ord og ikke-ord, som navn, lånord og liknende ble talt opp. Stopplisten på 1000 ord ga 52 antatte feiltreff av de første 500 tilslagene. Det var alminnelige ord (*holdning*, *opplevelse*), egennavn (*Norge*), utenlandske ord og sitatord (*as*, *new*), tegn og tekststrukturerende elementer (*s.*, *2.1.*, *ii.*) og rene tall. Stopplisten på 2000 ord ga 48 feiltreff på de 500 første tilslagene, og stopplisten med 5000 ord ga 125 feiltreff, jf. tabell 1 nedenfor.

Stoppliste	Allm.ord	Navn	Lån/sitat	Tegn	Tall	Fork.	Sum
1000 ord	28 = 5,6 %	7 = 1,4 %	8 = 1,6 %	5 = 1 %	4	0	52 = 10,4 %
2000 ord	7 = 1,4 %	18 = 3,6 %	10 = 2 %	11 = 3,4 %	1	1	48 = 9,6 %
5000 ord	7 = 1,4 %	65 = 13 %	27 = 5,4 %	22 = 4,4 %	2	2	125 =25 %

Tabell 1: Oversikt over antall feiltreff med tre forskjellige stopplister.

Tabell 1 sier noe om hvilken stoppliste som er best egnet. Stoppliste på 1000 ord tillot relativt mange allmennord (5,6 %), mens stopplistene på 2000 og 5000 ord tok med få allmennord (1,4 %). En stoppliste på 1000 høyfrekvente allmennord er sannsynligvis for lite. Stopplisten på 5000 ord tillot en del egennavn (13 %), lånord/sitatord (5,4 %) og tegn (4,4 %), mens stoppliste på 2000 ord bare ga 3,6 % egennavn og 2,0 % lånord og 3,4 % tegn.

Vi konkluderte dermed med at en stoppliste med de 2000 mest frekvente ordene i et stort allmennspråklig korpus gir best resultat ved automatisk kartlegging av akademiske ord. Det ville være bedre å komme lavere enn

9,6 % feiltreff totalt sett, men med automatisk metode er det fortsatt vanskelig.

Vi har ikke funnet en konkret og avgrensende definisjon av 'akademisk ord' som kunne legges til grunn. Når ordet *akademisk* i seg selv er vagt, er det problematisk å gi en definisjon av hva som er et akademisk ord, og det er viktig å være klar over de metodiske problemene det innebærer. Det er derfor viktig å anvende stilistisk og pragmatisk metode i vurderingen av resultatene. De ordene som i feilanalysen anses som allmenne, har altså blitt brukt i høy grad og med stor spredning i de akademiske tekstene uten at de intuitivt ble ansett som nødvendige i beskrivelse av akademisk arbeid. Dermed utgjør de en del av den akademiske stilen. Videre mener vi at en feilforekomst eller tvilsom forekomst av "uakademiske ord" på 1,2 % heller ikke kan gjøre stor skade på en slik liste. Den automatisk genererte ordlisten må renses manuelt for feiltreff før publisering, men det er klart en fordel at det arbeidet minimeres mest mulig

4.5. Resultater for akademiske enkeltord

Analyse av det akademiske DUO-korpuset renset med stoppliste på de 2000 mest frekvente allmennord ga etter vårt skjønn rimelig gode resultater. Som en kontroll av resultatlistene ble frekvensen i DUO sammenliknet med frekvensen i det allmennspråklige korpuset Leksikografisk bokmålskorpus (LBK) (Knudsen & Fjeld 2013). De ti mest frekvente innholdsordene på resultatlisten, ordnet etter ordklasse og med frekvenstallene i DUO og LBK er presentert i tabell 2, 3 og 4.

Det kan diskuteres i hvilken grad alle disse ordene er typisk akademiske, eller såkalt akademiske kjerneord. Intuitivt virker utvelgelsen svært god for substantiv og langt på vei for verb, mens adjektivene har flere allmenne ord som sikkert også brukes i ikke-akademiske sammenhenger. Substantivene har gjennomgående høyest frekvens, deretter verb, og på tredjeplass adjektiv. Adverbene er gjennomgående minst frekvent av de undersøkte ordene. Denne fordelingen er i samsvar med fordeling på ordklasse i norsk språk generelt. Det viser seg også at de oppslagsordene som er med i NAV, har betydelig høyere frekvens i DUO enn i LBK, som er sammensatt av forskjellige typer tekster. Sammenlikning mellom forekomst i DUO og i hele LBK-korpuset gir følgende resultater:

Substantiv	DUO	LBK
kapittel	2	21899
informant	3	1415
analyse	5	28778
problemstilling	7	52037
relasjon	10	1447
oppfatning	11	1082
aspekt	12	1987
funn	13	1386
kontekst	14	2070

Tabell 2: De ti mest frekvente substantiv i NAV med frekvensnummer i DUO og LBK.

Verb	DUO	LBK
undertrykke	8	3293
tolke	9	1302
belyse	21	2580
referere	28	2389
vektlegge	30	2922
analysere	35	2143
omhandle	37	2676
formidle	40	1626
fremstå	41	1820
tilhøre	44	823

Tabell 3: De ti mest frekvente verb i NAV med frekvensnummer i DUO og i LBK.

Adjektiv	DUO	LBK
teoretisk	17	1795
bevisst	18	1460
kulturell	26	1169
kvalitativ	36	3184
spesifikk	42	2558
ulik	56	2745
overordnet	83	2307
formell	95	1654
menneskelig	111	1318
kompleks	119	2488

Tabell 4: De ti mest frekvente adjektiv i NAV med frekvensnummer i DUO og i LBK.

Tallene som er understreket i tabellene, har lavere frekvens i LBK enn i NoWac, siden stopplisten er satt ved frekvens 2000 i det korpuset. Tabellene viser likevel at frekvenstallene er svært mye høyere i NAV enn i LBK, noe som viser at oppslagsordene stilistisk sett er mer typiske i akademiske tekster. Videre antar vi at det ikke bare er innholdsord (substantiv, verb og adjektiv) som er relevante for akademiske ordlister til opplæring av studenter. Resultatene viser nemlig at vår statistiske metode relativt sett fanger mange strukturord, særlig adverb, men også konjunksjoner, subjunksjoner, tekststrukturerende adverbialer og preposisjoner. Dette er ord som viser pragmatisk funksjon, som teksters logiske oppbygning eller argumentasjonsstruktur. I tillegg brukes de der det kreves i modifisert og vurdert framstilling, der man ofte må presentere sammenhengende tankerekker. Det er nettopp slik tekststruktur som er typisk for akademiske tekster, f.eks. å sette forskjellige analysedeler i sammenheng eller formulere kompliserte sammenlikninger og motsetninger eller redegjøre for forskjellen mellom kjent og ny innsikt. Spesielle adverbialer som *imidlertid, dessuten, herunder* brukes sjelden i dagligspråk, og er heller ikke nødvendig i alt fagspråk, men er pragmatisk sett nødvendige når akademiske overlegninger skal formuleres. Erfarne veiledere vet at slike ord og fraser ofte er mangelvare i mindre gode akademiske tekster.

AVL har adverbet *however* på 6. plass og SAO har preposisjonen *enligt* på 5. og konjunksjonen *samt* på 7. plass. Slike ord anvendes for å uttrykke begrunnelser, innrømmelser, forbehold eller forklarende tillegg, og det er typisk slike uttrykk man svært ofte trenger i en akademisk diskurs når man utfører det Martin (1976:92) kaller akademisk arbeid. *New Oxford Dictionary of English* definerer *however* med følgende grunnbetydning: "used to introduce a statement that constrast with or seems to contradict something that has been said previously", og det stemmer bra i vår sammenheng.

I tabell 5 nedenfor presenteres de ti mest frekvente strukturord i DUO sammenliknet med frekvens i LBK.

Adverb	DUO	LBK
ibid.	24	62151
hvorvidt	33	1910
ovenfor	55	1982
underveis	86	2138
delvis	92	1325
fremfor	198	2066
således	222	2140
innad	303	5091
ifra	397	2773
innledningsvis	464	5791

Tabell 5: De ti mest frekvente adverb i NAV med frekvensnummer i DUO og i LBK.

De ti mest frekvente strukturord i AVL: *6 however, 32 both, 40 thus, 130 therefore, 151 particularly, 172 indeed, 209 significantly, 210 generally, 242 highly, 251 relatively.* Og i SAO: *1 dock, 5 enligt, 15 utifrån, 26 kring, 28 därmed, 40 exempelvis, 50 endast, 59 såväl, 74 däremot, 77 dels.*

Både AVL og SAO har flere og mer frekvente strukturord enn NAV. Årsaken til det kan være at strukturordene i norsk ofte omskrives til fraseologiske enheter. Vi skal derfor ta for oss en type slike flerordsenheter spesielt.

4.6. Kartlegging av flerordsenheter i akademiske tekster

Siden et vokabular ikke bare består av enkeltord, men også av fraser og andre flerordsenheter, kan videre arbeid med fokus på det fraseologiske ved akademisk språk forbedre listene. Vi har utført en avgrenset fraseologisk analyse som muligens gir bedre grunnlag for å gi svar på spørsmålet om hva et akademisk vokabular er, ut over å kartlegge enkeltord ut fra en viss frekvens i akademiske tekster. Så vidt vi har kunnet se, har arbeidet med de nordiske akademiske ordbøkene og ordlistene hittil mest fokusert på enkeltord. Resultatene i 4.5. viser at enkeltord alene ikke gir nok informasjon til å skille akademiske uttrykksmåte fra den allmennspråklige. Studier av flerordsenheter har vært sentralt i nyere leksikalsk forskning. Flerordsenheter defineres på mange måter og ut fra forskjellige perspektiv og oppgaver i språket, vi bruker her den helt nøytrale betegnelsen

flerordsenheter (engelsk Multiword expressions), som langt på vei kan erstattes med «lexical bundles» (Biber et al. 1999:990), definert som «recurrent expressions, regardless of their idiomaticity, and regardless of their structural status». Paquot har bl.a. anvendt lexical bundles (på norsk leksikalske knipper) i en studie av leksikalsk overføring fra morsmål til innlærerspråk (Paquot 2013). Vi anvender den leksikografiske termen flerordenheter, fordi leksikalske knipper antyder at det er en tettere leksikalsk sammenbinding mellom ordene enn tilfeldig samforekomst.

Vi har forsøksvis anvendt manuell analyse av bigram og trigram av tekster i DUO og sammenliknet resultatet med samme analyse av skjønnlitterære tekster for å undersøke om vi fant typiske akademiske flerordsenheter eller leksikalske knipper. Ut fra våre foreløpige resultater ser det ut til at akademisk vokabular skiller seg fra allmennvokabularet særlig på flerordsnivå. Ved hjelp av et sett metoder for automatisk seleksjon av kollokasjoner vil vi seinere undersøke mer systematisk om dette stemmer. En måte å gjøre det på, er å bruke metoden beskrevet over for bigram og trigram. En annen måte er å bruke syntaktisk avhengighets-grammatikk (semantic dependency parsing) sammen med metoden over for å lage lister basert på syntaktiske forhold. Her vil vi velge ut bare én type som en indikasjon på frasenes akademiske funksjon eller verdi.

Svært vanlige fraser i de akademiske tekstene er preposisjonsuttrykk som *med henblikk på, under forutsetning av* osv, ofte kalt seriepreposisjoner, bestående av preposisjon + nomen + preposisjon (her kalt PNP). Grammatisk fungerer slike ledd som regel som adverbialer, og kan erstattes av det (f.eks. *under forutsetning av = dersom: i løpet av = mens*). Om man vil finne slike konstruksjoner i et korpus, er det ikke tilstrekkelig med ettordsanalyser, man må søke på bi- og trigram. Vi har derfor gjort en trigramanalyse av tekster fra LBK, fordi vi der også kan sammenlikne med frekvensen i forskjellige teksttyper. LBK består av 100 millioner løpeord og har merking som gjør det enkelt å velge ut delkorpus, bl.a. etter fagområde eller teksttype. Vi etablerte derfor et sakprosakorpus (50,3 millioner løpeord), et periodikakorpus (5,8 millioner løpeord) og et skjønnlitterært korpus (36,5 millioner løpeord) som delmengder av LBK.

I sakprosakorpuset fant vi blant de 500 mest frekvente trigrammene 29 PNP-er, og 21 av disse forekom blant de 226 mest frekvente. Også delkorpuset periodika hadde høy frekvens av PNP, jf. tabell 6 nedenfor.

Sakprosa	Periodika	Skjønnlitteratur
2 i forhold til	*6 i løpet av*	*7 ved siden av*
6 på grunn av	*10 på grunn av*	*46 på grunn av*
11 i løpet av	*12 i forhold til*	*59 i løpet av*
14 i forbindelse med	*16 i forbindelse med*	*257 på vei til*
20 i tillegg til	*22 i tillegg til*	*473 i nærheten av*
31 i form av	*60 ved universitetet i (?)*	
71 i henhold til	*80 til tross for*	
78 i motsetning til	*86 ved siden av*	
84 ved siden av	*155 i slutten av*	
85 på bakgrunn av	*161 i form av*	

Tabell 6: De ti mest frekvente PNP i sakprosa, periodika og skjønnlitteratur i LBK (markert med frekvenstall) (blant de 500 mest frekvente trigram i LBK).

Til sammenlikning har vi undersøkt trigram i skjønnlitteratur. Blant de 500 mest frekvente var det bare 5 PNP i skjønnlitteratur. Disse resultatene dokumenterer at seriepreposisjoner (preposisjon+substantiv+preposisjon) er mer frekvent i sakprosa og periodika, som er de tekstene som ligger tett opp til akademisk språk, enn i skjønnlitteratur. I hvert fall er både frekvens og type svært forskjellig i de to hovedgruppene. To av de som forekom i skjønnlitteratur, kan neppe regnes til typen seriepreposisjon, da de sannsynligvis beskriver konkrete forhold: *i nærheten av* og *på vei til*.

5. Oppsummering og videre forskning

Akademiske ordlister utarbeides som hjelpemidler for studenter og ferske akademikere. Formålet er først og fremst å gi studenter hjelp ved produksjon av akademiske tekster. Basis for NAV-ordlisten er et korpus av akademiske tekster produsert ved Universitetet i Oslo på over 300 millioner løpeord. Korpuset er analysert med såkalt redusert frekvens for å finne både frekvens og spredning av potensielle akademiske ord. Resultatene ble renset med stopplister av forskjellig størrelse for å ekskludere allmennord. Det viste seg at en stoppliste på de 2000 mest frekvente ordene i et stort internettkorpus ga resultater med færrest uønskede lemmakandidater. Vi vil seinere undersøke om en adaptert versjon av metoden til Gardner & Davis vil gi færre feiltreff enn metoden som er anvendt for SAO, slik at manuell rensing blir mindre arbeidskrevende.

En forsøksvis analyse av treordsenheter på resultatlisten viser at visse konstruksjoner med tekststrukturerende funksjon er tydelig mer frekvent i sakprosa og tidskriftsartikler enn i romaner. En omfattende analyse av trigram i DUO-korpuset ville være interessant som sammenlikning, f.eks. bigram av adjektiv + substantiv, verb + substantiv eller leksikalske knipper av leksikalske verb og objekt, i samsvar med flere gjennomførte flerordsanalyser. Også analyser av korpus med trebankstruktur, f.eks. INESS-korpuset, som er utviklet for utforsking av syntaks og semantikk i norsk språk, der også LBK er lagt inn og analysert, vil kunne gi sikrere og tydeligere resultater om forskjeller i flerordsenheter i akademisk språk i forhold til allmennspråk.

Litteratur

Biber, D., Johansson S., Leech, G., Conrad, S. & Finegan, E. (1999): *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Coxhead, A. (2000): A new academic word list. I: *TESOL Quarterly*, 34:2, 213-238.

Engberg, Jan (1998*): Introduktion til fagsprogslingvistikken*. Århus: Systime.

Gardner, D. & M. Davies (2013): A New Academic Vocabulary List. I: *Applied Linguistics* 4. <applij.oxfordjournals.org/content/early/2013/08/02/applin.amt015.full>

Guevara, Emiliano Raul (2010): NoWaC: a large web-based corpus for Norwegian. I: *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics, 1-7.

Hyland, K. & P. Tse (2007): Is there an ”academic vocabulary”? I: *TESOL Quarterly* 41:2, 235-253.

Knudsen, Rune Lain & Ruth Vatvedt Fjeld (2013): LBK2013: A balanced; annotated national corpus for Norwegian Bokmål. I: *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013*; May 22-24 2013 Oslo. NEALT Proceedings Series 19.

Martin, A. (1976): Teaching Academic Vocabulary to Foreign Graduate Students. *TESOL Quarterly* 10(1), 91-97

Molde, Bertil (1976): *Fackspråk.* Skrifter utgivna av Svenska språknämnden, 57. Lund

NAV = Norwegian Academic Vocabulary list. <www.tekstlab.uio.no:4000/>.

Paquot, Magali (2013): Lexical bundles and L1 transfer effects. I: *International Journal of Corpus Linguistics* Vol. 18, no. 3, 391-417.

Ralph, Bo (1981): Hur mycket fackspråk är fackspråk? I: *Svenskans beskrivning* 12. Umeå: Universitetet i Umeå.

Ribeck, Judy, Håkan Jansson & Emma Sköldberg (2014): Från aspekt till övergripande – en ordlista över svensk akademisk vokabulär. I: Fjeld, Ruth og Marit Hovdenak (red.): *Nordiske studier i leksikografi 12. Rapport fra Konferanse om leksikografi i Norden Oslo 13.-16- august 2013.* Oslo: Novus forlag, 370-384.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer & Helge Dyvik (2012): An open infrastructure for advanced treebanking. I: Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco (eds.): *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. Istanbul, Turkey, 22-29

SAO = Svensk Akademisk Ordlista- (spraakbanken.gu.se/ao/om.html).

Vinje, Finn-Erik & Helge Østbye (red.) (1981): Vanskelige ord i nyhetene (VON). En ordkunnskapsundersøkelse. (Upubl.).

Vinje, Finn-Erik (1982): *Journalistspråket.* Fredrikstad: Institutt for journalistikk.

West, M. (1953): *A General Service List of English Words*. London: Longman, Green and Co.

Ruth Vatvedt Fjeld
professor
Universitetet i Oslo
Institutt for lingvistiske og nordiske studier
Boks 1102
NO-0317 Oslo
r.e.v.fjeld@iln.uio.no

Arash Saidi
vitenskapelig assistent
Universitetet i Oslo
Institutt for informatikk
arashsa@student.matnat.uio.no