NyS

Titel:	Kvantitativ terminologi. <i>En sprogstatistisk tilgang til automatisk identifikation og udtrækning af teksttypiske og -begrænsede lemmata i fagsprogstekster</i>
Forfatter:	Bjarne Blom
Kilde:	<i>NyS – Nydanske Sprogstudier 30. Korpuslingvistik</i> , 2002, s. 100-122
Udgivet af:	Akademisk Forlag A/S
URL:	www.nys.dk

© NyS og artiklens forfatter

Betingelser for brug af denne artikel

Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed

Artiklerne i de ældre NyS-numre (NyS 1-36) er skannet og OCR-behandlet. OCR står for 'optical character recognition' og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Kvantitativ terminologi¹

En sprogstatistisk tilgang til automatisk identifikation og udtrækning af teksttypiske og -begrænsede lemmata i fagsprogstekster

BJARNE BLOM

1. DEN KVANTITATIVE TERMINOLOGIFORSKNING SOM ALTERNATIV OG SELVSTÆNDIG DISCIPLIN

I løbet af det seneste tiår har der været en så ganske markant tilvækst i arbejder, som specifikt omhandler automatisk identifikation og udtrækning af termer, at det er forsvarligt at tale om et decideret forskningsparadigme. Imidlertid udtrykker næsten alle disse arbejder en samlet tendens, nemlig terminologibehandling på det syntagmatiske plan.

Denne tendens kan afgjort tilskrives ét bestemt værk, nemlig John Justeson og Slava Katzs tekniske rapport fra IBM fra 1993, *Technical terminology: some linguistic properties and an algorithm for identification in text*. Heri viser Justeson og Katz, at der ud fra et empirisk materiale kan udledes en række regelmæssige syntagmatiske mønstre for flerordstermer, som efterfølgende kan identificeres i løbende tekst ved hjælp af en simpel algoritme.

Som følge af den intuitivt plausible tilgang til emnet og algoritmens høje grad af pålidelighed, blev rapporten i løbet af meget kort tid citeret ofte, og blev som konsekvens heraf gjort tilgængelig i tidsskriftet *Natural Language Engineering* i 1995.

I kraft af det store antal arbejder i 1990'erne, som havde identifikation af flerordstermer som fokuspunkt, foreligger der på dette felt af den kvantitative terminologi et både solidt teoretisk grundlag og en lang række empiriske undersøgelser.

Arbejder, som har terminologi på det paradigmatiske niveau, enkelt-ordstermer, som studieobjekt er få og spredte, hvorfor der ikke kan siges at eksistere nogen samlet teori herfor. Eftersom den kvantitative terminologiforskning må regnes som en decideret forskningsdisciplin (og det endda af nyere dato – ACL afholdt den første workshop på området i 1998) er det efter min mening nødvendigt, at arbejdsfeltet definerer sine egne arbejdsbetingelser. Den kvantitative terminologi arbejder i høj grad med en statistisk indfaldsvinkel til problemfeltet "terminologi", og må derfor anses som et teoretisk alternativ til den traditionelle terminologiforskning, som har været normsættende for måden, hvorpå begrebet *terminologi* opfattes, og til dels et metodologisk alternativ til den traditionelle korpuslingvistik, som har været normsættende for måden, hvorpå arbejde med tekstmateriale i elektronisk form foregår.

Terminologilærens grundlægger, Eugen Wüster, var ingeniør af profession, og netop ingeniørens stræben efter det klare, logiske, ordnede og strukturerede afspejles i Wüsters terminologiske arbejde. Der er ingen tvivl om, at Wüster er meget inspireret af sin tids idealvidenskab, fysikken, da han skriver sin doktordisputats (Wüster, 1931). Under alle omstændigheder overtager terminologilæren en del af de eksakte videnskabers instrumenter til præcisionsbeskrivelse, herunder *definitionen*, som er en af disciplinens helt centrale elementer. I terminologilæren fastlægges en *term* ud fra netop definitionen, og en "term" opfattes meget rigoristisk i direkte dikotomisk opposition til en "ikke-term".

I den videnskabelige litteratur findes der ikke noget objektivt belæg, som underbygger terminologilærens eksplicite postulat om, at en term er et ord, som er udstyret med særlige egenskaber, hvormed det adskiller sig fra andre ord, jf. f.eks. (Drodz, 1973:44-45), (Drodz, 1972:42), (Felber & Budin, 1989:2), (Picht & Draskau, 1985:97). En nærmere analyse af de videnskabelige definitioner af begreberne term og terminologi afslører, at det sproglige definerende udsagn, definiens, altid rummer mindst ét kategorialt karakteristikum, jf. (ISO/DIS, 1087:7) eller (de Bessé, 1990: 253), hvorved fastlæggelsen af forskningsobjekterne term og terminologi ikke er distinktiv, men tentativ. Sådanne definitioner kan naturligvis ikke tjene som genereringsgrundlag, da en computer kræver en præcis afgrænsning heraf, som de terminologiske definitioner ikke honorerer.

FIGUR 1.

Problemet med den videnskabelige definition som anvendt i terminologilæren. Som det fremgår af de repræsentative eksempler på definitioner af begreberne term og terminologi, rummer definitionerne mindst ét uspecifikt og kategorialt træk (markeret med understregning). Kilde for de tre definitioner: [ISO/ DIS, 1987:7], [ISO/DIS, 1987:7], [de Bessé, 1990].

Hvor terminologilæren altså arbejder onomasiologisk ud fra et ekspliciteret postulat om, at en term besidder distinktive træk, hvorved den adskiller sig fra dens logiske modsætningsværdi ikke-term, arbejder den kvantitative terminologi ud fra det lingvistiske tegns udtryk.

For den kvantitative terminologi er der store fordele forbundet med at udlede information om én bestemt tekst (eller en samling af domænebeslægtede tekster) frem for at generalisere i forhold til et termsystem, som er baseret på en repræsentationsmåde, som reflekterer et på forhånd afdækket vidensområde. I bedste fald er det praktiske terminologiarbejde dermed præsriptivt normstyret, i værste fald er det overflødigt, fordi resultatet af den terminologiske analyse er forhåndsbestemt og termsystemet at betragte som en facitliste.

Som følge af terminologilærens terminologiopfattelse kommer definitionen til at fungere som en ligning, ved hjælp af hvilken der etableres et absolut synonymt forhold mellem definiendum (term) og definiens (termens karakteristiske og distinktive træk) og et differensforhold mellem termen og alle andre begreber. På denne måde står og falder definitio-

nens autoritative forklaringskraft med, hvorvidt begrebet term overhovedet besidder sådanne distinktive træk (ikke mindst over for størrelsen ikke-term).

I figur 1 vises terminologinormeringens, dvs. den internationale standardiseringsorganisations, konsensusdefinitioner af begreberne *terminology* og *term* samt et eksempel på et af mange varianter over dette tema, som præger den teoretiske litteratur på området. En videnskabelig definition skaber rimelige forventninger om klarhed, modsigelsesfrihed og plausibilitet, men disse forventninger opfylder ingen af de anførte definitioner, eftersom der ikke i definiensbeskrivelsen skabes en lukket mængde af beskrivelsesfaktorer, hvorved beskrivelsesobjektet term på nogen måde afgrænses i forhold til alle denne verdens øvrige fænomener – specifikt: ord, som ikke er termer. Problemet opstår, fordi der i alle tre definitioner optræder mindst ét kategorialt træk, {"concept", "special language", "linguistic expression"}, {"terms", "system", "subject field"}, {"défini", "unité", "concept", "l'interieur", "domaine"}, hvorved i princippet et hvilket som helst begreb kan indsættes på definiendums plads og modsvare den samlede entropiske definiensværdi i de tre definitioner.

Der er to mulige årsagsforklaringer på, hvorfor definitionerne i figur 1 ikke fungerer i forhold til at ekskludere alle andre begreber fra de aktuelle: term og terminologi. Måske har en term vitterligt ingen konkrete formulerbare distinktive træk, hvormed den adskiller sig fra en ikke-term, hvilket vil have temmelig vide implikationer for det videnskabelige fundament for i al fald terminogilæren, idet grundlaget for at opretholde en meningsfuld distinktion mellem de to begreber dermed skrider. Det forholder sig dog nok nærmere således, at definitionen som eksakt beskrivelsesinstrument egner sig meget dårligt til gradsprototypiske og flerdimensionale fænomener som sproglige. Jeg hævder på ingen måde, at fagsprog ikke findes, blot at fagsproget som fænomen vanskeligt lader sig indfange i en definitionsbeskrivelse, fordi beskrivelsesinstrumentet fordrer en endimensionalitet og kerneprototypiskhed af beskrivelsesobjektet, som en lang række sproglige fænomener ikke kan honorere. Beskrivelsesinstrument og -objekt er ganske enkelt inkompatible.

I erkendelse af det forhold, at sprogets elementer er gradsprototypiske og flerdimensionale, jf. de overbevisende belæg for denne opfattelse i *prototypeteorien* (Rosch, 1975) og *fuzzy set-teorien* (Zadeh, 1974), foreslås en parametrisk kategorisering frem for definitionens logiske dikotomi, hvil-

ket falder bedre i tråd med opfattelsen af samlebegrebet terminologi som *skalare værdier*, som især de tyske fagsprogslingvister Lothar Hoffmann og Hartwig Kalverkämper har foreslået, jf. (Kalverkämper, 1990), (Hoffmann, 1976), og som har inspireret til begrebet *terminologicalness*.

Korpuslingvistikken har den meget sympatiske og tiltalende tilgang til lingvistisk analyse, at der arbejdes med autentiske sprogdata frem for konstruerede eksempler, hvor sidstnævnte tilgang rummer faren for, at eksemplerne cirkulært underbygger en opstillet lingvistisk teori. Empirien forekommer under alle omstændigheder langt mere overbevisende ved anvendelse af autentiske belæg, som John Sinclair fremhæver i sin klassiske Corpus Concordance Collocation: "One does not study all of botany by making artificial flowers." (Sinclair, 1991:6).

En række forskellige forhold har dog været direkte eller medvirkende årsag til korpuslingvistikken generelle udvikling. Således har erhvervslivet været tilbageholdende med at frigive virksomhedsrelaterede data, hvilket har bevirket en naturlig orientering imod sprogdata af en almindelig og generel natur, hvorfor der er indarbejdet en tradition for arbejde med LGP-korpora, mens der praktisk talt ikke eksisterer nogen teori for arbejde med LSP-korpora.

Chomskyanernes krav om *competence* som lingvistikken gyldige studieobjekt, som i følge denne teoretiske retning rettelig burde studeres ved hjælp af introspektion, miskrediterede i nogen grad korpuslingvistikken op gennem 1960'erne. Korpuslingvistikken måtte derfor retfærdiggøre sin videnskabelighed ved at kunne generalisere observationer foretaget på basis af et korpus i forhold til sprogsystemet, hvilket atter har afstedkommet repræsentativitetsbegrebet, således at det undersøgte korpus – opfattet som en stikprøve fra populationen "Sproget" – skal være repræsentativt.

Rammerne for denne artikel tillader ikke en nærmere gennemgang af den kritik, som fra forskellig side har været rejst imod repræsentativitetsbegrebet. Grundproblematikken kan imidlertid sammenfattes ved et citat fra Henning Spang-Hanssens Indføring I Sprogstatistik: "I sproglige problemstillinger er det ofte sagligt urimeligt at tale om populationer – hvad er populationen af "det engelske ordforråd" eller "alle nudanske sætninger"? (...) Derimod er det et led i lingvistisk forskning at operere med størrelser af modelkarakter og søge at vurdere empiriske resultater i forhold til en model (eller omvendt)." (Spang-Hanssen, 1985:B7).

Både sprogsystem og sprogbruger kan opfattes som subtile intervenierende variable for, at sproget overhovedet er i stand til at levere en stikprøve og dermed være repræsentativ, i streng statistisk forstand. Leksemer er ikke indbyrdes uafhængige, idet syntaktiske, stilistiske, pragmatisk og idiomatiske konventioner i en lang række tilfælde bevirker, at tilstedeværelsen af ét bestemt leksem dikterer tilstedeværelsen af et eller flere forudgående og/eller efterfølgende andre bestemte leksemer, f.eks. *value added tax* over for **value tax*, *she sang beautifully* over for **she sang beautiful* eller *strong tea* over for **powerful tea*. Leksemer har ikke lige mulighed for at blive udtrukket i en stikprøve, da der ikke eksisterer noget ensartet distributionsmønster for leksemer – ikke alle ord følger eksempelvis poissonfordelingen, og som Spang-Hanssen antyder, antager populationen "Sproget" i sig selv form af estimat.

En af de modeller man med fordel kan lægge nedover sine sproglige data, er den statistiske normalfordeling, som kendes som gaussfordelingen, og som afbildes som en klokkeformet kurve. En lang række fænomener følger denne fordeling – således også sproglige observationer. Ideen med gausskurven er, at langt de fleste måleobservationer (f.eks. ordfrekvenser i et korpus) falder inden for den centrale del – eller mere statistisk korrekt udtrykt: den signifikante del – af en række af forekomster (dvs. tekstens ordforråd). Der kan drages den behagelige lære af gaussfordelingen, at marginale tilfælde ganske vist forekommer, men at de netop forekommer *marginalt*. Målet er derfor at adskille den signifikante del af en teksts ordforråd fra den non-signifikante del, og til dette formål anvender man standardafvigelsen. Som sikkerhed for at afgrænsningen nu også er statistisk pålidelig, kan man anvende et konfidensniveau, som udtrykkes i antal standardafvigelser. For læseren som interesserer sig for de statistiske præmisser, kan jeg henvise til (Ruus, 1995) og (Blom, 2001). For alle praktiske formål er det nødvendigt at forudsætte, at den signifikante del af analyseteksten er identificeret, inden ordforrådet kan viderebehandles.

2. FAGSPROGSOPFATTELSE

Jeg vil foreslå et fagsprogssyn som tager sit udgangspunkt i en i fagsprogslingvistikken fremherskende opfattelse, nemlig:

1. For hvert fagligt domæne eksisterer der et begrænset antal eksperter, og dermed også tekstproducenter, set i forhold til det totale antal sprogbrugere i et bestemt sprogsamfund, (Hoffmann, 1976:170). Rent informationsteoretisk må det derfor antages, at termer er ord, som har en lav spredning over tekstsæt.
2. En af de typiske kohærensegenskaber ved fagsprog (LSP) er tendensen til at foretrække leksisredundans frem for stilistisk variation. I (typiske) fagsprogstekster vil man søge at matche ét begreb med ét sprogligt symbol i én bestemt diskurs frem for at sprede det samme begreb ud på et antal forskellige sproglige symboler ved anvendelse af synonymi, parafrase, pronominal henvisning, etc., (Engberg, 1998:29ff). Rent informationsteoretisk må det derfor antages, at termer er *indholdsord*, som optræder hyppigt i én bestemt tekst.

En del af denne problematik er naturligvis velbeskrevet ud fra og i forhold til poissonmodellen i den informationsteoretiske litteratur fra 1960'erne og til omkring midt-70'erne, f.eks. (Bookstein & Swanson, 1974; 1975), (Harter, 1975), (Stone & Rubinoﬀ, 1968), (Dennis, 1967), (Damerau, 1965). Det kan ikke desto mindre være ganske illustrativt at underbygge det intuitivt plausible i det anførte fagsprogssyn ved en simpel logisk test, hvor det iagttages, hvorvidt et bestemt ord enten optræder eller ikke optræder i en tekstsamlings forskellige delkorpora. Tekstsamlingen er den store (og den *er* stor – i omegnen af 200 millioner løbende ord) britiske tekstbank, Bank of English, som danner grundlag for Collins COBUILD English Dictionary.

TABEL 1 Spredningstest for "and", "human" og "spectrometry"

"and"		"human"		"spectrometry"	
bbc	1	bbc	1	bbc	0
npr	1	npr	1	npr	0
oznews	1	oznews	1	oznews	0
sunnov	1	sunnov	1	sunnov	0
times	1	times	1	times	1
today	1	today	1	today	0
ukbooks	1	ukbooks	1	ukbooks	0
ukephem	1	ukephem	1	ukephem	1
ukmags	1	ukmags	1	ukmags	0
ukspok	1	ukspok	1	ukspok	1
usbooks	1	usbooks	1	usbooks	1
usephem	1	usephem	1	usephem	0

Graden, hvori et givet ord er almindeligt forekommende kan aflæses af ordets spredning målt på tekstsæt. Der registreres udelukkende, hvorvidt et ord forekommer (værdien 1) eller ej (værdien 0). Kilde: Bank of English.

Som det fremgår af tabel 1, er den mellemtekstlige spredning forskellig for de tre undersøgte ord: "and", "human" og "spectrometry", således at de to førstnævnte ord forekommer i alle subkorpora i Bank of Englishs tekstbank, mens sidstnævnte ord har nulforekomst i 2/3 af de tolv subkorpora.

Tabel 2 viser, at der også er belæg for antagelsen om, at indholdsord optræder højt placeret på en rangsorteret frekvensliste i fagsproglige tekster, mens dette ikke er tilfældet for almensproglige tekster. Forskellen er markant – indholdsordene (og her ser jeg helt bort fra modal-, hjælpe- og kopulaverber) manifesterer sig slet ikke i LGP-korpusset, mens "person", "supply", "section" og "goods" er blandt de hyppigst forekommende ord i LSP-korpusset.

Alene disse ord ville tjene som ret håndfaste spor til klassifikation af teksten, i hvilken de forekommer: det er tydeligt, at der er en klar sammenhæng mellem ordklasse, frekvens og topikhed. Ud fra de foreliggende stikord ville man måske ikke entydigt kunne klassificere teksten som

Den Britiske Momslov, men ud fra en forlænget rangliste, ville det bestemt være muligt.

Det skal imidlertid nævnes, at LSP-korpora naturligvis kan rumme *falske positiver* – elementer, som forekommer i outputtet, mens som er irrelevante i forhold til genereringsformålet. Dette moment af støj bør nedbringes til at acceptabelt minimum (jf. diskussion i (Blom, 2001)).

TABEL 2. Sammenligning mellem almensprogligt korpus og fagsprogligt korpus.

LGP-korpus	LSP-korpus
1. the	1. the
2. be	2. of
3. of	3. be
4. and	4. a
5. a	5. to
6. he	6. in
7. in	7. or
8. that	8. and
9. to	9. any
10. have	10. by
11. it	11. that
12. i	12. for
13. to	13. person
14. for	14. as
15. they	15. supply
16. you	16. section
17. on	17. this
18. with	18. fall
19. no	19. which
20. she	20. goods

Kilde: British National Corpus og
The VAT Act 1994.

3. DATA

I erkendelse af at den opstillede algoritme er en præ-prototype udvælges en LSP-analysetekst, som ud fra såvel diskursive som tekstkonventionelle kriterier må regnes som hardcore, nemlig Den Britiske Momslov, The VAT Act 1994. Diskursorienteringen er esoterisk, skrevet af fagfolk til fagfolk, og ud fra tekststaffattelsesidealet tilstræbes der en høj grad af præcision i lovkonciperingen, forstået som bestræbelserne på at matche ét bestemt begreb med ét bestemt sprogligt udtryk for den aktuelle tekst. Problemerne med flertydige indholdsord må derfor antages at være absolut minimale, hvilket bekræftes af en analyse af den signifikante del af ordforrådet i VAT Act, som ikke optræder på den tekstspezifiske stopliste, hvilken har til formål at screene indholdsord med spatio-temporalt betydningsindhold. Her er under 1% af alle ordformer homografe, jf. (Blom, 2001:173).

Tekstens ordvolumen fylder godt 80.000 løbende ord, mens der er omkring 2.500 differente ordformer. Det suffikslemmatiserede² ordinventar udgør ca. 1.200 lemmata.

4. METODIK

I den informationsteoretiske litteratur fra 1960'erne og op til midten af 1970'erne redegøres for forholdet mellem indholdsord og funktionsord (eller *specialty words* og *non-specialty words* for at anvende den rette terminologi fra dengang) – og frekvens. Funktionsordene udgør en finit ordmængde, som har overvejende syntaktisk funktion, og bidrager til at skabe sammenhæng mellem elementerne fra den betydningsbærende del af en teksts ordforråd – indholdsordene. Funktionsordene er uegnede til at kunne udtrykke dele af en teksts centrale tematik, hvorfor man kan se bort fra disse. Indholdsordinventaret kan identificeres automatisk ved hjælp af en POS-tagger, som kan foretage sondringen indholdsord versus funktionsord med en høj grad af præcision.

Alle tokenvarianter af en bestemt grundbetydning, f.eks. *acquire*, som kan repræsenteres ved tekststrengene *acquire*, *acquires*, *acquired*, *acquiring*, *acquirer*, *acquirer's*, *acquirers*, *acquirers'*, *acquisition*, *acquisition's*, *acquisitions*, *acquisitions'*, kan indfanges automatisk ved at lemmatisere

analysetekstens ordforråd suffiksivt. På denne måde opnås en vis garanti for, at en lemmaform ikke registreres som non-signifikant, blot fordi den aktuelle tekstkonvention favoriserer bestemte bøjnings- eller afledningsmønstre frem for andre. I tilfældet *acquire* kunne man forestille sig grænseværdien 5, og at hver morfologisk variant af grundstammen *acquire* forekommer 4 gange i teksten, hvorved lemmaet som sådan ville udgå, til trods for at det forekommer 48 gange.

Når man har identificeret den centrale del af analysetekstens ordforråd – indholdsordslemmata, som forekommer med mindst den fastsatte grænseværdi, og som ikke forekommer på en kriteriebestemt stopliste – kan den egentlige identifikation af de tekstspecifikke ord begynde.

Dette statistisk set centrale lemmatiserede ordforråd underkastes en analyse for at konstatere, hvilke dele heraf, som dels er aktive tekstdeskriptorer, og dermed med en høj grad af sandsynlighed udtrykker dele af tekstens centrale tematik, dels nærer en særlig præference for analyseteksten. Ideen er at isolere den del af ordinventaret, som karakteriserer analyseteksten, således at netop denne tekst adskiller sig fra alle andre tekster. Lemmata som er aktive tekstdeskriptorer udtrykker gyldig – genuin – topikhed, mens lemmata som udviser en markeret præference for analyseteksten frem for andre tekster, udtrykker tekstens specifikke – partikulære – tematik. Fællesmængden af lemmata som honorerer begge kriterier, er tekstens *genuine partikulære tekstdeskriptorer*, eller forkortet: gpt'er.

Denne statistiske fagsprogsopfattelse bryder fuldstændigt med det traditionelle onomasiologiske terminologisyn, idet analyseteksten – defineret som en LSP-tekst ud fra kommunikative og diskursive faktorer frem for vidensindhold, hvilket vanskeligt lader sig afgrænse konsekvent og modsigelsesfrit – og ikke et forhåndsdefineret termsystem danner rammen for, hvilke ord (eller rettere lemmata), som opnår betegnelsen *term* (eller rettere gpt).

4.1 TEKSTINTERNT KRITERIUM: FREKVENNS OG SÆTNINGSDISPERSION

For at kunne afgøre, hvilke af VAT Acts statistisk signifikante lemmata (som ikke er stopord), som udtrykker dele af tekstens centrale tematik, må der opstilles en heuristik for, hvorved *topikhed* kendetegnes. Et topik

aktualiseres ved en række intersektionspunkter gennem en tekst, således at de samme indholdslemmata samoptræder inden for rammerne af et tekstsegment, som vi lader være sætningen. En sætning udtrykker en diskursdelmængde, hvis status som vellykket affattelse som betydningsenhed er underlagt helt bestemte sproglige, stilistiske, pragmatiske og domænebestemte konventioner. Problemet, hvorvidt en sætning er at regne som en troværdig velformet og veludført betydningsenhed, vil næppe være påtrængende i den aktuelle sammenhæng. Det må antages, at en LSP-forfatter er en kompetent sprogbruger, som er bekendt med de ovenfor nævnte affattelsesmekanismer for, hvorledes en sætning bliver vellykket, hvorfor spørgsmålet ikke skal forfølges nærmere. Der er under alle omstændigheder god ræson i at bryde en tekst ned i mindre bestanddele, hvorved gives et afgrænset analyserum, således at analysen bliver både overskuelig og håndterbar.

En afgørende forudsætning for, hvorvidt et indholdslemma udtrykker en grad af tekstens tematik er derfor først og fremmest lemmaets sætningsdispersion, hvilket afspejler lemmaets evne til at sprede sig ud over tekstens sætninger. Den tekstinterne dispersionsgrad kan anskues fra to synsvinkler. For det første: analyselemmaets *dispersionspotentiale*, hvilket vil sige analyselemmaets forekomsttal sat i forhold til tekstens maksimale antal sætninger. For det andet: analyselemmaets udnyttelse af sit *dispersionspotentiale*, hvilket vil sige det aktuelle antal sætninger, i hvilke analyselemmaet forekommer sat i forhold til tekstens maksimale antal sætninger.

Samler man disse synsvinkler i ét udtryk, indfanger man et dækkende billede af analyseordets sætningsdispersion. Denne ordiagttagelse er vertikal, fordi man så at sige følger lemmaets vej igennem teksten fra start til slut. Den vertikale ordiagttagelse afspejler analyselemmaets dominanskraft i teksten, og bygger på den grundtanke, at et ord, som forekommer i mange forskellige sætninger frem for at have tendens til at klumpe sig sammen i få bestemte tekstafsnit, er en topikindikator for hvor centralt et lemma er i forhold til teksten.

En anden faktor er spørgsmålet, hvorvidt et ord indgår i en *aktiv* topikrelation, hvorfor det er nødvendigt at supplere den vertikale ordiagttagelse med en horisontal ordiagttagelse – graden hvori et analyselemma samoptræder med andre specifikke indholdslemmata. En sådan ordiagttagelse kan bringes i stand ved at sammenligne forskellige samoptræden-

de indholdslemmatas tekstdominanskraft i forhold til analyselemmaets egen tekstdominanskraft. Sandsynligheden taler for, at samoptrædende indholdslemmata, hvis spredningsstruktur (i én bestemt analyse) ligger meget tæt på hinanden, indgår i det samme topik.

Et analyselemmas evne til at korrelere med andre forskellige specifikke indholdslemmata kan derfor udnyttes til at vurdere, hvorvidt analyselemmaet er en overvejende passiv eller aktiv tekstdeskriptor. Ud fra denne heuristik er topikhed en kvantificerbar størrelse – antallet af opnåede gyldige korrelativer er målestokken herfor. Læsere, som interesserer sig for de principielle frem for de strengt matematiske aspekter af rang_0-teknikken, kan med fordel springe afsnit 4.2 over.

4.2 DE VERTIKALE OG HORIZONTALTE ORDIAGT'TAGELSER: RANG_0-ANALYSE³

Et analyselemmas sætningsdispersionspotentiale sammenholdt med den aktuelle sætningsdispersion kan formaliseres ved følgende udtryk: (analyselemmaets forekomststal sat i forhold til tekstens samlede antal sætninger) som sættes i forhold til (tekstens samlede antal sætninger sat i forhold til antal sætninger, i hvilket analyselemmaet optræder) – altså spredningspotentialet sat i forhold til aktuel sætningsspredning. Denne formalisering kan udtrykkes i følgende formel:

$$\text{rang}_0 = ((x/x_n)/(y/y_n))$$

De anvendte variabler er: lemmaforekomststal for hele teksten (x), lemmaforekomststal for de aktuelle sætninger (x_n), sætningsforekomststal for hele teksten (y), sætningsforekomststal for de aktuelle sætninger (y_n).

Rang_0-teknikken har et vist slægtskab med *mutual information* (MI), jf. (Church & Hanks, 1990) og (Magerman & Marcus, 1990), en meget udbredt formalisme, hvorved foretages en sammenligning af sandsynligheden for, hvorvidt to ord forekommer mere hyppigt sammen end hver for sig, således at ordpar med en høj gensidig informationsværdi med stor sandsynlighed indgår i et kollokationsmønster, mens ordpar med lav gensidig informationsværdi med stor sandsynlighed er tilfældigt samoptrædende ord. Hvor MI foretager en sammenligning ord/ord i forhold til teksten, foretages ved den aktuelle metode en sammenligning

$\text{ord}_1/\text{ord}_2/..\text{ord}_n$ i forhold til sætninger i forhold til teksten, hvorved gives et mere nuanceret beskrivelse. MI i sin grundform beskriver bigrammer, og opererer derved udelukkende med *kookkurensværdier*, mens den aktuelle metode, kaldet *rang_0*, beskriver de i den eller de aktuelle sætninger forekommende indholdsord på én gang, og opererer derved med *korrelationsværdier*.

På basis af formelen genereres analyselemmaets *rang_0*-score, som er idealværdien, hvilken alle andre indholdslemmata (som er signifikante og non-stopord), som optræder i de samme sætninger som analyselemmaet, skal matche. På denne måde måler man analyseordets evne til at tiltrække sig andre indholdslemmata og derved skabe løbende intersektionspunkter gennem analyseteksten. Jo flere sådanne intersektionspunkter, desto stærkere topikrelation etableres der.

Det formelle udtryk rummer en statistisk skævvridning (bias) i forhold til utilsigtet favorisering af lavfrekvente lemmata, hvis frekvens/sætningsantal-ratio har relativ stor sandsynlighed for at konvergere mod værdien 1 og dermed medvirke til en favorabel *rang_0*-score. Denne statistiske begivenhed kan hindres i at indtræffe ved at introducere et kriterium for, hvilke samoptrædende indholdslemmata, som overhovedet tillades at indgå i *rang_0*-analysen. Derfor er kun lemmata, hvis forekomsttal befinder sig mindst en standardafvigelses afstand fra analyselemmaets forekomsttal, gyldige potentielle korrelativer til analyselemmaet.

For lemmata, som honorerer kravet om, at forekomstallet ligger inden for den tilladte afstand af analyselemmaets eget forekomsttal, udregnes en samlet værdi for antal tekstforekomster i forhold til antal sætningsforekomster over for antal aktuelle sætninger i forhold til det totale antal sætninger i hele teksten. For analyselemmaet, som underkastes *rang_0*-analyse, bliver frekvenstallet altid 1, fordi tæller og nævner er identiske, mens disse værdier varierer blandt de samoptrædende lemmata.

FIGUR 2

Princippet bag rang_0-analyse. Der beregnes en afstand fra analyselemmaets egen rang_0-score, som udgør sammenligningsintervallet, inden for hvilket samoptrædende indholdslemmats individuelle rang_0-scores skal ligge, for at de opnår status som gyldige korrelativer til analyselemmaet.

Tæller for sætningsantallet er altid konstant, fordi dette refererer til én og samme tekst, mens nævner er variabel, fordi sætningsantallet er forskelligt for de individuelle analyselemma, men identisk for alle samoptrædende lemmata i en given individuel rang_0-analyse. Derfor kan man påregne, at beregningen leder til vidt forskellige resultater for de samoptrædende lemmata, og der vil generelt forekomme en meget stor spredning af resultatværdierne. Denne spredning af resultatværdierne anvendes til at indfange de samoptrædende lemmata, hvis rang_0-score ligger tættest på analyselemmaets egen rang_0-score.

I min afhandling diskuterer jeg principperne for minimumantallet af korrelativer, som kan accepteres for, at et analyselemma er en aktiv tekst-deskriptor, jf. (Blom, 2001:184ff).

4.3 TEKSTEKSTERNT KRITERIUM: FREKVENSDOMINANS OG TEKSTDISPERSION

Algoritmens sidste modul kvantificerer kriteriet *partikulær teksttypiskhed* i form af en intuitivt set elegant formel, hvorved sammenlignes det samme ords dominansværdi i to forskellige tekstmængder, og tekstpræferen-

cen udtrykkes ved den såkaldte *coefficient of weirdness* (Ahmad, 1995). Formlen, som illustreres i figur 3, er konstrueret til belysning af et ords tilknytning til Tekst A, Tekst B eller begge tekster ved simpel division af ordets to relativfrekvenser i teksterne, og er konstrueret af Khurshid Ahmad og hans kolleger ved universitetet i Surrey, (Ahmad et al., 1994). Lader man Tekst A være en LSP-tekst og Tekst B være en modeltekst, som skal emulere et LGP-korpus, vil ord som tenderer mod Tekst A have en weirdness-score under 1, ord som tenderer mod Tekst B have en weirdness-score over 1, mens ord, som forekommer med lige stor vægt i begge tekster, vil have weirdness-scoren 1. Fællesmængden af ord er ganske almindeligt forekommende ord, enten de meget hyppige funktionsord, som forekommer i alle tekster, eller indholdsord, hvis betydningsindhold refererer til gradsmæssigt almindeligt udbredte og kendte fænomener eller foreteelser.

Underkaster man indholdslemmata i VAT Act, som efter rang_0-analyse har opnået det krævede antal korrelativer, weirdness-analyse, vil man kunne afgrænse den del af tekstens ordforråd, som fortrinsvis begrænser sig til netop VAT Act frem for alle andre tekster. Weirdnessformlen kan illustreres som følger:

FIGUR 3

$$\text{Weirdnesskoefficient} = \frac{F_n \text{ Tekst B}}{F_n \text{ Tekst A}}$$

Princippet bag weirdness-analyse. For hvert analyseord beregnes ordets relative frekvenstygde i to tekster, en LSP-tekst og et LGP-modelkorpus (f.eks. BNC). Weirdnessværdien $< 1,00$ indikerer, at ordet fortrinsvis forekommer i LSP-teksten. Weirdnessværdien $> 1,00$ indikerer, at ordet fortrinsvis forekommer i LGP-teksten. Weirdnessværdien 1,00 indikerer, at ordet forekommer er lige hyppigt i begge tekster.

Antager man, at det anvendte LGP-korpus i vid udstrækning er en virkelighedstro afspejling af gængse ordfordelinger, må det være forventeligt, at jo mere et ords weirdness-score går imod en uendelig talværdi, desto mere almindeligt forekommende er ordet (gruppen af unikke ord i Tekst B), mens jo mere et ords weirdness-score går imod talværdien nul, desto mere LSP-specifikt er ordet (gruppen af unikke ord i Tekst A). Ord, hvis weirdness-score befinder sig i intervallet 0 til 0,99 er de LSP-teksttypiske ord.

5. RESULTATER

Den modulære algoritme genererer 93 gpt'er. Det gælder, at disse med høj sandsynlighed tilhører VAT Acts karakteristiske ordforråd, og at deres forekomst heri får denne tekst til at adskille sig fra alle andre tekster. Det genererede ordinventar er VAT Act's *kvantitative terminologi* og er identificeret ud fra syntaktiske, ortografiske og statistiske kriterier. Sammenholdes gpt-listen med det fulde lemmaindeks for VAT Act, vil det formentligt undre et trænet øje, at man hverken finder act, tax, warehouse, duty, fiscal, Part, sub-paragraph, incorporate, appeal, refund, tribunal eller zero-rate blandt de genererede gpt'er. Sådanne lemmata ville efter alt at dømme blive godkendt som gyldige termer ud fra en traditionel terminologibehandling. Rent tematisk spiller ingen af disse lemmata dog nogen afgørende rolle i VAT Act, men vil ganske givet kunne henføres til en eller flere andre specifikke tekster, hvor de hver især er gyldige gpt-lemmata. De 93 lemmata gengives i Tabel 3 nedenfor.

TABEL 3. Liste over de genererede gpt'er.

accountant	consist	imprison	prevent	statement
acquire	council	inaccuracy	private	statistics
action	course	legislate	provision	statute
allocate	customs	liability	purpose	subject
amount	debt	liquidate	rate	subsection
appropriate	default	lower	reason	substitute
assess	describe	maximum	receiver	summary
asset	due	member	regard	supplement
bankruptcy	effect	notice	register	supply
business	enter	occupy	regulate	surcharge
cancel	equal	order	remove	treasury
capacity	excuse	output	resolve	treat
carry	exercise	paragraph	respect	trust
cease	expedience	pay	satisfy	understate
claim	falsify	penalty	schedule	value
commissioner	foregoing	per_cent	section	vat
conduct	furnish	person	serve	vehicle
consequence	goods	pound	service	
consideration	import	prescribe	specify	

De angivne ordformer udtrykker én bestemt repræsentationsform for en række paradigmatiske værdier (bøjnings- og afledningsformer).

Nogle af de angivne gpt'er har gradsrelevans for VAT Act i kraft af syntagmatiske egenskaber, typisk som del af et kollokationsmønster, f.eks. lemmaet *carry*, som ofte findes i sammensætningen "carry out a business", eller *member*, som udelukkende bidrager til en karakteristik af VAT Act i forbindelsen "member state". Derfor kan den opstillede algoritme med fordel anvendes som første identifikationstrin, hvis også flerordstermer ønskes udtrukket af teksten. Eksempelvis kunne de genererede gpt'er indgå som nøgleord i syntagmestrukturer, som udtrækkes vha. Justeson & Katz's algoritme, (Justeson & Katz, 1993; 1995). Syntagmatiske forekomster som rummer mindst ét af de genererede gpt'er, kan også lokaliseres ved anvendelse af N-gramstatistik (Blom, 1998a; 1998b).

Nogle af gpt'erne i tabel 3 ville formentlig ikke blive klassificeret som termer af en terminolog, hvilket dog udelukkende er en konsekvens af den omstændighed, at en terminolog identificerer en term ud fra onomasiologiske kriterier, hvilke ikke er umiddelbart forenelige med den

kvantitative terminologis korpusorienterede tilgang, hvor termer identificeres ud fra det lingvistiske tegns udtryksside og fra dets statistiske adfærd.

De 93 gpt'er er identificeret ud fra konsistente, håndgribelige samt kvantificer- og operationaliserbare statistiske parametre, hvorved en LSP-teksts (eller et LSP-korpus') topikstruktur kan afdækkes i form af et sæt stikord. Den kvantitativt terminologiske analyse kan tilmed foretages automatisk.

6. PERSPEKTIVER

Der er en række oplagte praktiske anvendelsesformål for den skitserede algoritme, herunder en systemimplementering, hvorved kan udtrækkes kvantitativ terminologi af en LSP-tekst, hvilket vil have meget stor anvendelsesværdi for erhvervssproglige studerende, translatører eller leksikologer, som vil kunne få et godt overblik over en fagteksts tematik og fagspecifikke ordinventar.

En anden oplagt anvendelsesmulighed for specifikt rang_0-teknikken er som delmodul i en *www-søgemaskine*. Det nævnes ofte, at informationssamfundets store paradoksproblem består i, at den enorme informationstilvækst, som kontinuerligt finder sted, medfører uoverskuelighed, fordi det er umuligt at frasortere uvæsentlig information. I forhold til Internet forholder problemet sig nærmere således, at vi i dag har *søgemaskiner* og ikke *findemaskiner*. Hermed mener jeg, at alle eksisterende søgemaskiner er baseret på boolsk logik, hvorved alle denne verdens webdokumenter søges matchet med en søgestreng – og inddeles i enten dokumenter, som matcher eller ikke matcher søgestrengen. På denne måde fremfindes udelukkende dokumenter, som *muligvis* er relevante i forhold til søgekriterierne, mens webbrugeren har behov for dokumenter, som *sandsynligvis* er relevante i forhold til søgekriterierne. Under alle omstændigheder forekommer det grundlæggende kontraintuitivt at søge efter information udtrykt i natursprog vha. matematisk sprog, idet man netop dermed indbygger informationssamfundets fundamentale paradoksproblem i selve søgemetodikken: at søgemaskinen fralægger sig ethvert ansvar i forhold til udvælgelsesprocessen, og sender sorteper videre til webbrugeren, som må bruge uforholdsmæssig lang tid til at vælge og vrage mellem et stort antal af fremfundne webdokumenter, hvoraf

en betragtelig andel vel og mærke ikke er relevante i forhold til den af webbrugeren ønskede specifikke teksttematik, men blot matcher søgekriterierne.

Rang_0-teknikken ville kunne forbedre hit-raten for fremfundne dokumenter, idet en betragtelig del af sorteringen er indbygget i rang_0-teknikken. Jeg vil i den kommende tid søge at finde samarbejdspartnere for en decideret kommerciel udnyttelse af rang_0-teknikken til netop sidstnævnte anvendelsesmulighed med henblik på at skabe den intelligente søgemaskine.

Bjarne Blom
Handelshøjskolen i Århus,
email: bjarneblom@hotmail.com

NOTER

1. Denne artikel er baseret på en række problemstillinger i min PhD-afhandling fra Syddansk Universitet.
2. Proceduren at henføre paradigmatiske forskellige ordformer til det samme grundlæggende morfem ved at se bort fra eventuelle suffikser, således at lemmatiseringsproceduren både omfatter bøjning og afledning (f.eks. kan følgende ordformer sammenfattes i lemmaet *tax*: tax, taxes, taxation, taxable), kan foretages automatisk.
3. Rang_0 refererer til det forhold, at analyselemmaet tildeles en neutral rangværdi, nemlig nul, hvorved dette lemma forhindres i at kunne indgå som korrelativ til sig selv.

LITTERATUR

- Ahmad, K. (1995): *Language engineering and the processing of specialist terminology*. www-postscript manuscript. Department of Mathematical and Computational Sciences, University of Surrey, Guildford.
- Ahmad, K. et al. (1994): What is a term? The semi-automatic extraction of terms from text. M. Snell Hornby, F. Pöschhacker & K. Kaindl (red.): *Translation Studies An Interdiscipline*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.
- Blom, B. (1998a): A statistical and structural approach to extracting collocations likely to be of relevance in relation to an LSP sub-domain text. *Proceedings of NODALIDA' 98*: 160-167.
- Blom, B. (1998b): A Method for automatic extraction of collocations likely to be of relevance in relation to an LSP sub-domain text. *Pre-final UDOG-rapport*. København.
- Blom, B. (2001): *Kvantitativ Terminologi - En algoritme til automatisk identifikation af teksttypiske lemmata i tekster med esoterisk diskursorientering*. PhD-afhandling. Institut for Erhvervssproglig Informatik og Kommunikation. Syddansk Universitet, Kolding.
- Bookstein, A. & Swansson, D.R. (1976): A decision-theoretic foundation for indexing. *Journal of the American Society for Information Science* 26: 45-50.
- Bookstein, A. & Swansson, D.R. (1975): Probabilistic models for automatic indexing. *Journal of the American Society for Information Science* 25: 312-318.

- de Bessé, D. (1990): La définition terminologique. J. Chaurand & F. Mazire (red.) *La Définition. Actes du colloque la définition*: 252-262. Paris: Larousses.
- Church, K.W. & Hanks, P. (1990): Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1): 22-29.
- Damerau, F.J. (1965): An experiment in automatic indexing. G. Schechter (red.) *American Documentation* (4): 283-289.
- Dennis, S.F. (1965): Dennis: The design and testing of a fully automatic indexing-search system for documents consisting of expository text. G. Schechter (red.) *Information Retrieval: A critical review*: 67-94. Thompson Books Co.
- Droz, S.L. (1973): *Deutsche Fach- und Wissenschaftssprache*. Wiesbaden: Brandstetter Verlag.
- Droz, S.L. (1972): Zum Gegenstand und zur Methode der Terminologielehre. *Muttersprache* 86(2): 109-117.
- Engberg, J. (1998): *Introduktion til fagsprogslingvistikken*. Århus: Systime.
- Felber, H. & Budin, G. (1989): Terminologie in Theorie und Praxis. *Forum für Fachsprachen-Forschung* 9.
- Harter, S.P. (1975): A probabilistic approach to automatic keyword indexing. Part 1: On the distribution of specialty words in technical literature, Part 2: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science* 26: 197-206, 280-289.
- Hoffman, L. (1976): *Kommunikationsmittel Fachsprache*. Berlin.
- ISO/DIS, 1087: *Terminology - vocabulary. Draft International Standard ISO/DIS 1087*. International Organization for Standardization, 1988.
- Justeson, J. & Katz, S. (1995): Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1: 9-27.
- Justeson, J. & Katz, S. (1993): *Technical terminology: Some linguistic properties and an algorithm for identification in text*. Technical Report RC 18906, IBM Research Division.
- Kalverkämper, H. (1990): Gemeinsprache und Fachsprachen - Plädoyer für eine integrierende Sichtweise. *Deutsche Gegenwartssprache. Tendenze und Perspektiven. Jahrbuch des Institut für deutsche Sprache* 1989: 88-133. Gerhard Stickel, Berlin - New York.
- Magerman, D. & Marcus, M. (1990): *Parsing a natural language using mutual information statistics*. Pennsylvania University, CIS Department, Ms.

- Picht, H. & Draskau, J. (1985): *Terminology: An introduction*. University of Surrey.
- Rosch, E. (1975): Basic objects in natural categories. *Cognitive Psychology* 8: 382-439.
- Ruus, H. (1995): *Danske kerneord, centrale dele af den danske leksikalske norm*. Københavns Universitet: Museum Tusculanums Forlag.
- Sinclair, J.M. (1991): *Corpus Concordance Collocation*. Oxford University Press.
- Spang-Hanssen, H. (1985): *Indføring i sprogstatistik*. Københavns Universitet.
- Stone, D.C. & Rubinoff, M. (1968): Statistical generation of technical vocabulary. *American Documentation*: 411-412.
- Wüster, E. (1931): *Internationale Sprachnormung in der Technik, besonderes in der Elektrotechnik*. 3. Udgave. Bonn.
- Zadeh, L.A. (1974): Fuzzy logic and its application to approximate reasoning. *Information Processing* 74: 591-594. Amsterdam: North-Holland.