Kvantitative metoder inden for korpuslingvistiske projekter – illustreret ved eksempler fra Den Danske Ordbogs korpus, Korpus 2000 og Korpus 90.

Jørg Asmussen
Det Danske Sprog- og Litteraturselskab
www.dsl.dk

Forelæsning holdt på ph.d.-metodekursus, 8.9.2004
Forskerskole Øst
Version 1.1 – 12.10.2004

Resumé

Indledningsvis gives et par generelle betragtninger om grammatikalitet og statistisk sandsynlighed. Derpå skildres korpuslingvistikkens fremkomst – og dermed empirismens genkomst – som et overvejende teknologisk betinget fænomen. Da store korpora ikke kan overskues med det blotte øje, hverken mht. komposition eller mht. fremtrædende tekstuelle regelmæssigheder i dem, udgør statistiske metoder et vigtigt redskab inden for korpuslingvistikken. Der illustreres derfor to prototypiske metoder, nemlig frekvensprofiler og log-likelihood til korpussammenligning og mutual information til fremfinding af fremtrædende tekstuelle strukturer. Endvidere diskuteres materialets beskaffenhed, som er en væsentlig forudsætning for kvaliteten af kvantitative analyser.

Til forelæsningen hører en præsentation, som er tilgængelig under `korpus.dsl.dk/staff/ja/papers/gradeast2004/presentation`. Der henvises til de enkelte slides i præsentationen med en angivelse af deres numre i margin til højre. 2

Indhold

1 Baggrund

1.1 Grammatikalitet og statistisk sandsynlighed

Grundlaget for lingvistik er antagelsen, at sprogets struktur kan beskrives – og i bedste fald forklares. Beskrivelsen kan ske i form af generaliserende regler: en given sproglig ytring vil da enten svare til de opstillede systemiske regler og dermed være *grammatisk*, eller ikke gøre det, og således være *ugrammatisk*. Det, der således beskrives, er sprogsystemet som sådant, *kompetensen*. Beskrivelsen kan også ske i form af sandsynlighedsangivelser for, at en bestemt struktur (eller et bestemt fænomen) optræder i sproget – en sådan beskrivelse tager sit afsæt i de sproglige frembringelser, *performansen*. Den første type beskrivelse betegnes gerne som *rationalistisk*,[1] idet den principielt kan nøjes med, at lingvisten bruger sin egen eller informanters kompetens som kilde til formuleringen af sprogsystemets regler, mens den anden hyppigt kaldes *empiristisk*,[2] hvor naturligt forekommende data er udgangspunktet for beskrivelsen. Begge metoder kan føre til de samme resultater. 3–4

Inden for rationalistisk sprogbeskrivelse gås der ud fra, at kompetensen, menneskets evne til sprog, er medfødt.[3] Dette begrundes med, at den sproglige stimulus i sig selv er for vag til at blive opfattet som sprog, hvis ikke man allerede har et latent, medfødt sprogligt system at hægte stimulus op på.[4] Inden for empiristisk sprogbeskrivelse går man ud fra, at sprogbrugeren ikke har en medfødt sproglig kompetens, som skal aktiveres, men derimod en medfødt evne til at associere, genkende strukturer og generalisere, jf. fx [Manning and Schütze, 1999]. Udstyret med disse evner behandles det sproglige input, så der opnås en færdighed i at forstå og selv frembringe sproglige ytringer – ikke på baggrund en medfødt kompetens, men *afledt af* det sproglige input. Tilsvarende forsøger den empiristiske sprogbeskrivelse at beskrive den sproglige struktur ud fra sproglige data. Data er her ofte – og ofte i mangel af andet – store skriftsprogskorpora. Da performansen ikke følger sprogsystemets generaliserede regler helt, ændres fokus i den empiristiske sprogbeskrivelse fra det binære enten-eller-princip til en kvantificerbar størrelse. Således kan der være større eller mindre sandsynlighed for at en 5–6 6

[1] [DDO: Hjorth et al., 2005] definerer *rationalisme* på flg. måde: "filosofisk retning der hævder at viden og erkendelse kan opnås udelukkende med fornuften og intellektet, uafhængig af sanserne og erfaringen".

[2] [DDO: Hjorth et al., 2005] definerer *empirisme* således: "filosofisk retning der hævder at viden og erkendelse kun kan opnås på grundlag af erfaringer, iagttagelser og eksperimenter".

[3] Den ofte anførte formel i denne sammenhæng lyder *tacit competence*, jf. [Chomsky, 1965].

[4] Formlen lyder *poverty of the stimulus*, jf. [Chomsky, 1986].

sproglig struktur eller et andet sprogligt fænomen forekommer.

Der er blevet ført en del skyttegravskrige mellem den rationalistiske og 7
empiristiske lejr. Således skal Chomsky ifølge [Leech, 1991] have fremført

> some sentences won't occur because they are obvious, others because they are false, still others because they are impolite

som et argument mod empiristisk sprogvidenskab. Indtager man standpunktet, at man kun beskriver det, man ser, men samtidig som kompetent sprogbruger per introspektion godt ved, at man faktisk kan frembringe ytringer, man ikke finder i sit empiriske materiale, så bliver ens beskrivelse af sproget ufuldstændig, måske endda skævvredet, idet man kan være tilbøjelig til at tillægge hyppige fænomener en større betydning for beskrivelsen end sjældne, hvilket Chomsky ifølge [Halliday, 1991] skal have illustreret ved at *I live in New York* jo så må tillægges en højere vægt i ens sprogbeskrivelse end *I live in Dayton, Ohio*.

Fra den empiristiske side fremføres gerne argumentet, at hvad der gælder for andre videnskaber, der beskæftiger sig med naturfænomener, må også gælde for lingvistikken, der beskæftiger sig med naturfænomenet sprog:

> One does not study all of botany by making artificial flowers. [Sinclair, 1991]

Selvom modsætningerne mellem "empiristerne" og "rationalisterne" ofte bliver fremført som en god historie, er der grundlæggende egentlig kun tale om en metodologisk gradsforskel, jf. [Manning and Schütze, 1999]. Ingen sprogempirist kan slukke for sin egen sproglige kompetens, når han gransker sit materiale, lige så lidt som en sprogrationalist kan se bort fra den faktiske sprogbrug i verden uden for hans eget hoved. Begge veje kan føre til samme mål.

Gradsforskellen betyder bl.a., at man inden for den empiristiske sprog- 8
videnskab i stedet for et grammatikalitetsbegreb snarere bruger et sandsynlighedsbegreb: der hersker sproglige regelmæssigheder, der gør det sandsynligt, at en bestemt ytring har en bestemt struktur, men det udelukker ikke, at andre strukturer kan forekomme og fungere i den sproglige kommunikationsproces. Ud fra sandsynlighedsbegrebet kan den stadige sproglige forandringsproces måske beskrives mere hensigtsmæssig end ud fra et mere rigidt grammatikalitetsbegreb, jf. [Manning and Schütze, 1999]. I takt med at metoder til opbygning og undersøgelse af store tekstkorpora er blevet forbedret, har empirisk orienteret sprogbeskrivelse vundet mere frem i de seneste år.

1.2 Korpuslingvistikkens fremkomst og udvikling

Fremkomsten af empirisk orienteret sprogbeskrivelse skyldes vel egentlig i mindre grad overvejelser om, hvilken teoretisk-metodologisk tilgang til sprogbeskrivelse der er den mest hensigtsmæssige, men er først og fremmest begrundet i en teknologisk udvikling, der har gjort det muligt at gennemføre kvantitative undersøgelser af store mængder sprogligt materiale. Korpuslingvistikken, som den forstås i dag, er utænkelig uden den informationsteknologiske udvikling, der har fundet sted især gennem de sidste 20 år. 9

Det område, hvor korpusbaseret sprogbeskrivelse meget hurtigt fik en vis udbredelse, var leksikografien; og mange af de første statistiske metoder har derfor også et primært leksikografisk sigte, jf. fx motivationen for [Church and Hanks, 1989] og [Church et al., 1991].

Inden for leksikografien skelnes principielt mellem to beskrivelsesprincipper, det *præskriptive* og det *deskriptive*.[5] Sidstnævnte tager sit udgangspunkt i sprogets faktiske brug, ikke i en idealiseret forestilling om, hvordan det egentlig burde bruges, og den er det fremherskende princip for bl.a. dansk-, tysk- og engelsksproget leksikografi i Europa. 10

Traditionelt har den leksikografiske metode beroet på indsamlingen af sprogbrugscitater, som typisk blev nedfældet på sedler. Disse udgjorde siden det materiale, det korpus, som leksikografen arbejdede ud fra. Fremgangsmåden blev fx anvendt under udarbejdelsen af Ordbog over det danske Sprog, Svenska Akademiens Ordbok, Oxford English Dictionary og Grimms Deutsches Wörterbuch. Metoden har den skavank, at indsamlingen ofte vil bære præg af, at filologen eller leksikografen har en større tilbøjelighed til at lægge mærke til det usædvanlige, det overraskende, det ukendte sprog, og registrere dét snarere end det forventelige. 11

Til leksikografiske formål er store maskinlæsbare tekstkorpora en mere neutral kilde, idet de normalt vil indeholde lange passager af ubearbejdet, autentisk tekst, dvs. også det forventelige sprog. Det første overvejende korpusbaserede ordbogsprojekt var [COBUILD: Sinclair et al., 1987], der blev søsat i 1980 som et samarbejde mellem University of Birmingham og og forlaget Collins. Her blev de klassiske seddelkasser afløst af computere, som store tekstmængder kunne gemmes på, og som hurtigt kunne opstille såkaldte konkordanser, hvori et bestemt ord kunne vises i alle de kontekster, det optrådte i i korpus uden det menneskelige filter, der tidligere var lagt imellem det sprog, man faktisk kunne læse og høre, og det, der så

[5] I praksis vil de fleste ordbøger både "beskrive og vejlede" brugeren, jf. fx [DDO: Hjorth et al., 2005] bd.1, p.8, altså være blandingsformer. Den noget diffuse betegnelse "proskriptiv" for visse blandingsformer forekommer også, jf. [Bergenholtz, 1998], men har næppe vundet udbredelse som et særligt leksikografisk beskrivelsesprincip.

siden blev registreret på ordsedlerne. I Danmark er et tilsvarende korpusbaseret ordbogsprojekt i øvrigt netop afsluttet, nemlig Den Danske Ordbog, [DDO: Hjorth et al., 2005], som blev udarbejdet på baggrund af Den Danske Ordbogs korpus i perioden 1991–2003.

Men også med anvendelsen af korpora støder man hurtigt ind i visse 12
praktiske begrænsninger. Dels kan det være vanskeligt at danne sig et præcist indtryk af, hvordan ens korpus er sammensat og hvilken slags sprog det afspejler: er det virkelig *sproget* som sådant? Dels kan antallet af belæg for et givet fænomen være problematisk: sålænge der kun er op til hundrede eller måske tohundrede forekomster af et ord, vil leksikografen være i stand til at læse alle eksemplerne og danne sig et indtryk af, hvilke forskellige anvendelser af et givet ord, der er på spil. Er der derimod tale om tusindvis af eksempler (*abundance*),[6] er det umuligt at læse dem og generalisere over dem inden for acceptabel tid. Her må der særlige midler til, der kan sammenfatte og strukturere de mange data. Disse særlige midler involverer forskellige statistiske metoder, der ofte er indbygget i gængse korpusværktøjer som fx WordSmith Tools eller Semaskop.[7]

1.3 Statistiske metoder i korpuslingvistikken

Umiddelbart falder de mest udbredte metoder inden for korpuslingvistikken 13
således i to grupper:

1. Metoder til undersøgelse af en teksts eller et korpus' beskaffenhed. Ønsket er at kunne beskrive meget præcist, hvad der er i et korpus, hvilke teksttyper det er sammensat af, og at sammenligne korpora og tekster med hinanden. Mere sofistikerede anvendelser af disse metoder er fx automatisk dokumentklassifikation eller indholdsresumering.

2. Metoder til fremfinding af bestemte sproglige strukturer. Ønsket er at finde både faste udtryk, fremtrædende samforekomster, syntaktiske strukturer m.v. En mere sofistikeret anvendelse af disse teknikker er fx sprogmodellering.

Fælles for disse metoder er, at de ikke tager udgangspunkt i en kvalitativ fortolkning af materialet, inden de kommer til anvendelse: med andre ord betragtes kun kvantificerbare enheder, der kan bestemmes entydigt, dvs. algoritmisk/automatisk. Vi skal i det følgende se eksempler på, hvordan metoder

[6] Også det modsatte af *abundancy*, nemlig *scarcity*, for få forekomster af et bestemt ord, er et problemkompleks inden for korpusleksikografien.

[7] WordSmith findes under `http://www.lexically.net/wordsmith` og Semaskop under `http://korpus.dsl.dk/e-resurser/semaskop.php?lang=dk`

fra disse to grupper kan anvendes: Metoder fra den første gruppe beskrives i afsnittene 2.2.1 og 2.2.2, mens en metode fra den anden gruppe vil blive beskrevet i afsnittene 2.2.3 og 2.2.4. I praksis kan metoderne godt vikariere for hinanden, dvs. at metoder fra den første gruppe godt kan anvendes til formål beskrevet under den anden gruppe og omvendt.

Vi vil ikke komme ind på korpusstatistiske metoder, der involverer en forudgående kvalitativ fortolkning af materialet, omend sådanne undersøgelser ellers er ganske udbredte. Eksempler herpå er undersøgelser af bestemte begrebers og begrebsdannelsers sproglige udtryk (fx *demokrati*, *terrorisme*, *dannelse* eller *politisk korrekthed*), holdninger, sådan som de måtte komme til udtryk i sproget (fx befolkningens holdninger over for *fremmede*, over for *EU* eller *homoseksuelles ligestilling*), eller sociolingvistiske undersøgelser, fx mænds over for kvinders sprog, lærer-elev- eller læge-patient-kommunikation. Sådanne undersøgelser tager ofte deres udgangspunkt i et indsamlet materiale, som der kvantificeres på, men det, der tælles, er ikke sproglige enheder, der kan bestemmes algoritmisk – kun gennem en menneskelig fortolkningsproces. Det er tvivlsomt, om sådanne undersøgelser kan kaldes empiriske i snæver forstand.

2 Anvendelser

2.1 Forudsætninger

2.1.1 Eksempelmateriale

I det følgende demonstreres et par basale korpusstatistiske anvendelser inden for de to grupper, som blev beskrevet i forudgående afsnit. Disse vil ikke primært tage udgangspunkt i et helt korpus, men derimod – for overskuelighedens og forståelsens skyld – også bero på et par mindre tekster, taget fra Den Danske Ordbogs korpus. Det, der skal bestemmes, er teksternes signifikante ord-inventar (jf. gruppe 1 i forudgående afsnit) og ord-samforekomster (jf. gruppe 2 i forudgående afsnit). Målinger foretages kun på entydigt identificerbare sproglige fænomener (her kun *ord*, jf. 2.1.2 ndf.). Når de fænomener, der skal behandles statistisk, ikke er entydigt identificerbare, fx fremkommer ved en fortolkningsproces, bør man være varsom med at antage, at de udsagn og fortolkninger, man fremsætter på baggrund af sine statistiske analyser, har generel gyldighed for sproget eller virkeligheden – de har det kun for det materiale, man har undersøgt!

Alligevel er korpuslingvister og -leksikografer tilbøjelige til at postulere en 14
vis generel udsagnskraft af deres resultater, gerne begrundet i anvendelsen af et stort og varieret – *balanceret* – tekstkorpus. Et sådant betragtes gerne som

en sprogprøve, der er repræsentativ for sproget som sådant: altså en "spand sprog" taget op til nærmere undersøgelse fra "sprogets hav" i forventningen om, at prøven er repræsentativ for helheden.

Korporas balancerethed eller repræsentativitet udgør ikke desto mindre et 15
stort metodisk problem for korpuslingvistikken, da afgørelsen om, hvorvidt et korpus er balanceret eller repræsentativt for noget, må bero på et subjektivt skøn, en fortolkning. Skal dette problem løses, bør der findes tilfredsstillende svar på i hvert fald de følgende sprøgsmål:

- Hvilke teksttyper skal et korpus være sammensat af for at være en statistisk anvendelig stikprøve for sproget som helhed?
- I hvilke mængdeforhold skal de forskellige teksttyper stå i forhold til hinanden?
- Hvilke tekstuelle egenskaber determinerer egentlig en bestemt teksttype?
- Hvordan vurderer (måler?) man, om en tekst har disse egenskaber eller ej?

Den Danske Ordbogs korpus (DDO's korpus), som blev samlet 1991-1993, 16
er det første større danske tekstkorpus, hvor man under udarbejdelsen har været meget bevidst om problemerne balancerethed og repræsentativitet, jf. [Norling-Christensen and Asmussen, 1998]. Den Danske Ordbogs korpus er på 40 mio. tekstord, hvoraf en delmængde på 28 mio. tekstord er gjort offentlig tilgængelig under betegnelsen *Korpus 90* (K90). Reduktionen på 12 mio. ord skyldes, at talesprogstekster og tekster, som er belagt med særlige brugsrestriktioner fra leverandørernes side, er blevet fjernet. Endelig er der sidenhen blevet udarbejdet endnu et almensprogligt korpus på 28 mio. tekstord, *Korpus 2000* (K2000), som en 10–15 år yngre pendant til Korpus 90.

For DDO's korpus blev der opstillet et komplekst beskrivelsesapparat for at få styr på teksternes forskellige karakteristika. Man var klar over, at helt objektive kriterier for, hvad der udgør et balanceret korpus, kan være vanskelige at opstille, så man valgte en pragmatisk tilgang i stedet. Man opstillede tre dikotomier, nemlig skriftsprog vs. talesprog, "professionelt" vs. "privat" sprog[8] og almensprog vs. fagsprog. Med udgangspunkt heri kunne alt tekstmateriale opdeles i otte forskellige klasser. For hver af disse klasser blev det undersøgt, hvilke og hvor mange forskellige tekstkilder der var tilgængelige,

[8]Eller *receptions-* hhv. *produktionssprog*.

og man stilede efter, at hver enkelt af de otte klasser skulle være velrepræsenteret i korpus, gerne så tæt på en ottendedel for hver af de klasser, som det var praktisk muligt.

Ud over disse tre grundlæggende dikotomier blev der til tekstbeskrivelsen anvendt et større inventar af karakteristika, som forudgår hver enkelt af de i alt 43.806 teksteksempler i korpus, jf. figur 1 fra [Norling-Christensen and Asmussen, 1998].

TextInfo	
TextID	Unambiguous identifier of the text sample — for citation purposes
Restrictions	
Anonymity	Proper names must be altered (A), or not (-), if cited
DD_Only	Text must only be used by The Danish Dictionary
TextTitle	Title of the text
VolTitle	Name of anthology, newspaper, magazine etc.
Publisher	Publishing house, broadcaster etc.
PublTime	
Day	{1, 2, ..., 30, 31}
Month	{1, 2, ..., 11, 12}
Year	{1983, 1984, ..., 1992, 1993}
Certainty	The year of publishing is known exactly (-), or not (?)
Location	E.g. book volume, newspaper section, page number
LangType	{general, specialized}
Expression	{written, spoken, and two intermediate types}
Aspect	{reception, production}
AgeRelation	{adult–adult, adult–juvenile, adult–child, ..., child–child}
Medium	{book, journal, radio, diary, ...} — 13 possible values
Genre	{novel, interview, essay, ...} — 131 possible values
GenreType	A reduced classification for statistical use — 17 values
Topic	{philosophy, geography, physics, ...} — 66 possible values
TopicType	A reduced classification for statistical use — 12 values
Group	Unambiguous identifier of a group of related text samples
Number	Serial number within the text group
Size	Number of tokens in the following text sample
UserInfo+	(one or more language users: author(s)/speaker(s))
UserID	Identifier referred to by speaker turns in the text
Surname	Surname of the language user
FirstName	First name of the language user
Sex	{male, female, unknown}
Born	{1880, 1881, ..., 1989, 1990}
Certainty	The year of birth is known exactly (-), or not (?)
BirthPl	Place of birth
Residence	Place of residence
Region	Dialectal region — 11 values
Education	Education of the language user
Occupation	Occupation of the language user
LangVar	Language variant {standard, regional}
Role	Communicative role of the language user, e.g. teacher, pupil

Figur 1: Tekstoplysninger knyttet til hver af de 43.806 tekster i DDO's korpus

Korpus 2000 anvender en væsentlig forenklet beskrivelse: dels er antallet af oplysningskategorier reduceret kraftigt, dels er værdimængderne, der bruges til at beskrive de enkelte kategorier, betydeligt mere overskuelige. Eksempelvis er oplysninger vedrørende sprogbrugernes uddannelse, erhverv

og herkomst ikke taget med, da disse kun giver mening, hvis mængden af mulige værdier, der kan beskrive en sprogbrug inden for en af disse kategorier er veldefineret og afgrænset, hvilket ikke er tilfældet for DDO's korpus. I øvrigt skal man have et overordentligt stort og konsekvent annoteret korpus, hvis man virkelig vil bruge det til at lave sociolingvistiske undersøgelser med, fx af typen "*hvad er forskellen på 'skolelæreres' og 'kontorassistenters' sprogbrug?*" Endvidere er værdimængderne for de forskellige oplysningstyper reduceret væsentligt, fx fra 131 forskellige genrer i DDO's korpus til kun tre i Korpus 2000.

Fra DDO's korpus skal der her bruges to skriftsproglige (S1 og S2) og to talesproglige (T1 og T2) teksteksempler til illustration af en række basale statistiske metoder, jf. tabel 1 på den følgende side. Der er tale om uddrag taget fra større tekster; uddragene er på godt 5000 ord[9] hver. De skriftsproglige er også indeholdt i K90 og findes af samme grund også i en grammatisk opmærket version, hvorimod talesprogsteksterne kun er med i DDO's korpus som "rene" tekster, men med tilhørende udførlige beskrivelser. 17–18

Alle teksterne foreligger i en forholdsvis ubehandlet version. Fra de skriftsproglige tekster er egentlige typografiske oplysninger som fx *kursiv*, *understreget* etc. fjernet. Til gengæld er der tilføjet lidt SGML-kode, fx `<p>` og `</p>`, der hhv. betegner begyndelsen og slutningen på et afsnit, {NL}, der angiver et tvungent linjeskift, `<note>` og `</note>`, der omgiver tekst, som oprindelig har stået i en fod- eller slutnote, samt en række andre – især i talesprogsgengivelsen –, som er nærmere beskrevet i [Norling-Christensen and Asmussen, 1998].[10]

2.1.2 Identificerbare sproglige enheder

Før man kan gå i gang med at tælle noget i sit tekstmateriale, skal dette "noget" defineres entydigt. Forholdsvist sikkert kan man segmentere tekstmaterialet i "ord" og "sætninger"; i denne fremstilling vil vi dog nøjes med udelukkende at se på ord.

Et ord – eller rettere et token – forstås som regel som en bogstavsekvens, der afgrænses af mellemrum i en skreven tekst. Mange af disse tokens vil umiddelbart kunne fortolkes som ord. Imidlertid er det et problem, hvilke tegn der må tælle med i et token. Hvilken status har fx punktum, binde- 19

[9] En mere præcis definition af, hvad vi her forstår ved et ord, følger på side 13.

[10] Teksterne er tilgængelige for egne eksperimenter under følgende web-adresser:
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/eksempeltekster/s1.txt`
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/eksempeltekster/s2.txt`
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/eksempeltekster/t1.txt`
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/eksempeltekster/t2.txt`

Fork.	Udtryk	Tekst	Type	Længde
S1	skrift	Kirsten Fink og Ole Terney: *Sådan reguleres genteknologi. Praksis og erfaringer*. Foreningen af Bioteknologiske Industrier i Danmark, 1988. Kapitel 1-3.	Teksten er i DDO's korpus registreret som en fagsproglig monografi om emnet biologi.	5511 ord
S2	skrift	Vita Andersen: *Petruskas laksko*. Gyldendal, 1989. Side 5-8, 12-17 og 22-26.	Teksten er i DDO's korpus registreret som en roman for børn.	5047 ord
T1	tale	Tekstuddrag fra *Projekt Bysociolingvistik* fra 1987.	Teksten er i DDO's korpus registreret som et gruppeinterview med en voksen interviewer og unge interviewede.	5621 ord
T2	tale	Fjernsynsinterview under titlen *Gamle mennesker fortæller*. DR TV, 1983.	DDO's korpus registrerer, at interviewet handler om historie. Der medvirker en speaker med et par replikker, ellers en interviewer og en interviewet.	5369 ord

Tabel 1: Teksteksempler til illustration af en række basale statistiske metoder

streg eller apostrof? Skal de betragtes som en del af et token, eller er de tokenadskillere?

En bindestreg brugt som tankestreg er givet en tokenadskiller, men en bindestreg i fx *banegård-center* vil man sandsynligvis tælle med som del af tokenet. Et punktum vil man umiddelbart måske betragte som en periodegrænse, men hvad så med punktum i forkortelser som fx *etc.* eller *ph.d.* eller et tal som *12.000,-* – og hvilken status har kommaet og bindestregen her? Hvilken status har sætningstegn i det hele taget? Skal de betragtes som selvstændige tokens – eller skal de blot ignoreres helt? Og hvis man betragter dem som tokens, må man da tælle dem med i sine statistiske undersøgelser?

Beslutter man sig for kun at tage hensyn til ordene i sit tekstmateriale og således ignorere sætningstegn, er det næste problem at bestemme, hvad der er

ord, og hvad der ikke er. Hvis *banegård-center* er ét ord, så er *banegårdcenter* sikkert også, men hvor mange ord er da *banegård center*? Eller hvad med *underskrive aftalen* og *skrive aftalen under*? Og hvordan kan man bestemme særskrevne ord i sit korpus med sikkerhed?

En anden overvejelse er, om man skal omsætte store bogstaver til små (eller omvendt). Gør man det, får man en mere overskuelig optælling af tokens, der er således ikke forskel på *man* og *Man* eller *hans* og *Hans*. Gør man det ikke, kan man måske lettere få et indtryk af, hvilke ord der oftest bruges som initialord i en periode.

Endelig er der spørgsmålet om, hvorvidt man bør normalisere ortografien 20–21
i sit korpus, om man fx skal beslutte, at *banegårdcenter* er normalformen, som har en række varianter, herunder ikke blot de allerede nævnte, men også dem, hvor et eller flere bogstaver er med stort. Normaliserede former behøver ikke "ødelægge" korpussets ortografiske autenticitet, idet normalformerne kan introduceres som et særligt lag – en normaliseret parallelform til hvert enkelt ord i korpus.

Til spørgsmålet om ortografisk normalisering hører også selve nedskriften af talesprogstekster. I DDO's korpus har man valgt at linearisere tekstmaterialet totalt, dvs. at replikker ikke kan overlappe hinanden i tid, jf. appendiks A.1 på side 32. En anden måde at transskribere talesprog på, er i form af partiturer (se appendiks A.2 på side 33), hvor der er én linje per deltager i samtalen, og hvor replikker principielt kan noteres overlappende.[11]

Hvilken *ortografi* vælger man? Som det ses af eksemplerne, har DDO's version *fader*, hvor BySocs web-version i stedet for bruger *far*, og dét selvom der i bund og grund er tale om samme kildemateriale! Det er meget tænkeligt, at BySoc på et tidspunkt – efter at DDO havde fået sin version – har ændret ortografien fra *fader* til *far*, dog heller ikke helt konsekevent, idet der stadigvæk er otte forekomster af *fader* i BySocs webversion.

Hvordan fortolker man det hørte egentlig rigtigt? Talestrømmen er generelt jo omsat til pænt afgrænsede ord, der overvejende følger retskrivningen, men hvad kan begrunde et *hjnå*? Og hvorfor staves præpositionen *inden for* så *indenfor*? Introduktionen af ordgrænser er endvidere problematisk, især hvis der sidenhen skal arbejdes videre med materialet på en statistisk ren måde: skal talestrømmen genkendes automatisk, fx inden for machine learning eller sprogmodellering, bør den ikke være præsegmenteret, idet netop segmenteringen burde være en del af lære- og sprogmodelleringsprocessen. En segmentation i foner (eller måske morfer) ville være en mere forsvarlig fremgangsmåde under transskriptionen, hvis man siden ønsker at udnytte materialet til disse

[11] Om der faktisk forekommer overlapninger i BySoc-notationen, kan ikke umiddelbart ses ud fra web-versionen.

anvendelser. Hvis man derimod blot ønsker at opstille frekvensprofiler og foretage kollokationsanalyser, som i denne fremstilling, kan ordopdelingen godt forsvares, idet man således forholdsvis umiddelbart vil kunne sammenligne resultaterne med dem fra skriftsprogstekster.

BySocs web-version opererer derudover med en ret nuanceret angivelse af pauselængder – der er tre gradueringer –, som i DDO's version er reduceret til én standardpause. Hvad kan disse pauser ækvivaleres med i skriftsprogstekster? En ækvivalering (eller i det mindste en stillingtagen til dette spørgsmål) er ønskelig, hvis man vil sammenligne talesprogs- med skriftsprogstekster.[12]

BySocs web-version har endvidere anonymiseret samtlige proprier, således at kun første bogstav bliver tilbage, mens de efterfølgende erstattes af et tilsvarende antal procenttegn (*Holmen* → *H%%%%%*).

Disse få eksempler blot få at give et indtryk af, at identifikationen af ord i en tekst i høj grad afhænger af de definitoriske valg, man træffer, og at man altid må regne med mulige uoverensstemmelser mellem konceptet af et fænomen og dets konkrete repræsentation i sit system.

Væstenligst er det, at definitionen af det, der skal afgrænses, er algoritmiserbar. For vores eksempeltekster fastlægger vi derfor følgende ord-defintion: 22

Ord — er de karaktersekvenser, som afgrænses af mellemrum (blanktegn, linjeskift eller tabulatorer). Tags, dvs. tokens omgivet af `<` og `>` eller {...} betragtes som ekstratekstuelle elementer og ignoreres helt. Der skelnes mellem store og små bogstaver.
Til teksten hører kun de ord der står mellem taggene `<Tekst ID=.*?>` og `</Tekst>`.

Følgende tegn ignoreres:

- punktum efterfulgt af mellemrum: konsekvensen er dog, at forkortelser med afsluttende punktum vil få fjernet punktummet
- udråbstegn, spørgsmålstegn, komma, semikolon efterfulgt af mellemrum
- bindestreg efterfulgt af mellemrum: konsekvensen er, at der ved sammensætninger, hvor sidsteleddet er angivet ved en bindestreg, vil den blive fjernet og førsteleddet identificeres som et selvstændigt ord; jf. fx *højt- eller lavtflyvende*
- flere på hinanden følgende bindestreger eller punktummer

[12] Det store antal af uløste metodiske problemstillinger ved gengivelse af talesprogstekster har været medvirkende til, at talesprog blev fjernet fra K90 og ikke indgår K2000.

- højre- og venstreparenteser
- dobbelte anførselstegn
- enkelte anførselstegn forudgået eller efterfulgt af mellemrum: mellemrummenes tilstedeværelse adskiller dette tegns funktion som anførselstegn fra dets funktion som apostrof

Når der er fastlagt kriterier for, hvordan man segmenterer teksten i ord (og fx perioder), kan man tilføje yderligere informationer til disse niveauer. Til hvert ord kan der således fx oplyses ordklasse, bøjningsoplysninger, grundform (gennem tagging) og syntaktisk funktion (gennem parsning). Teksterne i K90 og K2000 er både taggede og parsede, dvs. at S1 og S2 også foreligger i en sådan annoteret version. Da automatisk tagging og parsning af nedskrevne talesprogstekster som regel ikke giver gode resultater,[13] foreligger disse to tekster ikke i en morfosyntaktisk annoteret version. Figur 2 viser en periode fra S2 med morfosyntaktisk opmærkning. 23

```
<s s_id="WBNHKKZE" s_nr="471465" txt_id="GxCa" preom="0" bop="1" eop="1">
Traditionelt            [traditionel]                ADJ     NEU S IDF NOM   @>N
forædlingsarbejde       [forædlingsarbejde]          N       NEU S IDF NOM   @SUBJ>
har                     [have]          <mv>         V       PR AKT          @FMV
de                      [den]                        ART     nG P DEF        @>N
samme                   [samme]                      DET     nG nN NOM       @>N
mål                     [mål]                        N       NEU P IDF NOM   @<ACC
$,
men                     [men]           <co-acc>     KC                      @CO
må                      [måtte]         <aux>        V       PR AKT          @FAUX
bruge                   [bruge]         <mv>         V       INF AKT         @ICL-AUX<
metoder                 [metode]                     N       UTR P IDF NOM   @<ACC
$,
der                     [der]           <rel>        INDP    nG nN NOM       @SUBJ>
er                      [være]          <mv> <np-close> V    PR AKT          @FS-N<
mere                    [meget]         <aquant>     ADV     COM             @>A
tidskrævende            [tidskrævende]               ADJ     nG nN nD NOM    @<SC
$,
og                      [og]            <co-fin>     KC                      @CO
som                     [som]           <rel>        INDP    nG nN           @SUBJ>
ofte                    [ofte]          <atemp>      ADV                     @ADVL>
gør                     [gøre]          <mv> <np-close> V    PR AKT          @FS-N<
det                     [den]                        PERS    NEU 3S ACC      @F-<ACC
svært                   [svær]                       ADJ     NEU S IDF NOM   @<OC
eller                   [eller]         <co-oc>      KC                      @CO
umuligt                 [umulig]                     ADJ     NEU S IDF NOM   @<OC
at                      [at]                         INFM                    @INFM
overskride              [overskride]    <mv>         V       INF AKT         @ICL-A<
artsbarriererne         [art+barriere]  <compound>   N       UTR P DEF NOM   @<ACC
$.
</s>
```

Figur 2: Eksempel på morfosyntaktisk opmærkning i K2000/90

Som det ses, er grundstrukturen givet ved et starttag `<s>` og sluttag `</s>` for perioden. Til starttagget knytter der sig en række attributter, der dels identificerer perioden (`s_id`, `s_nr`), dels fortæller, hvilken tekst den stammer fra (`txt_id`). Endelig gives der nogle kontekstuelle oplysninger til intern brug,

[13] Dette var en yderligere årsag til, at talesprog blev elimineret fra K2000/90.

om hvorvidt perioden står i begyndelsen (bop) eller slutningen af et afsnit (eop), samt om der i selve korpus er udeladt en passage før denne periode (preom).

Inden for <s> og </s> står selve perioden, således at der kommer et ord hhv. et sætningstegn (sidste præfigeret af $) per linje. For hvert løbende ord/token i teksten (1. kolonne) gives der oplysninger om grundform (2. kolonne, i kantede parenteser), ordklasse (4. kolonne), bøjningsoplysninger (5. kolonne) samt syntaktisk funktion (6. kolonne, præfigeret af @). Endelig findes forskellige supplerende oplysninger af forskellig art i kolonne 3. Taggingen og parsningen er udført af Eckhard Bick, VISL-projektet ved Syddansk Universitet.[14]

2.2 Undersøgelser

2.2.1 Hyppighedsstatistik og frekvensprofiler

Når det er fastlagt, hvilke grafiske enheder der skal identificeres og indgå i statistikken, kan optællingen begynde.

Den mest primitive optælling er at se på, hvor mange ord teksten er 25
lang, dvs. bestemme antallet af løbende ord eller *tokens*. Dernæst kan man undersøge, hvor mange forskellige ord eller *types* teksten indeholder, hvorefter man kan beregne, hvor tit hver type bruges i gennemsnit i teksten (token/type-forhold, *TTF*). Tabel 2 viser resultaterne for de fire eksempeltekster.

Tekst	Tokens	Types	TTF
S1	5511	1822	3,02
S2	5047	1097	4,64
T1	5621	990	5,68
T2	5369	989	5,43

Tabel 2: Simple ordoptællinger i eksempelteksterne

Umiddelbart ser det ud til, at ordgentagelsen er størst i talesprogsteksterne, endda større end i en skønlitterær børnebog. Man fristes til at postulere, at en ukendt tekst med et token/type-forhold på over fem nok sandsynligvis er en talesprogstekst. Selv om denne form for statistik er yderst primitiv, viser den en tendens. Man skal dog huske på, at vi under vores orddefinition valgte at skelne mellem store og små bogstaver, og at BySoc-teksten

[14]Om baggrund og metode jf. [Bick, 2003b] og [Bick, 2003a]. En præcis beskrivelse af de anvendte tags og deres betydning findes på VISL's hjemmeside http://visl.sdu.dk/visl/da/info/tagset_da.html.

ikke er delt op i perioder, der begynder med stort. I talesprogsteksterne vil der derfor ganske enkelt være færre ord, der begynder med stort.[15]

Et bedre indtryk af en teksts kvantitative egenskaber får man, hvis man opstiller en såkaldt frekvensprofil. Dette er i sin mest primitive udgave en liste over samtlige forskellige ord*former* (types) i en given tekstmængde sorteret efter faldende[16] hyppighed. Toppen af disse profiler for de fire eksempeltekster er vist i appendiks B.1–B.4 på side 34–37.

Til sammenligning vises toppen for hele Korpus 90 i appendiks B.5 og for Korpus 2000 i figur B.6 på side 38–39.[17]

På de første 22 pladser i resultatet for S1 findes der udelukkende partikler og pronomener, et par former af hjælpeverberne (*er* på rang 3, *har* på rang 10) og verbalformen *kan* (rang 15). Alle er de i sig selv ret indholdstomme, og de kan næppe give et fingerpeg om fx tekstens indhold. Først på rang 23 optræder et ord med mere konkret indhold, nemlig substantivformen *mikroorganismer*.

Ser vi på T1-resultatet, er billedet endnu mere karakteristisk: partikler, pronomener og hjælpeverber dominerer listens top.[18] Først på rang 39 optræder et mere indholdsbærende ord, verbet *hedder*; første entydige substantiv er *år* på rang 58. Ud fra frekvenslistens top er det næsten umuligt at danne sig et indtryk af, hvad teksten fx handler om. Og også T2 er tilsvarende: det første egentlige indholdsord er substantivet *far* på rang 36.

S2 afviger lidt fra de øvrige tekster: blandt de ti hyppigste ord findes navnet *Petruska* allerede på rang 2, *sagde* på rang 4, *mor* på rang 8 og *Marie* på rang 10. Dette kan være et genrespecifikt træk, teksten er uddrag fra en roman, herunder selve begyndelsen, hvor hovedpersonerne introduceres.

Det er påfaldende, at talesprogsteksternes indhold eller tema kun vanskeligt kan bestemmes ud fra de tilhørende frekvenslisters top: dette kunne måske fortolkes som, at deiksis er meget dominerende i talesprog, og at talesprog bliver nærmest indholdstomt uden den situative kontekst. Denne fortolkning støttes desuden af token/type-forholdet.

Selvom der er små forskelle de enkelte tekster imellem, så er det næppe forbavsende, at de (næsten) udelukkende indeholder funktionsord i toppen – og det er som regel uvist, om forskelle i relativ hyppighed eller rang kan

[15] Indflydelsen af dette på statistikken burde undersøges nærmere.

[16] Mere præcist: *ikke-stigende* hyppighed.

[17] De komplette profiler er tilgængelige under følgende web-adresser:
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/analyseresultater/...`
`s1_fprofil.txt`, `s2_fprofil.txt`, `t1_fprofil.txt`, `t2_fprofil.txt`,
`k90_freq.txt.zip`, `k2000_freq.txt.zip` .

[18] Formen *så* på rang 5 kunne i princippet godt skyldes hyppig brug af præteritum af *se* i teksten, men vi anser dette for usandsynligt.

bruges som kvantitativt acceptabelt grundlag for en kvalitativ fortolkning. Vil man fx danne sig et mere konkret billede af tekstmængdens semantiske karakteristika, er det nødvendigt at filtrere de mange funktionsord fra. Det kan enklest ske ved at opstille en liste over såkaldte stopord: ord der ikke skal med i profilen. Listen kunne fx indeholde samtlige ord fra de lukkede ordklasser. I stedet for slet ikke at tage disse ord med kunne man betragte dem alle sammen som (varianter af) ét ord og tælle dem alle sammen med under dette: andelen af stopord i forhold til øvrige ord i ens tekstmængde kan således nemt bestemmes. Hvis vi anvendte denne fremgangsmåde på vores tekster, ville vi formodentlig se, at andelen af funktionsord er større i talesprogs- end i skriftsprogstekster. Noget tyder altså på, at en høj andel af funktionsord i et tekstmateriale er et tegn på, at det enten er talesprog, eller at det indeholder større mængder talesprog.

Men der er andre – og måske mere interessante – statistiske undersøgelser, som kan udføres på sprogets umiddelbare skriftlige repræsentation, fx kan gennemsnitlig ordlængde og ordlængdefordeling udsige noget om tekstmaterialet. Endvidere kan gennemsnitlig periodelængde og fordelingen på periodelængder være interessant.

For talesprogsrepræsentationen er der imidlertid forbundet visse problemer med dens egentlig unaturlige skriftsprogsrepræsentation. Det gælder både på ord- og periodeniveau.

Som vi har set, bruger man ved transskriptionen af talesprog ikke altid de almindelige sætningstegn. For BySocs vedkommende bruger DDO's korpus taggene `<replik> ... </replik>` for at markere hhv. begyndelsen og slutningen på en replik, mens BySoc selv anvender partiturnotationen. I andre gengivelser af talesprog, fx tv-interviews, jf. tekst T2, anvendes der i DDO's korpus ud over replikmarkeringerne almindelig interpunktion, sandsynligvis stiltiende indsat af transskriptøren. Talesprog er derfor vanskeligt at segmentere over ordniveau, da der mangler klare periodeafgrænsninger som i skriftsprogstekster:

- Skal man under en eventuel periodesegmentering af T2 betragte punktummer som periodegrænser eller skal man blot ignorere dem?
- Hvilken rolle spiller noterede pauser og brud – kan de ækvivaleres med periodegrænser?
- Hvordan forholder man sig til en længere monolog?

Besvarelsen af disse spørgsmål er ikke entydig, men kan have stor betydning for de kvantitative undersøgelsesresultaters udseende.

2.2.2 Statistisk sammenligning af tekster vha. log-likelihood

Hidtil har vores sammenligning af frekvensprofiler for de forskellige tekster været temmelig impressionistisk. Vi har koncentreret os om de hyppige tokens og prøvet at se, hvad der er i den ene tekst og ikke i den anden osv. Ved denne fremgangsmåde er der tale om enkeltiagttagelser, der snarere styres af vores egen prædisponerethed for at finde ting, vi synes er interessante, end af statistisk signifikans. Vi skal derfor se lidt nærmere på, om man ikke ad statistisk vej kan komme til mere tilforladelige udsagn om ligheder og forskelle tekster imellem.

Hvis vi kun betragter førstepladsen på de fire frekvenslister, ser vi, at S1 har typen *at* med 3,18% af samtlige tokens, mens S2 har *og* med 3,72%, T1 har *det* med 4,80% og T2 har *og* med 4,43%. Har dette en betydning? Kan der lægges en fortolkning i det? Lad os udvide spørgsmålet til at omfatte de tre mest frekvente types i hver af de fire tekster, ni forskellige types i alt. Tabel 3 viser disse types for de fire tekster og deres relative frekvens udtrykt i procent og deres rang i profilen i parentes; til sammenligning er deres frekvens og rang for hele K90 (som jo indeholder S1 og S2, men ikke T1 og T2) vist i sidste kolonne. 26

	S1	S2	T1	T2	K90
at	3,18 (1)	1,19 (16)	0,93 (22)	1,01 (21)	2,41 (3)
i	2,50 (2)	2,60 (3)	1,32 (14)	1,73 (9)	2,76 (2)
er	2,27 (3)	2,02 (5)	2,81 (4)	0,88 (24)	1,90 (4)
og	1,67 (6)	3,72 (1)	2,38 (7)	4,43 (1)	3,00 (1)
Petruska	0 (-)	2,64 (2)	0 (-)	0 (-)	0,00 (14103)
det	1,20 (11)	1,90 (6)	4,80 (1)	3,39 (2)	1,59 (6)
ikke	0,89 (18)	1,88 (7)	3,82 (2)	1,92 (8)	1,00 (15)
der	1,43 (9)	0,69 (26)	3,10 (3)	2,31 (6)	1,14 (12)
jeg	0,02 (1193)	1,49 (11)	2,40 (6)	3,32 (3)	0,52 (26)

Tabel 3: Sammenligning af de tre hyppigste types i eksempeltekster

Umiddelbart viser det sig, at typen *at* ligger på en betydelig lavere rang i de to talesprogstekster end i skriftsprogsteksterne, især S1. Samtidig er hyppigheden af *og* meget stor i T2. Hvis ellers de forskelle, vi kan observere i hyppigheden for *at* og *og*, ikke blot er tilfældige udsving... Hvis der ikke er tale om et tilfælde, så kunne observationen måske bruges til at støtte en formodning om, at talesprog (og børnebøger) bruger færre bisætninger (i hvert fald sådanne, der indledes med *at*) og færre infinitivkonstruktioner (i hvert fald *at*-infinitiver), og – især med baggrund i udbredelsen af *og* i 27

T2 – bruger mere koordination end subordination. Det, vi ikke må, er at tage vores umiddelbare observation, som den fremgår af tabellen ovenfor, som argument for vores formodninger. En sådan fremgangsmåde ville ikke være empirisk, men blot intuitiv: den vil kun kunne bruges til at bekræfte vores forudantagelser, men næppe til at afkræfte dem eller til at føre til nye indsigter. Desværre er fremgangsmåden dog ret udbredt, og ofte under den falske varebetegnelse *empirisk*.

Helt tilsvarende gælder for en række andre observationer, vi kan foretage:

- Typen *det* synes ret udbredt og højt placereret i T-teksterne, men åbenbart mindre brugt i S-teksterne. Men er denne observation tilstrækkelig til at støtte en antagelse om, at deiksis er mere udbredt i talesprog?

- Indikerer den frekvente brug af *jeg* i T2, at teksten er fra en person, som taler om sig selv? Eller er det tilfældigt, at *jeg* optræder højt placeret?

- Kan man konkludere, at T2 handler om noget fortidigt? Vi har konstateret frekvent brug af *jeg* og kunne derfor måske også forvente frekvent brug af former af *være*. Det viser sig, at *var* optræder på rang 4 med 3,15%.

- Handler S2 om en person ved navn *Petruska*? Eller er det blot et tilfælde, at *Petruska* optræder så højfrekvent? Intuititivt er det næppe tilfældigt, men hvad hvis der havde været tale om en *Christian* eller *Susanne*?

Endelig kan man med udgangspunkt i [Asmussen, 2004a] spørge, 28

- om frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok til, at man kan fortolke dem som indikatorer for sproglig forandring

- om *kambrium* med sikkerhed *ikke* kan fortolkes som indikator for en sproglig forandringsproces

- om hyppighedsforskelle for lemmaerne *bil*, *land*, *Danmark*, *cykel*, *hus* og *mand*, som kan konstateres mellem K90 og K2000, har konsekvenser for sammenlignende undersøgelser mellem de to korpora

- om der kan opstilles en komplet liste over samtlige lemmaer, der i deres udbredelse i de to korpora afviger så meget fra hinanden, at der næppe kan være tale om et tilfælde.

For at besvare disse spørgsmål har man brug for en metode, en test, der kan 29
afgøre, om de observerede forskelle blot er tilfældige, eller om de virkelig er statistisk signifikante. Hvis resultaterne er statistisk signifikante, kan vi med temmelig stor sikkerhed (typisk med 95% eller 99%) antage, at de ikke kan skyldes et tilfælde.

Der findes en række signifikanstests, som kan bruges, når man sammen- 30
ligner korpora, men ikke alle er lige gode til dette formål. En letlæst og mere generel, ikke specifik lingvistisk introduktion til nogle af de ofte anvendte tests gives i [Clegg, 1990]. En detaljeret gennemgang og afprøvning af de forskellige tests anvendelighed for tekst- og korpussammenligning gives i [Kilgarriff, 2001]. Hans konklusion er, at Mann-Whitney ranks-test er den mest velegnede. Da den imidlertid er lidt omstændelig at implementere, vil vi her introducere en anden, næsten lige så brugbar test, nemlig *log-likelihood*. Log-likelihood fremhæver overraskende kvantitative uoverensstemmelser mellem to korpora og giver især gode resultater for lav- og mellemfrekvente ord. Testen er ligeledes velegnet til fremfinding af nye ord hhv. termekstraktion, jf. [Daille, 1995]. Testen er nem at anvende,[19] den kræver blot antal forekomster af ord i de to (eller flere) korpora, man vil sammenligne med hinanden, samt det samlede antal løbende ord i hvert af de to korpora. Matematikken bag testen er kompleks og vil ikke blive behandlet her.

Grundantagelsen bag alle statistiske signifikanstests er, at der nok ikke er 31
nogen forskel på de data, man vil sammenligne. Denne grundantagelse kaldes som regel *nul-hypotese*. Hvis der ikke var nogen forskel på de to tekster eller korpora, man vil sammenligne, så betyder det i princippet, at ordene i dem burde optræde lige hyppigt (målt i forhold til korporaenes faktiske størrelse).

Hvis vi nu vil sammenligne udbredelsen af *at* i S1 (3,18%, rang 1) og T1 32
(0,93%, rang 22), hvilken hyppighed skal vi da vælge som den norm eller basis, vi kan relatere den anden hyppighed til? Er det hyppigheden i T1, vi skal betragte som basis, eller er det hyppigheden i S1?

Valget er umuligt, for vi kender netop ikke denne norm – derfor må vi 33
konstruere den. Det gør vi på baggrund af begge tekster[20] betragtet som et samlet hele. S1 var på 5511 ord (N_{S1}), T1 på 5621 ord (N_{T1}), S1 og T1 har tilsammen altså $N_{S1} + N_{T1} = 11132$ ord (N). I S1 forekommer *at* 175 gange

[19]En nærmere beskrivelse af fremgangsmåden findes i [Garside and Rayson, 2000]. Metoden har han også implementeret i en web-baseret *log-likelihood calculatator* (`http://lingo.lancs.ac.uk/llwizard.html`). Denne vil kunne bruges til at undersøge de ovf. stillede spørgsmål i forbindelse med topscorerne i vores eksempeltekster. På web-siden er der yderligere referencer vedr. log-likelihood og andre tests.

[20]Eller alle tekster, som måtte indgå i sammenligningen. Her nøjes vi dog kun med at sammenligne to tekster ad gangen.

(O_{S1}) og i T1 52 gange (O_{T1}), i S1 og T1 tilsammen altså $O_{S1} + O_{T1} = 227$ gange (O). Den samlede relative frekvens for *at* i de to tekster er derfor $\frac{O}{N} = \frac{227}{11132} = 0{,}02039$ svarende til 2,04%. Det er denne hyppighed, vi ophøjer til vores basisværdi.

Næste skridt er at beregne, hvor mange forekomster af *at* vi burde kunne 34
forvente i hhv. S1 og T1, hvis de svarede til vores konstruerede norm. Vi ved nu, at der er $N_{S1} = 5511$ ord i S1, og at den relative normfrekvens er $\frac{O}{N} = \frac{227}{11132}$. Altså burde vi kunne forvente

$$E_{S1} = \frac{N_{S1} \cdot O}{N} = \frac{5511 \cdot 227}{11132} = 112 \text{ forekomster af } at \text{ i S1} \qquad (1)$$

og tilsvarende

$$E_{T1} = \frac{N_{T1} \cdot O}{N} = \frac{5621 \cdot 227}{11132} = 115 \text{ forekomster af } at \text{ i T1,} \qquad (2)$$

men faktisk er der jo 175 hhv. 52.

Hvis vi skulle sammenligne flere tekster på én gang, skulle vi opstille 35
en tilsvarende formel og udregning for hver af dem. Normalt vil man dog generalisere dette i én formel. Vi husker, at $O = O_{S1} + O_{T1}$ og $N = N_{S1} + N_{T1}$. Hvis vi indsætter dette i formel 1 og 2 ovf., får vi

$$E_{S1} = \frac{N_{S1} \cdot (O_{S1} + O_{T1})}{(N_{S1} + N_{T1})} \qquad (3)$$

hhv.

$$E_{T1} = \frac{N_{T1} \cdot (O_{S1} + O_{T1})}{(N_{S1} + N_{T1})} \qquad (4)$$

Vi kan se, at for hver tekst i, vi tilføjer i sammenligningen, vil vi skulle lægge dens observerede antal forekomster O_i af det aktuelle ord til i parentesen over brøkstregen, og tekstens længde N_i skal lægges til i parentesen under brøkstregen – der skal laves så mange additioner de to steder, som der er tekster i sammenligningen. Desuden skal der for hver tekst oprettes endnu en formel, der gælder specielt for denne. Dette kan udtrykkes mere generelt på følgende måde:

> Det forventede antal forekomster af et ord i en given tekst (E_i) findes ved at bestemme denne teksts længde N_i og gange den med summen af alle i sammenligningen involverede teksters observerede antal forekomster for dette ord divideret med summen af alle involverede teksters længde.

På formel:

$$E_i = \frac{N_i \cdot \sum_i O_i}{\sum_i N_i} \tag{5}$$

Det, vi mangler nu, er at finde et mål for, hvor overraskende meget de faktiske forekomsttal af *at* i de to tekster afviger fra den konstruerede norm: dette gøres med log-likelihood (som vi her simplificeret kalder L) beregnet efter flg. generelle formel: 36

$$L = 2 \cdot \sum_i O_i \cdot \ln\left(\frac{O_i}{E_i}\right) \tag{6}$$

I vores tilfælde, hvor vi kun sammenligner de to tekster S1 og T1 med hinanden, svarer det til at beregne log-likelihood efter følgende formel:

$$L_{\mathrm{S1,T1}} = 2 \cdot \left(O_{\mathrm{S1}} \cdot \ln\left(\frac{O_{\mathrm{S1}}}{E_{\mathrm{S1}}}\right) + O_{\mathrm{T1}} \cdot \ln\left(\frac{O_{\mathrm{T1}}}{E_{\mathrm{T1}}}\right)\right) \tag{7}$$

Jo højere L-værdi, desto mere statistisk signifikant er forskellen mellem de to forekomsttal. Det er fastlagt, at hvis L er større end eller lig med 3,8 – så er der 95 procents sandsynlighed for, at hyppighedsforskellen på det givne ord i de to tekster ikke skyldes et tilfælde ($p \geq 0,95$). Er L 6,6 eller større, er der endda 99 procents sandsynlighed herfor ($p \geq 0,99$). 37–38

Appendiks C.1 på side 40 viser toppen af en sammenligning af samtlige ord i S1 og T1 sorteret efter faldende L-værdi. Et plus (+) i en af kolonnerne "≫S1" og "≫T1" viser, at pågældende ord er signifikant ($p \geq 0,99$) overrepræsenteret i pågældende tekst, sammenlignet med den anden. Oversigten viser med andre ord de types, hvis frekvens er signifikant forskellig i S1 og T1 – så signifikant, at man med meget stor sandsynlighed kan sige, at forskellen ikke skyldes nogen tilfældighed. Det overlades til læseren at overveje,

- hvordan resultatet kan fortolkes
- hvorvidt det kan bruges til at be- eller afkræfte nogle af vores tidligere fortolkningsforsøg
- hvorvidt det siger noget om talesprog vs. skriftsprog generelt
- hvilke begrænsninger en sammenligning af to relativt vilkårlige tekster har.

En sammenligning af to tekster vi altid være præget af visse vilkårligheder: således gælder mulige fortolkninger udelukkende denne ene sammenligning

med netop de involverede tekster, men har ikke nogen generel relevans for sproget som sådant; det er derfor metodisk uforsvarligt ud fra sammenligningen af S1 med T1 at konkludere noget som helst generelt om skriftsprog over for talesprog, selvom log-likelihood-undersøgelsen afslører nogle træk, man traditionelt vil forbinde med skrift- hhv. talesprog. Desuden er det ikke muligt på baggrund af en sådan sammenligning at udsige noget absolut om den ene af de to involverede tekster: de mest signifikant afvigende ord for den ene af teksterne (altså fx *jeg*, *så*, *det*, *ikke*, *var*, *sådan*, *jaer* m.fl.) er ikke de mest signifikante for denne tekst som sådan, men kun for denne tekst set i relation til S1! Vil man lave log-likelihood-undersøgelser, som udpeger afvigende ord for en tekst som sådan, helt generelt, må man sammenligne med en fastlagt sproglig norm, fx i form af et stort, balanceret korpus, der kan antages at være nogenlunde repræsentativ for fx skrift- eller talesproget (eller begge varianter) som helhed. En sådan norm er oplagt K90, hvorfor vi i det følgende vil sammenligne vores eksempeltekster med K90 og ikke umiddelbart indbyrdes. Appendiks C.2–C.5 på side 41–44 viser de 35 mest signifikant forskellige ord fra hver af sammenligningerne af eksempleteksterne S1, S2, T1 og T2 med K90.[21]

Konklusioner overlades til læseren.[22] Man bør i særdeleshed overveje, hvad resultaterne for T1 afspejler, nemlig snarere indiosynkrasier (for ikke at sige tilfældigheder) ved notationen end noget som helst generelt om talesprog...

Endelig viser tabel 4 log-likelihood-værdierne for udvalgte ord i K90 og K2000, jf. [Asmussen, 2004a]. 39

Oversigten delvis støtter fortolkningerne givet ovenfor, nemlig

- at frekvensforskellene for *mobiltelefon*, *benchmarking* og *biltelefon* faktisk er sikre nok, men man bør ikke nødvendigvis fortolke dem som tegn på sproglig forandring, men kun som tegn på en forskel på K90 og K2000
- at *kambrium* med sikkerhed *ikke* kan fortolkes som tegn på en sproglig forandringsproces, ej heller som sikker indikator på en forskel på K90 og K2000
- at hyppighedsforskelle for lemmaerne *bil*, *land*, *Danmark*, *cykel*, *hus* og *mand*, som kan konstateres mellem K90 og K2000, er statistisk sig-

[21] De komplette resultater er tilgængelige under følgende web-adresser:
`http://korpus.dsl.dk/staff/ja/papers/gradeast2004/analyseresultater/...`
`ll_k90_s1.txt.zip`, `ll_k90_s2.txt.zip`, `ll_k90_t1.txt.zip`, `ll_k90_t2.txt.zip`.

[22] Grunden til, at S1 har en overrepræsentation af tal, herunder to årstal, ligger i, at (næsten) alle tal i K90 er præfigeret med et $-tegn, således skrives *1987 $1987* i K90. Ifølge den her anvendte orddefinition er der derfor ikke tale om samme ord.

lemma	K2000	K90	L	overrepræsenteret ($p \geq 0,99$)
mobiltelefon	1486	59	1607	i K2000
benchmarking	34	0	40	i K2000
biltelefon	9	51	33	i K90
kambrium	0	4	4	nej
bil	8353	10364	265	i K90
land	28204	21455	769	i K2000
Danmark	30677	22217	1168	i K2000
cykel	1343	1773	69	i K90
hus	8146	12016	840	i K90
mand	24612	29878	639	i K90

Tabel 4: L-værdier for udvalgte ord i K2000 over for K90

nifikante. Hvis man antager, at disse ord bør have en rimelig stabil udbredelse i sproget, også over lidt længere tidsrum, så har det konsekvenser for sammenlignende undersøgelser mellem de to korpora

- at der bør opstilles en komplet liste over samtlige ordformer, der i deres udbredelse i de to korpora afviger så meget fra hinanden, at der næppe kan være tale om et tilfælde, for at vurdere de to korporas ensartethed. I appendiks C.6 på side 45 vises toppen af denne liste.[23]

Resultatet viser umiddelbart to forhold: (1) at der er tidsbetingede forskelle på de to korpora, og (2) at der må være påfaldende forskelle i sammensætningen.

Ad 1) Ord som *EU/EF*, *internettet*, *Dansk_Folkeparti* samt årstallene *1996*, *1997*, *1998*, *2000*, *2001* er entydigt tidsbestemte, og det overrasker derfor ikke, at de optræder blandt de ord, der mest signifikant adskiller de to korpora fra hinanden.

Ad 2) Pronomener som *du/dig*, *hun/hende*, *han/ham*, *jeg/mig*, verber som *siger*, *havde/har*, *var/er*, subjunktionen/infinitivmærket *at*, adjektivet *danske* samt de fleste andre ord, som optræder på listen i appendiks C.6, og som ikke allerede er nævnt under (1) ovf., burde ud fra en ren intuitiv, introspektiv betragtning ikke optræde blandt de ord, der adskiller de to korpora mest signifikant

[23]Den fuldstændige liste er tilgængelig under `http://korpus.dsl.dk/staff/ja/papers/gradeast2004/analyseresultater/ll_k90_s1.txt.zip`

> fra hinanden. Man burde kunne gå ud fra, at de var konstant udbredte i to korpora, der i deres sammensætning skulle være rimelig identiske eller netop sammenlignelige. At de alligevel optræder med så høje L-værdier, tyder stærkt på, at de to korpora er sammensat mere forskelligt, end det måske ser ud til ved en ren overfladisk sammenlignende betragtning; og det viser ligeledes nødvendigheden af at udvikle statistiske metoder, der kan sikre en ensartet opbyging af to korpora, der principielt kun må adskille sig i tidsdimensionen.[24]

Som det ses, kan log-likelihood bruges til at lave mere tilforladelige sammenlignende undersøgelser af tekster/korpora, end man kan ved blot at sammenligne type-hyppigheder ud fra fx frekvensprofiler. Især ved enkelttekster eller meget homogent opbyggede korpora er log-likelihood en meget enkel og velfungerende metode. Ved sammenligning af store korpora, som består af mange meget forskelligartede tekster, vil log-likelihood være tilbøjelig til at betragte tekstuelle særheder, fx meget specielle ord, som kun optræder i en enkelt tekst i hele korpus, som særlige for hele dette korpus sammenlignet med det andet – forudsat selvfølgelig, at det andet korpus ikke indeholder en tekst med præcis de samme særheder. Således kan vi iagttage, at *Petruska* kun har to forekomster i K2000 mod 127 i K90; den er dermed ifølge log-likelihood stærkt overrepræsenteret i K90 (L=161). I S2, som jo er indeholdt i K90, er der 133 forekomster af *Petruska*. At der er færre forekomster i K90 end i S2 selv, hænger sammen med, at visse tekstsekvenser er blevet fjernet fra S2,[25] inden den blev indlemmet i korpus, men det må antages, at samtlige 127 forekomster af *Petruska* i K90 hidrører fra denne ene tekst, S2. Hvis S2 af en eller anden tilfældig årsag ikke havde været med i K90, havde *Petruska* heller ikke været et signifikant adskillende ord for K90 og K2000: én tekst får således markant indflydelse på hele korpus. Man kan betragte dette som en svaghed ved log-likelihood, nemlig at testen ikke tager højde for ordenes fordeling over hele materialet, altså spredningen eller *dispersionen*. Andre metoder, især Mann-Whitney ranks-test, tager højde for dette, og kan derfor være mere velegnede, jf. [Kilgarriff, 2001].

Vi har her kun beskæftiget os med frekvensprofiler og sammenligninger på enkeltord, men metoderne lader sig selvfølgelig også anvende dels på andre sproglige enheder end ord, fx lemmaer, ordklasse-angivelser, syntaktiske

[24] Et forslag er at definere såkaldte *invariante tekstuelle træk*, der måske kan tjene som særlige pejlemærker ved balanceringen af et korpus i forhold til et andet, jf. [Asmussen, 2004b].

[25] Det drejer sig om overskrifter, tekst i parenteser samt visse replikker, der kan virke forstyrrende på den automatiske morfosyntaktiske opmærkning af materialet.

angivelser, bogstaver, stavelser, foner, morfer etc., dels på grupper af sådanne enheder.

Grupper af sproglige enheder kan enten optræde i en fastlagt rækkefølge eller i vilkårlig orden inden for en nærmere defineret kontekst. Grupper af sproglige enheder kaldes ofte n-grammer, hvor n angiver gruppens længde i antal enheder.[26]

Ved at opstille frekvensprofiler for forskellige n-grammer i et sprog vil man kunne få et indtryk af gældende kombinatoriske regelmæssigheder. Vælger man som sin sproglige enhed fx n-grammer over syntaktiske annotationer, vil man kunne få et indblik i sprogets syntaks.

2.2.3 Struktur og kollokation

Frekvensprofiler over n-grammer er dog ikke altid særligt sigende: de viser 46–48
ganske vist det, der er hyppigt, men dette er ikke nødvendigvis ensbetydende med, at det også er strukturelt signifikant – og interessant. Problemet kan illustreres med et lille eksempel fra Korpus 2000's nuværende brugerinterface.[27] Ud over at det er muligt at søge på både enkelte ord/lemmaer og grupper af ord, lemmaer eller ordklasser, kan der opstilles enkle frekvensprofiler med udgangspunkt i et givent lemma. Indtastes et ord i brugerinterfacet, kan man i det derpå følgende skærmbillede aktivere funktionen *nabo-ord* (*hyppige ord i naboskabet*), og man får da opstillet en simpel frekvensprofil, jf. figur 3.[28] En nabo er her defineret som et ord, der daner et (diskontinuert) bigram[29] med det givne, dvs. det indtastede, lemma ved at det står 1 eller 2 pladser enten til højre eller til venstre for dette.

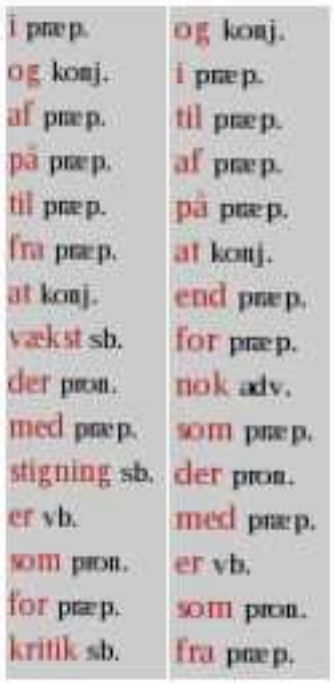
Listerne over de hypigste højre naboer for de to undersøgte lemmaer *kraftig* og *stærk* er igen domineret af funktionsord; og der er umiddelbart ikke den store forskel at se, når man sammenligner listerne: de fleste af ordene står åbenbart hyppigt til højre for både *kraftig* og *stærk*, kun *vækst*, *stigning* og *kritik* optræder åbenbart ret frekvent efter *kraftig* (og sandsynligvis sjældnere efter *stærk*); omvendt forekommer *end* og *nok* åbenbart hyppigere til højre for *stærk*. I det følgende vil vi demonstrere en metode til at opspore iøjnefaldende samforekomster af ord i et korpus: vi vil finde de højre naboer, der netop er typiske for *kraftig* hhv. typiske for *stærk*. Samforekomster af

[26] Ordpar er følgelig 2-grammer og kaldes også bigrammer.

[27] På `http://www.korpus2000.dk` – brugerinterfacet vil blive afløst af et nyt i løbet af 2005.

[28] Frekvensoplysningerne er fjernet fra figuren. I K2000-interfacet bruges kun forenklede kvantitative oplysninger, hvis beregning er beskrevet i [Asmussen, 2004b].

[29] At bigrammet er diskontinuert betyder, at de to ord ikke behøver at stå umiddelbart ved siden af hinanden, men at der også kan stå et antal ord imellem dem.



i præp.	og konj.
og konj.	i præp.
af præp.	til præp.
på præp.	af præp.
til præp.	på præp.
fra præp.	at konj.
at konj.	end præp.
vækst sb.	for præp.
der pron.	nok adv.
med præp.	som præp.
stigning sb.	der pron.
er vb.	med præp.
som pron.	er vb.
for præp.	som pron.
kritik sb.	fra præp.

Figur 3: De 15 hyppigste højre naboer i K2000 til lemmaerne *kraftig* (t.v.) og *stærk* (t.h.)

denne art vil ofte lede opmærksomheden hen på kandidater for faste udtryk, idiomer e.l.

Der findes en hel række statistiske metoder, som kan detektere sproglige 49
strukturer, herunder grupper af signifikant samforekommende sproglige enheder. En af disse metoder er *mutual information*, som vi vil se nærmere på i det følgende. Mutual information er ikke nødvendigvis den mest velegnede metode til dette formål, men måske den, der er lettest at forstå.

Mutual information er et informationsteoretisk mål. Vi vil ikke gå nærmere ind på dets informationsteoretiske baggrund,[30] men blot nøjes med at nævne, at informationsteorien går ud på matematisk at beskrive håndtering af information i tekniske systemer, fx inden for telekommunikation og it.

Mutual information beror på den hypotetiske forventning, at alle ord eller andre relevante sproglige størrelser i et korpus kommer i vilkårlig, tilfældig rækkefølge, at der med andre ord ikke er tilbagevendende mønstre eller regelmæssigheder, at alt er jævnt og tilfældigt fordelt. Det rent hypotetiske udgangspunktet ('nul-hypotesen') er dermed det, at sproget slet ikke har no-

[30]Mere om denne findes fx i [Manning and Schütze, 1999].

gen struktur: vi (lader som om vi) forventer ganske enkelt ikke at finde noget interessant.

Lad os se nærmere på en sådan strukturløs "tekst". Lad os antage, at tekstens omfang i antal ord er n. Lad os endvidere antage, at antallet af forekomster af et bestemt ord w_1 er x. Hvis vi nu peger på et tilfældigt ord i denne tekst, er sandsynligheden p for, at det faktisk er en forekomst af w_1, vi peger på: 50–51

$$p(w_1) = \frac{x}{n} \tag{8}$$

Lad os endvidere antage, at der er et andet ord w_2 med y forekomster i hele teksten, så er sandsynligheden for tilfældigt at udpege dette w_2: 52

$$p(w_2) = \frac{y}{n} \tag{9}$$

Hvad er da sandsynligheden $p(w_1, w_2)$, at vi først tilfældigt udpeger et w_1 og i næste forsøg tilfældigt udpeger et w_2? I denne sammenhæng er det i øvrigt ligegyldigt, om vi i blinde peger én gang til på teksten eller blot rykker fingeren til det efterfølgende (eller forudgående ord) – tilfældigheden er den samme. Den er ifølge almindelig sandsynlighedsregning (jf. fx kast med to terninger): 53–54

$$p(w_1, w_2) = p(w_1) \cdot p(w_2) = \frac{x \cdot y}{n \cdot n} \tag{10}$$

Dette gælder principielt for enhver strukturløs "tekst".

2.2.4 Identifikation af kollokationer vha. mutual information

Lad os nu se på de faktiske forhold i teksten; det, vi faktisk kan observere. Vi tager samtlige bigrammer w_1, w_2 og bestemmer deres faktiske antal z i teksten. Sandsynligheden for, at vi i teksten tilfældigt udpeger ordparret w_1, w_2 er da: 56

$$p(w_1, w_2) = \frac{z}{n} \tag{11}$$

Næste skridt består i at sammenholde det *faktisk observerede* med vores hypotetiske forventning, som den kommer til udtryk i formel 11. Det gør vi ved at beregne kvotienten K mellem den faktisk observerede sandsynlighed for, at vi finder et w_1, w_2, O, med den antagede sansynlighed for at finde et sådant ordpar, E:

$$K = \frac{O}{E} \tag{12}$$

Hvad fortæller denne kvotient os? Lad os tage et ikke-sprogligt eksempel for at illustrere dette. Over en tiårsperiode har det vist sig, at 4 ud af 10 studerende dumpede til en bestemt eksamen; vi forventer derfor, at den statistiske

sandsynlighed for at dumpe til denne eksamen er $\frac{4}{10} = 0,4$ (eller 40%) næste gang, den afholdes. Vi kan også kalde denne forventning (E) hypotetisk, for vi aner jo ikke, hvordan det faktisk kommer til at gå næste gang. Ved næste eksamen viser det sig så, at 6 ud af 10 dumper: dette er den faktiske observation (O) ved denne konkrete eksamen. Kvotienten bliver i dette tilfælde:

$$K = \frac{O}{E} = \frac{0,6}{0,4} = 1,5 \tag{13}$$

Sandsynligheden for at dumpe til denne eksamen var altså 1,5 gange større end forventet på baggrund af de hidtige erfaringer. Hvis der omvendt kun havde været 2 ud af 10, der var dumpet til denne eksamen, havde kvotienten været:

$$K = \frac{O}{E} = \frac{0,2}{0,4} = 0,5 \tag{14}$$

– og sandsynligheden for at dumpe til denne eksamen dermed kun det halve af den forventede.

Tilbage til mutual information: hvis vi i $K = \frac{O}{E}$ indsætter de oprindelige sandsynlighedsudtryk i stedet for O og E, får vi:

$$K = \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)} \tag{15}$$

– og K fortæller os da, hvor mange gange oftere vi observerer ordparret w_1, w_2 i den faktiske tekst end i en fuldstændig strukturløs, der havde svaret til vores hypotetiske forventning. Dette er – næsten – mutual information. Da mutual information som nævnt er et informationsteoretisk mål og information kvantificeres i bits (antal af nuller eller et-taller), angiver man K i bits. Omregningen sker ved at tage logaritmen med grundtallet 2 af K. Denne omregning er ikke nødvendig i vores sammenhæng, men for en god ordens skyld skal vi her anføre den helt nøjagtige formel for mutual information (I):

$$I(w_1; w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)} \tag{16}$$

Figur 4 viser mutual information[31] anvendt til fremfinding af "typiske"[32] 57–58
højre naboer i K2000.

Eksemplet viser, at mutual information egner sig til fremfinding af fremtræ- 59

[31] I K2000-systemet anvendes mutual information faktisk på en lidt modificeret måde, der primært sørger for at lavfrekvente kollokater til en vis grad elimineres; mere om dette i [Asmussen, 2004b].

[32] Betegnelsen *typisk* er faktisk en fortolkning, ufortolket burde det hedde *overrepræsenteret i forhold til antaget strukturløshed*.

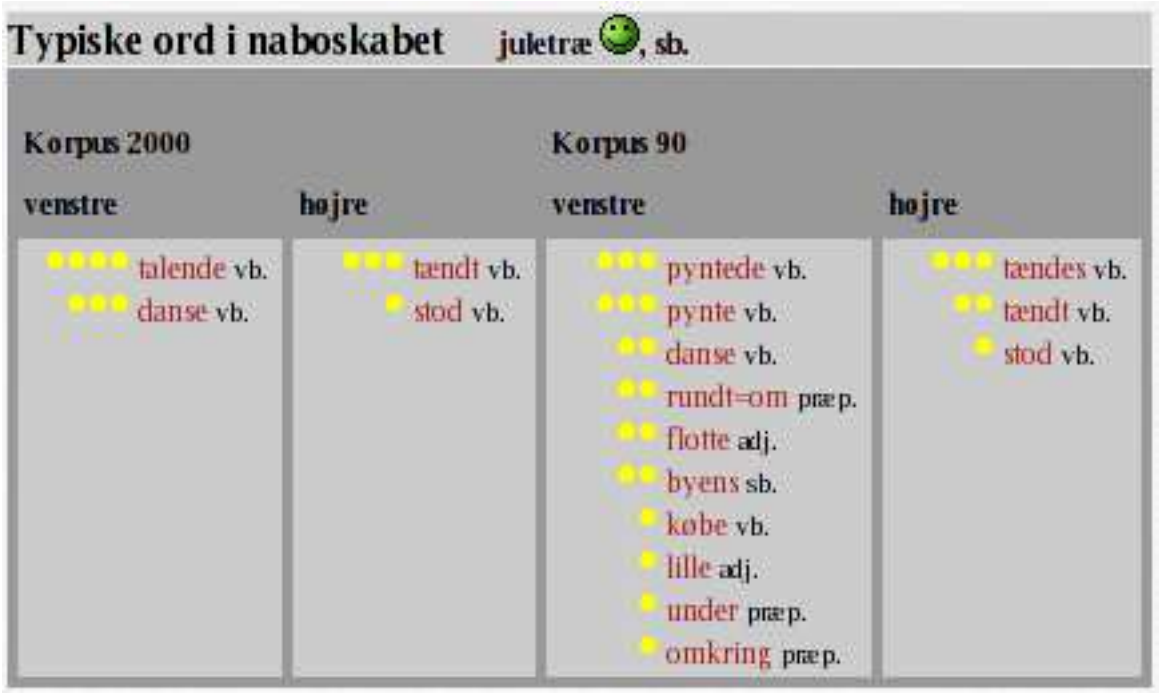
Figur 4: De mest fremtrædende højre naboer i K2000 til lemmaerne *kraftig* og *stærk*, fundet vha. mutual information

dende strukturer i et tekstmateriale. En ulempe ved mutual information kan være, at metoden giver lavfrekvente kollokationer høj vægt. Som det fremgår af figur 5 bliver *talende* bestemt som det mest fremtrædende ord til venstre for lemmaet *juletræ*, alene fordi *talende juletræ* optræder et antal gange i én tekst i K2000. Vil man eliminere den slags tilfældigheder, bør mutual information udvides med et mål, der kan tage højde for den undersøgte sproglige enheds jævne fordeling over hele korpus (dispersion).

3 Konklusion

Kvantitative korpusanalyser kan betjene sig af et væld af metoder, hvoraf der her kun blev demonstreret ganske få for at give et indtryk af, hvilke overvejelser der ligger til grund med hensyn til tekstmaterialets beskaffenhed og undersøgelsens formål. 61

Ved korpus- eller tekststatistiske undersøgelser er det således vigtigt, at



Figur 5: De mest fremtrædende naboer til lemmaet *juletræ*, fundet vha. mutual information

- udvælge tekstmaterialet kritisk og vurdere, hvorvidt det i det hele taget er egnet til undersøgelsen
- definere de tekstuelle enheder, der skal kvantificeres, på en entydig og konsekvent måde
- vælge statistiske metoder og tests ud fra en viden om, hvad de faktisk gør
- være varsom med generaliseringer for sproget som helhed
- holde resultaterne op imod introspektivt vundne: er der forskel, er der mindst lige så stor grund til at tvivle på ens empiriske som på ens introspektive evner.

A Appendiks: Transskriptioner af talesprog

A.1 Lineariseret nedskrift af talesprog fra DDO's korpus

```
<Tekst ID=HQUa>
<replik id=INT>
<p>din broder som bor her hva- hvad hedder han</p>
</replik>
<replik id=MD1>
<p>Steen</p>
</replik>
<replik id=INT>
<p>hn {pause} og det er ham der er syvogtyve?</p>
</replik>
<replik id=MD1>
<p>otteogtyve</p>
</replik>
<replik id=INT>
<p>otteogtyve {pause}</p>
</replik>
<replik id=MD1>
<p>jeg har også en anden broder der er femogtyve {pause}</p>
</replik>
<replik id=INT>
<p>nå</p>
</replik>
<replik id=MD1>
<p>men han er {tøven} han er en uge indenfor soldaterne
{pause} han er lige blevet ind for soldaterne, ikke?</p>
</replik>
<replik id=INT>
<p>hjnå</p>
</replik>
<replik id=MD1>
<p>altså {tøven} {uf} hvad hedder det {latter}</p>
</replik>
<replik id=MD2>
<p>{uf} genindkaldt</p>
</replik>
<replik id=MD1>
<p>genindkaldt</p>
</replik>
<replik id=INT>
<p>njå {pause}</p>
</replik>
<replik id=MD1>
<p>så til ham {tøven} kan vi ikke {uf} foreløbig så</p>
</replik>
<replik id=INT>
<p>hvor {tøven} i {tøven} Søværnet?</p>
</replik>
<replik id=MD1>
<p><usikker>næe det er lige</usikker> ovre på Holmen {pause}
i Søværnet {pause} lige der hvor jeg er {latter} {pause}</p>
</replik>
<replik id=INT>
<p>jaer. jaer. men plejer I at tænke på Holmen som lige
derovre {pause}</p>
</replik>
<replik id=MD1>
<p>{pause} jaer {pause}</p>
</replik>
<replik id=INT>
<p>hvor- hvor- hvordan kommer jeres fader på arbejde?</p>
</replik>
```

A.2 Partitur af talesprogsnedskrift fra webversionen af BySoc

```
1>             mm ff                                                S%%%%
2>
3>
A>tten B%%%%%%       din bror som bor her hva- hvad hedder han       mm f og det
K>
-------------------------------------------------------------------------------
1>                            otteogtyve                jeg har også en anden bro
2>
3>
A> er ham der er syvogtyve ?            otteogtyve fff
K>
-------------------------------------------------------------------------------
1>r der er femogtyve f    men han er~ ff han er en uge indenfor soldaterne f han
2>
3>
A>                     nå
K>
-------------------------------------------------------------------------------
1> er lige blevet ind for soldaterne ik'      altså~ f (fu) hvad hedder det (lat
2>
3>
A>                                    hjnå
K>
-------------------------------------------------------------------------------
1>ter)                  genindkaldt        så til ham~ kan vi ikke (uf) foreløbi
2>     (uf) genindkaldt
3>
A>                                 nå~ ff
K>
-------------------------------------------------------------------------------
1>g så                      [ næ det er lige ] ovre på H%%%%% f i S%%%%%%% f li
2>
3>
A>     hvor~ f i~ S%%%%%%%?                                ja~
K>
-------------------------------------------------------------------------------
1>ge der hvor jeg er (latter) ff
2>
3>
A>    ja~                        men plejer I at tænke på H%%%%% som lige derovr
K>
-------------------------------------------------------------------------------
1>     # ja~ f                                                  han~ tager cykl
2>
3>
A>e ff         hvor- hvor- hvordan kommer jeres far på arbejde ?
K>
-------------------------------------------------------------------------------
```

B Appendiks: Frekvensprofiler

B.1 De 30 hyppigste types i S1

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	at	175	3,18	175	3,18
2	i	138	2,50	313	5,68
3	er	125	2,27	438	7,95
4	af	124	2,25	562	10,20
5	for	97	1,76	659	11,96
6	og	92	1,67	751	13,63
7	til	91	1,65	842	15,28
8	en	84	1,52	926	16,80
9	der	79	1,43	1005	18,24
10	har	68	1,23	1073	19,47
11	det	66	1,20	1139	20,67
12	med	62	1,13	1201	21,79
13	de	59	1,07	1260	22,86
14	om	55	1,00	1315	23,86
15	kan	54	0,98	1369	24,84
16	som	54	0,98	1423	25,82
17	man	52	0,94	1475	26,76
18	ikke	49	0,89	1524	27,65
19	på	46	0,83	1570	28,49
20	den	44	0,80	1614	29,29
21	fra	39	0,71	1653	29,99
22	ved	37	0,67	1690	30,67
23	mikroorganismer	36	0,65	1726	31,32
24	et	33	0,60	1759	31,92
25	organismer	32	0,58	1791	32,50
26	I	29	0,53	1820	33,02
27	Det	28	0,51	1848	33,53
28	gensplejsede	28	0,51	1876	34,04
29	arbejde	23	0,42	1899	34,46
30	f.eks	22	0,40	1921	34,86

B.2 De 30 hyppigste types i S2

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	og	188	3,72	188	3,72
2	Petruska	133	2,64	321	6,36
3	i	131	2,60	452	8,96
4	sagde	112	2,22	564	11,17
5	er	102	2,02	666	13,20
6	det	96	1,90	762	15,10
7	ikke	95	1,88	857	16,98
8	mor	95	1,88	952	18,86
9	på	89	1,76	1041	20,63
10	Marie	88	1,74	1129	22,37
11	jeg	75	1,49	1204	23,86
12	så	70	1,39	1274	25,24
13	med	66	1,31	1340	26,55
14	hun	65	1,29	1405	27,84
15	du	63	1,25	1468	29,09
16	at	60	1,19	1528	30,28
17	var	55	1,09	1583	31,37
18	en	53	1,05	1636	32,42
19	de	51	1,01	1687	33,43
20	skal	44	0,87	1731	34,30
21	til	44	0,87	1775	35,17
22	Jeg	39	0,77	1814	35,94
23	af	38	0,75	1852	36,70
24	Det	38	0,75	1890	37,45
25	Og	36	0,71	1926	38,16
26	der	35	0,69	1961	38,85
27	den	34	0,67	1995	39,53
28	laksko	33	0,65	2028	40,18
29	kan	31	0,61	2059	40,80
30	dem	30	0,59	2089	41,39

B.3 De 30 hyppigste types i T1

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	det	270	4,80	270	4,80
2	ikke	215	3,82	485	8,63
3	der	174	3,10	659	11,72
4	er	158	2,81	817	14,53
5	så	144	2,56	961	17,10
6	jeg	135	2,40	1096	19,50
7	og	134	2,38	1230	21,88
8	har	121	2,15	1351	24,03
9	var	114	2,03	1465	26,06
10	vi	84	1,49	1549	27,56
11	de	81	1,44	1630	29,00
12	sådan	76	1,35	1706	30,35
13	en	75	1,33	1781	31,68
14	i	74	1,32	1855	33,00
15	jaer	65	1,16	1920	34,16
16	men	65	1,16	1985	35,31
17	altså	59	1,05	2044	36,36
18	på	59	1,05	2103	37,41
19	du	57	1,01	2160	38,43
20	hun	56	1,00	2216	39,42
21	til	54	0,96	2270	40,38
22	at	52	0,93	2322	41,31
23	noget	45	0,80	2367	42,11
24	den	43	0,76	2410	42,87
25	han	39	0,69	2449	43,57
26	hvad	39	0,69	2488	44,26
27	også	38	0,68	2526	44,94
28	havde	36	0,64	2562	45,58
29	da	35	0,62	2597	46,20
30	fordi	35	0,62	2632	46,82

B.4 De 30 hyppigste types i T2

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	og	238	4,43	238	4,43
2	det	182	3,39	420	7,82
3	jeg	178	3,32	598	11,14
4	var	169	3,15	767	14,29
5	så	149	2,78	916	17,06
6	der	124	2,31	1040	19,37
7	han	107	1,99	1147	21,36
8	ikke	103	1,92	1250	23,28
9	i	93	1,73	1343	25,01
10	jo	93	1,73	1436	26,75
11	Ja	78	1,45	1514	28,20
12	havde	68	1,27	1582	29,47
13	en	66	1,23	1648	30,69
14	de	65	1,21	1713	31,91
15	den	65	1,21	1778	33,12
16	til	64	1,19	1842	34,31
17	da	62	1,15	1904	35,46
18	vi	60	1,12	1964	36,58
19	med	59	1,10	2023	37,68
20	for	55	1,02	2078	38,70
21	at	54	1,01	2132	39,71
22	sådan	53	0,99	2185	40,70
23	på	48	0,89	2233	41,59
24	er	47	0,88	2280	42,47
25	du	46	0,86	2326	43,32
26	hun	45	0,84	2371	44,16
27	kan	45	0,84	2416	45,00
28	også	45	0,84	2461	45,84
29	men	43	0,80	2504	46,64
30	af	33	0,61	2537	47,25

B.5 De 30 hyppigste types i K90

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	og	869097	3,00	869097	3,00
2	i	798346	2,76	1667443	5,76
3	at	697824	2,41	2365267	8,17
4	er	551448	1,90	2916715	10,08
5	en	461363	1,59	3378078	11,67
6	det	459195	1,59	3837273	13,26
7	til	438550	1,51	4275823	14,77
8	af	418063	1,44	4693886	16,22
9	på	403009	1,39	5096895	17,61
10	med	349058	1,21	5445953	18,81
11	for	341448	1,18	5787401	19,99
12	der	331042	1,14	6118443	21,14
13	den	316095	1,09	6434538	22,23
14	de	295784	1,02	6730322	23,25
15	ikke	289993	1,00	7020315	24,25
16	som	276337	0,95	7296652	25,21
17	har	250419	0,87	7547071	26,07
18	var	228441	0,79	7775512	26,86
19	et	215132	0,74	7990644	27,60
20	om	201698	0,70	8192342	28,30
21	så	169439	0,59	8361781	28,89
22	Det	165856	0,57	8527637	29,46
23	han	162726	0,56	8690363	30,02
24	kan	156812	0,54	8847175	30,56
25	sig	151320	0,52	8998495	31,09
26	jeg	150918	0,52	9149413	31,61
27	fra	134389	0,46	9283802	32,07
28	vi	113174	0,39	9396976	32,46
29	man	113090	0,39	9510066	32,85
30	men	107076	0,37	9617142	33,22

B.6 De 30 hyppigste types i K2000

rang	type	$f_{absolut}$	f_{pct}	$\sum f_{absolut}$	$\sum f_{pct}$
1	og	819070	2,86	819070	2,86
2	at	812410	2,83	1631480	5,69
3	i	765861	2,67	2397341	8,37
4	er	617041	2,15	3014382	10,52
5	en	480728	1,68	3495110	12,20
6	det	448365	1,56	3943475	13,76
7	på	423419	1,48	4366894	15,24
8	til	423392	1,48	4790286	16,71
9	af	397390	1,39	5187676	18,10
10	for	343001	1,20	5530677	19,30
11	der	339838	1,19	5870515	20,48
12	med	333565	1,16	6204080	21,65
13	den	316447	1,10	6520527	22,75
14	de	314822	1,10	6835349	23,85
15	har	294858	1,03	7130207	24,88
16	som	280009	0,98	7410216	25,86
17	ikke	261957	0,91	7672173	26,77
18	et	227398	0,79	7899571	27,56
19	om	197127	0,69	8096698	28,25
20	var	176829	0,62	8273527	28,87
21	kan	157893	0,55	8431420	29,42
22	Det	156627	0,55	8588047	29,97
23	fra	140876	0,49	8728923	30,46
24	sig	136396	0,48	8865319	30,93
25	han	135119	0,47	9000438	31,41
26	så	130888	0,46	9131326	31,86
27	jeg	124235	0,43	9255561	32,30
28	vi	119385	0,42	9374946	32,71
29	man	118223	0,41	9493169	33,12
30	skal	109610	0,38	9602779	33,51

C Appendiks: Log-likelihood-undersøgelser

C.1 Ord, der adskiller S1 og T1 mest signifikant fra hinanden

rang	type	≫S1	≫T1	L
1	jeg		+	174,1
2	så		+	145,4
3	det		+	128,9
4	ikke		+	109,4
5	var		+	100,3
6	sådan		+	82,6
7	jaer		+	81,9
8	af	+		78,3
9	at	+		72,8
10	altså		+	71,9
11	du		+	71,2
12	hun		+	69,9
13	vi		+	65,1
14	for	+		56,7
15	mikroorganismer	+		44,8
16	organismer	+		39,4
17	noget		+	36,0
18	der		+	34,7
19	men		+	34,7
20	gensplejsede	+		34,1
21	her		+	33,6
22	Det	+		32,0
23	naej		+	30,4
24	som	+		29,7
25	han		+	29,2
26	om	+		28,7
27	hedder		+	27,8
28	f.eks	+		26,1
29	havde		+	25,8
30	arbejde	+		25,4

C.2 Ord, der adskiller S1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S1	L
1	mikroorganismer		+	295,2
2	f.eks		+	229,6
3	organismer		+	227,2
4	1987		+	199,8
5	gensplejsede		+	197,2
6	udsætning		+	177,5
7	miljøstyrelse		+	132,5
8	1		+	110,7
9	EPA		+	98,8
10	Marcker		+	95,4
11	Bacillus		+	85,4
12	klasse		+	83,3
13	bakterie		+	77,4
14	K12-kolibakterien		+	71,5
15	genteknologi		+	70,4
16	10		+	70,2
17	2		+	70,2
18	forsøg		+	68,3
19	3		+	68,0
20	Kjeld		+	66,9
21	subtilis		+	66,2
22	1986		+	65,5
23	K12		+	65,5
24	anvendes		+	64,6
25	bakterier		+	62,7
26	organismen		+	60,4
27	retningslinier		+	59,1
28	amerikanske		+	58,5
29	NAS-rapporten		+	57,2
30	sygdomsorganismer		+	55,9

C.3 Ord, der adskiller S2 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫S2	L
1	Petruska		+	1932,9
2	Marie		+	1022,0
3	mor		+	557,1
4	sagde		+	454,1
5	laksko		+	452,6
6	lakskoene		+	273,6
7	Petruskas		+	173,6
8	Mor		+	168,8
9	du		+	120,7
10	børnehave		+	105,8
11	osse		+	92,7
12	hun		+	89,3
13	far		+	80,0
14	gaver		+	77,4
15	børnehaven		+	77,0
16	Maries		+	76,1
17	drillenissen		+	57,9
18	dig		+	57,9
19	jeg		+	56,5
20	lyserøde		+	56,3
21	vågnede		+	55,1
22	Bvadr		+	53,7
23	spurgte		+	48,3
24	så		+	46,9
25	numsen		+	46,3
26	spøgelset		+	45,1
27	sko		+	44,0
28	chokoladefrøer		+	43,4
29	løb		+	42,8
30	at	+		42,5

C.4 Ord, der adskiller T1 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫T1	L
1	jaer		+	1096,4
2	naej		+	436,2
3	sådan		+	342,7
4	ikke		+	272,2
5	altså		+	256,8
6	det		+	230,8
7	så		+	222,0
8	njå		+	199,3
9	hn		+	197,3
10	jeg		+	194,0
11	Nyboder		+	189,3
12	najåer		+	182,5
13	njaer		+	182,5
14	jamen		+	171,3
15	njaåer		+	165,7
16	njåer		+	165,7
17	jae		+	163,6
18	najå		+	148,8
19	der		+	128,5
20	hedder		+	105,9
21	fru		+	104,9
22	derovre		+	101,1
23	nåer		+	98,6
24	vi		+	97,7
25	du		+	89,9
26	hnn		+	81,9
27	Suensonsgade		+	81,9
28	Tyttebær-Maja		+	81,9
29	hvad		+	81,7
30	moder		+	80,3

C.5 Ord, der adskiller T2 og K90 mest signifikant fra hinanden

rang	type	≫K90	≫T2	L
1	Ja		+	435,2
2	jeg		+	348,5
3	jo		+	341,6
4	så		+	248,7
5	sådan		+	207,3
6	var		+	205,9
7	Rodskov		+	149,7
8	ja		+	132,7
9	da		+	116,4
10	hm		+	112,5
11	han		+	112,1
12	sæbe		+	95,6
13	far		+	93,4
14	huske		+	85,1
15	det		+	79,7
16	havde		+	78,6
17	dernede		+	71,2
18	Tindrup		+	65,7
19	bedstemor		+	64,2
20	at	+		61,4
21	du		+	60,5
22	flyttelæsset		+	56,1
23	oppe		+	54,4
24	som		+	51,5
25	bedstefar		+	50,9
26	henne		+	50,9
27	der		+	49,9
28	nede		+	49,6
29	gangbrættet		+	49,2
30	persillesovs		+	48,0

C.6 Log-likelihood-sammenligning af K2000 og K90

rang	type	≫K2000	≫K90	L
1	EU	+		11126
2	du		+	10887
3	$	+		10076
4	var		+	7849
5	hun		+	6731
6	mio.	+		6612
7	at	+		6379
8	havde		+	6004
9	$1999	+		5788
10	$1998	+		5698
11	ham		+	5299
12	siger	+		5161
13	$1997	+		4909
14	EF		+	4394
15	hende		+	4227
16	%	+	+	4146
17	han		+	3251
18	jeg		+	3243
19	NN		+	3075
20	mill.		+	3027
21	Du		+	2938
22	$1996	+		2877
23	dig		+	2876
24	danske	+		2847
25	$2000	+		2821
26	pct.	+		2789
27	direktør	+		2742
28	Hun		+	2689
29	har	+		2663
30	Internettet	+		2642
31	mig		+	2601
32	EUs	+		2514
33	$2001	+		2481
34	Dansk=Folkeparti	+		2474
35	Ytring	-	+	2448

Litteratur

[Asmussen, 2004a] Asmussen, J. (2004a). Korpus 2000 – til hvilken nytte? Muligheder og grænser for empiriske sprogundersøgelser. In Duncker, D., editor, *Studier i Nordisk 2002-2003*, Copenhagen/København. Selskab for Nordisk Filologi.

[Asmussen, 2004b] Asmussen, J. (under udgivelse 2004b). Towards a methodology for corpus-based studies of linguistic change. Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 Corpora of Danish. In Archer, D., Rayson, P., and Wilson, editors, *Corpus Linguistics Around the World*. Rodopi, Amsterdam.

[Bergenholtz, 1998] Bergenholtz, H. (1998). Deskriptiv, proskriptiv og præskriptiv leksikografi. In Fjeld, R. V. and Wangensteen, B., editors, *Normer og Regler. Festskrift til Dag Gundersen 15. januar 1998*, pages 233–246.

[Bick, 2003a] Bick, E. (2003a). A CG & PSG hybrid approach to automatic corpus annotation. In *Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003) held in conjunction with the Corpus Linguistics 2003 Conference. UCREL technical paper no. 17*, Lancaster. UCREL, Lancaster University.

[Bick, 2003b] Bick, E. (2003b). Morfosyntaktisk opmærkede corpora for dansk. In *9. Møde om Udforskningen af Dansk Sprog 10.–11. oktober 2002*, Århus. Aarhus Universitet.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, Massachusetts.

[Chomsky, 1986] Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger, New York.

[Church et al., 1991] Church, K. et al. (1991). Using Statistics in Lexical Analysis. In Zernik, U., editor, *Lexical Acquisitition. Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, New Jersey.

[Church and Hanks, 1989] Church, K. and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting*, Vancouver.

[Clegg, 1990] Clegg, F. (1990). *Simple Statistics. A course book for the social sciences*. Cambridge.

[COBUILD: Sinclair et al., 1987] COBUILD: Sinclair, J. et al., editors (1987). *Collins COBUILD English Language Dictionary*. Collins.

[Daille, 1995] Daille, B. (1995). Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical Report 5, Lancaster.

[DDO: Hjorth et al., 2005] DDO: Hjorth, E., Kristensen, K., Lorentzen, H., Trap-Jensen, L., Asmussen, J., et al., editors (2003-2005). *Den Danske Ordbog 1-6*. DSL & Gyldendal, Copenhagen/København.

[Garside and Rayson, 2000] Garside, R. and Rayson, P. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6, Hong Kong.

[Halliday, 1991] Halliday (1991). Corpus studies and probabilistic grammar. In Aijmer and Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman, London.

[Kilgarriff, 2001] Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

[Leech, 1991] Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer and Altenberg, editors, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman, London.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 2nd (2000) edition.

[Norling-Christensen and Asmussen, 1998] Norling-Christensen, O. and Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos. Afrilex Series*, 8:223–242.

[Sinclair, 1991] Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.