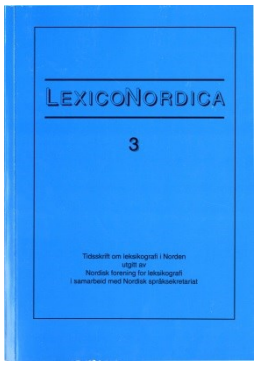
LexicoNordica

Titel:	Genbrug af korpora
Forfatter:	Henrik Holmboe
Kilde:	LexicoNordica 3, 1996, s. 49-57
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



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 LexicoNordica (1-16) 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.

Henrik Holmboe

Genbrug af korpora

The word *corpus* has at least two meanings of interest to us; an older meaning: "A body or complete collection of writings", and a more recent one: "A limited collection of written or spoken material upon which a linguistic analysis is based". In this paper we show that the older meaning is approximately 2000 years old, while the more recent meaning is hardly 50 years old. Corpus in the old sense and corpus in the recent sense of the word establish two quite different scientific scenarios as for use and reuse. We assume that the concept of a text bank as a repository of texts in machine readable form will gain importance; from here the user can select and compose corpora of his own for specific and focused linguistic purposes. Such texts being more widespread and more easily accessible will give rise to the demand for precise declarations of corpora and for standardisation of corpora and of the electronic software tools used for the linguistic analysis of text corpora.

Betydninger af ordet *korpus*

Ordet *korpus* havde som filologisk og sprogvidenskabeligt fagudtryk én betydning, førend først og fremmest den datamatiske lingvistik og forskellige andre datamatstøttede filologiske og sprogvidenskabelige studier medførte mindst én ny betydning. Det er i regelen klart ud fra tradition og situationel kontekst, hvilken af de mulige betydninger af korpus der er den på stedet aktuelle, men lad os ikke glemme, at der er flere betydninger. Den eller de nye betydninger af ordet har ikke fortrængt den ældre betydning, men eksisterer ved siden af den. Ligeledes er det sandt, at der er lingvister og filologer – og jeg regner mig selv heriblandt – der arbejder med forskellige opgaver, hvor snart den ældre betydning, snart en af de yngre betydninger er relevant.

Den ældre betydning

Hvis vi går til de store danske ordbøger, finder vi kun sparsom hjælp til en fastlæggelse af betydning og betydningsudvikling. Videnskabernes selskabs ordbog (VSO) har ikke ordet, hverken under **corpus** i bindet fra 1793 eller under **korpus** fra 1820. ODS s.v. corpus henviser til korpus, som vi finder i pågældende bind, bind XI, som er fra 1929: den fjerde af de anførte betydninger er med "kors", der står for "nu

ubrugeligt": *† samlingsværk, samlingsbind*, med følgende eksempel: "alle de Forordninger, som fra Tid til anden ere udkomne .. skal sankes tilsammen, og bringes i et Corpus." Eksemplet stammer fra Kiøbenhavnske nye Tidender om lærde Sager fra 1742. ODS-supplementets seddelsamlinger rummer også et enkelt Holberg-citat, men ikke noget fra 19. årh. Ordet er naturligvis et lærd og filologisk ord, og når der ikke bringes flere belæg i ODS, og når danske ordbøger fra 19. århundrede ikke har denne brug med, kunne det hænge sammen med, at folk, der professionelt brugte ordet, udtrykte sig på latin eller i al fald ikke på dansk. Og dette fører mig til også at nævne den femte og sidste betydning i ODS, nemlig *korpus* som navn på en skriftstørrelse, således kaldet, fordi den først anvendtes ved trykningen af den romerske lovsamling Corpus iuris civilis fra 6. årh. i 1583.

På klassisk latin kender vi ordet *corpus* i bl. a. betydningen – og jeg henholder mig til Lewis & Short (1958) – "a whole composed of parts united", herunder om bøger; i sit skrift om grammatikere bruger Sueton (75–150 e.Kr.) *corpus* om en bestemt udgivelse af ni volumina og begrunder det udtrykkeligt med, at disse ni hørte sammen. Også Cicero (106–43 f.Kr.) og Livius (59 f.Kr–17 e.Kr.) kan citeres for belæg, og her fra sammenhænge, der ikke er grammatiske traktater, man kunne næsten fristes til at sige sammenhænge fra det almindelige sprog, altså ikke-LSP. Da corpus på latin i den specielle betydning, vi efterlyser her, er et lærd ord, er det nærliggende at undersøge, om det skulle være et oversættelseslån fra græsk, kort sagt, kan det græske ord *sʏma* [så:ma] bruges i samme betydning? Liddell & Scott (1953) angiver "a body of writings" med kun ét belæg, nemlig et Cicerobrev (ad Att. II,1,3). Hvis dette virkelig er ældste belæg, må det betyde, at opfattelsen af en komplet samling af skrifter om et bestemt emne som et begreb er en romersk nydannelse; hvad har man kaldt det i Alexandria eller i Athen? måske ingenting; det ville være konsekvent, hvis begrebet først stammer fra romersk tid.

Den yngre betydning

Ser vi ordet **Korpus** efter i ordbøger for nogle moderne sprog, finder vi en ret god overensstemmelse fra sprog til sprog.

Begyndende med tysk konstaterer vi, at Trübner (1939–1957) slet ikke har ordet med; hos Klappenbach (1973) finder vi den latinske betydning: "Gesamtheit von Texten, Schriften: das K. der altdeutschen Urkunden." og tilsvarende i Dudens Bedeutungswörterbuch (Duden 1970), der i udgaven fra 1970 har: "vollständiges gedrucktes Werk, in

dem Urkunden, Gesetze o. ä. gesammelt sind (meist die Antike oder das Mittelalter betreffend)." Bemærkelsesværdigt er det, at Dudens Das große Wörterbuch der deutschen Sprache i 6 bind fra 1978 (Duden 1978) noterer: "Sammlung einer begrenzten Anzahl von Texten, Äußerungen o. ä. als Grundlage für sprachwissenschaftliche Untersuchungen." og anfører et citat fra 1974.

Opslag i Le Robert électronique, 1989 giver tilsvarende: "Recueil de pièces, de documents concernant une même discipline." og derefter som betydning to med et belæg fra 1961: "Ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique."

I den seneste udgave af The Oxford English Dictionary. Second Edition on Compact Disc 1992 finder vi også betydningen "A body or complete collection of writings or the like", og så tidligt som fra 1956 også betydningen: "The body of written or spoken material upon which a linguistic analysis is based."

	Ældre betydning	Yngre betydning	Første belæg
Klappenbach: Wörterbuch der deutschen Gegenwartssprache, 2. Aufl. 1973	Gesamtheit von Texten, Schriften		
Duden: Das große Wörterbuch der deutschen Sprache, 1978		Sammlung einer begrenzten Anzahl von Texten, Äußerungen o.ä. als Grundlage für sprachwissenschaftliche Untersuchungen	1974
Le Robert életronique, 1989	Recueil de pièces, de documents concernant une même discipline	Ensemble limité des éléments (énoncés), sur lesquels se base l'étude d'un phénomène linguistique	1961
The Oxford English Dictionary. Second Edition on Compact Disc 1992	A body or complete collection of writings or the like	The body of written or spoken material upon which a linguistic analysis is based	1956

Denne betydning af samling af et begrænset sprogligt materiale som grundlag for sprogvidenskabelige undersøgelser er den nye betydning af *korpus*, som i modsætning til den gamle ikke stiller forventning om totalitet; den gamle betydning ekspliciterer til gengæld intet om et fokuseret formål; formålet er blot at have alle dokumenter af en speciel kategori samlet på en minimalt systematisk måde. Belægget for den nye betydning kan som anført for tysks vedkommende dateres til begyndelsen af 70-erne, og noget tilsvarende – måske lidt senere – vil formentlig være muligt for dansks vedkommende. Riber Petersen

(1984) noterer dog ikke ordet. Dansk Sprognævns seddelsamlinger har først et belæg fra 1990, og det samme gælder Den danske Ordbogs materiale. Blinkenberg/Høybye (1975) har betydningen i 3. udg. af dansk-fransk ordbog fra 1975 både under **corpus** og **korpus** (Tekstudvalg for filologiske undersøgelser); men om dette beror på observationer af dansk eller på en tilbageoversættelse af det franske *corpus*, er nok usikkert; (jeg vil selv vove den påstand, at der er tale om en tilbageoversættelse.) Blinkenberg/Høybye (1984) har denne betydning med i fransk-danske ordbog fra 1984, 2. reviderede og forøgede udgave; men ikke i første udgave fra 1964. I norsk er betydningen registreret fra 1978, og i svensk fra 1986.

Brug og genbrug af korpora

Når vi taler om brug og genbrug af korpora, må vi gøre os klart, at der er tale om to forskellige ting, alt efter om vi mener korpora i den ældre eller i den nyere betydning af ordet.

Først om den ældre betydning: Her er korpus en totalitet af et materiale. Sammensætningen af et sådant korpus er dikteret af, at man er færdig med at sammensætte sit korpus, når der ikke er mere materiale. Eksempler herpå er velkendte. Det danske runekorpus er alle danske runeindskrifter, Søren Kierkegaard-korpus eller H.C. Andersen-korpus er det samlede værk, og tilsvarende for udenlandske forhold: Shakespeare-korpus og Corpus inscriptionum latinarum er totaliteten. Dette giver en videnskabelig sikkerhed, idet indsamlingen af materiale og udnyttelsen af materialet i forskellige undersøgelser er principielt adskilte. Der er ikke tale om, at det pågældende korpus er sammensat med henblik på en eller flere på forhånd besluttede eller forudsete undersøgelser. Formålet med indsamlingen af et korpus kunne således være rent musealt. Dette betyder ikke, at korpusset ikke er etableret videnskabeligt eller uden axiomatik, men etableringsfasen er selvberoende og vil som sit fornemste formål have at leve op til forventningen om først totalitet og derefter autenticitet. Sekundære hensyn, bl.a. til anvendelighed og søgbarhed, kan tilsige, at man afviger fra autenticiteten. Man kan således vælge på denne måde at få et større antal tokens subsumeret under kun én type. Som konkret eksempel kan jeg anføre mit tidligere arbejde med de danske runeindskrifter. I en vis fase af korpusetableringsprocessen traf vi det valg at opretholde den autentiske forskel på 16-tegns og 24-tegns futharken, i en anden fase valgte vi at suspendere forskellen og anvende en form for standardrunenordisk, hvilket i denne fase betød, at tokens fra de to

skriftsystemer kunne sammenfattes under én og samme type. Et andet valg væk fra det autentiske var brugen af en standardlakune til gengivelse af en lakune af en hvilken som helst størrelse. Afgørende er det imidlertid, at alle senere brugere af sådanne korpora ikke er præjudicerede i deres undersøgelser og i deres undersøgelsesmetoders tilhørsforhold til den ene eller anden sprogvidenskabelige skole. Brugeren må selvfølgelig affinde sig med det autenticitetsniveau, der er lagt i korpus, men derefter står brugeren helt frit. Også til at foretage udvalg af tekster fra dette korpus for at etablere et udgangspunkt for fokuserede lingvistiske undersøgelser. Det er almindelig sprogbrug at tale om at bruge et sådant korpus, og at bruge det igen og igen; jeg er meget i tvivl om, hvorvidt man kan tale om at genbruge et sådant korpus.

Den nye betydning af ordet *korpus* etablerer en helt anden laboratoriesituation. Sigtet er her at foretage en lingvistisk undersøgelse for at nå et eller andet resultat. Udgangspunktet er ønsket om at opnå en eller anden indsigt, og med dette for øje tilrettelægger man etableringen af et korpus. Sigtet, hensigtsmæssigheden styrer, bestemmer næsten på forhånd, hvad der er støj, og hvad der er signal, og dette kan være såre godt og legitimt og videnskabeligt velbegrundet; man sørger for så at sige at rense præparatet, inden man undersøger det. Man kan sikre, at den samme overordnede hypotese er til stede både i etableringen af eksaminand og i eksaminationen. Men denne tæthed medfører også en vis lukkethed. Et sådant korpus kan naturligvis helt uproblematisk genbruges, hvis genbrugeren deler undersøgelsesformål og overordnet hypotese med det projekt, der oprindelig tilvejebragte det pågældende korpus, men i det omfang, dette ikke er tilfældet, vil korpusset være mindre velegnet til det ændrede formål. Den nye bruger kan så vælge at revidere korpusset, tilføje nye tekster og udelade andre mindre relevante, og dermed i realiteten etablere et nyt korpus.

Korpora og tekstbanker

På denne måde kan allerede etablerede korpora, både i den ældre og den yngre betydning, komme til at fungere som tekstbanker, hvorfra brugere kan forsyne sig, når de sammensætter egne korpora til fokuserede formål. Jeg forestiller mig, at netop denne adfærd vil blive mere og mere udbredt, efterhånden som flere og flere enkelttekster bliver tilgængelige i digitaliseret form. Når vi i de forløbne 20–30 år har oplevet, at enkelte korpora i den nye betydning har fået en vis status, hænger det nok bl. a. sammen med, at elektroniske korpora var

nye, sjældne og dermed uprøvede; selve fremkomsten af dem trak disciple til; den eller de, der havde sammensat disse korpora, blev med det samme anset for også at være eksperter i at udnytte dem – og også i stand til at inspirere og lære andre kunsten; der groede videnskabelige miljøer ud af disse korpora under de rette omstændigheder. Naturligvis har jeg Sture Allén og hans miljø i Göteborg i taknemmelig erindring.

Men som før anført, adgangen til korpusmateriale er blevet meget lettere, adgang til forholdsvis simple, men effektive værktøjer til en vis analyse af store tekstmængder er tilsvarende blevet lettere, og mange flere benytter sig i dag af det i det daglige arbejde. I dag er efterslæbet ikke manglende adgang til maskinlæsbare tekster eller værktøjer, men derimod en manglende viden om teorier og metoder inden for kvantitativ lingvistik og sammenhængen mellem kvantitative og kvalitative problemstillinger.

Deklaration af korpustekster

Ikke mindst i dette perspektiv er det på sin plads at overveje, hvilke mindstekrav til varedeklaration af korpustekster man kunne ønske sig. Som regel rummer et korpus oplysning om, hvad det indeholder, og det er jo godt. Ofte savner man derimod oplysning om autenticitets-niveauet. Oplysninger om repræsentationsform, anvendte tegnsæt og koder kan man undertiden finde, og lang erfaring har belært mig om, at de sjældent står til troende. Og dette betyder ganske konkret, at mindre befarne kan komme til på grundlag af eksisterende elektroniske tekster fra forskellige kilder evt. suppleret med egne bidrag at etablere egne korpora, der ganske vist har en præcis indholdsoversigt, men som ikke har et ensartet autenticitetsniveau, og som rummer en blanding af forskellige repræsentationsformer. Det er let at påvise, at mange senere analyser vil være upræcise fra fødselen, hvis autenticitet og repræsentationsform er inegale. Jeg vil derfor gerne genopfriske nogle gamle filologiske dyder og formulere dem som tre gyldne regler for opstilling af korpora af enhver art, nemlig vedr.

- indhold
- autenticitet
- repræsentationsform

disse tre skal være veloplyste. Er de ikke det, vil det nedsætte værdien af deriverede undersøgelsesresultater, også selv om disse ikke sættes ind i nogen sammenhæng uden for selve det aktuelle projekt.

Standarder for etablering af korpora

Et korpus rummer lingvistisk viden om sig selv, og denne viden kan vi analysere frem på forskellig måde. Denne korpusinterne viden skal kunne ekspliciteres på en reproducerbar måde, og alene derfor skal de gyldne regler gerne være overholdt. Imidlertid vil et korpus også kunne karakteriseres relativt i forhold til en række fikspunkter, der ligger uden for korpusset selv. Denne viden kommer kun frem, hvis vi kan referere til en veletableret viden uden for vort korpus, og en sådan reference vinder betragteligt i værdi, hvis mål og vægt i korpusset korresponderer med det, der er anvendt i de korpuseksterne fikspunkter. Som eksempel kan jeg nævne det velkendte, at en tekst består af en lille mængde meget højfrekvente ord og en stor mængde lavfrekvente ord. Detaljerne herom kan teksten selv meddele efter en hensigtsmæssig analyse. Det er også rigtigt, at de moderat frekvente og lavfrekvente ord kan anskues på en skala, der i midten har de stabile ord og på fløjene henholdsvis neologismer og obsolete ord, ord i knop og ord, der visner. Men ingen tekst kan i isolation meddele denne information om sig selv, ingen analysemetode kan levere dette.

Hensynet til ønsket om at kunne sammenligne analyseresultater fra forskellige korpora har formuleret et ønske om standarder, og helst internationale, for etablering af korpora. At argumentere imod standarder som sådanne lader sig næppe gøre og skal ejheller forsøges her. Men det må være på sin plads at kræve, at områdets standarder bliver detaljerede og veldokumenterede for at kunne tage hensyn til den mangfoldighed af fænomener, som vil vise sig i forskellige korpora. Og erfaring viser, at detaljerede standarder er svære at opfylde og overholde. Er standarderne ikke detaljerede, risikerer man, at en overholdelse af standarder medfører en lavere grad af autenticitet; og i så fald bliver standarder et tvivlsomt gode.

Standardisering af softwareværktøjer til korpusanalyse

En anden bestræbelse med samme overordnede sigte vil være lettere at gennemføre, nemlig ønsket om, at korpora skal analyseres med værktøjer, der opfylder visse standarder. På dette felt er man allerede godt igang. Ønsket om standardiserede informationsteknologiske værktøjer er nævnt i den danske regerings Handlingsplan som udtrykt i forskningsministerens Redegørelse til Folketinget om "Info-samfundet år 2000" og IT-politisk handlingsplan 1995. (Fra vision til handling). Tilsvarende er der i EU's telematikprogram under Language Engi-

neering reserveret ressourcer til udvikling af og udbredelse af softwareværktøjer til lingvistiske formål, herunder også behandling af korpora. Parallelt hermed igangsættes et opklaringsarbejde og en evalueringsproces netop med det formål at identificere eventuelle de facto standarder og om muligt få sammensat et repertoire af værktøjer, som kan anbefales som standardværktøjer og måske få endelig status af at være officielle standarder på området. Standardværktøjerne skal opfylde en række specifikationer, og der er hermed mulighed for, at den enkelte bruger til specielle formål kan udvikle egne værktøjer, der som udgangspunkt opfylder disse specifikationer og dermed kan være certificerede inden for et vist område samtidig med, at de opfylder en række specielle formål, som måske endnu ikke er omfattet af nogen standard.

I forbindelse med standarder for og standardisering af korpora har jeg udtrykt et vist forbehold med hensyn til store forventninger på kort sigt. Med andre ord: På dette punkt er der meget lang vej igen. Jeg har slet ikke et sådant forbehold med hensyn til standardisering af elektroniske analyseværktøjer. Årsagen er ikke bare, at vi her kan imødese en succes på kort sigt, men en grundlæggende opfattelse af, at det er videnskabeligt korrekt at standardisere måleredskaberne snarere end patienten. En del arbejde er allerede i gang, alle er tilsyneladende vel tilfredse med konceptet om at tilvejebringe standardiserede elektroniske analyseværktøjer, og argumenterne imod er næppe til at få øje på. Arbejdet er ikke let eller trivielt, men jeg må anse det for at være et overordentligt vigtigt sprogvidenskabeligt anliggende, at analyserne af den stadigt stigende mængde af elektronisk tilgængelige tekster resulterer ikke bare i output, men i viden, ny viden om sprogene og om sproget.

Litteraturliste

Andersen, Ingeborg/Henrik Holmboe 1983: *Konkordans over de danske runeindskrifter – Transskription.* Sprog Og Mennesker 5. Århus.

Andersen, Ingeborg/Henrik Holmboe 1983: *Konkordans over de danske runeindskrifter – Translitteration.* Sprog Og Mennesker 6. Århus.

Blinkenberg, A/Høybye, P. 1975: *Dansk-fransk ordbog*. 3. rev. og forøg. udg. København.

Blinkenberg, A/Høybye, P. 1984: *Fransk -dansk ordbog*. 2. rev. og forøg. udg. København.

Duden 1970: *Bedeutungswörterbuch.* Mannheim.
Duden 1978: *Das große Wörterbuch der deutschen Sprache.* Mannheim.
Fra vision til handling. Info-samfundet år 2000. Forskningsministerens Redegørelse til Folketinget om "Info-samfundet år 2000" og IT-politisk handlingsplan 1995. København.
Klappenbach, Ruth/Steinitz, W. 1973: *Wörterbuch der deutschen Gegenwartssprache.* 2. durchges. Aufl. Berlin.
Lewis, Ch. T. & Short, Ch. 1958: *A Latin Dictionary.* Oxford.
Le Robert électronique 1989= *Le Robert électronique*. Paris.
Liddell, H. G./Scott, R. 1953: *A Greek-English Lexicon.* 9. Ed. Oxford.
ODS = Dahlerup, V.: *Ordbog over det danske Sprog.* København 1919–1954.
The Oxford English Dictionary. Second Edition on Compact Disc. Oxford 1992.
Riber Petersen, Pia 1984: *Nye ord i dansk 1955–75*. København.
Trübner 1939–1957: *Deutsches Wörterbuch.* Berlin.
VSO = *Dansk Ordbog udgivet under Videnskabernes Selskabs Bestyrelse.* København 1793–1905.