Samtaler i korpusformat: Repræsentation af talesprog i LANCHARTs korpus-infrastruktur

Philip Diderichsen & Torben Juel Jensen
Københavns Universitet

Abstract

LANCHART-korpusset udgøres dels af optagelser indsamlet i forbindelse med dialektologiske og sociolingvistiske projekter i 1960'erne, 1970'erne og 1980'erne, dels af optagelser af samtaler indsamlet af Sprogforandringscentret på Københavns Universitet mellem 2005 og 2015. Geografisk dækker korpusset en række lokaliteter bredt fordelt i Danmark samt danske udvandrersamfund i Argentina, Canada og USA. Korpusset er i TextGrid-format, hvilket muliggør en direkte kobling mellem transskriptionerne og lydoptagelserne samt fleksibel annotation af ord og længere tekstpassager. Korpusset er for nylig blevet relanceret i en ny søgeinfrastruktur baseret på Corpus Workbench (CWB) og den brugervenlige søgegrænseflade Korp, som udover hurtige og fleksible søgninger udmærker sig ved at være open source software der frit kan udvides med ny funktionalitet. Indlæsning af korpusdata i konkordansværktøjer som Korp kræver data i lineært format, hvilket medfører særlige problemstillinger i forhold til samtaledata, hvor der ofte forekommer overlap mellem talerne. I artiklen diskuterer vi disse problemstillinger og præsenterer vores løsning i form af en ny partiturvisning, der viser taledataene med lydsporet synkroniseret til transskriptionen.

Nøgleord: korpusinfrastruktur, talesprogsdata, samtaledata, annotation

1. Indledning

LANCHART-korpusset, som huses af Sprogforandringscentret på Københavns Universitet, består af lydoptagelser med tilhørende transskriptioner af danske samtaler indsamlet i perioden fra 1960'erne til i dag. Korpusset er blevet udviklet med henblik på at undersøge variation og forandring i det danske talesprog og indeholder i øjeblikket over 12 millioner ord.

Selvom de fleste talesprogsforskere formentlig vil være enige om at talesproget på mange måder adskiller sig fundamentalt fra skriftsproget, begynder de fleste forskningsprojekter der omhandler talesprog, paradoksalt nok med at producere en skriftsproglig repræsentation af de talesproglige data, en transskription, som analyserne herefter i vidt omfang baserer sig på. Det sker af gode, praktiske grunde, men indebærer ikke desto mindre i flere henseender en reduktion af talesprogets kompleksitet. Talesprogsdata er således i udgangspunktet lydlige begivenheder, der ofte er tidsligt overlappende med hinanden, men af hensyn til brugerens overblik og mulighed for analytiske annoteringer er det ofte ønskeligt med en skriftlig repræsentation, og af hensyn til korpusværkstøjer er det ofte nødvendigt at konvertere denne til et simpelt lineært format.

I artiklen vil vi indledningsvis præsentere indholdet i LANCHART-korpusset. Herefter vil vi fokusere på de problemstillinger det indebærer når talesprogdata skal repræsenteres skriftligt i et korpus, og de løsninger vi har valgt for LANCHART-korpusset. I relation hertil præsenterer vi den Korp-baserede infrastruktur som for nylig er blevet udviklet til at søge i LANCHART-korpusset. Infrastrukturen inkluderer en partiturvisning der synkroniserer lydsporet med transskriptionen, og som dermed giver brugeren umiddelbar adgang til den oprindelige optagelse fra den skriftsproglige repræsentation.

2. LANCHART-korpussets indhold

LANCHART står for *LANguage CHAnge in Real Time*, og LANCHART-korpusset blev etableret med en bevilling fra Danmarks Grundforskningsfond til et center for sociolingvistiske sprogforandringsstudier på Københavns Universitet. Projektet, der løb i perioden 2005-15, blev ledet af Frans Gregersen. Formålet med dette projekt var at undersøge dels hvordan og hvorfor dansk talesprog har forandret sig i det 20. århundrede (i praksis primært perioden fra 1970'erne til 2010), dels hvordan det enkelte individs sprogbrug ændrer sig gennem livet. Formålet er afgørende for hvordan korpusset er sammensat. LANCHART-

 Nordlyd 47.2: 77–89, *Struktur, ideologi og mangfald*, redigert av Ragni Vik Johnsen, Carola Kleemann, Øystein A. Vangsnes & Maud Westendorp. Publisert ved UiT Noregs arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.7084

korpusset er således indsamlet så det er muligt at undersøge sprogforandring i virkelig tid (*real time*), dvs. ved at sammenligne sprogbrugen på to eller flere forskellige tidspunkter, og et særlig karakteristikon er at en stor del af de informanter hvis tale indgår i korpusset, er optaget flere gange med en længere periode (20-30 år) imellem (se nedenfor).

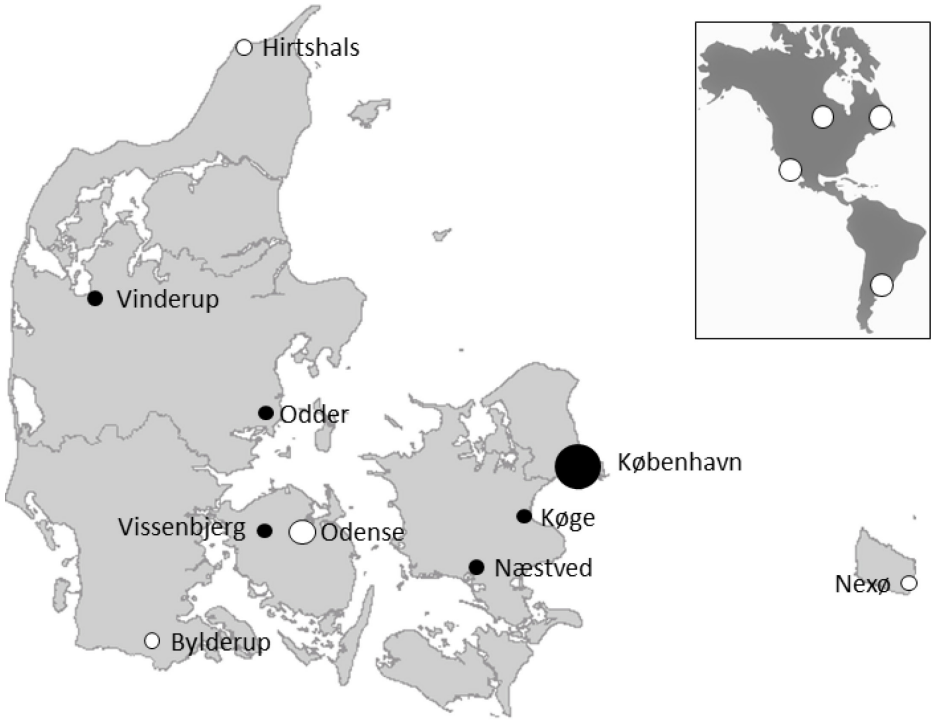
LANCHART-projektets dataindsamling bestod for en stor dels vedkommende i at genoptage informanterne fra en række dialektologiske og sociolingvistiske projekter som blev afviklet i slutningen 1970'erne eller 1980'erne. Projekterne fandt sted i den lille stationsby Vinderup i Vestjylland (Kristensen 1977, 1980), Odder i Østjylland, nær Danmarks næststørste by Århus (Nielsen & Nyberg 1992, 1993), det regionale centrum Næstved på Sydsjælland (Kristiansen 1991; Jørgensen & Kristensen 1994) samt København (Gregersen et al. 1991); se Figur 1. Ud over at genoptage et bredt udvalg af de gamle informanter blev der på hver lokalitet inkluderet et antal yngre informanter fra 8. og 9. klassetrin. Endelig blev der i begrænset omfang genoptaget informanter fra et projekt om flersprogede børn i Køge lidt syd for København (Jørgensen 2003). Derudover blev data fra et tidligere sociodialektologisk projekt fra Vinderup på Fyn (Pedersen 1994) samt ældre data fra starten af 1970'erne fra København inkluderet i korpusset. Informanterne blev udvalgt efter det kriterium at de var født og opvokset i området, og de havde for størstedelens vedkommende dansk som førstesprog. Det sidste gælder dog ikke informanterne fra Køge, hvor en stor del havde tyrkisk minoritetsbaggrund, og i København blev der i de nye optagelser også inkluderet en gruppe af flersprogede informanter i den yngste aldersgruppe. Vi henviser til Gregersen (2009) og Gregersen & Kristiansen (2015) for en nærmere beskrivelse af denne del af LANCHART-korpusset.

I perioden efter Sprogforandringscentret i 2015 blev indlejret på Københavns Universitet som en del af Institut for Nordiske Studier og Sprogvidenskab, er LANCHART-korpusset blevet udvidet med data fra flere nye projekter. Det drejer sig for det første om to projekter der udvider den geografiske dækning af Danmark. Det ene af disse er LaPUR (*Language and Place. Linguistic Variation in Urban and Rural Denmark*), som bidrager med optagelser med skoleelever fra landsbyen Bylderup i Sønderjylland og fra den multietniske Odense-bydel Vollsmose (Quist 2020). Det andet er Dialekt i periferien, som inkluderer optagelser med skoleelever og deres forældre og bedsteforældre fra Hirtshals i Nordjylland, Nexø på Bornholm og Bylderup, dvs. samme lokalitet som også indgår i LaPUR (Maegaard 2020). Derudover er LANCHART-korpusset blevet forøget betydeligt med optagelserne fra projektet Danske stemmer i USA og Argentina. Denne del af korpusset (også kaldet CoAmDa – *Corpus of American Danish*) udgøres af optagelser med danske udvandrere og deres efterkommere i USA, Canada og Argentina, optaget i perioden 1963-2015 (Kühl et al. 2017).

Informanterne i korpusset er altovervejende optaget under sociolingvistiske interview af 1-3 timers varighed, nogle også under gruppesamtaler med eller uden interviewer. En del af informanterne er som tidligere nævnt optaget to gange med 20-30 års mellemrum.[1] Det samlede korpus indeholder p.t. mere end 12 millioner ord produceret af 1435 forskellige talere i 1362 transskriberede optagelser. Der er for informanter og samtaler registreret en række baggrundsdata, først og fremmest køn, socialklasse (arbejderklasse vs. middelklasse baseret på uddannelsesniveau og jobtype; jf. Gregersen 2009:9-11), fødeår (1870-2005), førstesprog, lokalitet og optagetidspunkt (1966-2015).

Designet af LANCHART-korpusset gør det muligt at undersøge sprogforandring gennem såvel panelstudier (samme informanter følges over tid) som trendstudier (grupper af informanter med samme alder på optagetidspunktet sammenlignes på forskellige optagetidspunkter). Samtidig er data indsamlet fra et bredt spektrum af lokaliteter i Danmark, hvilket gør det muligt at undersøge geografisk betingede forskelle i sprogbrugen, og hvordan sprogforandringer spreder sig geografisk over tid.

1 Det drejer sig først og fremmest om en stor del af informanterne født i perioden 1942-1973 fra Vinderup, Odder, Næstved og København, i alt knap 200 informanter.



Figur 1: LANCHART-korpussets optagelokaliteter. De sorte pletter markerer det oprindelige LANCHART-projekt mens de hvide markerer senere projekter.

3. Dataformat og dataprocessering

Da LANCHART-korpusset alene inkluderer talesprog, er de primære data naturligvis lydoptagelserne af samtalerne samt metadata om optagelser og talere. For at kunne tilføje annotationer, søge i dataene og analysere dem med korpuslingvistiske metoder er det imidlertid nødvendigt at producere en skriftsproglig repræsentation af talesproget. LANCHART-korpusset består således i praksis alene af de optagelser fra de ovennævnte projekter som er blevet transskriberet.

Transskriptionen er for alle projekter sket efter Sprogforandringscentrets udskrivningsmanual, hvilket indebærer at transskriptionen i udgangspunktet følger dansk ortografi som den fastlægges i Retskrivningsordbogen (RO). For ord der ikke findes i Retskrivningsordbogen, følges Den Danske Ordbog (DDO), eller, hvis der er tale om ord der kun forekommer i bestemte dialekter, i den relevante dialektordbog, primært Bornholms Ordbog, Jysk Ordbog og Ømålsordbogen (BO; ØMO; JO); i praksis drejer det sidstnævnte sig om ganske få ord. Transskriptionen afviger dog fra den ortografiske norm ved at der ikke er sat interpunktion, ligesom den mht. ordstilling samt brug af affikser og artikler følger taleren også i tilfælde hvor det strider mod den ortografiske norm (fx "lillat hår" vs. standarddansk "lilla hår", "en hus" vs. "et hus" og "æ kirke" vs. "kirken"). Derudover er der særlige konventioner for repræsentation af talesprogsfænomener som selvafbrydelser, tøvelyde, minimalrespons mv. Udskrivningsmanualen er tilgængelig på Sprogforandringscentrets hjemmeside: https://dgcss.hum.ku.dk/online-ressourcer/lanchartkorpusset/.

Den ortografisk baserede transskriptionskonvention er valgt på baggrund af at konventionen dels skal kunne anvendes ens af mange forskellige udskrivere og med et rimeligt tidsforbrug dels skal danne grundlag for søgninger og optællinger på tegnniveau (dvs. fra morfemniveau og opefter). Til gengæld repræsenterer den langt fra den fulde variation der findes på udtryksniveau i talesproget. Det har derfor i valget af dataformat været afgørende at lydoptagelsen er umiddelbart tilgængelig ud fra transskriptionen, dvs. koblet til transskriptionen via tidskoder sådan at man i forbindelse med analyser kan tage udgangspunkt i

lydoptagelsen frem for transskriptionen i tilfælde hvor det er hensigtsmæssigt eller nødvendigt, fx ved analyse af fonetisk variation.

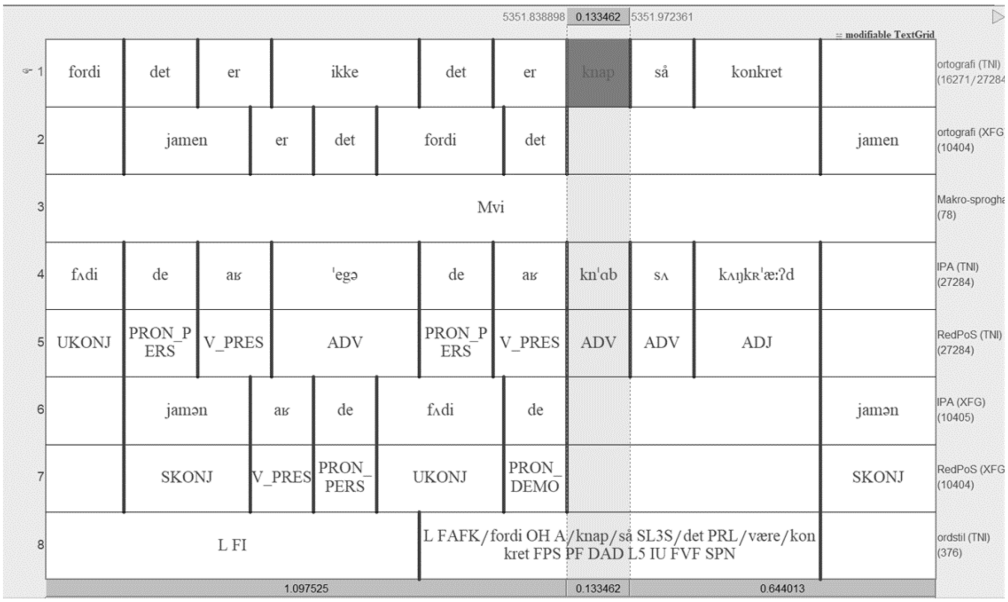
Talesprogsdata er grundlæggende lydbegivenheder der foregår i tid, og i modsætning til i skriftsproget sker det ofte at en sprogbrugers produktion sker samtidig med en andens, ligesom der kan forekomme samtidige lydlige begivenheder der er relevante for forståelsen af de sproglige ytringer.[2] Det var derfor også nødvendigt at dataformatet kan afspejle denne kompleksitet visuelt på en måde der er overskuelig for brugeren af korpusset. Derudover var det i valget af dataformat vigtigt at det er fleksibelt i forhold til annotationernes skopus, dvs. det skulle være let at tilføje analytiske annotationer uanset om de knytter sig til ordniveauet, sætningsniveauet eller længere passager som fx samtalegenrer. Endelig var det vigtigt at flere personer eller grupper kan arbejde med samme samtale samtidig, dvs. analysere og tilføje annotationer uafhængigt af hinanden – som sidenhen samles så alle annotationer til sidst er tilgængelige i samme fil.

Af disse grunde blev TextGrid-formatet valgt som det grundlæggende dataformat. Et TextGrid er en tekstfil organiseret i et antal opmærkningslag kaldet tiers, som er opdelt i intervaller der hver især er defineret ved to tidsværdier (i sekunder; xmin og xmax). Til hvert interval kan knyttes en tekst der beskriver det tilsvarende interval i den lydfil TextGrid-filen relaterer sig til. Det kan være et transskriberet ord, men det kan også være analytisk information som ordklasse eller sætningstype. En TextGrid-fil kan ved hjælp af programmet Praat (Boersma & Weenink 2023) vises i partiturformat, dvs. således at de forskellige tiers vises samtidig og alignet med lydfilen under hinanden på en intuitiv måde. Dette kan være to taleres overlappende tale eller intervaller i et tier med transskription og et tier med ordklassetagging.

Figur 2 viser et LANCHART-TextGrid i Praat. Her ses i alt otte tiers; yderst til højre vises tierenes navne, og øverst vises xmin- og xmax-værdierne for det markerede interval, som indeholder ordet "knap". For de tiers der knytter sig til en enkelt taler, vises et unikt taler-id (i dette tilfælde TNI eller XFG) i parentes efter tiernavnet. Dette gælder for de to tiers der indeholder transskription af talen (ortografi), og de tilhørende tiers der indeholder annotation vedrørende udtale (IPA)[3] og ordklasse (RedPoS for "Reduced Part of Speech", da det er en reduceret (simplificeret) version af de meget detaljerede PAROLE-ordklassetags; jf. Keson 1999), samt tieret "ordstil" der indeholder analytisk annotation relateret til ordstilling i ledsætninger hos informanten (og ikke, som man måske kunne tro, en klassificering af stil). Tieret "Makro-sproghandling" indeholder derimod information om diskurskontekst der knytter sig til interaktionen, og det er derfor ikke knyttet til en specifik taler (der er således intet taler-id efter tiernavnet). Tallene i parentes under tier-navnene angiver antallet af intervaller i de respektive tiers. Som det fremgår, svarer skopus for annotationen i IPA- og RedPoS-tierene en-til-en til ordinddelingen i ortografi-tierene mens intervallerne i Makro-sproghandling- og ordstil-tierene har et bredere skopus, dvs. de dækker over flere intervaller (ord) i ortografi-tierene. Formatet er således 100% fleksibelt i forhold til hvad annotationen knyttes til; det blev dog af praktiske grunde, primært af hensyn til den efterfølgende brug af korpusværktøjer, besluttet at ortografi-tierenes intervalinddeling er den grundlæggende, således at enhver intervalgrænse i de øvrige tiers falder sammen med en intervalgrænse i et ortografi-tier. Det indebærer også at ordniveauet er den mindste enhed for samtlige annoteringer der overføres til korpussøgeværktøjet (se næste afsnit).

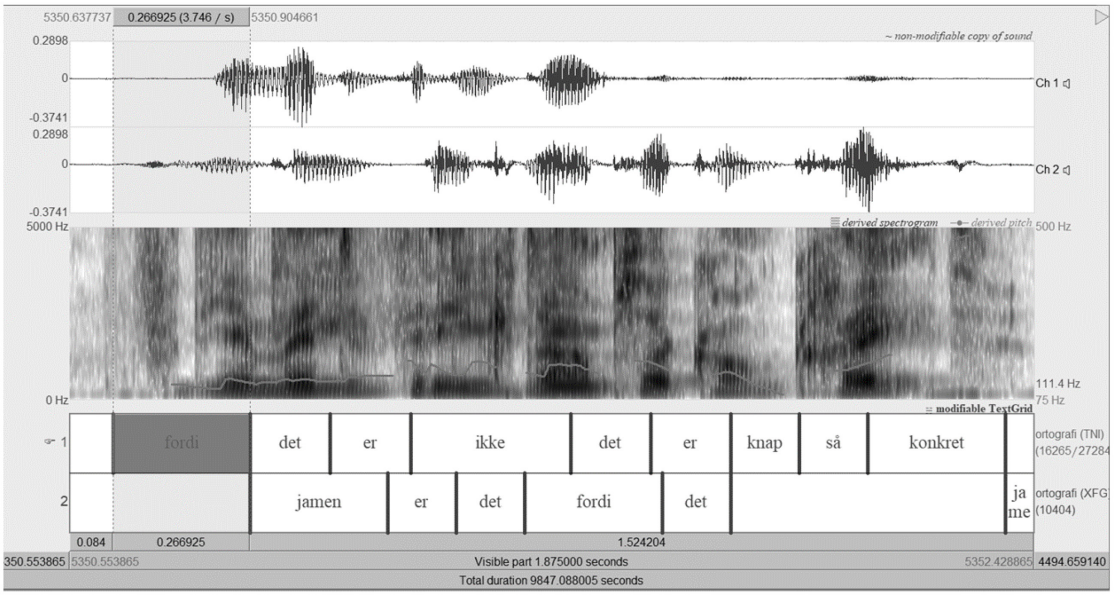
2 Det er åbenlyst at der også er andre modaliteter end tale som er relevante for forståelsen af en samtale, ikke mindst det visuelle, men der foreligger, med undtagelse af det såkaldte CLARIN-delkorpus med københavnske gymnasieelever optaget i 2010, kun lydoptagelser fra de projekter der danner grundlag for LANCHART-korpusset.

3 LANCHARTs IPA-annotering er ikke baseret på den faktiske udtale i den enkelte optagelse, men er ligesom PoS (ordklasse og bøjningsform)-taggingen foretaget automatisk (algoritmisk) ud fra transskriptionen. Lydskriften repræsenterer således en relativt distinkt standardudtale (også for optagelser af dialekttalende), og kan primært bruges til at identificere fonetiske kontekster i samtalerne, som så kan analyseres ud fra lydoptagelsen. Såvel PoS- som IPA-annotering er foretaget med programmet Phonix udviklet af Peter Juel Henrichsen (Henrichsen 2009, 2011).



Figur 2: LANCHART-TextGrid vist i partiturformat i Praats editor (ikke alle tiers i TextGridet er vist).

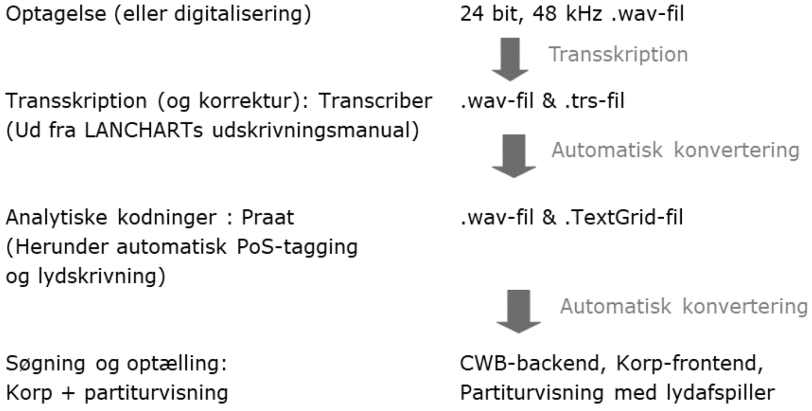
Et TextGrid kan åbnes i Praat sammen med den tilhørende lydfil, som så vises i form af en lydbølge per kanal og et spektrogram (se Figur 3), og Praat stiller en lang række muligheder for akustiske analyser (og i øvrigt også for manipulation af signalet) til rådighed, hvilket er en stor fordel ved dataformatet. Derudover kan TextGridene bearbejdes i Praat via et indbygget programmeringssprog, hvilket er meget anvendeligt til helt eller delvist at automatisere forskelige former for analytisk opmærkning (fx ved at sætte koder ind i et analytisk tier hver gang bestemte ordformer forekommer i ortografi-tieret).



Figur 3: LANCHART-TextGrid med lydbølger og spektrogram.

Transskriptionen kunne principielt godt være foregået i Praat, men det viste sig hurtigt at det ikke var hensigtsmæssigt af hensyn til tidsforbruget. Optagelserne blev derfor udskrevet i programmet Transcriber (Barras et al. 1998; https://trans.sourceforge.net/) og efterfølgende konverteret til Praats TextGrid-format.

Som nævnt er alle intervaller i et TextGrid defineret ved tidskoder, hvilket indebærer at der i transskriptionen principielt skulle indsættes en tidskode før og efter hvert enkelt ord. Det ville dog have været en alt for omfattende opgave for et så stort korpus som LANCHART at gøre det manuelt ud fra lydfilen, så udskriverne satte i praksis kun tidskoder ind før og efter hver ytring (dvs. ved talerskift) samt ved længere pauser eller andre former for ophold i talen (tøvelyde, hørbar vejrtrækning, latter o.l.). I forbindelse med konverteringen til TextGrid-format blev de resterende intervalgrænser mellem ordene indsat automatisk, baseret på antallet af stavelser i de enkelte ord. Ordenes intervalgrænser er således, bortset fra den første og sidste grænse i en ytring, ikke præcise i forhold til lyden.



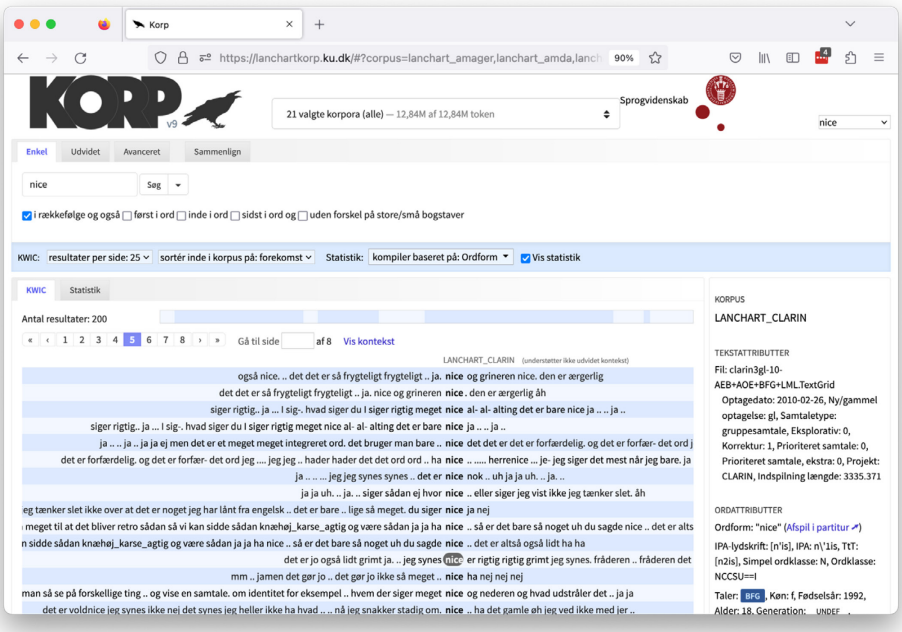
Figur 4: Dataformater og dataprocessering i LANCHART.

På trods af at Praat stiller en lang række meget brugbare værktøjer til rådighed og derfor blev valgt som projektets værktøj til annotation, er det, som det vil blive beskrevet i næste afsnit, ikke velegnet som korpussøgeværktøj for et korpus af LANCHARTs størrelse og med en så bred brugergruppe. Efter de forskellige analytiske kodninger er foretaget, bliver TextGrid-filerne derfor konverteret til et format der kan indlæses i korpussoftwaren Corpus Workbench (CWB, Evert & Hardie 2011), der fungerer som back-end (dvs. database) for webgrænsefladen Korp (Borin et al. 2012) (se Figur 4). Der kan ikke foretages yderligere annoteringer i CWB/Korp-systemet – det kan alene bruges til søgninger og optællinger – men såfremt der ønskes yderligere annoteringer af en fil, kan de foretages i TextGrid-versionen af filen i Praat, hvorefter CWB/Korp-korpusset automatisk bliver opdateret med de(t) nye eller reviderede analysetier(s)[4].

4. Fra synkroniseret til lineariseret visning og tilbage igen

For at kunne søge på tværs af mange samtaler må dataene indlæses i et søgeværktøj der egner sig til formålet, og her er det oplagte valg korpuslingvistikkens arbejdshest, konkordansværktøjet – dvs. selve det værktøj der stiller søgehits op i rækker centreret omkring søgeordet/-ordene. Konkordansværktøjet indgår i ethvert korpussøgeværktøj, således også i Korp, se Figur 5.

4 Det er værd at nævne at opdateringen af CWB/Korp-infrastrukturen netop er fuldautomatiseret. Der kører hver nat programmer på de relevante servere der tjekker om der er kommet nye data i korpus og i givet fald automatisk indlæser de nye data i CWB og Korp.



Figur 5: Konkordans af ordet *nice* i Korp.

Her melder problematikken omkring skriftlig repræsentation af samtaler sig som det konkrete og lavpraktiske problem at konkordansværktøjer som udgangspunkt er udviklet til endimensionelle sproglige sekvenser, dvs. tekst, transskriberet (ene)tale eller andet der forløber lineært i tid uden flere samtidige, overlappende delsekvenser. Men overlap er som bekendt udbredt i samtaler. For at kunne repræsentere dette i et korpussøgeværktøj må man enten transformere samtaledataene til et lineært format der er kompatibelt med en klassisk konkordansvisning, eller gribe til en helt anden type resultatvisning. Vi har valgt at gøre begge dele. Samtaledataene lineariseres så man kan søge i dem og vise dem på klassisk vis, og som et supplement til dette har vi udviklet et subsidiært værktøj der viser dataene i partiturformat, jf. forrige afsnit.

Lineariseringen bygger på et simpelt princip hvorved turdele ordnes efter starttidspunkt. Det indebærer at sekvenser fra forskellige talere lægges i rækkefølge efter hinanden uden hensyntagen til hvordan de evt. måtte overlappe, jf. Figur 9 nedenfor. Turdele er de talesekvenser som de oprindelige Transcriber-transskriptioner er opdelt i af udskriverne – oftest sekvenser på et dusin ord eller mindre, som ofte men ikke altid svarer til hele taleture (og derfor benævnes *turdele*, ikke *ture*).[5] Lineariseringen i Korp indebærer i øvrigt at de forskellige taleres turdele placeres efter hinanden uden nogen markering af overgangen fra den ene (talers) turdel til den næste(s)[6]. De usynlige overgange mellem turdelene gør det ekstra relevant med en supplerende partiturvisning, jf. nedenfor.

5 At turdelene ikke altid svarer til hele taleture, skyldes at der, som tidligere nævnt, også er indsat grænser ved længere pauser eller andre former for ophold i talen inden for den enkelte taletur. Ved overlappende tale er begge taleres tale transskriberet i Transcriber inden for samme tidsgrænser, dvs. i samme tidsinterval: først den ene talers tale, og herefter den anden talers tale markeret med spidse parenteser for at angive at det sker samtidig med den første taler.

6 De enkelte taleres ofte lange tavse passager *efter* de har haft turen (og typisk mens en anden taler), vises ikke i det lineariserede format. Dette er valgt for ikke at give et falsk indtryk af at der er en lang pause i samtalen efter hver tur, hvilket der oftest ikke er. Pauser *indenfor* de enkelte turdele – som ofte også vil være pauser i samtalen som sådan –

Til at fremsøge og vise de lineariserede data har vi valgt Korp da det er et intuitivt og brugervenligt værktøj med mange nyttige og avancerede funktioner, der desuden er tilgængeligt som open source og under aktiv udvikling af en gruppe softwareudviklere ved forskellige nordiske sproginstitutioner, herunder selvfølgelig især Språkbanken i Göteborg (Borin et al. 2012).

Indlæsning af annoterede talesprogsdata i en korpusdatabase som CWB frembyder – foruden den nævnte linearisering – den udfordring at dataene skal bringes i en form hvor alle annotationer kan knyttes til enkeltord (tokens). Det skyldes at CWB's inputformat er forholdsvis simpelt struktureret med tokens som den grundlæggende enhed med et token pr. linje og annotationer ud for hvert token, jf. Figur 6.

```
<corpus id="lanchart_clarin">

<text samtaler_dato="2010-02-26" samtaler_projekt="32" samtaler_samtaletype="gruppesamtale" ...>
<sentence>
...
to    [t'o&#720;&#704;]          t\'1o\:f\?g          [t2o:!]     NUM ...
tre   [tR'&#230;&#720;&#704;]    t\rc\'1\ae\:f\?g    [tr2z:!]    NUM ...
...
</sentence>
</text>
</corpus>
```

Figur 6: Eksempel på tokenbaseret dataformat til indlæsning i CWB. Bemærk at "corpus", "text" og "sentence" er mærker der traditionelt bruges til at strukturere *skriftlige* korpusdata hierarkisk (således at et korpus består af et antal tekster, der hver består af et antal sætninger). Vi har bibeholdt de traditionelle mærker selv om der i vores tilfælde ikke er tale om tekster, men samtaler, og ikke sætninger, men tur(del)e. I det citerede forekommer ordene "to" og "tre", her med lydskrivning i tre formater (html, Praat og SAMPA) samt ordklasseannotering.

I Figur 6 knytter alle annotationerne sig til et enkelt af de to ord "to" eller "tre", hvilket er uproblematisk. Der er imidlertid en mængde annotationer i TextGrid-filerne der ikke er knyttet til enkeltord, men derimod til længere sekvenser af ord, fx annotationer der knytter sig til analysen af ledstilling i ledsætninger. Den slags annotationer skal for at kunne indlæses i CWB omkodes således at hvert ord i en annoteret sekvens er annoteret med den givne kategori. Ganske vist er det muligt i CWB's dataformat at annotere sekvenser ved at omgive dem med mærker (eksempelvis "<NP>", "<kommentar>", "<ledsætning>" eller andet) - men formatet tillader ikke at sådanne mærker krydser hinanden, hvilket fx opmærkningen af ledsætninger ville gøre, jf. eksempel 1 nedenfor, hvor en "<ledsætning>"-opmærkning begynder i én "<sentence>"-opmærkning, men slutter i den næste (husk at mærket "sentence" reelt markerer turdele, ikke sætninger; passagen er blevet til to turdele fordi taleren holder en lang pause mellem ordene "stå" og "i").

(1) <sentence>[..] det kan jo ikke være meningen <ledsætning>man skal stå</sentence> <sentence>[..] i tolden og vise test</ledsætning> det ved jeg sgu ikke</sentence>

I vores transformation af data fra TextGrid-format til CWB-format omkoder vi sekvensannotationer til en variant af den såkaldte BIO-opmærkning (Ramshaw & Marcus 1999; Jurafsky & Martin 2023: kap. 8) – BIO for "Begin", "In", "Out" – der markerer om et givet token er i begyndelsen, indeni eller udenfor en given sekvensannotation. I vores variant er omkodningen lavet ved at sætte et løbenummer på en given annotationskategori som præfiks og koden "I" eller "E" som suffiks. "I" og "E" har henholdsvis betydningen "in" og "end", og markerer således om det givne ord er inde i (eller udgør starten af) en given annoteret sekvens eller om det udgør det sidste ord. Eksempelvis vil sekvensen "deres eget sprog", der

vises derimod vha. "pauseprikker": en prik (".") for pauser på under 200 ms, to prikker ("..") for længere pauser på under 1000 ms og tre prikker "…" for pauser på 1000 ms og derover.

oprindeligt er annoteret samlet med kategorien "fynsk" i annotationslaget (tieret) "Fynsk intonation", blive omkodet på følgende måde:

Før omkodning:

Ortografi	taler	..	deres	eget	sprog	nogle	gange
Fynsk intonation			fynsk				

Efter omkodning:

Ortografi	taler	..	deres	eget	sprog	nogle	gange
Fynsk intonation			1_fynsk_I	2_fynsk_I	3_fynsk_E		

Figur 7: Omkodning fra sekvensopmærkning til BIO-opmærkning.

I Figur 8 er ledsætningen fra eksempel (1) vist med BIO-opmærkning for at illustrere hvordan denne type annotation løser problemet med overlappende sekvensannotationer.

meningen	man	skal	stå	</sentence>	<sentence>	i	tolden	og	vise	test	det
	I_leds_1	I_leds_2	I_leds_3			I_leds_4	I_leds_5	I_leds_6	I_leds_7	E_leds_8	

Figur 8: BIO-opmærkning af ledsætningen i eksempel (1). Bemærk hvordan denne type opmærkning er token-baseret og dermed frit kan overlappe med "<sentence>"-opmærkningen.

Med disse transformationer kan vores TextGrid-data indlæses i CWB og fremsøges og vises i konkordanser i Korp. Men konkordansformatet kan ikke stå alene. Det illustreres af det følgende eksempel (2), hvor det tydeligt fremgår hvor dårligt det lineære konkordansformat af og til afspejler den faktiske interaktion. Eksemplet er en konkordanslinje fra en gruppesamtale (prikker markerer pauser: jo flere, des længere, jf. ovenfor).

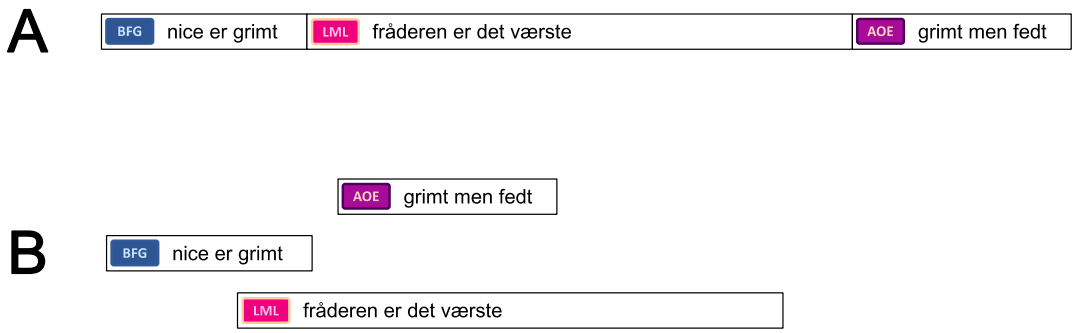
(2) jeg synes nice er rigtig rigtig grimt jeg synes . fråderen .. fråderen det er det værste ord .. og det er altså også gået lidt af mode igen nu det det er grimt grimt men det det er derfor det er fedt

Der diskuteres sprog, og ordene *nice* (DDO: 'dejlig, god eller på anden måde tiltalende') og *fråderen* (DDO: 'noget der er meget lækkert, smager godt eller ser godt ud') kritiseres. Først og fremmest er det svært at se hvornår turen skifter. Den relevante information findes som annotationer til dataene, men i den rene, lineære skriftrepræsentation af talen er turskiftene ikke synlige. Vi kan her markere turskiftene ved at sætte talerkoderne ind i tekststrengen:

(2') [BFG:] jeg synes nice er rigtig rigtig grimt [LML:] jeg synes . fråderen .. fråderen det er det værste ord .. og det er altså også gået lidt af mode igen nu [AOE:] det det er grimt grimt men det det er derfor det er fedt

Det giver sekvensen struktur, men som det ses i Figur 9 nedenfor (hvor de enkelte taleture er forkortet af hensyn til overskueligheden), er det slet ikke nok til at gøre repræsentationen retvisende. Tre talere kommer på banen. Lineært set ser det velordnet og kohærent ud: Ytringerne fra BFG, LML og AOE kunne, som i A-versionen, sagtens have ligget pænt efter hinanden som perler på en snor, med BFG's mishagsytring om *nice* først, LML's ditto om *fråderen* som et relevant næste indlæg og AOE's ytring som en reaktion på LML's. Det gør de bare ikke.

I virkeligheden er der tale om en noget mere konkurrencepræget, overlappende sekvens hvor AOE's ytring, der reelt er en reaktion på BFG's ytring om *nice*, drukner i LML's ytring om *fråderen*, der dækker den både i tid og lydstyrke. Det der ordnet lineært ser ud som A nedenfor, forløber i virkeligheden som i B. Konkordansformatet (A) er selvsagt ikke nogen særlig tro repræsentation af den faktiske samtalesekvens i den slags tilfælde, og partiturformatet (B) er derfor et meget væsentligt supplement til standardformatet.

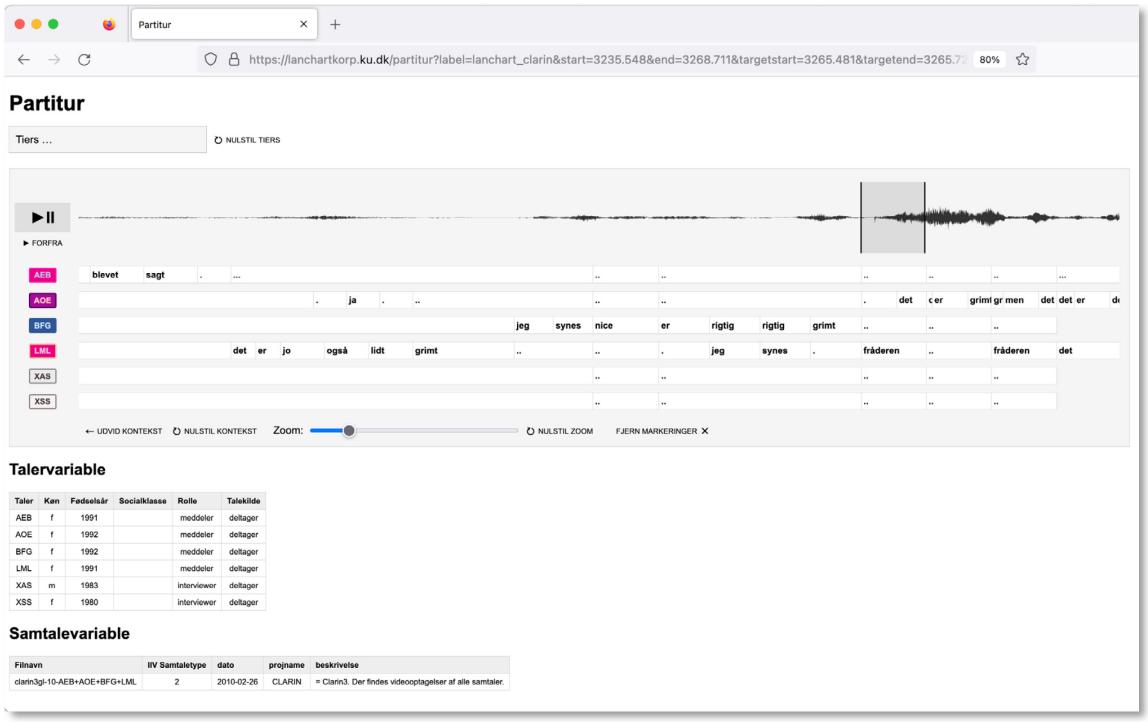


Figur 9: Eksempel på lineariseret (mis)visning vs. partiturvisning.

En anden helt central grund til at udvikle en særlig partiturvisning er at det tidssynkroniserede format giver mulighed for at vise dataene synkroniseret med lydsignalet og således altså afspille lyden med annotationer næsten som man kender det fra Praat.

Skærmbilledet i Figur 10 viser LANCHARTs partiturvisning. Den er en lokalt udviklet tilføjelse til Korp i form af en selvstændig webapplikation der modtager de nødvendige søgeparametre fra Korp (bl.a. start- og sluttidspunkt for det givne søgematch samt et filnavn på den fil matchet forekommer i). Dette sker når brugeren trykker på et link der genereres af information tilknyttet hvert token som annotationer (linket med teksten "Afspil i partitur ↗" kan anes til højre i figur 5).

Partiturvisningen er udformet som en lydafspiller med en play/pause-knap, hvor lydsignalet vises som en lydbølge synkroniseret med transskriptionerne af de forskellige talere.



Figur 10: LANCHARTs partiturvisning.

Partiturvisningen har forskellige features, der løbende vil blive udbygget og ændret efter behov af Sprogforandringscentret. Under knappen "Tiers ..." gemmer der sig en menu hvor man kan klikke diverse annotationslag (svarende til tiers i de tilgrundliggende TextGrids) til og fra. Play/pause-knappen og knappen "Forfra" under den kan betjenes fra tastaturet vha. hhv. mellemrumstasten (space) og skift+mellemrumstast (shift+space). Den grå markering på lydbølgen repræsenterer et stykke der kan afspilles separat; det gøres ved at klikke på markeringen. Partiturvisningen er designet til automatisk at markere en sekvens svarende til en given start- og sluttidskode, typisk det matchende ord i en konkordanssøgning. Men man kan lave så mange yderligere markeringer man vil, med musen og også trække dem længere eller kortere (og til sidst helt væk) eller flytte dem i deres helhed. Samtlige markeringer kan fjernes med et klik på knappen "Fjern markeringer". En zoomfunktion gør det muligt at gøre visningen af partituret bredere eller smallere, og venstrekonteksten kan udvides vha. en knap. Under selve partituret vises to tabeller med metadata, en med talervariable og en med samtalevariable.

Den nemme adgang til lydsporet bag transskriptionerne gør det muligt at verificere lydlige forhold ved en given ordforekomst langt hurtigere og smidigere end før, og partiturformatet kan afdække væsentlige samtalestrukturelle forhold som i eksemplet ovenfor.

5. Andre visninger

Partiturvisningen med lyd er et væsentligt supplement til konkordansvisningen – og den er samtidig *proof of concept* for andre visninger. Muligheden for at gå til en token-koblet URL fra et hvilket som helst token i korpusset kan udnyttes til visninger hvor kun fantasien sætter grænser. Oplagte eksempler kunne være en visning af det givne token med video (evt. i form af en udvidelse af partiturvisningen) eller en grafisk visning af data fra sociale medier, og i forbindelse med fx gamle tekster vil en faksimilevisning centreret omkring det givne token være oplagt. Da Korp-projektet som nævnt er open source, er det endvidere muligt at udvide selve Korp med ny funktionalitet og nye datavisninger. Denne fleksibilitet agter vi at udnytte i den videre udbygning af LANCHART-korpusset.

6. Søgning på baggrund af metadata

Vi har i denne artikel fokuseret på grundlæggende problemstillinger i relation til repræsentation af samtaledata. I forbindelse med Korp-implementeringen af LANCHART-korpusset har vi fokuseret på auditive og visuelle visninger af søgeresultater. En anden og lige så vigtig del af korpusudviklingsarbejdet har naturligvis ligget i at effektivisere søgninger i korpusset på baggrund af de metadata der er beskrevet i afsnit 2.

Selvom vi har omtalt LANCHART-korpusset som ét korpus, er det således opdelt i 21 subkorpusser på baggrund af hvilket projekt (og dermed også hvilken lokalitet) optagelserne stammer fra, og man kan som bruger vælge at søge enten i hele korpusset eller i et eller flere af disse subkorpusser. Ligeledes kan man via Korp-grænsefladen specificere sine søgninger i forhold til metadata som talerens køn, socialklasse, fødeår og/eller optagetidspunktet for samtalen, eller man kan efterfølgende sortere eller filtrere søge-resultaterne på baggrund af disse metadata.

7. Andre korpusser: LANCHART-infrastrukturen som projekthotel

Ligesom der er gode muligheder for at udvikle og udvide den Korp-baserede korpusinfrastruktur, er vi også interesserede i at udnytte de muligheder der er indbygget i Korp for at tilføje flere korpusser. LANCHART-projektet har fra starten indoptaget og samordnet data fra forskellige projekter som en slags projekthotel, og det har vi i sinde at blive ved med således at sammenlignende søgninger og undersøgelser kan udføres på nye data og datatyper – fx fra sociale medier, som antydet ovenfor.

8. Bliv bruger!

Vores version af Korp med LANCHART-korpusset er tilgængeligt for forskere via login, som vi udleverer ved henvendelse. Af såvel juridiske som etiske hensyn er der dog en række betingelse man som bruger skal skrive under på at overholde. Dette kan man læse nærmere om på Sprogforandringscentrets hjemmeside (https://dgcss.hum.ku.dk/online-ressourcer/lanchart-korpusset/).

Referencer

Bornholmsk ordbog (BO). 1908. København: Det Kgl. Danske Videnskabernes Selskab.

Ømålsordbogen - en sproglig-saglig ordbog over dialekterne på Sjælland, Lolland-Falster, Fyn og omliggende øer (ØMO). 1992–. København: Institut for Dansk Dialektforskning.

Retskrivningsordbogen (RO). 2012. 4. udg. København: Dansk Sprognævn.

Den danske ordbog (DDO). u.å. København: Det Danske Sprog- og Litteraturselskab. Tilgængelig på https://ordnet.dk/ddo

Jysk Ordbog (JO). u.å. Århus: Peter Skautrup Centret for Jysk Dialektforskning, Aarhus Universitet. Tilgængelig på http://jyskordbog.dk

Barras, Claude, Edouard Geoffrois, Zhibiao Wu & Mark Liberman. 1998. Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. I *First International Conference on Language Resources and Evaluation (LREC)*. 28.–30. maj, 1998, pp. 1373–1376.

Boersma, Paul & David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.09. Tilgængelig på http://www.praat.org/

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. I *Proceedings of LREC 2012*, pp. 474–478.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. I *Proceedings of the Corpus Linguistics 2011 conference*, pp.

Gregersen, Frans. 2009. The data and design of the LANCHART study. *Acta Linguistica Hafniensia* 41: 3–29. https://doi.org/10.1080/03740460903364003

Gregersen, Frans, Jon Albris & Inge Lise Pedersen. 1991. Data and design of the Copenhagen study. I *The Copenhagen study in urban sociolinguistics,* redigeret af Frans Gregersen & Inge Lise Pedersen. C. A. Reitzels Forlag, København.

Gregersen, Frans & Tore Kristiansen. 2015. Indledning. Sprogforandring i virkelig tid. I *Hvad ved vi nu - om danske talesprog,* redigeret af Frans Gregersen & Tore Kristiansen. Sprogforandringscentret, København.

Henrichsen, Peter Juel. 2009. The CBS Text-to-Speech Workbench. *Working Paper / Internationale Sprogstudier og Videnskabsteknologi No. 2009-1*, Tilgængelig på https://research-api.cbs.dk/ws/portalfiles/portal/58999553/2009_1.pdf.

Henrichsen, Peter Juel. 2011. Program Phonix. DGCSS' redskab til fono-morfo-syntaktisk annotation.

Jurafsky, Dan & James H. Martin. 2023. *Speech and Language Processing*, 3. (draft) udg. Tilgængelig på https://web.stanford.edu/~jurafsky/slp3 (tilgået 23.3.2023).

Jørgensen, Jens Norman. 2003. Bilingualism in the Køge project. *International Journal of Bilingualism* 7(4): 333–352. https://doi.org/10.1177/13670069030070040101

Jørgensen, Jens Normann & Kjeld Kristensen. 1994. *Moderne sjællandsk*. C.A. Reitzel, København.

Keson, Britt. 1999. *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus.* Det Danske Sprog- og Litteraturselskab, Tilgængelig på https://korpus.dsl.dk/documentation/PAROLE-dokumentation/paroledoc_da.pdf (tilgået 23.08.2023).

Kristensen, Kjeld. 1977. Variationen i vestjysk stationsby-mål. *Dialektstudier* 4(1): 29–109.

Kristensen, Kjeld. 1980. Situationsafhængig sprogbrug hos vestjyske skoleelever. *Danske Folkemål* 22(2): 29–124.

Kristiansen, Tore. 1991. *Sproglige normidealer på Næstvedegnen*. Ph.d.-afhandling, Københavns Universitet, København.

Kühl, Karoline, Jan Heegård Petersen, Gert Foget Hansen & Frans Gregersen. 2017. CoAmDA. Et nyt dansk talesprogskorpus. *Danske talesprog* 17: 131–160.

Maegaard, Marie. 2020. Introduction: Standardization as Sociolinguistic Change. I *Standardization as Sociolinguistic Change,* redigeret af Marie Maegaard, Malene Monka, Kristine Køhler Mortensen & Andreas Candefors Stæhr. Routledge, New York. https://doi.org/10.4324/9780429467486

Nielsen, Bent Jul & Magda Nyberg. 1992. Talesprogsvariation i Odder kommune. I. Lokalsprog og rigsmål i sociolingvistisk belysning. *Danske Folkemål* 34: 45–202.

Nielsen, Bent Jul & Magda Nyberg. 1993. Talesprogsvariation i Odder kommune. II. Yngre og ældre rigsmålsformer i sociolingvistisk belysning. *Danske Folkemål* 35: 249–348.
Pedersen, Inge Lise. 1994. Linguistic Variation and Composite Life Modes. I *The Socioloinguistics of Urbanization. The Case of the Nordic Countries,* redigeret af Bengt Nordberg. de Gruyter, Berlin/New York. https://doi.org/10.1515/9783110852622.87
Quist, Pia. 2020. Sprog og sted: En undersøgelse af sproglig variation i forstaden og landsbyen. *Danske talesprog* 20: 175–194.
Ramshaw, Lance A. & Mitchell P. Marcus. 1999. Text Chunking Using Transformation-Based Learning. I *Natural language processing using very large corpora,* redigeret af Armstrong Susan, Kenneth Church, Isabelle Pierre, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2390-9_10