Psyke & Logos, 1988, *9*, 337-356

TALEGENKENDELSE OG TALESYNTESE

Peter Holtse & Peter Molbæk Hansen

Inden for rammerne af en generel model for sproggenkendelse og sprogproduktion diskuteres nogle principielle problemer i forbindelse med genkendelse og syntese af tale. Der argumenteres for, at talegenkendelse og -syntese er integrerede dele af datalingvistik og såkaldt »naturligt sprog«, og parallelliteten mellem genkendelse og syntese søges demonstreret.

1. Indledning

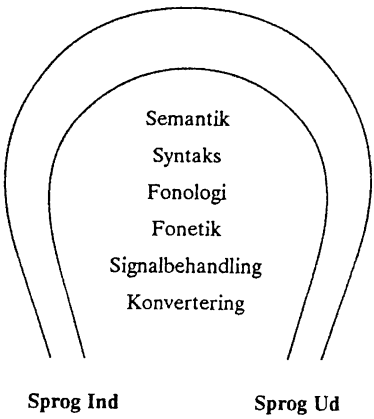
Talesyntese og talegenkendelse er som rene videnskaber relativt gamle, eftersom de har været dyrket ved sprogvidenskabelige, elektroniske og enkelte psykologiske forskningslaboratorier mange steder i verden, længe før nogen kunne se nytten heraf. I dag er *taleteknologi* især ved tekniske højskoler og lignende, noget man venter sig en del praktiske anvendelser af, mens det mange steder stadig kniber med at få de mere basale og generelle datalingvistiske problemer inddraget.

Taleteknologien er således inde i en rivende udvikling, hvor mange forestiller sig, at maskinerne er lige på nippet til at tale og høre så godt som noget menneske. Desværre er mange tidlige forhåbninger om nyttige og smarte apparater allerede blevet gjort til skamme, og det er så småt ved at blive alment erkendt, at der ikke blot er tale om teknologiske problemer, men at taleteknologien er stærkt afhængig af resultater inden for især sprogvidenskabelig forskning.

Dette bliver umiddelbart forståeligt, hvis vi ser på en generel model for behandling af sprog og tale. I figur 1 er vist den såkaldte »hesteskomodel« af sprogbehandlingen. I denne model kommer tekst eller tale ind forneden til venstre, passerer op gennem en eller flere analysefunktioner - signalbehandling, fonetik, syntaks etc. - for derefter at flyde ned gennem den højre gren af hesteskoen og komme ud forneden til højre som skreven tekst eller som lyd i en højttaler.

En af pointerne ved denne model er naturligvis, at vi gerne vil forsøge at illustrere de generelle problemer ved behandling af naturligt sprog i datamaskiner, uanset hvilken medieform sproget optræder på. Endvidere giver den os lejlighed til at klassificere forskellige praktiske måder at udforme specielt

et tale input/output system på. Vi vil i den forbindelse se bort fra alle de tilfælde, der udelukkende omfatter skrevet sprog, selv om genkendelse af bogstaver på papir formentlig på mange måder minder om genkendelse af talt sprog.



Figur 1:

Skematisk fremstilling af hesteskomodellen. Modellen illustrerer, hvorledes sprog, i skriftlig eller talt form, kommer ind forneden til venstre for efter en passende bearbejdelse oppe i hesteskoen at komme ud forneden til højre.

Hesteskomodellen er også velegnet til at afkræfte et udbredt dogme, der - let forenklet - kan formuleres således:

Automatisk talegenkendelse er betydeligt vanskeligere at virkeliggøre end automatisk talegenerering ud fra tekst, fordi input til et talegenkendelsessystem - det rå akustiske signal - er en uhyre righoldig blanding af forskelligartede informationer, af hvilke mange er irrelevante for identificeringen af sproglig struktur i snævrere forstand. Man kan tænke på alle de informationer, vi som menneskelige talegenkendere »hører bort fra« eller i hvert fald (afhængigt af situationskonteksten) mere eller mindre ubevidst »frasorterer« som ikke-sproglig information, nemlig de dele af signalet, der tillader os at identificere den talendes køn og hans/hendes stemmekarakteristik (der i øvrigt sætter os i stand til at identificere en person, vi kender), at afgøre om den talende er forkølet, snøvler, læsper etc., at erkende eksistensen af og måske oven i købet identificere et dialektpræg, at danne os en mening om den talendes humør, at koncentrere os om en bestemt persons ytringer, hvis flere taler samtidig, og meget andet, altsammen måske tilmed i støjfyldte omgivelser på en trafikeret gade i stormvejr. Hertil kommer det akustiske signals kontinuerte karakter.

Heroverfor står - stadig ifølge denne opfattelse - tekst som et rimeligt overskueligt medium indeholdende diskontinuerte strenge af letgenkendelige bogstaver, mellemrum og interpunktionstegn, der netop er beregnet til at symbolisere det sprogligt relevante og ikke mere. Tværtimod er mængden af information reduceret således, at mange sprogligt klart kodificerede udtryksmidler som intonationsmønstre og fremhævende trykplaceringer normalt ikke udtrykkes - og dårligt nok kan udtrykkes - i skrift.

Denne opfattelse beror efter vor mening mere på de praktiske og historiske betingelser, forskningen inden for de to områder hidtil har haft, end på teoretiske overvejelser over de kognitive aspekter ved de to former for strukturidentifikation.

Det er således noget af et paradoks, at informationsrigdom på inputsiden (som i det akustiske signal) skulle være hæmmende og informationsfattigdom (som i tekst) fremmende for en fyldestgørende strukturfortolkning.

Når forskningen inden for tekst-til-syntetisk-tale efter manges opfattelse er »længere fremme« end talegenkendelsesforskningen, er det efter vor mening kun tilsyneladende, og det kan forklares ved, at man hidtil inden for tekst-til-tale-området stort set har opereret med idealiseret input, nemlig strenge af diskrete tegn repræsenterende korrektstavede og -interpunktuerede tekster. Selve det store problem at omsætte det visuelle mønster, en tekst udgør, til en streng af tegn har normalt ikke været opfattet som hørende under tekst-til-tale området. Der er gjort betydelige fremskridt på dette område i de senere år, men selv om man nu ved hjælp af såkaldte tekstscannere stort set kan få almindelig trykt tekst i en bestemt eller nogle få veldefinerede typografiske satser omsat til tegnstrenge, er dette dog stadig ikke problemfrit. Og der er et meget langt skridt til at kunne identificere mange typografiske satser for slet ikke at tale om håndskrift med dens ikke-diskrete karakter og dens myriader af individuelle træk.

Inden for talegenkendelsesforskningen har selve diskretiseringen og normaliseringen af det rå input været det helt centrale problem. Inden for Talegenerering har man (hidtil) kunnet springe dette område over, fordi man i form af deldefinerede ASCII-strenge har kunnet simulere resultatet af en rå »tekstgenkendelse«.

Vi vil i det følgende prøve at gennemgå de forskellige problemer, der rejser sig for såvel talesyntese - specielt tekst-til-tale - som talegenkendelse på de forskellige niveauer i hesteskoen og i denne forbindelse understrege parallelliteten, der hvor den findes.

1.1. Forskning og udvikling i Danmark

Arbejdet med udvikling af syntetisk tale i videste forstand har først og fremmest været koncentreret omkring det nuværende *Institut for Almen og Anvendt Sprogvidenskab* ved Københavns Universitet *(IAAS),* hvor man har haft et nært samarbejde med *Teleteknisk Forskningslaboratorium (TFL)* især

vedrørende tekst-til-tale og generelt om syntese teknologi. Derudover har man på *Elektroniklaboratoriet* ved Danmarks Tekniske Højskole arbejdet med talekodning og tekstscanning.

Talegenkendelse har, bortset fra nogle meget tidlige forsøg med mønstergenkendelse på Elektroniklaboratoriet, været dyrket siden 1981 i samarbejde mellem *Center for Taleteknologi* ved Aalborg Universitetscenter *(CTT), JTAS* og *IAAS*. Man har i dette samarbejde først og fremmest udviklet teknologi til statistisk baseret enkeltordsgenkendelse, men der har samtidig været arbejdet seriøst på metoder til fonetisk baseret genkendelse.

2. De Input/output-nære dele af modellen

I dette afsnit omtales nogle af de problemer, der knytter sig til modellens allernederste dele, d.v.s. mediekonvertering og rene signalbehandlingsmetoder.

2.1. Talekodning og mønstergenkendelse

De simpleste talende og lyttende systemer er sådanne, som kun bevæger sig i de to nederste lag af modellen, i.e. som efter konvertering (f.eks. analog-til-digital for at få lyden ind i datamaskinen) straks går gennem signalbehandlingslaget til konvertering på outputsiden. Altså med en slags kortslutning vandret gennem hesteskoen.

Systemer af denne type kaldes talekodningssystemer. Kategorien dækker faktisk størstedelen af, hvad der i dag kaldes syntetisk tale. Det er f.eks. den slags syntesemaskiner, der »taler« fra en benzinstander eller er indbygget i de mere avancerede telefonapparater.

De signalbehandlingsmetoder, der anvendes i sådanne syntesemaskiner, kan være endog meget sofistikerede med analyse, datakompression og resyntese, typisk byggende på *Linear Predictive Coding (LPC)*, og implementeret i hurtige signalprocessorer. Men fælles for alle disse metoder er, at de arbejder ud fra generelle signalbehandlingsprincipper og uden egentlig udnyttelse af det faktum, at de behandler menneskelig tale.

På basis af hurtige signalprocessorer findes ligeledes allerede udviklet en del talegenkendelsessysstemer. De fungerer alle på den måde, at maskinen skal gennem en indlæringsperiode, hvor den præsenteres for en passende mængde eksemplarer af det talemateriale, den skal kunne orientere sig i. Genkendelsen foregår så som en simpel sammenligning af hvert indkommende akustisk mønster med de netop lagrede forlæg. Sådanne systemer kan normalt håndtere fra en snes op til et par hundrede ord samtidig.

For de mere simple systemers vedkommende kan denne indlæring foregå lige før man skal bruge systemet og gerne som noget, der er indbygget i apparatet. Brugerne gentager simpelthen ordmaterialet et par gange i den indbyggede mikrofon.

I den senere tid er man imidlertid kommet langt med metoder, der bygger på yderst komplekse statistiske betragtninger i beslutningsprocessen. Meget tyder således på, at man ved brug af såkaldte skjulte Markovkæder, kan håndtere et væsentligt større ordforråd, i.e. adskillige tusinde samtidige ord. Ulempen ved disse metoder er, at indlæringsprocessen er så beregningskrævende, at der kræves mange timers kørsel med vektorakcelleratorer og andre hurtige datamaskiner for at lære apparatet et nyt sæt termer.

For både talekodningssyntese, de simple mønstergenkendelsessystemer og de avancerede statistiske genkendelsessystemer gælder det, at de arbejder med et helt fast ordforråd. For genkendelsessystemernes vedkommende er det oven i købet som regel nødvendigt med pauser mellem ordene.

På tekst-til-tale området, altså syntese fremstillet på grundlag af tekst, svarer konvertering på inputsiden til den ovenfor omtalte tekstscanning, der foretager transformering af visuel tekst til ASCII-strenge. Signalbehandlingsdelen ville her svare til, at tekstmønstre blev transformeret direkte (f.eks. ved tabelopslag) til lydmønstre eller andre tekstmønstre, men praktiske tekst-til-tale-systemer inkluderer altid i det mindste nogen fonetisk viden og bevæger sig derfor aldrig udelukkende i de to nederste lag. (Bemærk at en proces som *tekst ind -> konvertering -> tekst ud* i denne model ville svare til det, der almindeligvis kaldes tekstbehandling).

3. Fonetiske problemer

Som fonetiske problemer betragtes alt vedrørende analyse eller produktion af konkrete lydsignaler samt relationerne mellem rent lydlige niveauer og de højere lingvistiske niveauer. Afsnittet vedrører altså modellens »fonetik« og delvis også »fonologi«.

3.1. Segmentering af lydsignalet

Et af de fundamentale problemer i behandlingen af talt sprog ved hjælp af datamaskiner er spørgsmålet om segmentering, i.e. opdeling af talen i håndterbare mindre enheder.

Set fra et datamaskinelt synspunkt er det et praktisk problem, men de bedste resutlater opnås helt klart, når segmenteringen udføres i enheder, der har en vis relation til de enheder, mennesker tilsyneladende opererer med som praktiske sproganvendere.

Talen er som bekendt en kontinuert strøm uden eksplicit markering af f.eks. ordgrænser, som vi finder dem i skreven tekst. Dette meget vanskelige og principielle problem er naturligvis grunden til, at praktisk tilgængelige talegenkendelsessystemer alle fungerer på basis af isolerede ord, i.e. med kunstige pauser mellem de enkelte ord. Selv om vi på grundlag af forskellige fonologiske/lingvistiske kriterier kan beskrive talen ved hjælp af fonemer,

d.s.v. enheder, der erfaringsmæssigt svarer til skriftens bogstaver, er der kun sjældent en entydig korrespondens mellem et bestemt akustisk tidsafsnit og et givet fonem.

Den fonetiske udfordring ved talegenerering bliver derfor at få diskontinuerte fonemsekvenser (som f.eks. kan være afledt fra tekst) omsat til de bedst mulige kontinuerte akustiske »manifestationer«, mens udfordringen ved talegenkendelse bliver at få foretaget den rigtige diskretisering af den kontinuerte talestrøm.

Først og fremmest kan det i en del tilfælde være vanskeligt overhovedet at finde egentlige segmentgrænser i det akustiske billede. (For eksempel kan man i et ord som »fjordreje« stort set kun trække en grænse omkring afslutningen af f'et, mens resten af ordet består af kontinuerte overgange). Et mindre problem er det, at en del sproglyd, f.eks. p, t og k, typisk består af flere distinkte akustiske segmenter svarende til de enkelte faser i lydenes produktion.

Når de enkelte segmenter er blevet afgrænset, eller i hvert fald nogenlunde lokaliseret, kommer så vanskelighederne med at relatere bestemte fonemer til konkrete akustiske segmenter. Også dette kan volde vanskeligheder på grund af det fænomen, der traditionelt kaldes koartikulation, d.v.s. at udtalen af en given sproglyd er bestemt af de konkrete fonetisk-fonologiske omgivelser, den befinder sig i.

Der er en tendens til at betragte koartikulationsfænomener som en form for »støj«, der bevirker, at »fonemerne« afviger fra deres ideale udtale. Dette er imidlertid på ingen måde tilfældet. Snarere er det en styrke ved den kode, vi bruger til at transmittere segmental information, at alle oplysninger om en bestemt lyd ikke er lokaliseret inden for en bestemt tid, men i stedet er bredt ud over flere akustiske segmenter. Det er sandt, at det resulterende signal bliver mere komplekst, men dette skyldes, at der i praksis sendes en del redundant information svarende til, at vi f.eks. i et ord som »spytte« allerede under s'et får oplysninger om vokalens runding til forskel fra et ord som »spilde«, hvor vokalen er urundet.

For et talegenkendende system betyder dette princip på den ene side, at systemet må håndtere et meget kompliceret og stærkt varieret inputsignal. På den anden side betyder det også, at systemet burde kunne gøres mere robust, for så vidt det kan bringes til at udnytte redundansen i koden.

Erfaringer med syntetisk tale viser, at mennesker i stor udstrækning netop udnytter bl.a. koartikulationsredundansen. F.eks. forbedres forståeligheden af helt syntetisk genereret tale mærkbart ved bedre koartikulationsregler, mens helt manglende koartikulation, svarende til at de enkelte lydsegmenter blot splejses efter hinanden, stort set gør talen uforståelig.

Det har været den traditionelle opfattelse, at mennesker i første omgang udfører analysen eller genkendelsen af et talesignal som en bottom-up procedure. Dette er naturligvis den tankegang, der ligger bag den traditionelle fonetiske transskription betragtet som en objektiv registrering af udtalen.

Meget tyder imidlertid på, at både segmentering og primær identifikation af sproglydene bygger på information fra højere lingvistiske niveauer, i.e. top-down procedurer, i meget større omfang end hidtil antaget.

På denne måde sikrer vi os i første omgang, at vi får opløst mulige tvetydige tolkninger af talesignalet. F.eks. kan man forestille sig, at den fonetiske sekvens [sbrúdåtobág] uden nogen form for top-down analyse kan have i hvert fald fem forskellige tolkninger afhængigt af omgivelserne iøvrigt:

1. sprut og tobak
2. sprutter tobak
3. sprut Otto Bak
4. sprut og tog Bak
5. sprutter tog Bak

3.2. Metoder til segmentering og klassificering

Håndsegmentering, d.v.s. manuel markering af grænser mellem akustiske segmenter, har været praktiseret i forskningsøjemed i meget lang tid, og der er en omfattende litteratur om disse problemer. Der findes en udmærket oversigt i Lehiste (1970), og Fant (1962) har mange principielle overvejelser. Endelig findes der en del undersøgelser af, hvorledes trænede fonetikere bærer sig ad med at analysere og tolke f.eks. lydspektogrammer, se f.eks. Zue & Cole (1979).

Automatisk segmentering er en nødvendighed ved talegenkendende systemer, for så vidt de overhovedet udnytter fonetisk baseret viden. Generelt bygger de simplere segmenteringsalgoritmer på en løbende analyse af talesignalet i et antal beskrivende parametre, f.eks. stemthed, antal nulgennemgange, total energi, eller energi i udvalgte frekvensområder. På grundlag af disse parametre markeres diskontinuitet over bestemte grænseværdier.

Nært koblet til segmenteringen er en primær klassifikation, der kan være rettet mod traditionelle fonetiske kategorier, som f.eks. Weinstein et al. (1975). Men der har også været gjort forsøg med at gå direkte fra akustiske spektra til leksikalske opslag (Klatt (1980)), hvorved man i nogen grad kan undgå en detaljeret fonetisk analyse.

Et af problemerne med segmentering og klassifikation, som man har været bekendt med siden starten af den manuelle analyse, er, at der ofte findes mere end én korrekt tolkning. Problemer af denne art er oplagte kandidater til løsning ved hjælp af ekspertsystemer, se f.eks. de Mori & Laface (1980) eller Erman & Lesser (1980).

Men også selve den basale segmenteringsteknik er søgt raffineret, netop fordi talesignalet i virkeligheden indeholder flere samtidige segmenteringsmuligheder. F.eks. anvender Glass & Zue (1987) hierarkisk clustering på grundlag af auditivt vægtede spektra til segmentering og klassifikation. Stra-

tegien minder om »scale space filtering« (Witkin (1984)), men er mindre beregningskrævende. Space scale filtering har iøvrigt været søgt appliceret på dansk af Tinggaard Nielsen og Tejlgaard Pedersen (1987), dog kun som analyse - ikke anvendt på egentlig genkendelse.

Det sidste nye er, at man har forsøgt sig med neurale netværk som fonetiske analysatorer, f.eks. Elman & Zipser (1988). Det er endnu for tidligt at bedømme anvendeligheden af disse metoder, der indtil videre har været forsøgt på meget begrænsede datamængder.

3.3. Klassifikation og normalisering

Et særligt problem for talegenkendelse, som ikke har nogen oplagt parallel i hidtidig praksis inden for talesyntese er det, man kunne kalde klassifikations- eller normaliseringsproblemet.

I forbindelse med klassificeringen støder man på en af talegenkendelsens mest principielle vanskeligheder, nemlig spørgsmålet om, hvorledes vi bærer os ad med at identificere en given sproglyd absolut.

Som det vil være mange bekendt, beror den fonetiske kvalitet af sproglydene på relationerne mellem taleorganernes nederste tre eller fire resonanser eller såkaldte formanter. Det har man udnyttet i lang tid til fremstilling af syntetisk tale, hvor praktiske erfaringer hurtigt har vist, hvorledes man har skullet indstille formanterne for at få noget forståeligt ud af den konkrete syntesemaskines stemme. De fleste har været tilfredse med at syntetisere en enkelt eller højst nogle få forskellige stemmekvaliteter, og man har derfor ikke behøvet at bekymre sig så meget om forskelle mellem stemmer, selv om de fleste har været klar over, at f.eks. kvindestemmer var »vanskelige« at syntetisere.

Nu viser det sig imidlertid, at det er yderst vanskeligt at forudsige, med hvilken fonetisk kvalitet en given konfiguration af formanter vil blive perciperet af en lytter. På grundlag af forholdet mellem formanterne kan man identificere grove kategorier i retning af fortunge-bagtunge eller snæveråben. Det ser imidlertid ud til, at den nøjagtigere bestemmelse foregår ud fra en slags talerspecifikt koordinatsystem, som vi i øjeblikket ikke er i stand til at beskrive nøjagtigt.

Det blev ret tidligt klart, at en vigtig parameter i korordinatsystemets fastlæggelse i hvert fald er den generelle placering af formanterne gennem en hel ytring, i.e. normaliseringen foregår ikke kun indenfor den enkelte lyd (se f.eks. Ladefoged & Broadbent (1957)). På grundlag af viden om placeringen af i hvert fald yderpunkterne i en given talers vokalrum kan man opbygge en slags statistisk model af den pågældendes udtale (se Dismer (1980) for en oversigt). Desværre er den slags modeller af ringe praktisk betydning, fordi en ny taler jo udmærket kan forstås, også selv om man ikke har hørt et repræsentativt udsnit af alle vedkommendes vokaler.

En interessant iagttagelse er det imidlertid i denne forbindelse, at de fle-

ste talere i virkeligheden er ret unøjagtige med udtalen af deres vokaler - en unøjagtighed som lytteren normalt ser bort fra, men som man kan provokeres til at være opmærksom på (Fairbanks & Grubb (1961) har nogle data for amerikansk engelsk). For et talesyntesesystem er dette fænomen en bekvemmelighed: Det er ikke vigtigt, hvor nøjagtig udtalen af de enkelte lyde er ramt, når blot de passer nogenlunde. For et talegenkendende system betyder det imidlertid, at vi må gå ud fra som givet, at udtalen i en del af inputmaterialet simpelthen vil være forkert! Taleren er nemlig vant til, at det ikke er nødvendigt at være særlig omhyggelig.

Det blev nævnt ovenfor, at det her omtalte problem inden for genkendelse ikke i praksis har nogen parallel, når det drejer sig om syntese. Dette hænger sammen med, at man i synteseforskningen hidtil har været tilfreds, hvis man blot har kunnet få en enkelt »individuel« form for syntese til at lyde nogenlunde naturligt. På længere sigt bliver systematiske former for »afnormalisering« og generering af mange individuelle syntesetyper svarende til forskellige stemmer etc. givetvis også en udfordring for synteseforskningen.

Den menneskelige lytter håndterer formentlig sådanne problemer ved ikke snævert at fiksere på den akustiske analyse men snarere inddrage information fra andre sproglige niveauer - i.e. viden om mulige ord på den pågældende plads i sætningen etc. Det vil sige en top-down analyse.

At denne strategi er den naturlige, fremgår også af, at det kræver lang tids træning at lytte »fonetisk«, hvilket altså vil sige, at man prøver at se bort fra information fra højere niveauer. (Det er tvivlsomt, om den trænede fonetiker i virkeligheden analyserer rent bottom-up. Det er faktisk tænkeligt, at man, når man lærer at transskribere fonetisk, blot lærer at erstatte et naturligt sprogs referenceramme med et kunstigt defineret sprog med unaturligt mange »fonemer«).

4. Den midterste del af hesteskoen

Generelt kan det hævdes, at de midterste dele af hesteskoen udviser de største ligheder mellem de to inputaspekter, tale og tekst. Dette hænger sammen med, at det er i dette område, de i traditionel forstand »rene« lingvistiske discipliner fonologi, morfologi og syntaks er relevante for begge forskningsgrene. Uanset beskaffenheden af input er opgaven i den midterste del af hesteskoen i venstre side at tilordne de fra inputenden kommende størrelser til de sproglige strukturer, de antages at repræsentere, og i højre side at »manifestere« strukturerne i passende overfladesekvenser.

På dette område er der - eller bør der være - en vekselvirkning mellem såvel

(1) almene lingvistiske teorier inden for de tre nævnte discipliner
(2) datalogisk og datalingvistisk teori og praksis

(3) kognitionsforskning i det omfang denne skiller sig klart ud fra (1) og (2).

Den tværvidenskabelige orientering inden for taleteknologien, som her er skitseret, og som hesteskomodellen illustrerer, er af relativt ny dato, og inden vi uddyber ovennævnte punkter, vil vi skitsere vor opfattelse af den historiske udvikling inden for området, der har ført frem til den aktuelle situation. Vi vælger tekst-til-tale området som eksempel, fordi behovet for tværvidenskabeligt samarbejde vel nok *i praksis* er mere følelig på dette område end på talegenkendelsesområdet, hvor bestræbelserne som nævnt, bortset fra de rene forsøgsmodeller, har været mere koncentreret om de nederste ender af hesteskoen. Men der er ingen tvivl om, at dette behov om få år vil være mindst lige så stort, når det gælder genkendelse - en udvikling der her i landet kan ses i det før nævnte samarbejde mellem bl.a. *CTT* og *IAAS* og mellem *TFL* og *IAAS*.

4.1. Tekst til tale

Det er karakteristisk for det hidtidige forløb af tekst-til-tale-forskningen, at den i hvert fald i Europa først og fremmest har været drevet af telekommunikationsingeniører og i de fleste tilfælde har haft klart applikationelt sigte (f.eks. hjælpemidler for syns- og talehandicappede).

Dette forhold har haft betydning på to vigtige punkter:

For det første er man i mange tilfælde gået den kortest mulige vej og har i vidt omfang ignoreret - eller ikke været klar over eksistensen af - relevante lingvistiske faktorer, som har stor indflydelse på udtaleforhold, men som i de fleste sprog ikke uden videre lader sig uddrage af almindelig ortografi (ganske særligt tryk- og intonationsforhold). Når dette har ført til tålelige og til tider imponerende resultater, skyldes det givetvis, at man er tilbøjelig til at være ovebærende over for syntetisk tale, dels ubevidst fordi syntetisk tale - i hvert fald når den er bedst - *forståelsesmæssigt* ligger på linie med naturlig tale under dårlige betingelser, idet det der mangler i kvalitet kan opfattes som »støj«, dels mere eller mindre bevidst under indtryk af, at det under alle omstændigheder er noget af en bedrift at omsætte bogstavsekvenser til forståelige akustiske signaler, uden at menneskelige taleorganer har været indblandet i processen.

For det andet har den teknisk orienterede tilgang til problemerne betydet, at resultatet af tekst-til-tale-teknikken har varieret temmelig meget afhængigt af, hvilket specifikt sprog man arbejdede med. Dette skyldes, at en mekanisk transformation af tekst til syntetisk tale uden inddragelse af dyberegående lingvistisk viden simpelthen er lettere at etablere for nogle sprog end for andre. Det er således *fra et rent teknisk synspunkt og uden dyberegående analyse* langt simplere at transformere tekst til syntetisk tale i sprog som finsk, italiensk og fransk, end det er i f.eks. dansk, tysk, engelsk eller hol-

landsk, alene af den grund at der i de førstnævnte sprog er langt større isomorfi mellem det grafemiske og det fonemiske system på enkeltordsniveau. Det er da også betegnende, at et af de første tekst-til-tale-systemer af tålelig kvalitet var finsk.

Udviklingen af de første tekst-til-tale-systemer var naturligvis forbundet med megen empirisk og eksperimentel forskning inden for akustisk fonetik og inden for digital signalbehandling samt signalgenerering, d.v.s. talesyntese i snævrere forstand (emner som er for specielle til, at vi kan komme nærmere ind på dem i denne sammenhæng), og det at denne forskning tilsyneladende ret hurtigt gav forbløffende gode resultater, i hvert fald for de »nemme« sprog, jfr. ovenfor, skabte igennem 1970'erne en vis optimisme blandt specialister med hensyn til mulighederne for inden for en overskuelig fremtid at skabe den fuldendte »højtlæsemaskine«, hvis output ikke var til at skelne fra menneskelig tale.

Igennem den sidste halve snes år er man imidlertid i stigende grad blevet opmærksom på, at mange af de kvalitative mangler ved de første systemer ikke udelukkende - og ikke først og fremmest - var et spørgsmål om manglende finpudsning af tekniske detaljer, men om at man ikke havde taget nævneværdigt hensyn til den dybereliggende sproglige strukturs betydning for talen.

At integrering af lingvistisk information af mere abstrakt art end bogstavsekvenser er nødvendig for følelige forbedringer af tekst-til-tale systemer erkendes i dag af de fleste eksperter (cf. Fisher (1984) og Molbæk Hansen (1985)), og det er nu ved at blive almindeligt, at ingeniører, fonetikere og lingvister finder sammen i mere eller mindre veletablerede forskergrupper, der i vidt omfang er grundforskningsorienterede.

Vi skal i det følgende uddybe det indhold, vi mener, der bør være i en sådan forskning, og dermed vende tilbage til punkt (1) - (3) ovenfor.

4.2. Teori og praksis i de midterste dele af hesteskoen

Idealet i enhver form for simulering af kognitive processer må være at sørge for ikke alene at få et givet input (hvad enten det er tekst eller tale) konverteret til et output, der er isomorft med det tilsvarende menneskelige »output«, men også at få denne konverteringsproces til at være isomorf med den tilsvarende kognitive proces.

Selvom dette ideal langfra er nået og næppe inden for en overskuelig fremtid kan nås, må idealet efter vor mening aldrig slippes af syne. Men problemerne med at tilnærme sig idealet er mange og store, ikke mindst fordi vor viden om den kognitive strukturering af sprogligt materiale må siges at være ringe (vi føler os ikke kompetente til at bevæge os ind på sprogpsykologiens område, men det er klart, at resultater fra dette forskningsområde må være yderst relevante i denne sammenhæng). Det næstbedste må være at integrere moderne lingvistisk teori i processerne, så godt det lader sig gøre, og selv

om der på det lingvistiske område er megen spekulativitet og mange teoridannelser, er det formentlig umagen værd at forsøge at integrere i hvert fald de lingvistiske begreber og strukturprincipper, der er rimeligt veldefinerede og formaliserede eller formaliserbare.

Af sådanne strukturprincipper, der uanset skoledannelse må siges at være alment anerkendte, kan nævnes i hvert fald to:

(1) Sproglig struktur fremkommer ved, at leksikalske enheder - på udtryksiden morfofonemer og/eller fonemer, på indholdssiden morfemer og/eller ord - fra et principielt finit korpus, et »leksikon« - sættes sammen efter bestemte regler.
(2) Der er en hierarkisk opbygning af det materiale, en sproglig ytring består af: morfofonemer og fonemer opbygger udtrykssiden af indholdsenhederne, og disse - frie og bundne morfemer - indgår i ord, disse i sætningskonstituenter, disse i sætninger, og sætninger i større diskursenheder.

På hvert af disse niveauer vil der være større eller mindre affinitet mellem bestemte udtaleforhold og bestemte enheder: således vil en sætnings opbygning af lydsegmenter være et samspil mellem de i ordene indgående morfemers udtryksbyggesten - (morfo)fonemer - og fonologiske regler i sproget, mens en sætnings prosodiske struktur (dens tryk- og intonationsmønster) i vidt omfang vil være bestemt af dens syntaktiske struktur og dens semantiske indhold.

Set fra et processeringssynspunkt vil der være to vigtige opgaver for et automatisk system: 1) *identificering af leksikalske enheder* og 2) *strukturfortolkning af de identificerede enheder.* Med hensyn til 1) er forholdene her noget forskellige for talt og skrevet input, idet segmenteringsproblemet, der af let forståelige grunde er nært forbundet med identifikationen, i hvert fald i praksis er meget større for talt input end for skrevet input (jfr. afsnittet om segmentering), eftersom der ved skrevet input foreligger en - som oftest pålidelig - grovsegmentering i form af, at der er mellemrum og interpunktionstegn, og i form af at i hvert fald trykskrift er diskontinuerte strenge af tegn.

Men selv ved en korrekt segmentering af de minimale »udtryksenheder« (bogstaver og lyde) er der ved begge former for inputmedium et andet problem: der er ofte flertydighed med hensyn til identifikationen af leksikalske enheder. Denne form for flertydighed kan som regel kun resolveres i forbindelse med strukturfortolkningen, og i automatiske systemer er det derfor almindeligt, at de to opgaver - identifikation og strukturfortolkning - udføres ved hjælp af en samlet algoritme - en såkaldt parser. Vi vil imidlertid af fremstillingsmæssige grunde i det følgende behandle identifikationsproblematikken isoleret.

4.3. Leksikalsk identifikation

Ved enhver genkendelse af sprog, uanset om der er tale om talt eller skrevet input, er der, efter at den primære analyse er overstået, behov for at relatere de fundne enheder til en liste over mulige morfemer og ord, et såkaldt leksikon. For så vidt der er tale om input på maskinlæsbar form, hvad der som regel vil være tilfældet ved maskinoversættelse eller et tekst-til-tale system, er problemet af beherskede dimensioner. I tilfælde, der samtidig omfatter en mediekonvertering, f.eks. fra lyd til maskinlæsbar repræsentation, stiller sagen sig noget anderledes, eftersom konverteringen nødvendigvis må implicere en fortolkning med heraf følgende muligheder for at introducere nye fejl.

4.3.1. Fejltyper

Principielt kan man, ved leksikalske opslag med lyd eller anden fejlbehæftet repræsentation som input, skelne mellem to typer fejlmuligheder i den leksikalske søgeproces: Syntagmatiske fejl og paradigmatiske fejl.

Paradigmatiske fejl er sådanne tilfælde, hvor den primære analyse giver et forkert bud på identiteten af enheden i inputsignalet, f.eks. i form af en identificeret vokalkvalitet, der afviger fra talerens intenderede vokalkvalitet.

Som syntagmatiske fejl betragtes de tilfælde, hvor den primære analyse identificerer for få primære enheder i forhold til den ideale repræsentation, f.eks. oversprungne konsonanter eller sammentrukne stavelser. Principielt må for mange identificerede primære enheder også betragtes om en syntagmatisk fejltype. Denne fejltype vil dog i forbindelse med talegenkendelse hyppigst hidrøre fra oversegmentering eller ikke fuldført arbejde i den primære analyse og burde derfor være af marginal interesse på dette område. Inden for tekst-til-tale er det derimod en hyppig fejltype, når det gælder identifikationen af morfemer i ord, jfr. nedenfor.

En særlig form for syntagmatisk fejl er forkert rækkefølge, hvor de enkelte segmenter i og for sig alle er til stede, men blot ikke på deres normale pladser.

Når det drejer sig om identifikation af udtryksenheder, kender vi ved input i skreven form alle disse fejltyper som »trykfejl« eller til nød som læsefejl, hvis analyse stort set har været betragtet som irrelevant for tekstanalysen.

Ved input i talt form har der på samme måde været en overbevisning om, at det principielt kunne lade sig gøre at lave en ideal fonetisk transkription alene på grundlag af det lydlige signal. Og hvis bare man kunne lære sprogbrugerne at lade være med at sige ordene forkert, skulle det hele nok komme til at hænge sammen. Der er imidlertid en voksende erkendelse af, at i hvert fald udtalefejlene er så faste bestanddele af normalt talesprog, at det er nødvendigt at tage hensyn til dem ved den leksikalske søgning.

Hverken syntagmatiske eller paradigmatiske fejl kan altså alene betragtes som symptomer på dårlig primær analyse. Selv ikke den bedste primære analyse kan finde segmenter, der ikke er blevet udtalt - et fænomen der optræder meget hyppigt på dansk, ligesom identifikationen af de enkelte segmenter aldrig kan ventes at nå til større nøjagtighed end den, hvormed taleren har produceret lydene.

4.3.2. Leksikonnets udseende

Med ovenstående betragtninger in mente kan vi sige en del om, hvorledes leksikonnet ved genkendelse af talt sprog må være opbygget. For det første ved vi, at det ikke er nødvendigt at have en fuldt korrekt fonetisk analyse af et givet ord for at kunne finde det via den fonetiske indgang. For det andet ved vi, at hele genkendelsen af et ord styres af meget andet end blot den fonetiske genkendelse.

Alle sprogbrugere kender fænomenet med at kunne huske f.eks. rytmen i et ord, men iøvrigt ikke kunne komme på ordet. Meget tyder på, at netop rytme eller trykforhold kan være meget væsentlige søgeparametre. Samtidig ved vi, at man sagtens kan finde et ord, der er blevet udtalt forkert. Vi kan oven i købet rette udtalen og sige, hvordan det burde have været udtalt.

Man har i psykofysiske eksperimenter, se f.eks. Pisoni et al. (1985), kunnet påvise, at et ords bekendthed påvirker den hastighed, hvormed ordet genfindes fra leksikonnet. Således genkaldes hyppigt forekommende ord lettere end sjældne ord, ligesom ord, der allerede har været brugt i samtalens løb, genkaldes lettere end helt nye ord. Endvidere ser det ud til, at også ordenes semantiske indhold kan påvirke et givet ords villighed til at dukke frem. Herudover er man begyndt at gøre sig overvejelser over, hvor nødvendig en detaljeret fonetisk analyse egentlig er, for at et givet ord kan genkendes (se f.eks. Larar (1986)), og det viser sig, at ved blot nogenlunde store ordforråd er den fonologiske udnyttelsesgrad (i.e. forholdet mellem de teoretisk mulige fonemkombinationer i det pågældende sprog og de fonemkombinationer, der faktisk findes udnyttet som ord i sproget) så lav, at man i mange situationer kan genkende ord på grundlag af meget lidt fonetisk information.

Alt i alt kan man sige, at hele den leksikalske genkaldelse er stærkt styret af, hvad der i den givne situation er mest sandsynligt ud fra lydlige, syntaktiske og semantiske kontekstforhold. Dette svarer godt til den almindelige iagttagelse, at folk hører det, de venter at høre!

I det foregående har vi især berørt problemerne ved leksikalsk identifikation af *minimale udtryksstørrelser* og især i forbindelse med talt input. Her skal gives et par eksempler på identifikationsproblemet i forbindelse med *minimale indholdsssstørrelser* ved tekstinput.

Et klasissk eksempel fra dansk - som ofte har været fremdraget i anden sammenhæng - er ordet »trækvinde«, som udover de to oplagte leksikalske

»segmenteringer« »træ-kvinde« og »træk-vind-e« også ud fra et rent identifikationelt synspunkt kunne repræsentere sekvensen »træk-vin-de«, idet det forhold, at den leksikalske enhed »de« (pers. pro. 3. pers. plur.) normalt kun kan forekomme som monomorfemisk ord, strengt taget ikke har noget med identikationen af selve strengen at gøre.

Nu indrettes algoritmer til morfologisk analyse ganske vist oftest sådan, at muligheden »træk-vin-de« vil blive afvist allerede ved identifikationen, idet de »leksikonopslagsprocedurer«, der undersøger mulighederne for identifikation, som regel vil behandle oplysninger af typen »kan kun optræde som monomorfemisk ord« som leksikalske egenskaber, der indgår i bedømmelsen af, om morfemet kan accepteres som en mulighed på den pågældende plads i input; men i mange tilfælde er det ikke oplagt, hvilke egenskaber ved en leksikalsk enhed, der på denne måde kan eller bør »leksikaliseres«:

I ord som »skole« og »sømand« er de ikke-trivielle fortolkninger hhv. »sko-le« og »søm-and« således vanskelige at afvise på leksikalsk grundlag: der er hverken morfosyntaktiske eller morfosemantiske kriterier for ikke at fortolke sådanne ord som sammensatte, jfr. eksistensen af »sko-børste« og »lyng-le« samt af »søm-kasse« og »avis-and« (selvom der ved en nærmere analyse af sammensætninger på dansk kan opstilles visse regler for morfosemantiske restriktioner i kombinabiliteten af leksemer i sammensatte ord (se f.eks. Bauer (1978)), er det givet, at der vil være et meget stort antal potentielle sammensætninger, der kun ved inddragelse af pragmatiske kriterier kan siges at være umulige eller ikke-eksisterende).

I et tekst-til-tale system er den pragmatisk »rigtige« fortolkning af sådanne ord ofte uhyre vigtig for genereringen af den korrekte syntetiske »udtale«, jfr. den stærkt forskellige udtale af de forskellige strukturer af de nævnte ord.

Som et morfologisk eksempel på strukturel flertydighed af den art, der skyldes, at samme morfemsekvens kan repræsentere forskellige strukturer på højere niveau - og dermed sammenhørende forskelige udtaler - kan nævnes en ordform som *undergraver.* Den almindeligste brug af denne ordform vil være som nutidsform af verbet »at undergrave«. Men lad os se på to andre mulige fortolkninger.

Ved fortolkningen »person der driver undergravende virksomhed« vil ordet blive udtalt med stød på tredje stavelse og kan tillægges denne morfosyntaktiske struktur (her udtrykt ved hjælp af såkaldt »labelled bracketing«):

$$[\,[\,[\,\textbf{under}\,]_{PRÆFIKS}\,[\,\textbf{grav}\,]_{VB.ROD}\,]_{VERBAL}\,[\,\textbf{er}\,]_{SUFFIKS}\,]_{NOMINALSTAMME}$$

mens det fortolket som »graver i underordnet stilling« (f.eks. ved større kirkegård) udtales uden stød og kan tildeles strukturen:

$$[\,[\,\textbf{under}\,]_{PRÆFIKS}\,[\,[\,\textbf{grav}\,]_{VB.ROD}\,[\,\textbf{er}\,]_{SUFFIKS}\,]_{NOMINALSTAMME}\,]_{NOMINALSTAMME}$$

(Man kunne hævde, at der er tale om to forskellige men homografe præfikser »under1« og »under2«, der forbinder sig med hver sin stammetype; eksemplet ville så være et eksempel på den førstnævnte form for flertydighed (forskellige morfemsekvenser); men problemet for en automatisk strukturfortolkning ville stort set være det samme).

Fra det syntaktisk-semantiske område kan nævnes eksemplet, *han stod på bussen,* med tryk på »stod«, og betydningen »han befandt sig stående på bussen« (f.eks. på bussens tag) eventuelt tildelt en syntaktisk (overflade)struktur som denne:

$$[\,[\,\textbf{han}\,]_{SUBJ}\,[\,\textbf{stod}\,]_{FINIT\ VERB}\,[\,[\,\textbf{på}\,]_{PRÆPOS}\,[\,\textbf{bussen}\,]_{STYRELSE}\,]_{ADVERBIAL}\,]_{S}$$

henholdsvis uden tryk på »stod« og med den almindeligere betydning »han steg ind i bussen«, eventuelt tildelt en syntaktisk (overflade)struktur som

$$[\,[\,\textbf{han}\,]_{SUB}\,[\,[\,\textbf{stod}\,]_{VERB}\,[\,\textbf{på}\,]_{ADVERB.}\,]_{KOMPL.FIN.VB.}\,[\,\textbf{bussen}\,]_{OBJ}\,]_{S}$$

Det blev nævnt ovenfor, at de fleste lingvistiske teorier antager, at »indholdsleksikonnet« er morfembaseret, med andre ord at indholdet er morfemer (groft sagt de mindste sproglige enheder med en fast korrespondens mellem form (fonologisk/grafematisk) og indhold). Dette er givetvis også det mest tilfredsstillende fra et deskriptivt lingvistisk synspunkt.

Når det drejer sig om at simulere kognitive strukturer og processer, er det derimod ikke så oplagt at operere med et sådant strengt morfembaseret leksikon.

For det første er der ingen garanti for, at menneskelig sprogkompetance er særlig isomorf med en eller anden deskriptiv analyse.

For det andet er det sandsynligt, at visse af de fænomener som ovenfor er beskrevet som problemer for leksikalsk identifikation i et morfembaseret automatisk system er »artifacts«, der ikke afspejler noget reelt i den menneskelige sprogprocessering: Som danske læsere fornemmer vi det på ingen måde som et problem at identificere ord som »sømand« og »skole«, og dette kunnne tyde på, at vort internaliserede »leksikon« i det mindste ikke udelukkende indeholder størrelser, der ud fra en lingvistisk analyse er monomorfemiske. Det viser sig da også i praksis, at man kan simplificere tekst-til-tale systemer ved at inkorporere *in casu* både »sø«, »mand« og »sømand« i leksikon.

For det tredie er det sandsynligt, at forskellige mennesker er forskelligt indrettede på dette som på andre områder, eller måske snarere at den enkelte sprogbruger kan anlægge forskellige strategier under forskellige betingelser. Det er umuligt på grundlag af sådanne overvejelser at uddrage konkrete retningslinier for design af automatiske systemer, men sådanne overvejelser opfordrer i det mindste til at være udogmatisk.

4.4. Parsing

Som nævnt indrettes automatiske sprog-input/output systemer ofte sådan, at såvel identifikationen af leksikalske enheder som strukturtildeling af det identificerede materiale foretages af en helhedspræget algoritme, en såkaldt parser, og konstruktionen af parsere har været et meget diskuteret tema inden for datamatisk lingvistik. Datalogiske, lingvistiske og kognitive problemstillinger krydses inden for dette område:

Fra et datalogisk synspunkt er man interesseret i at designe velstrukturerede og effektive algoritmer.

Fra et lingvistisk synspunkt er man interesseret i, at disse algoritmer så nøje som muligt skal afspejle lingvistiske teorier om formen og indholdet af leksikon og grammatik. Dette kan til en vis grad gøres ved, at man - som i de såkaldte deklarative systemer - holder strukturbeskrivelsen og leksikonnet klart adskilt fra selve algoritmen, ligesom statiske data holdes adskilt fra kode også i andre edb-programmer.

Fra et kognitivt synspunkt kan man være interesseret i at skabe algoritmer, der er isomorfe med den menneskelige processering af sprogligt materiale, og et sådant synspunkt kombineret med datalogiske synspunkter afspejler sig i såkaldte procedurale systemer, hvor den lingvistiske »viden« og processeringen af et sprogligt input netop ikke holdes adskilt (f. eks. såkaldte ATN-parsere (augmenterede transitionsnetværk)).

En let læst oversigt på dansk over parsere og deres indretning uden nævneværdig vægt på kognitive synspunkter foreligger i Erlandsen (1983). Antologien King (1983) giver også et glimrende overblik over forskellige parserstrategier.

Selvom vor viden på området »menneskelig processering af sprogligt input« er yderst begrænset, findes der i dag i det mindste en vis hypotesedannelse om, hvordan vi som lyttere til tale og som læsere af tekst processerer »inputtet«, f.eks. er der visse ting, der tyder på, at den menneskelige »parsing« foregår »deterministisk« på den måde, at vi ved at læse eller lytte nogle få ord frem straks i langt de fleste tilfælde vælger den rigtige strukturering. Visse parsere, se f.eks. Sampson (1983) bygger på sådanne hypoteser og er indrettet på en måde, der afspejler denne processeringsstrategi.

Der har været en tendens til, at man ved parserteori og -design har koncentreret sig om skrevet tekst som input (og i øvrigt om syntaks som det interessanteste område). Specielt når det gælder tekst-til-tale og talegenkendelse er denne bias uheldig: der findes adskillige fænomener inden for netop disse områder, der på grund heraf er blevet mere eller mindre overset. F.eks. vil man ved tekstbaseret morfologisk analyse være tilbøjelig til at ignorere systematiske udtaleforskelle mellem ord med identisk morfologisk struktur og identisk stavemåde af de i ordet indgående morfemer: Således skelnes der f.eks. i en algoritme til morfologisk analyse af danske ord (lemmatisering, se Ruus (1977)) ikke mellem nutidsformer med og uden stød, jfr. »kommer«

vs. »hæmmer«, en forskel det naturligvis ikke er specielt vigtigt at identificere ved tekstuel lemmatisering, men som er af stor betydning i talen.

5. Den øverste del af hesteskoen

Det har været påpeget flere gange i det foregående, at alt tyder på, at menneskers behandling af sprog under normale omstændigheder foregår på samtlige hesteskomodellens niveauer, og at det kræver en betydelig abstraktionsgrad af en sprogbruger at isolere et enkelt beskrivelseslag. Desværre er vor viden om, hvorledes sproget faktisk fungerer ganske afgjort mest konkret i de nederste lag, hvilket først og fremmest afspejler sig i, at praktisk talt ingen syntese- eller genkendelsessystemer når op over syntaks-niveau i modellen, mens semantikken stort set overlades til en slags styrende applikationsorienteret komponent af meget ringe teoretisk interesse.

Vi kan kun håbe på, at den voksende interesse for datalingvistiske emner vil betyde en bedre integration af også modellens øverste lag.

6. Afsluttende betragtninger

For nu at sætte de foregående afsnit i relation til hesteskomodellen og til den aktuelle situation inden for tale-input/output området kan man sige følgende:

En rimelig kvalitet af talegenerering og -genkendelse kræver vidtgående hensyntagen til de sproglige strukturer, talen eller teksten repræsenterer.

Denne hensyntagen *kan i et vist omfang* opnås ved, at man simulerer strukturelle fænomener på højere niveauer ved at referere til deres manifestationer på lavere niveauer. Ved tekst-til-tale kan man således simulere leksikalsk betingede fænomener uden at operere med noget egentligt leksikon ved at anvende kvasifonologiske regler med den ønskede effekt: f.eks. kunne man for dansk anvende og algoritmisere et stort sæt substitutionsregler af nedenstående type (hvor lydskrift er omsluttet af [. . .]

»lidt« −> [led]
»midt« −> [med]
»skidt» −> [sgid]
»indsæt en sammensætningsgrænse i stillingen« »o r s k e _ l e v« (jævnfør »torskelever«)

o.s.v. En sådan fremgangsmåde har været anvendt i flere hidtidige tekst-til-tale systemer. (Bemærk at denne fremgangsmåde i hesteskomodellen betragtes som signalbehandling).

Tilsvarende kan man i et vist omfang simulere syntaktisk betingede leksikalske entydiggørelser ved at anvende sådanne regler, f.eks.

hul –> {hål} i stillingen efter »mellemrum + e + t + mellemrum« (jævnfør »et hul«)

ellers

hul –> {hu:?l} (jævnfør »en hul røst«)

Man kunne givetvis også finde eksempler på, at semantisk-pragmatisk betingede syntaktiske entydiggørelser kunne simuleres på tilsvarende måde, men det må være klart på baggrund af de allerede givne eksempler, at jo større niveauforskel, der er mellem de sproglige fænomener, det drejer sig om, og de enheder de udtrykkes i, desto mere arbitrært og lingvistisk-kognitivt uinteressant bliver det at søge problemerne løst på denne måde. I hesteskomodellen svarer disse »lappeløsninger« til, at man så at sige kortslutter venstre og højre side af hesteskoen eller forbinder venstre og højre side med vandrette linier på lavere niveauer.

Vi opfatter det som langt vigtigere på forskningens nuværende stade at sikre, at de sproglige fænomener, der udspiller sig på forskellige niveauer, udtrykkes i termer, der er kommensurable med deres tema, ikke mindst fordi det synes at gælde generelt, at flertydighed på lavere niveauer kun kan entydiggøres systematisk ved inddragelse af viden fra højere niveauer, hvad ovenstående eksempler viser.

En sådan samlet betragtning vil i den yderste konsekvens betyde, at vi kan begynde at tale om tale*forståelse* i stedet for blot om *genkendelse,* svarende til, at vi kommer hele vejen rundt i hesteskomodellen.

REFERENCER

BAUER (1978): L. Bauer: *The Grammar of Nominal Compounding.* Odense University Press, 1978.

DE MORI & LAFACE (1980): Renato de Mori and Pietro La Face: »Use of fuzzy algorithms for phonetic and phonemic labeling of continuous speech«. IEEE PAMI 2, 2, 1980, pp. 136-148.

DISNER (1980): Sandra Ferrari Disner: »Evaluation of vowel normalization procedures«. *J.Acoust.Soc.Am.* 67, 1, 19809, p. 253-261.

ELMAN & ZIPSER (1988): Jeffrey L. Elman and David Zipser: »Learning the hidden structure of speech«. *J.Acoust.Soc.Am.* 83, 4, 1988, pp. 1615-1626.

ERLANDSEN (1983): Jens Erlandsen: »En introduktion til parsing og parseren GESA«. *Skrifter om Anvendt og Matematisk Lingvistik* 10, 1983, p. 101-157.

ERMAN & LESSER (1980): Lee D. Erman and Victor R. Lesser: »The Hearsay-II speech understanding system« in *Trends in Speech Recognition.* Wayne A. Lea (ed.), Prentice-Hall, 1980, pp. 361-381.

FAIRBANKS & GRUBB (1961): Grant Fairbanks and P. Grubb: »A psychological investigation of vowel formants«. *Journal of Speech and Hearing Research* 4, 1961, p. 203-219.

FANT (1962): Gunnar M. Fant: »Descriptive analysis of the acoustic aspects of speech«. *LOGOS* 5, 1, 1962, pp. 3-17.

FISHER (1984): William M. Fisher: »Headaches, Moustaches and Coaches - or why it's Difficult to Build a Good Reading Machine«. *Speech Technology* 2, 3, 1984.

GLASS & ZUE (1987): »Acoustic segmentation and classification« preprint 1987, pp. 6.

KING (1983): M. King: *Parsing Natural Language.* Academic Press, 1983.

KLATT (1980): Dennis H. Klatt: »Scriber and LAFS: Two new approaches to speech analysis« in *Trends in Speech Recognition.* Wayne A. Lea (ed.), Prentice-Hall, 1980, pp. 529-555.

LADEFOGED & BROADBENT (1957): Peter Ladefoged and D.E. Broadbent: »Information conveyed by vowels«. *J.Acoust.Soc.Am.* 29, p. 98-104.

LARAR (1986): J.N. Larar: »Lexical access using broad acoustic-phonetic classifications«. *Computer Speech and Language* 1, 1986, p. 47-59.

LEHISTE (1970): Ilse Lehiste (ed.): *Suprasegmentals* MIT Press, 1964, 194 pp.

MOLBÆK HANSEN (1985): Peter Molbæk Hansen: »Kanel, model, adel, sadel - og andre vanskeligheder ved at lære en maskine at læse højt«. *Teleteknik* 4, 1985, p. 198-207.

PISONI ET AL. (1985): David B. Pisoni, Howard C. Nusbaum, Paul A. Luce and Louisa M. Slowiaczek: »Speech recognition, word recognition and the structure of the lexicon«. *Speech Communication* 4, 1985, op. 75-95.

RUUS (1977): Hanne Ruus: »Ordmekanik«. *Skrifter om Anvendt og Matematisk Lingvistik* 3, 1977, p. 79-106.

SAMPSON (1983): G.R. Sampson: »Deterministic Parsing«. *Parsing Natural Languages* (M. King (ed.)), Academic Press, 1983, p. 91-116.

TINGGAARD NIELSEN & TEJLGAARD PEDERSEN (1987): Kim Tinggaard Nielsen og Jens Tejlgaard Pedersen: *Automatisk talegenkendelse* specialeafhandling DIKU 85-1-1, 1987, 130 pp.

WEINSTEIN ET AL. (1975): Clifford J. Weinstein, Stephanie S. McCandless, Lee F. Mondshein and Victor W. Zue: »A system for acoustic-phonetic analysis of continuous speech«. *IEEE ASSP* 23, 1, 1975, pp. 54-67.

WITKIN (1984): A.P. Witkin: »Scale space filtering: A new approach to multi-scale description«. *IEEE ICASSP* 1984.

ZUE & COLE (1979): Victor W. Zue and R.A. Cole: »Experiments on spectrogram reading«. *IEEE ICASSP* 1979.