

Notat / Memo

BRAGE

Brukerdialoger med naturlig tale



Standard for ortografisk annotering av spontan tale i BRAGE

Notat nr./Memo No: Annotering nr 2

Dato/Date: 2003-04-30

Status: Utkast

Ingunn Amdal, Knut Kvale, Telenor R&D

Versjonshistorie

VERSJON	DATO	KOMMENTAR/ENDRINGER
1.0	2003-02-13	Forslag til standard for merking av spontan tale i BRAGE basert på SpeechDat med noen utvidelser. Dialog-annotering ikke fullført
2.0	2003-04-30	Oppdatert standard for merking av spontan tale i BRAGE etter erfaring med merking av MOBEL. Dialog-annotering utsatt og derfor tatt ut.

Innhold

1	Innledning	3
1.1	Retningslinjer for standarden	3
1.2	Bakgrunn	4
1.3	Ikke med i standarden	5
2	Valg av ordformer	6
2.1	Standard	6
2.2	Anbefalt	6
2.3	Begrunnelse	7
3	Pauser og støy fra taler	8
3.1	Standard	8
3.2	Anbefalt	8
3.3	Begrunnelse	9
4	Støy fra omgivelser	10
4.1	Standard	10
4.2	Anbefalt	10
4.3	Begrunnelse	10
5	Disfluenser	11
5.1	Standard	11
5.2	Anbefalt	11
5.3	Begrunnelse	11
6	Prosodi	12
6.1	Standard	12
6.2	Anbefalt	12
6.3	Begrunnelse	12
7	Fil-format	13
7.1	Standard	13
7.2	Anbefalt	13
7.3	Begrunnelse	14
8	Krav til merkeverktøy	15
8.1	Standard	15
8.2	Anbefalt	15
8.3	Begrunnelse	15
9	Oppsummering av bruk av skilletegn og andre spesielle tegn	16
10	Vedlegg	17
10.1	Instruksjonar for ortografisk annotering i BRAGE	17
10.2	SpeechDat merkefil (SAM-format)	23
10.3	Dialektregioner fra SpeechDat	25

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

1 Innledning

Standard for ortografisk annotering av spontan tale i BRAGE-prosjektet tar for seg:

- Valg av ordformer (uttalevariasjon)
- Pauser og støy fra taler (ikke-verbal og/eller ikke-vokal)
- Støy fra omgivelser (ikke-menneskelig)
- Disfluenser
- Prosodi
- Fil-format
- Krav til merkeverktøy

Erfaringer fra første runde av MOBEL-annotering er tatt hensyn til. Dette er i tråd med erfaringer fra "Spoken Dutch Corpus" (CGN) hvor de rapporterer tre iterasjoner med forslag til standard og erfaringer fra annotering.

Dialogmerking og merking av multimodale data er ikke tatt med i denne standarden.

I det følgende er courier font brukt for å vise eksempler på hvordan det skal merkes. Eksempelene er omslutta av hermetegn (som ikke skal være med i merkinga): " dette er et eksempel på merking i Brage ".

1.1 Retningslinjer for standarden

Det er en avveining mellom hva vi ønsker å modellere og teste, og hvor mye resurser det tar å annotere relevant talemateriale. Vi har identifisert følgende retningslinjer for valg av standard:

- **Konsistent:** Annoteringen skal følge de samme prinsippene konsekvent, slik at like situasjoner blir annotert likt.
- **Enkelt:** Det skal være mulig å annotere konsistent etter standarden for folk som ikke er eksperter på lingvistikk/taleteknologi
- **Effektivt:** Annoteringen bør utføres med et rimelig tidsforbruk.

Ut fra kravene om konsistent, enkel og effektiv annotering, er *ortografisk merking valgt som standard*. Ortografisk annotering betyr vanlig norsk rettskriving (bokmål og nynorsk). Ortografi er den enkleste og mest kjente symbolske form for å representere tale og derfor et naturlig valg.

Den ortografiske annoteringen vil seinere kunne være underlag for eventuell finere annotering (fonetisk, prosodisk, o.s.v.). Det er derfor viktig at den ortografiske annoteringen er utført nøyaktig og konsistent.

Standarden er delt i to; en "må ha" i delkapittel "standard" og en "kjekt å ha" i delkapittel "anbefalt". Vi har tatt med begrunnelser for endel av valgene så vi slipper å repetere samme diskusjon flere ganger.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

Annoteringen må verifiseres! Det kan gjøres ved å la de som annoterer kontrollere hverandre (slik f.eks CGN gjør det).

1.2 Bakgrunn

Vi bruker SpeechDat som utgangspunkt for å definere BRAGE annoteringsstandard. En viktig grunn til å foretrekke SpeechDat er at denne allerede er brukt for norsk og vi har "norsk" tolkning av standarden samt erfaringer med den. En annen er at SpeechDat er forholdsvis edruelig i forhold til hvor mye som skal merkes. SpeechDat er ikke spesielt tilpasset spontan tale og det er derfor anbefalt noen tillegg.

I Tabell 1 har vi lista et utvalg relevante organisasjoner med tilhørende web-side (besøkt 2003-04-11).

Erfaringene fra CGN (Spoken Dutch Corpus) er også tatt med som underlag for denne standarden. De identifiserer fire mål for merkingen:

1. CGN følger i stor grad internasjonale standarder (EAGLES, SpeechDat, LDC, CHILDES)
2. Transkripsjonen skal være ordrett og derfor inneholde minimalt med tolkning.
3. Talen skal deles opp i relativt små deler, 3 sekunder eller mindre.
4. Den resulterende ortografiske transkripsjonen skal være brukbar for målgruppa; tale og språk-teknologer, lingvister, fonetikere o.s.v.

EAGLES anslår at transkribering tar 3-10 ganger varigheten av talen alt etter hvor spontant det er snakka.

Akronym	Web-adresse
EAGLES (Expert Advisory Group on Language Engineering Standards)	http://lingue.ilc.pi.cnr.it/EAGLES96/home.html
SpeechDat	http://www.speechdat.org/
CGN (Spoken Dutch Corpus)	http://lands.let.kun.nl/cgn/ehome.htm
LDC (Linguistic Data Consortium)	http://www ldc.upenn.edu/readme_files/switchboard.readme.html http://www ldc.upenn.edu/Projects/SPINE/index.html
CHILDES (Child Language Data Exchange System)	http://childes.psy.cmu.edu/
DAMSL (Dialog Act Markup in Several Layers)	http://www.cs.rochester.edu/research/cisd/resources/damsl/
SAMPA (Speech Assessment Methods Phonetic Alphabet)	http://www.phon.ucl.ac.uk/home/sampa/home.htm
BAS (Bavarian Archive for Speech Signals)	http://www.phonetik.uni-muenchen.de/Bas/
AGTK (Annotation Graph Toolkit)	http://agtk.sourceforge.net/

Tabell 1 Web-adresser til relevant informasjon for annotering

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

1.3 Ikke med i standarden

Her lister vi effekter/problemer som er vurdert, men ikke tatt med i standarden:

Semantisk sammenheng mellom ord

I NRK-databasen ble sammenheng mellom ord markert med "_". for eksempel "Karl_Glad" og "Ringerike_Grand_Prix"

Vi har ikke tatt dette med i BRAGE-standard.

Argumenter for å ha "_" mellom ord som henger sammen:

- Vi vil da reservere mellomrom som skille mellom leksikalske entiteter. Dette gjør at annoteringen kan benyttes til å generere ordlister.
- Språkmodellering?

Argumenter mot å ha "_" mellom ord som henger sammen:

- Vanskelig å gjøre konsistent, de som annoterer må vurdere hva som er leksikalske/semantiske enheter. En ting er navn, hva med "for eksempel" "i dag" og liknende begreper som består av flere ord?
- Strider mot prinsippet å følge vanlig norsk rettskriving.

2 Valg av ordformer

2.1 Standard

Detaljer med eksempler er beskrevet i instruksjonen til merkerne, se vedlegg kapittel 10.1.

Hovedpunkter:

- Annoteringa skal være ordrett, dvs inneholde ALLE orda som forekommer i talen.
- Annoteringa skal være ortografisk, dvs med vanlig norsk rettskriving, bokmål og nynorsk.
- Alle dialekter og vanlige stilistiske varianter er normal uttale av norsk. Hvis et ord har valgfrie skrivemåter i ordlista, skal den skrivemåten som likner mest på det som ble sagt brukes.
- Vi bruker Norkompleks (både bokmål og nynorsk) som referanseordliste. I tillegg må det lages ei applikasjonsspesifikk ordliste der det er nødvendig (for eksempel egennavn, ved medisinsk diktering eller utenlandske ord). Det må lages ei ordliste over alle ord som forekommer i databasen.
- Vi lager ikke uttalevarianter som ikke finnes i vanlig norsk rettskriving
- **Store og små bokstaver:** Det brukes normalt små bokstaver, store bokstaver brukes kun for egennavn, akronym og staving. (Dvs ikke for første ord i setning.)
- **Staving:** Staving skrives med store bokstaver. Når bokstavene blir uttala på vanlig måte, f.eks F uttala "eff" og C uttala "se", skriver vi " F " og " C ". Men hvis F blir uttala "fffff", skal det merkes " /f/ ".
- **Akronym (forkortinger):** Vanlige akronym skrives i ett ord slik som f.eks " SAS ". Dette gjelder både for uttale "sass" og "ess a ess". For sjeldne forkortinger brukes staving. Alle akronym som skrives i ett ord må finnes i ordlista (evt legges inn i den applikasjonsspesifikke ordlista).
- **Tall:** Alle tall skal skrives ut med ord som i vanlig norsk rettskriving.
- **Utenlandske ord:** Hvis orda ikke har norsk skrivemåte, så skriv ordet på original-språket (ordet må da legges inn i den applikasjonsspesifikke ordlista).

2.2 Anbefalt

Uttaleform(er) bør være inkludert i ordlista. Disse bør bruke SAMPA.

Punktum, utropstegn eller spørsmålstegn kan brukes for å markere slutt på setning. Andre skilletegn (som komma, kolon, semikolon og tankestrek) skal ikke brukes.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

2.3 Begrunnelse

Vi lager ikke uttalevarianter som ikke finnes i vanlig norsk rettskrivning.

- Argumenter for å lage uttalevarianter som ikke finnes i vanlig norsk rettskrivning:
 - o Anbefaling 3 fra EAGLES
 - o Spontan tale vil ha mer uttalevariasjon
 - o Vil lette arbeidet med å lage fonetisk transkripsjon før trening av talegjenkjennerne
- Argumenter mot å lage uttalevarianter som ikke finnes i vanlig norsk rettskrivning:
 - o Ingen av de brukte standardene (SpeechDat, CGN, LDC) gjør det
 - o Norske grafem-til-fonem-regler er ikke en-til-en. Det vil være vanskelig å gjøre en slik "halvfonetisk" transkripsjon konsistent
 - o Vil ta lengre tid å merke.
 - o Krever erfarne folk med fonetisk skolering til å annotere
 - o Vanskelig å definere kriterier for valg av uttalevarianter som skal merkes, f.eks :
 - Hvilke uttalevariasjoner skal transkriberes og hvilke skal ikke?
Dersom kriteriet er at uttalevarianter er "vanlig", (må en definere hvor ofte denne uttalevarianten skal forekomme for at den skal regnes som "vanlig"?)
 - Dersom en skal bruke fonetiske og diakritiske symbol, må settet av disse defineres, og en må fastsette hvilket detaljeringsnivå skal en følge. Dvs det må defineres hvor fonetisk nøyaktig skal en transkribere.
 - Dersom en skal annotere "halv-fonetisk" eller "kvasi-lydskrift" med vanlig bokstaver, må det defineres hvordan dette skal utføres og deretter hvordan det tolkes av lesere etterpå.
Eksempel: Ordet "morgenen" kan bli uttala /m O: O rn =n/ som med SAMPA-notasjon er ganske entydig. Men med bokstaver kan dette skrives f.eks "måårnn", noe som åpner for tolkninger
 - o Gjør automatisk kontroll av brukte ordformer umulig

Vi merker heller ikke reduserte former av samme grunn (f.eks uttalen "ække" skal transkriberes "er ikke")

For uttalevarianter som ikke er vanlig i bruk, brukes " *ord " som for feiluttale, se kapittel 5. Merkerne oppfordres til å skrive kommentarer i loggen for slike ord.

3 Pauser og støy fra taler

3.1 Standard

Vi bruker samme merker som i SpeechDat:

- [fil] - "filled pause". Fylte pauser er nølelyder som f.eks "ææh", "øh" og "mm".
- [spk] - "speaker noise". Støy som talaren lager, som f.eks smatting, hosting, kremting, latter, sukking eller høglytt pusting (både inn og ut pust).

Se vedlegg kapittel 10.1 for nærmere forklaring med eksempler.

Disse merkene vil normalt være mellom ord, hvis ikke skal merket settes før ordet som blir påvirket av støyen. For eksempel kan "Når går bussen fra Drag -host- voll?" merkes som " når går bussen fra [spk-p] *Dragvoll "

3.2 Anbefalt

I tillegg anbefaler vi å merke ikke-fylte pauser:

- [sil] - tydelig pause i talestrømmen, men uten nølelyd.

Anbefalt utvidelse av støymerkene:

- [fil] deles opp i
 - o [fil-e] for nølelyd som er hovedsaklig vokal-liknende
 - o [fil-m] for nølelyd som er hovedsaklig nasal-liknende
 - o [fil] kan brukes hvis ingen av kategoriene passer
- [spk] deles opp i
 - o [spk-f] for pust og "frikativ-liknende" støy
 - o [spk-p] for smatt, latter, host og "plosiv-liknende" støy
 - o [spk] kan brukes hvis ingen av kategoriene passer

3.3 Begrunnelse

Oppdeling av [fil] følger anbefaling fra EAGLES-1997.

Oppdeling av [spk] i tre kategorier er færre enn det som er "vanlig" (i LDC for eksempel). Det er ingen grunn til å kategorisere spesielle støylyder som forekommer kun få ganger og det er vanskelig å være konsistent med mange alternativer. Erfaringene fra SpeechDat er at det er mye pust, derfor er denne skilt ut som egen kategori.

I LDC-manual for WSJ er det anbefalt å merke om støyen kommer før eller etter ord: *"If a non-speech event overlaps with a spoken lexical item, the descriptor should be placed next to the lexical item it co-occured with and the character, ">" or "<" should be appended or prepended to the descriptor depending on whether it is placed to the left or right of the co-occurring lexical item."* Vi holder på SpeechDat-standardten med å merke før ordet som blir påvirket.

I LDC-manual for WSJ er det anbefalt å kun merke ikke-fylte pauser som er påfallende : *" Only conspicuous pauses are to be marked ... "*. De anbefaler bruk av punktum.

Ikke-fylt pause kunne blitt representert med " ... " for å følge SAMPA. Vi har valgt å bruke " [sil] " for å reservere punktum for avslutning av setning. Vi får da også samme type merking for fylte og ikke-fylte pauser.

4 Støy fra omgivelser

4.1 Standard

Vi bruker samme merker som i SpeechDat:

- avbrudd av tekniske årsaker: " ~ord ", " ord~ " eller " ~ord~ "
- [sta] - "stationary noise". Stasjonær bakgrunnstøy med tilnærma konstant amplitude over tid, som f.eks bil-støy, trafikk-støy, kanal-støy eller konstant bakgrunnstale (fest eller offentlige rom som f.eks flyterminaler).
- [int] - "intermittent noise". Kortvarig eller rykkvis støy kan være et brak (f.eks ei dør som blir slått igjen, eller ei dørklokke), periodisk (f.eks telefonringing eller hunde-bjeffing) eller støy med varierende styrke over tid (f.eks musikk). Andre situasjoner som bør merkes med [int] er: ujevn bakgrunnstale, unge-skrik, krysstale og papir knitring.

Se vedlegg kapittel 10.1 for nærmere forklaring med eksempler.

4.2 Anbefalt

Ved å bruke start "[merke/]" og stopp "[/merke]" kan vi få merke varigheten til støy samt skille på støy som bare påvirker endel av ytringen og ikke hele.

Anbefalt utvidelse:

- bruk "ord0 [sta/] ord1 ord2 ord3 [/sta] ord4" for stasjonær støy samtidig som flere ord
- samme med [int/] og [/int] (men [int] vil oftest være mellom ord)

[sta] og [int] alene vil da bety støy som ikke varer lenger enn neste ord.

4.3 Begrunnelse

Erfaringene med SpeechDat var at det kan være behov for å merke varigheten til bakgrunnsstøy, derfor dette tillegget. Men det kan være vanskelig å avgjøre når støyen starter og slutter. Det kan derfor bli tidkrevende å merke start og slutt i taleopptak med varierende bakgrunnsstøy.

Bruk av start og stopp er i tråd med LDC-manualen for WSJ: "*If a phenomenon is noted throughout, or co-occurs with, more than one lexical item, then the phenomenon's descriptor is to be used in the following notation to bound the lexical items it spans: [descriptor/] word word ... word [/descriptor].*"

5 Disfluenser

5.1 Standard

Vi bruker samme merker som i SpeechDat:

- " *ord " for feiluttale
- " ** " for uforståelig ord

Se vedlegg kapittel 10.1 for nærmere forklaring med eksempler. Merkerne oppfordres til å skrive kommentarer i loggen for disse tilfellene.

5.2 Anbefalt

Anbefalt tillegg:

- " -(manglende del av ord)uttalt del av ord " eller " uttalt del av ord (manglende del av ord)- " ved avbrudd av taler. For eksempel: " sta(sjon) - ". Når en fjerner parentesene skal ordet fortsatt følge vanlig rettskrivning.
- " *ord " skal fortsatt brukes for annen feiluttale av ord
- " ** " skal fortsatt brukes hvis det ikke er åpenbart hvilket ord taleren vil si
- merke repetisjon/korreksjon ved å sette "<>" rundt ord(ene) som blir repetert/korrigert slik at det som er igjen er en meningsfylt ytring. Merke for repetisjon kan inneholde ett eller flere ord.
Eksempel: " når går <toget nei> bussen fra Trondheim "

5.3 Begrunnelse

Merking av repetisjoner og korreksjoner er nyttig for f.eks dialoganalyse. En kan også enkelt fjerne <> ved behov. Men merking av repetisjon kan ta for mye tid evt være vanskelig å gjøre konsistent. Skjønn må brukes på om det er relevant for applikasjonen.

Merking av repetisjon/korreksjon er i tråd med LDC-manual for WSJ: " Words which are verbally deleted - replaced with other words by the subject later in the utterance - are to be enclosed in angle brackets, "<>":
The plane dropped <quickly> <uh> precipitously into the boiling ocean below "

Etter erfaringer med MOBEL åpner vi for at parentesene kan omslutte flere ord.

6 Prosodi

6.1 Standard

I BRAGE-standarden for ortografisk merking tar vi ikke med prosodi. Tidspunkt for hver setning (ytring) skal være merket, enten ved å dele opp i ei fil per setning eller merke av tidspunkt mellom setningene.

6.2 Anbefalt

Som definert under 3.2 anbefaler vi at pauser uten nølelyd (ikke-fylte pauser) merkes med " [sil] "

Ved bruk av skilletegn åpner vi for bruk av både punktum, spørsmålstegn og utropstegn der det er naturlig.

6.3 Begrunnelse

Følgende kunne vært aktuelt å merke:

- Prominens og/eller fokus
- Frase-grenser: "|" "||"

CGN rapporterer at slik merking tar 40 ganger varigheten av talen og vi velger derfor å la være å merke dette før vi vet om vi har bruk for det.

LDC-manual for WSJ anbefaler å merke "Emphatic Stress" (med utropstegn) og "Lengthening" (med kolon).

7 Fil-format

7.1 Standard

Navne-konvensjone for merkefiler og talefiler:

- NNNN-YYYY-MM-DD_hh_mm_ss

Hvor NNNN er taler-id, YYYY-MM-DD er dato og hh_mm_ss er klokkeslett.

Merkefilene bør være i et ikke-proprietært format (dvs ikke Word f.eks). ASCII-tekst i SAM-format er OK. Se eksempel på SAM-merkefil som brukt i SpeechDat i vedlegg kapittel 10.2.

Minimumsinfo i tillegg til ortografisk merking:

- Databasenavn
- Relevant info om filformat på tilhørende filer (f.eks lydfilformat)
- Opptakstidspunkt
- Taler-id
- Kjønn
- Aldersgruppe
- Dialektregion (vi følger SpeechDat, se vedlegg kapittel 10.3)
- Merker-id

7.2 Anbefalt

XML-format (for eksempel fra MUST eller AGTK)

Tilleggsinfo:

- Opptaksforhold f.eks telefon (fast eller mobil), bruk av håndsett eller håndfri osv
- Omgivelser for opptak, f.eks bakgrunnsstøy (togstasjon, stille rom osv)
- Beskrivelse av type tale eller dialog, f.eks opplesing av manuskript, diktering, WoZ, publikumstjeneste osv
- Kommentar fra den som har annotert talen. For eksempel om det er spesielt mye støy eller om taleren snakker veldig fort el.l.

7.3 Begrunnelse

Vi har valgt å følge MUST-standard siden den er lagt inn i Tabulib. Legger til taler-id som i SpeechDat-standard.

- SpeechDat: A1 NNNN CC. NO F
- MUST: YYYY-MM-DD_hh_mm_ss

Det er anbefalt av EAGLES å ta med kommentarene de som har annotert har skrevet om talen de har merka. LDC har en rubrikk for gradering fra 1-5 av DIFFICULTY, TOPICALITY, ECHO, STATIC, BACKGROUND og en for kommentarer i fri tekst.

8 Krav til merkeverktøy

8.1 Standard

Krav til merkeverktøy:

- Sjekk av om merking følger ordlista.
Nye ord skal kontrolleres av andre før det godkjennes i "offisiell" ordliste
- Sjekk av om lovlige støymerker er brukt
- Merkelogg

Både CGN og BAS (merking av SmartKom) bruker verktøyet PRAAT. Dette bør vi også vurdere.

8.2 Anbefalt

Merkeverktøyet bør kunne sjekke om valgt ord finnes i leksikonet. Dette bør helst skje under merking, men bør i det minste være en enkel prosedyre som merkerne kan gjøre for hver arbeidsøkt. (CGN har lagt inn stavesjekk i PRAAT).

Støymerker bør være lett tilgjengelig, enten som knapper eller i en meny. I det minste bør ei liste over støymerker være lett tilgjengelig.

Merkerne må skrive ned kommentar for hver taler. Dette bør lagres elektronisk, men bør være så enkelt at merkerne virkelig gjør det. Dette vil være kommentarer til selve merkingen, kommentarer som er relevant om talen skrives inn i merkefilen. Eksempler på ting som skal kommenteres i merkeloggen er: nye ord som skal legges inn i ordlista, kommentarer til feiluttale (ord merka " *ord ") og tvilstilfeller i merkinga.

8.3 Begrunnelse

Jo mer som er automatisert, jo enklere er det å være konsistent.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

9 Oppsummering av bruk av skilletegn og andre spesielle tegn

Følgende tegn kan brukes i tillegg til bokstaver:

- . ? ! Punktum, spørsmålstegn og utropstegn kan brukes som skilletegn
- // Brukes for staving med fonetisk uttale
- [] Brukes ved støymerker
- <> Brukes ved repetisjoner/korreksjoner
- Brukes ved avbrudd av taler
- ~ Brukes ved signalavbrudd
- * Brukes ved feiluttale " *ord " og uforståelig ord " ** "

Det skal normalt brukes kun små bokstaver og vi bruker "ISO8859-1 character set".

Store bokstaver brukes kun for egennavn, akronym og staving.

10 Vedlegg

10.1 Instruksjonar for ortografisk annotering i BRAGE

I instruksjonen er courier font brukt for å vise eksempel på korleis det skal merkes. Eksempla er omslutta av hermetegn (som ikkje skal være med i merkinga): " dette er et eksempel på merking i Brage ".

Der tilrådingar er tatt med i tillegg til standard, er det skreive i teksten.

10.1.1 Framgangsmåte

Bruk hodetelefonar og lytt minst 2 gonger til kvar fil

1. Trinn 1: Lytt på kva for ord som blir sagt. Ved avvik mellom eventuelt merkeforslag og det som er sagt må merkeforslaget endrast slik at teksten representerar det som vart sagt. Skriv ned alle orda du høyrer og merk:

- repetisjonar og korreksjonar: " <ord1 ord2 > " (tilråding)
- feiluttale: " *ord "
- forståelig, men ufullstendig ord: " -(manglande del av ord)uttalt del av ord " (tilråding)
Pass på at dette blir et lovlig ord viss du fjerner parentesane!
- uforståelig ord: " ** "
- signalavbrudd: " ~ord "

2. Trinn 2: Lytt etter støy. Merk støyen med støy-merkene:

- [spk], [fil], [int], [sta]
- [spk-f], [spk-p], [spk], [fil-e], [fil-m], [fil], [int], [sta] (tilråding)
Sett inn start " [merke/] " og stopp " [merke/] " viss støyen varer over fleire ord
- pausar: " [sil] " (tilråding)

Grunnen til at du skal lytte minst 2 gonger til kvar fil er at det er lett å overhøyre støy og andre ikkje-språklege detaljar når du ikkje lyttar spesielt etter det.

10.1.2 Hovedprinsipp

Merkinga er ortografisk, dvs med vanleg norsk rettskriving, bokmål og nynorsk.

Det brukast normalt små bokstaver, store bokstaver brukast berre for navn, akronym og staving. (Dvs ikkje for første ord i setninga.)

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

Du skal skrive alle orda du høyrer, samt merke feiluttale, støy og signalavbrudd.

Tenk deg at du er den perfekte talegjennar!

Ein slik talegjennar vil prøve å klassifisere lydane til ord i ordlista eller merke i støymerke-lista. Han vil i tillegg kunne detektere ord som er galt uttala og ord som ikkje er i ordlista (perfekt "OOV-rejection").

Ein talegjennar må lære dei ulike uttalevariantane og dialektane i språket, f.eks at ordet "hodet" kan uttalast som "hue", "huggu", "huvvu", "haue", "haude", "hode" eller "hodet". Desse uttalevariantane skal merkast med ordet "hodet" (bokmål-setning) eller "hovudet" (nynorsk-setning).

Talegjennaren må også vera såpass intelligent at han forstår og vel rett skrivemåte på ord som uttalast likt, men som skrivast forskjellig, som f.eks hjem-gjem, sei-seg eller hvert, vært, (hus-)vert.

Alle dialektar og vanlege stilistiske variantar er normal uttale av norsk. Viss eit ord har valfrie skrivemåtar i ordlista, så bruk den skrivemåten som liknar mest det som vart sagt.

I tvilstilfelle lyt du berre ta ei avgjerd som best du kan. Ikkje gruble for lenge, men noter spesielle problem i merkeloggen, så kan andre lytte på det seinare.

10.1.3 Trinn 1: Skriv alle ord som blir sagt

Fyrst skal du konsentrere deg om det som blir sagt og sjekke om det er likt med forslaget som eventuelt kjem på skjermen.

Dette kapittelet forklarar korleis du skal skrive orda du høyrer. Med ord meiner vi både vanlege ord, staving, navn og tall, så f.eks B og TRE er ord. Tilslutt blir merking av feiluttale og signalavbrudd forklart.

Vanlege ord

Skrivemåten

Alle ord og bøyde former av desse som finst i ordlistene, kan brukast. Skriv den ordforma som liknar mest det som blir sagt.

Viss ord ikkje fins i ordlista, så noter ordet saman med talar- og setningsnummeret i loggen.

Kva er rett uttale?

Regelen for rett uttale er: "Viss du trur at talaren ville meint at orda han sa var rett uttale, så er det rett!"

Det betyr at **ALLE DIALEKTAR OG STILISTISKE VARIANTAR ER RETT UTTALE!**

Eksempel:

- På østnorsk kan ordet "morgenen" bli uttala "måårnn", men du skal skrive: "morgenen"
- Seier talaren "se rei i mårra i oschlo", skal det skrivast: "ser deg i morgen i Oslo". Legg merke til at eigennavn skal ha stor forbokstav.
- Seier talaren "ække du i barn a", skal det skrivast: "er ikke du i baren da".

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

- Seier talaren "hu schlo sei i hue" eller "ho slo sæ i haue", skal det skrivast "hun slo seg i hodet".
- Om ein bergensar uttalar ordet "mannen" som "maan", så skal du skrive: "mannen".

Alle eksempla ovnanfor viser normale uttalevariantar i norsk som ein perfekt automatisk talegjenkjenner burde kunne klassifisere til ord i ordlista.

Vi må skilje mellom vanlege uttalevariantar for eit ord og at det blir uttala eit anna ord:

- Ordet BAREN uttala "barn", er ein vanleg uttalevariant av "baren" og vi skal skrive "baren" (sjølv om uttala likna meir på eit anna lovleg norsk ord).
- Setningen HAN TOK EN BUSS uttala "han tok en brus", er ein lesefeil av ordet "buss" som gjev ein meningsfylt setning, og skal difor merkast: "han tok en brus".

Du må altså bruke *samanhengen* for å tolke kva ord som er sagt.

Du skal skrive ALLE orda du høyrer

Det betyr at også omstart eller valg av andre ord enn det som står i merkeforslaget, må skrivast.

Vi tilrår at repetisjonar og korreksjonar omsluttast med tegna " < > ".

Eit par eksempel:

- Viss talaren seier: "Bussen nei toget fra nei til Trondheim drar om to timer" - så må alt dette skrivast:
" <bussen nei> toget <fra nei> til Trondheim drar om to timer".

Viss talaren seier: "Norvæst bris fæm te sju mete pæschekunn", skal dette skrivast slik: "nord vest bris fem til sju meter per sekund".

"Uvanlege" ord

Staving

Staving skrivast med store bokstaver. Når bokstavane blir uttala på vanleg måte, f.eks F uttala "eff" og C uttala "se", skriv vi " F " og " C ". Men viss F blir uttala "ffffff", skal det merkast " /f/ ". (Dette er kanskje slik nokon lærde å uttale bokstavane på barneskula).

Merk at bokstaven "J" kan skrivast både " jod " og " J ".

Eksempel:

- Staving av namnet "Jan-Ole", kan vera både slik:
" J A N O L E ", og slik:
" jod A N bindestrek O L E ".

Merk at viss bindestrek blir uttala som bindestrek, så skal ordet " bindestrek " skrivast. Ei anna vanleg uttale av bindestrek er " strek ".

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

Akronym (forkortingar)

Akronym skrivast med store bokstaver. Vanlige akronym der stavinga i seg sjølv oppfattast som ord, f.eks TV, WC, PC, NSB skal skrivast som ord: " NSB ". Dvs utan mellomrom mellom bokstavane, men med store bokstaver. Ein del forkortingar kan uttalast både som eit ord og som enkelt-bokstaver, slik som f.eks SAS som kan seiast "sass" eller "ess a ess". Det er likevel det same omgrepet og dette kan sjåast på som to uttalevarianter og skal difor skrivast " SAS " i begge tilfella.

Bøyde former av akronyma skrivast med bindestrek, som i vanleg rettskriving, f.eks " PC-en ". Samansetningar må enten finnast i ordlista eller leggst til, f.eks " EU-direktiv ".

For sjeldne forkortingar brukast staving. Alle akronym som skrives i eit ord må finnes i ordlista (evt leggst inn i den applikasjonsspesifikke ordlista).

Viss forkortinga NSB blir lese med dei orda forkortinga representerar, skal sjølvstaga orda skrivast: " Norges statsbaner ".

Viss ei forkorting som A B S blir lese som "Alfa Bravo Sigrid", så skriv " Alfa Bravo Sigrid "

Vanlege forkortingar som "osv" og "dvs" kjem dei fleste til å lesa opp med orda som forkortinga representerar, og desse orda skal då skrivast: " og så vidare " og " det vil si ". Men blir bokstavane lese, må det skrivast " O S V " og " D V S ".

Ein del forkortingar kan tolkast og lesast på mange måtar, f.eks "ms" kan lesast som " millisekund ", " meter per sekund ", " sekundmeter ", " multippel sklerose " eller " M S ".

Tall

Alle tall skal skrivast ut med ord etter vanlege norske rettskrivingsreglar:

- Talord under hundre skal skrivast i eitt ord: f.eks " tjuetale ".
- Jamne hundretal, tusental osv blir normalt delte opp i to ord: "to hundre ", " sekstifire tusen ", " sju hundre og femtetre millionar ".
- Årstall skal skrivast i eitt: "nittenhundreogtrettiåtte"
- "203" skal dermed skrivast " to hundre og tre ", (ikkje " tohundreogtre ")

Vi skil mellom "ny" og "gamal" teljemaåte:

- "53" kan skrivast både som: " femtetre " eller som: " treogfemti "

Vi skil mellom "gamal" og "ny" uttale av talla 7, 20, 30 og 40:

- " 7 " kan skrivast både " sju " og " syv "
- "20" kan skrivast både " tjue " og " tyve "
- "30" kan skrivast både " tretti " og " tredve "
- "40" kan skrivast både " førti " og " førr "

Comment [A11]: Dette er svar på MOBEL-spørsmål om "IS-ledd" som blir uttala /i: es led/ (?). Jeg foreslår at det legges til i ordlista. Alle bøyde former skal jo være med i Norkompleks. Switchboard (LDC) bruker syntax " C E O -s " og " I D -ed ", vet ikke om jeg synes det er noen god ide.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

Men for alle andre tall skal vanleg rettskriving brukast, så f.eks 16 skal skrivast "seksten" (sjølv om det blir uttala både "seisn" og "sekstn"), og 17 skal skrivast "syttten" (sjølv om det som oftast blir uttala "søtten").

Navn

Navn skrivast med stor forbokstav som i vanleg rettskriving. Det gjer at vi sidan lett kan sjå forskjell på navn og andre ord. Viss navnet ikkje finst i ordlista, skriv du navnet i merkeloggen og gå vidare.

Utanlandske ord

Det vil sannsynlegvis vera utanlandske ord i talen, spesielt engelske. Viss orda ikkje har norsk skrivemåte, så skriv ordet på original-språket.

Eksempel:

- "Fair" skrivast "fair".
- "Gear" skal skrivast "gir".

Viss ordet ikkje finst i ordlista, skriv du det i merkeloggen og gå vidare.

Feiluttale

Vi skil mellom tre typer feiluttale:

1. Feilen er ikkje grovare enn at ordet (eller orda) likevel er forståeleg
2. Talaren avbryt seg sjølv, men ordet er likevel forståeleg
3. Feiluttala er heilt uforståeleg

Forståeleg tale

Ord som verkeleg er uttala feil, men som likevel er fullt forståeleg, skal merkast med * framfor.

Eksempel:

- Viss "buss-stasjon" blir uttala "bussjon", må det merkast " *buss-stasjon ".

Feilen i eksempelet er at stavinga "sta" ikkje har vorte uttala. Dette er ikkje ei vanleg dialekt-uttale eller ein uttale-stil, men ein uttalefeil. Det same gjeld ord som blir avbrote av host o.l.: For eksempel kan "Når går bussen fra Drag -host- voll?" merkes som " når går bussen fra [spk-p] *Dragvoll ".

Avbrudd av talar

Viss talaren avbryt seg sjølv midt i ordet, men det likevel er forståeleg skal dette merkast som " -(manglande del av ord)uttalt del av ord " eller " uttalt del av ord (manglande del av ord)- ". Utan parentesane skal ordet fortsett følge vanlig rettskriving.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

Eksempel:

- Viss talaren seier: "æ ska te sta- eh togstasjon' ", skal dette skrivast " jeg skal til sta(sjonen)-[fil-e] togstasjonen ".

Uforståeleg tale

Heilt uforståeleg tale skal merkast med to stjerner: " ** " med operom før og etter.

I nokre tilfeller kan opplesaren seia ord du ikkje skjønar, men som likevel er eit ord, f.eks viss ein trønder seier "sjø". Oppfattar du ordet, så skriv du " sjø ". Viss du ikkje oppfattar ordet, skriv du " ** ".

Signalavbrudd

Nokre filer kan ha vorte avkorta fordi opplesaren startar før opptaket er sett i gang eller på grunn av feil i opptak-maskina. Signalavbrudd skal merkast med tilde-symbolet: ~ inntil det ordet som er avkorta (ikkje mellomrom) :

- Avkorting i starten av ordet: " ~ord "
- Avkorting i slutten av ordet: " ord~ "
- Avkorting i både starten og i slutten av ordet: " ~ord~ "

Merk at ~ symbolet skal stå heilt inntil ordet som er avkorta (ikkje mellomrom)!

Viss ordet er så avkorta at du ikkje forstår det som er att, skal du berre merke med " ** " (ingen ~).

10.1.4 Trinn 2: Merk støy

Når alle orda er skrive inn, skal du lytte på nytt og konsentrere deg om støy.

Merk berre støy som er tydeleg høyrbar. Bruk skjønn! (Bør ein talegjenkjenner registrere denne støyen?)

Vi skil mellom 4 slag støy:

- [fil] - "filled pause". Fylte pauser er nølelyder som f.eks "ææh", "øh" og "mm". Denne kategorien er tilrådd oppdelt i to underkategoriar:
 - o [fil-e] for nølelyd som er hovedsakleg vokal-liknande
 - o [fil-m] for nølelyd som er hovedsakleg nasal-liknande
 - o [fil] kan brukast viss ingen av underkategoriane passar
- [spk] - "speaker noise". Støy som talaren lagar, som f.eks smatting, hosting, kremting, latter, sukking eller høglytt pusting (både inn og ut pust). Denne kategorien er tilrådd oppdelt i to underkategoriar:
 - o [spk-f] for pust og "frikativ-liknande" støy

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

- o [spk-p] for smatt, latter, host og "plosiv-liknande" støy
- o [spk] kan brukast viss ingen av kategoriane passar
- [sta] - "stationary noise". Stasjonær bakgrunnstøy med tilnærma konstant amplitude over tid, som f.eks bil-støy, trafikk-støy, kanal-støy eller konstant bakgrunnstale (fest eller offentlege rom som f.eks flyterminalar).
- [int] - "intermittent noise". Kortvarig eller rykkvis støy kan være et brak (f.eks ei dør som blir slått igjen, eller ei dørklokke), periodisk (f.eks telefonringing eller hunde-bjeffing) eller støy med varierende styrke over tid (f.eks musikk). Andre situasjoner som bør merkes med [int] er: ujamn bakgrunnstale, unge-skrik, krysstale og papirknitring.

Det er kun desse seks symbola som er lovlege å bruke for å merke støy. Det skal vera operom mellom symbola og orda rundt. Merk støyen der han er. Støymerka [fil] og [spk] vil normalt vera mellom ord, når slik støy er inni ordet sett merket før ordet som bli råka. Merket [int] kan vera vanskeleg å plassere - bruk skjønn. For støy som varar over fleire ord (normalt for [sta]) bruk merker for start [/sta] og stopp [sta/]. Det er i tillegg tilrådd å merke påfallande pausar med [sil].

Comment [Ai2]: Er dette et bra svar på MOBEL-problemene?

Eksempel:

- Ei setning med stasjonær støy i starten, eit dørmell innimellom, ei pause før siste ord og eit lettelsens pust til slutt:
"/[sta] her var [int] det mykje støy [sta/] men no vart det visst stilt [sil] endeleg [spk-f]".

Ikkje bruk tid på finlytting for å lokalisere støyen: Sett heller støymerket litt for tidleg i setningen.

10.2 SpeechDat merkefil (SAM-format)

Kapittel 1.7 "Label files" fra fil design.doc som følger med SpeechDat-distribusjonen:

The orthographic transcriptions are provided in SAM format label files. These are text files containing information about the speaker, recording and annotation, as well as the orthographic transcription itself. The information is provided in a fixed set of fields, listed in Table 1.

SAM Mnemonic	Description	Example (Comment)
LHD:	Label format name and version	SAM, 5.10
TYP:	Label file type	Orthographic
DBN:	Database name	SpeechDat_Norwegian_Fixed_Network
VOL:	Database volume ID	FIXED1NO_01
SES:	Session number	0000
SHT:	Prompt sheet number	31004031
CMT:	Comment line	*** This is a comment
DIR:	Signal file directory	\\FIXED1NO\\BLOCK00\\SES0000
SRC:	Signal file name	A10000A1.NOA

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

CCD:	Corpus code	<i>A1</i>
CRP:	Corpus repetition	(not used, always empty)
BEG:	Beginning of label sequence	<i>0</i>
END:	End of label sequence	<i>14834</i> (final sample number)
REP:	Recording place	<i>TELENOR, KJELLER, NORWAY</i>
RED:	Recording date	<i>09/Dec/1996</i>
RET:	Recording time	<i>10:20:00</i>
SAM:	Sampling frequency	<i>8000</i>
SNB:	Number of bytes per sample	<i>1</i>
SBF:	Sample byte order	(not used, always empty)
SSB:	Significant bits per sample	<i>8</i>
QNT:	Quantization	<i>A-LAW</i>
NCH:	Number of channels	<i>1</i>
SCD:	Speaker code	<i>0000</i>
SEX:	Speaker sex	<i>F</i>
AGE:	Speaker age	<i>32</i>
ACC:	Speaker accent	<i>17</i>
ARC:	Telephone area code	<i>638</i> (3 first digits in telephone number)
REG:	Calling region	<i>22</i>
ENV:	Calling environment	<i>OFFICE</i>
NET:	Calling network	<i>FIXED</i>
PHM:	Telephone handset	<i>TOUCHTONE</i>
EXP:	Labelling expert	<i>CT</i>
SYS:	Labelling system	<i>Annot13</i>
LBD:	Label body starts	
LBR:	Recording prompt	<i>0, 14834,,,, slutt</i>
LBO:	Orthographic labels	<i>0,, 14834, [sta] slutt [spk]</i>
EXT:	Line extension	(for lines that exceed 80 characters)
ELF:	End of label file	

Table 1SAM mnemonics used in label files

In addition to the standardized mnemonics, the construction “CMT: LANG:” is used to identify the language variant (Bokmål or Nynorsk) used in the prompt sheet. An example of a complete label file given in Table 2.

```

LHD: SAM 5.10
TYP: orthographic
DBN: SpeechDat_Norwegian_Fixed_Network
VOL: FIXED1N0_01
SES: 0000
SHT: 31004031
CMT: LANG: BOKMÅL
CMT: *** Speech file information ***
DIR: \FIXED1N0\BLOCK00\SES0000
SRC: A10000A1.NOA
CCD: A1
CRP:
BEG: 0
END: 14834
REP: TELENOR, KJELLER, NORWAY
RED: 09/Dec/1996
RET: 10:20:00
CMT: *** Speech data coding ***

```


Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

```
SAM: 8000
SNB: 1
SBF:
SSB: 8
QNT: A-LAW
NCH: 1
CMT: *** Speaker information ***
SCD: 0000
SEX: F
AGE: 32
ACC: 17
CMT: *** Recording conditions ***
ARC: 638
REG: 22
ENV: OFFICE
NET: FIXED
PHM: TOUCHTONE
CMT: *** Labelling session information ***
EXP: CT
SYS: annot13
LBD:
CMT: *** Label file body ***
CMT: LBR, begin, end, gain, min, max, prompt
CMT: LBO, begin, centre, end, transcription
LBR: 0, 14834,,, slutt
LBO: 0,, 14834, [sta] slutt [spk]
ELF:
```

Table 2 Example of a complete label file

10.3 Dialektregioner fra SpeechDat

Fra svarskjemaet som ble sendt til oppleserne i SpeechDat:

Dialektregioner

Vi ønsker å finne den regionen som du synes passer best med din nåværende dialekt. Hvis du ikke er sikker på hvilken region dialekten din tilhører, kan du bruke den regionen du bodde i da du ble 15 år. Legg merke til at grensene for dialektregioner og fylker ikke nødvendigvis er sammenfallende. Region 21 går for eksempel over 4 fylker, mens region 17 er en del av ett fylke. Legg også merke til at region 24 kan brukes dersom du ikke har norsk språkbakgrunn.

Notat / Memo: Standard for ortografisk annotering av spontan tale i BRAGE

01 Fra Finnmark fylke: Vardø, Vadsø, Hammerfest, Alta, Hasvik, Kvalsund, Måsøy, Nordkapp, Porsanger, Lebesby, Gamvik, Berlevåg, Båtsfjord, Sør-Varanger.

02 Fra Finnmark fylke: Kautokeino, Loppa, Karasjok, Tana, Nesseby.
Fra Troms fylke: Lyngen, Storfjord, Kåfjord, Skjervøy, Nordreisa, Kvænangen.

03 Alle kommuner i Troms fylke unntatt Lyngen, Storfjord, Kåfjord, Skjervøy, Nordreisa, Kvænangen.

04 Fra Nordland fylke: Narvik, Tysfjord, Lødingen, Tjeldsund, Evenes, Ballangen, Bø, Øksnes, Sortland, Andøy.

05 Fra Nordland fylke: Bodø, Gildeskål, Beiarn, Saltdal, Fauske, Skjerstad, Sørfold, Steigen, Hamarøy, Røst, Værøy, Flakstad, Vestvågøy, Vågan, Hadsel, Moskenes.

06 Fra Nordland fylke: Dønna, Nesna, Hemnes, Rana, Lurøy, Træna, Rødøy, Meløy.

07 Fra Nordland fylke: Sømna, Brønnøy, Vega, Vevelstad, Herøy, Alstadhaug, Leirfjord, Vefsn, Grane, Hattfjelldal.

08 Fra Nordland fylke: Bindal. Fra Nord-Trøndelag fylke: Namsos, Fosnes, Flatanger, Vikna, Nærøy, Leka. Fra Sør-Trøndelag fylke: Hemne, Snillfjord, Hitra, Frøya, Ørland, Agdenes, Rissa, Bjugn, Åfjord, Roan, Osen. Fra Møre og Romsdal fylke: Kristiansund, Averøy, Frei, Gjemnes, Tingvoll, Sunndal, Sunndal, Rindal, Aure, Halså, Tustna, Smøla.

09 Fra Nord-Trøndelag fylke: Steinkjer, Meråker, Stjørdal, Frosta, Leksvik, Levanger, Verdal, Mosvik, Verran, Namdalseid, Inderøy, Snåsa, Lierne, Røyrvik, Namsskogan, Grong, Høylandet, Overhalla. Fra Sør-Trøndelag fylke: Trondheim, Orkdal, Melhus, Skaun, Klæbu, Malvik.

10 Fra Sør-Trøndelag fylke: Oppdal, Rennebu, Meldal, Røros, Holtålen, Midtre Gauldal, Selbu, Tydal.

11 Fra Møre og Romsdal fylke: Molde, Rauma, Nesset, Midsund, Aukra, Fræna, Eide.

12 Fra Sogn og Fjordane fylke: Selje. Fra Møre og Romsdal fylke: Ålesund, Vanylven, Sande, Herøy, Ulstein, Hareid, Volda, Ørsta, Ørskog, Norddal, Stranda, Stordal, Sykkylven, Skodje, Sula, Giske, Haram, Vestnes, Sandøy.

13 Alle kommuner i Sogn og Fjordane fylke unntatt Selje, Sogndal, Balestrand, Leikanger, Aurland, Vik, Lærdal, Årdal, Luster.

14 Fra Sogn og Fjordane fylke: Sogndal, Balestrand, Leikanger, Aurland, Vik, Lærdal, Årdal, Luster.

15 Fra Hordaland fylke: Voss, Granvik, Ulvik

16 Alle kommuner i Hordaland fylke unntatt Bergen, Askøy, Fjell, Sund, Os, Voss, Granvin, Ulvik.

17 Fra Hordaland fylke: Bergen, Askøy, Fjell, Sund, Os.

18 Alle kommuner i Rogaland fylke.

19 Alle kommuner i Vest-Agder fylke.

20 Alle kommuner i Aust-Agder fylke.

21 Alle kommuner i Oppland og Hedmark fylker. Alle kommuner i Telemark fylke unntatt Porsgrunn, Skien, Bamble, Kragerø, Siljan, Nome. Alle kommuner i Buskerud fylke unntatt Kongsberg, Øvre Eiker, Nedre Eiker, Drammen, Lier, Røyken, Hurum.

22 Alle kommuner i Oslo og Akershus fylker samt Rømskog fra Østfold fylke.

23 Alle kommuner i Vestfold fylke. Alle kommuner i Østfold fylke unntatt Rømskog. Fra Telemark fylke: Porsgrunn, Skien, Bamble, Kragerø, Siljan, Nome. Fra Buskerud fylke: Kongsberg, Øvre Eiker, Nedre Eiker, Drammen, Lier, Røyken, Hurum.

24 Utenlandsk språkbakgrunn, ingen av regionene over.

