

# Retningslinjer for transkripsjon av stortingsforhandlingene

Andrea, Håvard og Per Erik

## Formål og litt om dataene

Formålet med prosjektet er å transkribere lyden fra stortingsforhandlingene i to stortingssesjoner. Denne transkripsjonen skal brukes til å trene opp talegjenkjenningssystemer. Forhandlingene skal transkriberes ortografisk på bokmål eller nynorsk, ettersom hvilken målform de enkelte representantene bruker. For mer bakgrunnsinformasjon om prosjektet, se prosjektbeskrivelsen.

Vi kjører stortingsforhandlingene gjennom [Googles automatiske transkripsjonssystem](#). Den manuelle transkripsjonen vil bestå i å rette denne automatiske transkripsjonen, ikke transkribere fra grunnen av.

Det finnes allerede referat av stortingsforhandlingene, skrevet av stenografer på Stortinget. Selv om disse ligger tett opp til det som blir sagt i stortingssalen, har stenografene redigert teksten en del slik at den skal fungere som et lett lesbart referat. Dermed kan ikke referatene brukes slik de er til talegjenkjenning.

Referatene er imidlertid en nyttig kilde for oss på flere måter. For det første kan vi bruke referatene til å rette transkripsjonen fra Google automatisk. For det andre kan referatene brukes til å hente ut metadata som kan være nyttig for transkribørene å kjenne til, f.eks. hvilke representanter som snakker og hvilke målformer de bruker. Vi forsøker å ekstrahere disse dataene automatisk, men av og til vil det være nødvendig for transkribørene å konsultere referatene manuelt.

## Arbeidsflyt

Denne arbeidsflyten er utarbeida før prosjektets start. Det er sannsynlig at den vil bli revidert ettersom vi ser hva som fungerer best.

Arbeidsflyten inneholder følgende etapper:

1. Lydfilen transkriberes automatisk med Google Speech-to-Text
2. Disse transkripsjonene "flettes" med teksten fra de offisielle stortingsreferatene, som gir en del forbedringer
3. Per Erik deler den automatiske transkripsjonen med en transkribør, som retter den i Språklabben.

4. En annen transkribør reviewer transkripsjonen for å luke ut eventuelle feil.
5. Om revieweren finner noe, går fila tilbake til transkribøren for retting.
6. Når fila er retta, laster Per Erik fila ned og legger den i korpusmappa.
7. Dersom vi endrer transkripsjonsrutiner e.l., kan det hende at transkripsjonen må gjennom en andre retterunde på et seinere tidspunkt.
8. Filene transkriberes i såkalt *spoken domain*, se under. Konverteringsskript kjøres på transkripsjonene, slik at korpuset også foreligger i *written domain*.

Dette regnearket vil være et viktig verktøy. I andre kolonne står navnet på alle filene som skal transkriberes. Tredje kolonne inneholder en link med et regneark med metadata: dato, hvilke representanter snakker, hva er deres målform osv.

Når Per Erik har kjørt preprosesseringa på en fil og den er klar til transkripsjon (etappe 1), fører han opp datoen i feltet *ferdig preprosessert*. En transkribør kan da ta den (etappe 2). Hun skriver da brukernavnet sitt i kolonnen i *Manuell transkripsjon: brukernavn*, og datoen i *Manuell transkripsjon: dato*. Det er viktig å sjekke metadatafila, som det er link til i regnearket, for å finne representantnavnet. Dermed kan man finne ut hvilken målform representanten bruker i dette dokumentet. Av og til kan det hende at det mangler informasjon om representanter og målform, eller at transkribøren mistenker at denne informasjonen er feil. Da kan hun sjekke referatene, som ligger [her](#).

Når transkripsjonen er klar for review (etappe 4), legger transkribøren inn dato i feltet *Klar til review*, og gir beskjed til den som skal reviewe. Revieweren gir beskjed til transkribøren om det er noen rettinger som må til (etappe 5). Slik fortsetter man til reviewer er tilfreds med transkripsjonen. Da legger han inn dato i kolonnen SBU (ser bra ut). Per Erik laster så fila ned i korpusmappa (etappe 6).

Transkribørene loggfører tilfeller de lurer på. Disse diskuteres på transkripsjonsmøter, og avgjørelsene derfra føres inn i retningslinjene under. I de fleste tilfeller er det ok, og lurt, at transkribørene diskuterer seg imellom underveis. Men det er viktig at alle større avgjørelser (alt som er mer generelt enn "Hvordan bør vi transkribere dette ordet på denne dialekta") tas på transkripsjonsmøtet og loggføres.

# Transkripsjonsregler

## Segmentering, tidskoding og annotasjon av talere

### Segmentering

Transkripsjonene skal segmenteres i helsetninger. Googles talegjenkjenning kommer ikke med forslag til segmentering, så dette må gjøres manuelt.<sup>1</sup>

- 0:06:55 0 Stortinget mottok mandag meddelelse fra statsministerens kontor om at statsminister Erna Solberg vil møte til muntlig spørretime
- 0:07:02 0 statsministeren er til stede og vi er klare til å starte den muntlige spørretimen

Ytringer som ikke utgjør en setning, men som fungerer som en selvstendig grammatisk enhet, skal også utgjøre et selvstendig segment:

- 0:06:43 0 takk
- 0:10:40 0 Trond Helleland
- 0:12:29 0 så synes ikke

Ikke alle tilfeller er like åpenbare som disse. For å avgjøre hva som utgjør et segment, bruker vi følgende prinsipper:

1. Syntaks har forrang. I tilfeller der det er entydig ut fra syntaksen hvor setningsgrensa skal gå, benytter vi oss ikke av de andre kriteriene. I *Forlanger noen ordet? Så synes ikke* skal det være en setningsgrense mellom *ordet* og *så*, også dersom det ikke er noen pause mellom disse orda eller om intonasjonen ikke indikerer noen setningsgrense.
2. Om syntaksen tillater det, setter vi setningsgrenser på pauser. Setningsgrenser som ikke er på pauser, gjør tidskoding mer komplisert.
3. Om regel 1 og 2 tillater flere muligheter, bruker vi intonasjon og pragmatikk til å avgjøre. Det er ofte klart fra intonasjonskurven og pauser hva som fungerer som en enhet, og grammatisk og pragmatisk informasjon kan også være relevant i en del tilfeller. Under vil vi gi retningslinjer for hvordan vi segmenterer i en del spesifikke konstruksjoner.<sup>2</sup>

---

<sup>1</sup> I Språklabben er transkripsjonene segmentert i bolker på ett minutt.

<sup>2</sup> Regel 2 kom først til midt i februar 2020, så dette prinsippet er ikke fulgt i en del tidlige transkripsjoner.

## Koordinasjon og konjunksjoner

Koordinerte helsetninger segmenteres også som ett segment:

0:12:27 **E** vi kan nå kanskje konstatere at utgangspunktet for spørsmålet til representanten Tajik var feil og jeg forstår i oppfølgingsspørsmålet at man tar innover seg det

Men, noen ganger blir selvstendige ytringer innledet av konjunksjoner. Det er ikke alltid like åpenbart om den konjunksjonsinnledete setningen skal skilles ut som eget segment eller om den skal være del av en koordinert helsetning. Om syntaksen tillater flere alternativer (regel 1), og det er en pause mellom de konjugerte helsetningene (regel 2), bruker vi først og fremst fonologiske kriterier (regel 3). En indikasjon på at setningen skal skilles ut, er en tydelig intonasjonskurve.

Der de fonologiske kriteriene ikke strekker til, bruker vi diskurskriterier (eksplisitt temaskifte o.l.):

0:17:29 **E** derfor betyr det også at det er viktig å være klar over det

0:17:32 **E** og særlig for de som har vært lenge vekk fra arbeid så oppleves det også som en sosial utfordring

0:13:43 **H** men jeg vil gå tilbake til det som var noe av innledningen i hovedspørsmålet og det var handel

I det første av eksemplene over tyder adverbialet *“særlig for de som har vært lenge vekk fra arbeid”* på at vi ikke har å gjøre med standard setningskoordinasjon, men heller en ny, selvstendig helsetning som presiserer den forrige. Derfor er det naturlig å sette et segmentskille der. (I slike tilfeller må man nok i noen grad bruke skjønn.) I det andre eksempelet viser *“tilbake til det...”* at vi har et tematisk brudd i teksten, og det er ikke naturlig å ta dette som et konjunkt i forrige setning.

Hvis ikke noen av disse kriteriene holder, er det snakk om koordinasjon (altså ett segment).

## Elliderte ledd, oppramsing

Innledningen av møter i stortinget er svært formaliserte, og man kommer ofte over ytringer som denne, og tilsvarende:

- 0:08:28 **O** det foreligger to permisjonssøknader
- 0:08:32 **O** fra representanten Geir Jørgen Bekkevold om velferdspermisjon lørdag sjuende oktober
- 0:08:37 **O** fra Høyre stortingsgruppe om sykepermisjon for representant Michael Tetzschner fra og med sjuende oktober og inntil videre

Vi velger å dele ytringen opp i tre segmenter her, siden vi kan analysere det hele som at verbet og subjektet er implisitt – eller ellidert – i de to “fra”-segmentene (regel 1).

Altså:

- *det foreligger to permisjonssøknader* | (det foreligger en permisjonssøknad/en permisjonssøknad foreligger) *fra representanten Geir Jørgen Bekkevold om ...*<sup>3</sup>

Og et tilsvarende eksempel:

- 0:06:18 **O** følgende innkalte vararepresentanter tar nå sete
- 0:06:22 **O** for Buskerud fylke Tone Heimdal Brataas
- 0:06:24 **O** for Oppland fylke Ivar Odnes
- 0:06:26 **O** for Rogaland fylke Øystein Langholm Hansen

Oppramsing av saksliste o.l, segmenterer vi også ut.

- 0:27:15 **B** rapporten definerer tre målsetninger for det norske engasjementet
- 0:27:20 **B** en å støtte USA en NATO-alliert som ble angrepet
- 0:27:24 **B** to å bidra i kampen mot terror
- 0:27:28 **B** og tre å bidra til å bygge en stabil og demokratisk afghansk stat

Her er det kanskje mulig å argumentere for en annen syntaktisk analyse. Vi mener likevel at slike ledd så tydelig fungerer som selvstendige enheter at vi velger å segmentere dem ut. Dette vil typisk også stemme overens med pauser og intonasjon (regel 2 og 3).

Språket i Stortinget er, som nevnt, gjerne svært formelt og mye av det som ytres er rene opplesninger fra et manus. Da kan man ende opp med lange, oppramsende setninger som likevel både syntaktisk og fonologisk fungerer som en enhet:

---

<sup>3</sup> I retningslinjene indikerer vi segmentgrenser med “|”.

0:27:28 **V** naturmangfoldloven og Bernkonvensjonen stiller som vilkår for felling for det første at det ikke må true bestandens overlevelse for det andre at det må foreligge fare for skade av et visst omfang og alvorlighetsgrad og for det tredje at det ikke finnes andre tilfredsstillende løsninger

0:06:06 **M** til tross for lavere renter til tross for lav kronekurs eksporten fra fastlandsbedriftene svekkes

0:05:33 **M** og en veldig viktig forutsetning for at det skal kunne skje det er at vi får en solid vekst i eksporten men så langt så ser vi ingen tegn til det tvert imot

### Sitater

I sitater gjelder vanlige regler for segmentering. Helsetninger segmenteres dermed ut:

0:21:25 **V** i uttalelsen heter det og jeg siterer

0:21:28 **V** kravet om bestandsregulering gjelder også innenfor sona og uavhengig av at det må foreligge skadepotensial på husdyr og tamrein

0:21:35 **V** komiteens flertall vil understreke at bestandsmål fastsatt av Stortinget er det klart overordnede vedtak

0:21:41 **V** sitat slutt

Ofte innledes sitatet med at representanten sier "sitat" og avsluttes med ordene "sitat slutt":

1:14:53 **S** i helga for eksempel så uttalte finansministeren til Dagbladet om sin egen lukrative gullpensjon

1:14:58 **S** sitat

1:14:59 **S** det er fullstendig umusikalsk at politikerne på Stortinget skal sitte å gi seg sjøl så gode vilkår

1:15:04 **S** sitat slutt

Noen sitater forblir integrert i det aktuelle segmentet, som følgende eksempel:

2:14:42 **M** i dag så er samenes rettigheter jo grunnlovsfestet i paragraf hundre og åtte som jo da slår fast at den samiske folkegruppe og det er da jeg siterer det uttrykket har rett til å sikre og utvikle sitt språk sin kultur og samfunnsliv for den samiske folkegruppen

### Reparering

Ofte gjentar talere et ord i setningen, begynner på setningen flere ganger o.l. Slike tilfeller skal ikke segmenteres ut, men være del av samme segment:

- *jeg jeg må tenke litt på det*
- *hvis vi skal om vi skal stemme over dette må vi gjøre det nå*

### Omstart

Setninger som ikke fullføres, skal imidlertid være selvstendige segment:

- *hva skulle jeg | la oss komme til saken*

Det er ikke alltid trivielt å dele talespråk inn i setninger, så her vil det trolig være nødvendig å bruke skjønn i en del tilfeller.

### Annotasjon av talere

Transkripsjonen skal annoteres med hvem som taler. Metadatafila inneholder en liste over talere, men merk at denne er automatisk generert og kan være unøyaktig. Derfor er det av og til nødvendig å se på referatene. Om det ikke er mulig å finne ut hvem som taler, annoterer man transkripsjonen med **unknown**.

Merk at bokmålsbrukere skal transkriberes på bokmål og nynorskbrukere på nynorsk. Denne fila skal inneholde informasjon om dette. Dersom det ikke står, eller om informasjonen i fila virker usannsynlig, må man sjekke målformen til denne taleren i referatet.

### Språklige og ortografiske retningslinjer

Stortingsforhandlingene skal transkriberes ortografisk, som vil si at alle ordene (eller tokenene) i transkripsjonen skal være normerte ord på enten bokmål, om taleren er bokmålsbruker, eller nynorsk, om taleren er nynorskbruker. Det overordnede prinsippet for transkripsjonen er at vi normerer på ordnivå, ikke på setningsnivå. Vi gjør en ord-for-ord-transkripsjon der ordene følger normen for ortografi innenfor den aktuelle målformen.

- *så gode løsninger utenfor rettsapparatet kan forebygge konflikter og sørge for mer **solid** og **rimelige løsninger** for forbrukeren*

I eksempelet over ser vi at adjektivet (*solid*) og substantivet (*løsninger*) ikke kongruerer. Siden vi normerer på ordnivå, og taleren ytrer ordet *solid*, som er et normert ord, transkriberer vi det som *solid* og ikke *solide*. Vi retter altså ikke opp kongruensen (mer om dette i [avsnitt om kongruens](#)).

## Transkripsjonsdomene

Selv om vi har valgt ortografisk transkripsjon, kan transkripsjonen være mer eller mindre lydnær. Man skiller mellom “spoken domain” (SD) og “written domain” (WD).

I SD gjengir man talen så presist som mulig innafor rammen av normen. Tall vil bli skrevet med bokstaver, klokkeslett vil bli gjengitt slik de blir sagt, og ikke i et standardformat (*halv sju*, ikke *18.30*), og forkortelser blir ikke brukt (med mindre forkortelsen faktisk blir uttalt som en forkortelse, se [dette avsnittet](#)). SD gjengir språklig variasjon i transkripsjonen, som i mange tilfeller kan være en fordel. Samtidig vil sluttproduktet til taleteknologiske verktøy sjelden være SD. Som oftest ønsker man standard forkortelser, tall, datoer osv. Med SD mister man også informasjon om at, f.eks., *halv sju* og *atten tretti* er to måter å si samme ting på, nemlig *18.30*. Dette er informasjon som er nyttig for mange sluttprodukter trent på materialet.

Forskjellige brukergrupper vil trolig ha ulike preferanser når det gjelder transkripsjonsdomene. Firmaer som har en sterk norskspråklig kompetanse vil kunne utnytte en mer differensiert SD-transkripsjon og lage etterprosesseringsverktøy som konverterer fra SD til WD. Internasjonale firmaer med lite eller ingen norskspråklig kompetanse vil derimot kunne foretrekke WD-transkripsjon, fordi de da kan unngå språkspesifikk etterprosessering.

I stortingstranskripsjonene transkriberer vi i SD. I tillegg kommer vi til å utvikle etterprosesseringsverktøy som konverterer fra SD til WD. Merk at Google-transkripsjonen er i WD, så det innebærer noe ekstraarbeid å rette til SD.

## Normering

### Nynorsknormering

Vi normerer hardt til nynorsknormalen. Det betyr at vi konsekvent velger normerte nynorske former. Det er viktig at transkripsjonen er konsistent.

Der det er valgfrihet i normalen velger vi det ordet som ligger lydlig nærmest uttalen. Til illustrasjon har man ved normeringen av det bokmålsnære ordet *driftstilskuddet* valget mellom de nynorske ekvivalentene “drifttilskotet” eller “driftstilskottet”. Vi



velger “driftstilskottet” i dette tilfellet, fordi det ligger lydlig nærmest uttalen – både *driftstilskuddet* og *driftstilskottet* har kort vokal.

#### Nokon og nokre

Vi transkriberer alltid til “nokon” når talerne sier *noen*. Der taleren sier *nokre*, eller noe som ligger nært opptil, skriver vi “nokre”.

#### Bokmålsnormering

Bokmålsnormeringen følger de samme prinsippene som nynorsknormeringen. På samme måte som for nynorsk, ønsker vi bare å ha normerte bokmålsord i bokmålstranskripsjonen (BT), og vi bruker bakstrektranskripsjon ved forekomster av unormerte ord. En oversikt over de konverterte ordene finnes i denne listen.

#### Unormerte leksikalske ord

I transkripsjonen vil en ofte støte på ord som ikke er normerte i talerens målform. Med unormerte ord menes innholdsord som ikke finnes i den aktuelle målformen (for funksjonsord, se [her](#)). For å avgjøre hvorvidt et ord er unormert eller ei, lener vi oss på [Ordboka](#) og [NAOB](#). Et slikt eksempel kan være når en nynorskbruker bruker bokmålsord som *beslutning* istedenfor *avgjerd*. Disse tilfellene er litt kinkige for et korpus som dette, som skal brukes som treningsmateriale for talegjenkjenning. På den ene siden vil man at transkripsjonen, så langt som mulig, skal samsvare med det et tale-til-tekst-system skal produsere, og et slikt system bør ikke produsere unormert tekst. På den andre siden er det ikke så gunstig å transkribere, f.eks., *beslutning* som *avgjerd*, fordi disse orda er lydlig veldig forskjellige. Talegjenkjenningssystem bruker transkripsjoner til å utlede forholdet mellom lydbilde og tekst, og om det ofte ikke er samsvar mellom lyd og transkripsjon, kan det gå ut over presisjonen til talegjenkjenninga.

På grunn av disse motstridende hensynene har vi valgt å gjengi både den unormerte, uttalte formen i slike tilfeller samt en normert semantisk ekvivalent av den unormerte formen. Vi bruker bakstrek mellom den unormerte og den normerte formen, slik: **unormert\normert**. Det er viktig at denne bakstrektranskripsjonen er konsistent (alle tilfeller av *beslutning* i nynorskdelen av korpuset skal transkriberes som *avgjerd*). Derfor føres alle tilfeller av bakstrektranskripsjon opp i [denne lista](#).

Ofte vil en unormert form i én målform være normert i den andre målformen. Til illustrasjon ville ordene *beslutning* (bm) og *symje* (nn), blitt transkribert på følgende måte for henholdsvis en nynorskbruker og en bokmålsbruker:

Nynorskbruker:

- *beslutning\avgjerd*
- *symje*

Bokmålsbruker:

- *beslutning*
- *symje\svømme*

For å avgjøre hvilken normert form vi skal velge i bakstrektranskripsjonen, der hvor det finnes flere alternativer, diskuterer vi oss fram til den mest egnede formen og holder oss til den. Vi lener oss på [Ordboka](#) og, i tilfeller hvor vi må normere til nynorsk, [Språkrådets administrative ordliste](#) (SAO). [Språkrådets kanselliste](#) kan også være til hjelp og. Det kan også være til hjelp å kikke på teksten i [referatet](#). Den mest egnede formen er den formen som betydningsmessig ligger nærmest den unormerte formen. Dersom det ikke finnes en form som åpenbart er nærmest den unormerte formen betydningsmessig, velger vi den formen som er semantisk mest "nøytral", altså som kan brukes i flest domener. Til illustrasjon har en ved normeringen av *beslutning* i følge SAO valget mellom *avgjerd* og *vedtak*. *Avgjerd* ligger semantisk nærmere *beslutning*, og er i tillegg mer domenenøytralt enn *vedtak*. Derfor normerer vi konsekvent *beslutning* til *avgjerd*.

#### Dialektord

En kommer også over ord som ikke er normert i noen av målformene, typisk dialektord (*bekkalokk*, *undikk*, *jaseleg*). Hvis det er tydelig at ordet er en dialektal uttalevariant av et normert ord, transkriberer vi kun den normerte formen:

- /kønn/ blir *korn*
- /fjedl/ blir *fjell*

Hvis det derimot er tydelig at det ikke er en dialektal uttalevariant, men heller et eget, unormert dialektord, transkriberer vi både den unormerte formen og en semantisk ekvivalent normert form<sup>4</sup>:

- /bekkalokk/ blir *bekkalokk\kumlukk*
- /undikk/ blir *undikk\underbukse*
- /jaseleg/ blir *jaseleg\skvetten*

Det er ikke alltid åpenbart om et ord er å regne som et unormert dialektord eller en dialektal uttalevariant av et normert ord. Dette er ofte en skjønnsmessig vurdering. Merk at vi regner for eksempel apokopering i dialekter som har dette trekket som en dialektal uttalevariant (se dette [avsnittet](#)). Dersom en kommer til at ordet ikke er en dialektal uttaleform, men et unormert ord, bruker vi bakstrek.

Når vi skal bakstreke et unormert ord, vil vi så langt det er mulig skrive et ord som er normert i den motsatte målformen før bakstreken. Dette gjør vi for å være så konsekvente som mulig i hvordan vi velger å gjengi det som sies ortografisk, selv når

---

<sup>4</sup> Disse finnes i listen over unormerte ord.

det er unormert. Hvis ordet som sies ligger lydlig nært den normerte formen i den motsatte målformen, bruker vi den skrivemåten. For eksempel, hvis vi transkriberer på bokmål så er ordet /møtji/ såpass ulikt *mye* at vi velger å bakstreke. Videre er /møtji/ såpass likt *mykje* på nynorsk at vi velger nynorskformen før bakstrek og transkriberer ordet slik: *mykje\mye*.

For å avgjøre om ordet som skal bakstrekkes kan skrives slik det er normert i den motsatte målformen kan vi tenke på om det uttalte ordet hadde blitt godkjent på motsatt målform. Hvis vi hadde transkribert på den andre målformen i utgangspunktet, og vi hadde valgt å skrive ordet med sin normerte form, uten ytterligere markering, skriver vi den målformens normerte form foran bakstreken. Dersom vi kommer til at vi ikke hadde godtatt det unormerte ordet på motsatt målform heller, men hadde valgt å bakstreke da også, skriver vi en mer lydnær form. Ved valg av stavemåte for den unormerte formen bruker vi skjønn og norske rettskrivingsregler. Ha i mente at vi transkriberer ortografisk – vi ønsker ikke en fonetisk transkripsjon. Til hjelp kan en også bruke bl.a. [Norsk Ordbok](#) og [NAOB](#).

Det kan være vanskelig å avgjøre hvordan ord som ikke finnes i noen av målformene bør staves. I tilfeller der vi vet hva ordet betyr, men ikke får opp forslag på stavemåte i noen av de nevnte ressursene, bruker vi skjønn og norske rettskrivingsregler.

For eksempel, hvis vi transkriberer på bokmål og taleren sier /stæn/, en dialektal variant av presensformen av å *stå*, velger vi å bakstreke fordi /stæn/ er såpass forskjellig fra *står*. Når vi tenker på om vi kan velge å skrive den normerte formen i motsatt målform foran bakstreken, kommer vi til at vi ikke kan det. /stæn/ er såpass forskjellig fra den normerte formen i nynorsk også (*står*) at vi i nynorsktranskripsjonen også hadde valgt å bakstreke. Da må vi velge en annen skrivemåte. I dette tilfellet velger vi skrivemåten *stend* for /stæn/, fordi det er en skrivemåte som ligger nært norsk rettskriving og som vil favne flere lignende uttalevarianter av /stæn/, som alle vil få en konsekvent skrivemåte før bakstrek. Transkripsjonen til slutt blir *stend\står*.

Noen eksempler på unormerte uttaleformer som vil kunne komme inn under samme skrivemåte: *stend\står* for forekomster som /stende/, /stæn/, /stenn/ etc., *logge\ligget* (bm) eller *logge\lege* (nn) for forekomster som /lågge/, /låge/, /logge/ etc.

Hvis vi hører hva som blir sagt, men ikke forstår hva ordet betyr, tagger vi ordet som [UHØRBART]. Dette fordi vi da ikke kan gi det unormerte ordet en plausibel normert oversettelse. Vi gjetter ikke.

## Funksjonsord

Med funksjonsord mener vi alle pronomen, determinativ, preposisjoner, subjunksjoner, konjunksjoner og hjelpeverb (*burde, få, kunne, måtte, skulle tore/tørre, ville, ha, skulle, ville, være, bli*). Av adverbene regner vi spørreordene, negasjon, samt ordene *her, der, herre/herne, derre/derne, da/då, når* og *nå/no* som funksjonsord.

Vi normerer alltid funksjonsord, uten bruk av bakstrekranskripsjon. Grunnen til at vi har valgt å ikke bruke bakstrekranskripsjon for funksjonsord, er at disse er så hyppige og har stor grad av variasjon. Å ha en bakstrekranskripsjon i slike tilfeller ville innebære et betydelig merarbeid som ville gå utover størrelsen på korpuset. Det ville også være vanskelig å få konsistens, siden variasjonen er så stor.

Vi bruker *stjernetegn* (\*) for å markere tilfeller hvor uttalen av et funksjonsord, etter transkribørens skjønn, i stor grad avviker fra den normerte formen. Siden målet med stjernetranskripsjonen er å markere for fremtidige brukere av materialet at uttalen i stor grad avviker fra den normerte formen, er det rimelig å anta at ordene markert med stjerne i en del tilfeller vil bli sortert bort. Derfor ønsker vi å bruke minst mulig stjernetranskripsjon, og vi bruker ikke mye tid på å avgjøre om et ord skal ha stjernetranskripsjon eller ikke. Er vi i tvil, bruker vi ikke stjerne og går videre.

Vi tillater mange ulike realiseringer – det vil si vi markerer dem ikke med stjerne – av funksjonsord fordi de er høyfrekvente former som vi ikke ønsker at skal potensielt falle bort hos de som benytter seg av materialet. Eksempler er alle realiseringer av 1. person entall pronomen *jeg/eg*, alle realiseringer av nektingsadverbet *ikke/ikkje*, m.m.. Men, visse realiseringer som har et særlig avvik mellom normert og uttalt form vil bli transkribert med stjerne.

I de tilfellene det er klart at det uttalte ordet er et helt annet ord enn den normerte formen, og ikke bare en uttalevariant, stjernemerker vi (i nynorsktranskripsjonen: /man/ blir *ein*\*). I tilfeller der man kan argumentere for at det er en uttalevariant, men det likevel er veldig stor lydlig forskjell, bør man også stjernemerke (/koss'n/ blir *hvordan*\*). I mange tilfeller vil det være åpenbart at vi har å gjøre med et annet ord, som i eksemplene /koss/ for *hvilken*\*, /tå/ for *av*\* og /me/ for *vi*\*. Men i andre tilfeller er det ikke like åpenbart. Vi er bevisst vage her fordi det ikke er praktisk mulig å ha brukbare, helt presise kriterier i disse tilfellene.

Merk her at vi ikke tar hensyn til syntaks – så lenge formen er normert, skriver vi den ut. Det vil si at om en informant bruker en objektsform i subjektsposisjon (f.eks *dem* istedenfor *de*) skriver vi det som blir sagt:


- *dem gjør et meget viktig arbeid*


## Unntak

Det er likevel et svært hyppig unntak til regelen om at funksjonsord ikke skal ha bakstrektranskripsjon: På bokmål er tallorda *tyve* og *tredve* ikke normerte, på tross av at de hyppig forekommer muntlig. Det gjelder også bruk av det gamle tellesystemet, f.eks. *treogåtti* istedenfor *åttitre*. Siden disse tilfellene er så hyppige og siden det er et entydig forhold mellom unormert og normert form, har vi valgt å bruke bakstrektranskripsjon også for dem, altså *tyve\tjue* og *treogåtti\åttitre*.


## Bruk av andre språk eller en annen målform

Om en representant siterer noe eller sier noe på et annet språk enn norsk, skal dette transkriberes kun på det aktuelle språket, uten bakstrektranskripsjon. Ordene markeres med valgt språk, som oftest Engelsk (Storbritannia), i ctrl+i-dialogvinduet. I verktøyet blir disse segmentene markert i grått:

0:05:46  the Speaker is on an official visit to Norway and the discussions among us cover issues such as Arctic cooperation energy and trade

0:05:55  thank you again for visiting Norway

Det samme prinsippet gjelder for bruk av annen målform enn ens egen: dersom en representant siterer noe med en annen målform enn den representanten selv bruker, skal dette transkriberes på målformen til sitatet.

0:10:20  forslag tjuetre om at dei marine ressursane høyrer fellesskapet til og forslag nummer tjueseks om at de marine ressursene tilhører fellesskapet og skal komme kystsamfunnene til gode

Hvis det ikke er snakk om et direkte sitat, men heller en løs gjengivelse eller indirekte tale, skal utsagnet transkriberes på representantens oppgitte målform.

## Bruk av fremmedord

Egennavn skriver vi uten å markere at det er på et annet språk: *Buckingham Palace*, *Svenska Akademien*. Dersom egennavnet inngår i et sitat på et annet språk, blir navnet også markert med det aktuelle språket: *America needs Medicare for all*.

Lånord som ikke er normert i Ordboka, men som står i NAOB, skriver vi uten å markere at ordene kommer fra et annet språk (*backing*, *happy*, *common sense*, *drop-out-er*, *backup*). Dette er ord som ofte, men ikke alltid, har en fornorsket uttale og norsk morfologi.

Hvis ordet ikke står i NAOB, markerer vi at ordet er fra et annet språk (*awkward*, *satisfy*). Det hender at ord av denne typen får norsk morfologi. I de tilfellene bruker vi bakstrek: *satisfye\tilfredsstille*, *awkward\pinlige*.

Fagtermer som er så spesialiserte at de ikke står i ordboka, f.eks. *schwa* eller *replikant*, får heller ikke bakstrekstranskripsjon, og de markeres ikke ytterligere.

### Uttrykk med ulikt antall ord på bokmål og nynorsk

I tilfeller der uttrykk har ulikt antall ord i de to målformene bruker vi også bakstrekstranskripsjon. Vi bruker underscore (\_) for å markere at ordsekvensen henger sammen:

- *foregå* > *foregå\gå\_føre\_seg*
- *ivareta* > *ivareta\ta\_hand\_om*
- på tross av > *på\_tross\_av\tross*

## Bøying og kongruens

### Unormert bøying

Vi velger alltid normerte bøyingsformer. Det hender at folk bøyer ord på en unormert måte. For eksempel kan det hende at man velger "galt" kjønn på substantiv, som /vedtakane/ (istedenfor *vedtaka*) på nynorsk eller /kjevlen/ (istedenfor *kjevlet*) på bokmål og nynorsk, eller at dialekten har apokopering (/kast/ for verbet *kaste* og /kist/ for substantivet *kiste*). Siden transkripsjonen vår skal følge normen, må vi i disse tilfellene velge en normert form i transkripsjonen. Eksemplene over blir transkribert henholdsvis *vedtaka*, *kjevlet*, *kaste* og *kiste*.

### Sterke verb

Sterke verb har ingen endelse i preteritum i begge målformer, og ingen endelse i presens i nynorsk. I tillegg kan sterke verb i begge målformer ha ulik vokal i infinitiv, preteritum og presens perfektum. De individuelle tempusformene til sterke verb er ikke bøyingsformer, men selvstendige former. Sterke verb følger derfor retningslinjene for leksikalske ord, ikke for unormert bøying. Vi bruker derfor bakstrekstranskripsjon ved forekomster av sterke verb der uttalevarianten avviker i stor grad fra verbets normerte form: *stende\står*, *gjenge\går*, *sym\svømmer*(BT), *kjem\kommer*(BT), *seie\si*(BT). Merk at dette ikke gjelder for innskuddsvokal i presens av sterke verb: /kjeme/ for *kjem* transkriberer vi *kjem*.

### Andre unntak:

Noen av j-verbenes former avviker i så stor grad fra bokmålsformene, og dermed også ofte i uttalemåten, at vi bakstreker disse. Dette forekommer stort sett i presens:

*velger\vel*, men for noen verb også i andre verbformer: *selger\sel* -- *solgte\selde* -- *solgt\selt*.).

For noen uregelrette svake verb (stort sett i presens) bakstreker vi også: *set\setter*, *eig\eier*.

Ved visse andre forekomster av verb som enten mangler en ekvivalent på motsatt målform (*spise\ete*), eller der uttaleformen avviker i stor grad fra den normerte formen (*auka\økt*) bruker vi også bakstrek.

## Kongruens

Vi retter ikke kongruens (jf. overordnet prinsipp om å normere ordnivå, ikke setningsnivå). Vi ønsker en troverdig gjengivelse av det som blir sagt, og det er dermed lite hensiktsmessig å legge til kongruens der det er tydelig at taleren ikke bruker det. Så lenge den brukte formen er en normert form, skriver vi den ut, selv om dette vil føre til at kongruensen mellom ordene ikke stemmer. Dersom representanten (nynorskbruker) sier: /eg er sterk ueinig i at dei burde bli prioritert/, transkriberer vi det som: *eg er **sterk ueinig** i at **dei** burde bli **prioritert***. Merk at det i en del tilfeller vil falle bort endelser som en konsekvens av fonologiske prosesser. Det er for eksempel vanlig at e-en faller bort der neste ord starter på en vokal, som i /offentli instanser/. Vi har fonologiske prosesser i bakhodet, og skriver ut den normerte formen i slike tilfeller der vokaler faller bort: *offentlige instanser*. Det er en fordel at talegjenkjenningssystemer lærer at en normert form, *offentlige*, kan ha ulike realiseringer, som /offentli/ og /offentlie/.

## Navn

Firmanavn, produktnavn, stedsnavn, personnavn o.l. skal normalt følge den offisielle stavingen: *Rema 1000*, *E6*, *Jette F. Christensen*. Det gjelder også for stor forbokstav, tall, tegnsetting og orddeling. Produkt- og firmanavn som har utradisjonell bruk av stor forbokstav, anførselstegn og orddeling, skal med andre ord ha det også i vår transkripsjon: *iPhone*, *OsloMet*, *AF Gruppen* og *Presley's Mat Gleden*.

Om et navn uttales på en måte som ikke samsvarer med offisiell rettskriving, skal vi likevel bruke den offisielle rettskrivinga i transkripsjonen. F.eks. skal *Majorstua*, *Kystruta* (for /Kystruten/), *Morgonbladet* transkriberes som henholdsvis *Majorstuen*, *Kystruten* og *Morgenbladet*. Ved forekomster av navn som har både en bokmåls- og en nynorskvariant, som for eksempel noen partinavn (*Arbeidarpartiet* og *Arbeiderpartiet*), velger vi den formen som svarer til den aktuelle målformen.

Navn på bilmerker, telefontyper o.l. brukes ofte til å referere til objekter: *en Audi*, *en iPhone* etc. I denne bruken kan navnene bøyes i bestemthet og tall på norsk. I slike

tilfeller bruker vi den offisielle praksisen for store bokstaver for det navnet, men legger til endelsen, altså *Audien* og *iPhonen*.

I en del tilfeller er et navn forleddet i et sammensatt substantiv, f.eks.

*Samsung-telefonen* eller *Kystruten-saken*. Slike substantiver kan [etter normen](#) enten skrives med liten forbokstav og ingen bindestrek eller stor forbokstav og bindestrek: *Samsung-telefonen* eller *samsungtelefonen*. I transkripsjonen velger vi konsekvent bindestrek og stor bokstav, altså *Samsung-telefonen*.

## Orddanning

Sammensetninger er frekvente i norsk. Vi godtar derfor sammensatte leksikalske ord i transkripsjonen, og transkriberer ord som *hensynta*, *oppfølge*, *imøtekoma*, *framstå* og *umogleggjere* uten å markere dem på noe vis. For detaljer om tegnsetting ved visse sammensetninger, se [dette avsnittet](#).

Sammensatte ord bestående av unormerte former som har normerte ekvivalenter

Vi ønsker å normere hele ordet der det lar seg gjøre. Ved forekomster av sammensatte ord som inneholder unormerte former som har normerte ekvivalenter normerer vi hele ordet. Til illustrasjon vil vi i nynorsktranskripsjonen (NT) transkribere /rettighetsutvalget/ som *rettighetsutvalget\rettsutvalet*. Dette gjør vi fordi førsteleddet *rettighet* har en klar nynorsk ekvivalent: *rett*. I dette og lignende tilfeller er det altså førsteleddet som trigger bakstrekstranskripsjonen. Merk at forekomsten av /utvalget/ alene ikke trigge bakstrekstranskripsjon i NT, fordi det eksisterer en nynorsk ekvivalent som er lydlig lik nok: *utvalet*.

Sammensatte funksjonsord forekommer også en del i tale. Som nevnt velger vi alltid normerte former for funksjonsord, noe som innebærer at vi deler opp ordene der det er nødvendig for at formene blir normerte. Til illustrasjon blir de sammensatte preposisjonene /utforbi/, /innigjennom/ og /utifra/ transkribert *ut forbi*, *inn igjennom* og *ut ifra*. Ved forekomster av preposisjoner hvor en kan velge mellom å skrive i flere ord eller å sammenskrive, som ved *ut over* og *utover*, velger vi det alternativet som er lydlig nærmest det representanten sier.

## Tall

Vi skriver tall med bokstaver.

Tall under 100 skriver vi i ett ord:

- *førtifire*
- *nittini*



Tall som blir uttalt på den gamle tellemåten transkriberer vi i ett ord i sin lydnære form før bakstrek, og i sin normerte form etter bakstrek:

- *femogseksti\sekstifem.*

Vi bruker også bakstrektranskripsjon ved forekomster av gamle tallord: *tyve\tjue*, *tredve\tretti*

Årstall skrives også i ett ord:

- *totusenogtretten*
- *tjuefjorten*

Tall over 100 skriver vi i flere ord:

- *hundre og seksten*
- *to hundre og førtien*
- *ett tusen fire hundre og niogseksti\sekstini*

## Tegnsetting

Hovedregelen er at vi ikke bruker tegnsetting i transkripsjonen. Punktum, komma, utropstegn og spørsmålstegn bruker vi ikke.

Der normeringen har tankestrek mellom sifre eller ord som angir strekning, periode og omfang (m.m), bruker vi mellomrom:

- *det vil ta tre fire dager* (ikke "tre–fire")

Det samme gjelder for tilfeller der normeringen har andre tegn:

- *tjuesytten tjueatten* (ikke "tjuesytten/tjueatten" (for 2017/2018))
- *åtte førtisju* (ikke "åtte:førtisju (for 8:47))
- *atten tretti* (ikke "atten.tretti" (for 18.30))
- *fire komma seks* (ikke "fire,seks" (for 4,6))

## Unntak

Vi bruker bindestrek der [normeringen](#) har det: *Helse- og omsorgsdepartementet*, *mødre- og barnedødeligheten*, *EØS-saker*. I sammensetninger der førsteleddet består av flere ord bruker vi bindestrek mellom hvert ledd: *elite-og-folket-diskusjonen*, *hundre-og-fem-åringer*, *grå-stær-pasient*. Det samme gjelder dersom sammensetningen består av et eller flere fremmedspråklige ord eller et eller flere unormerte ord som ikke har en normert ekvivalent, i tillegg til et normert sisteledd: *brain-drain-ordning*, *reset-initiativ* (men merk at dersom *brain drain*, eller *reset* forekommer alene, markerer vi dette med Språk: Engelsk (Storbritannia)(jf. [dette avsnittet](#))). Der førsteleddet er et navn bruker vi bare bindestrek mellom de to siste

leddene: *Stoltenberg II-regjeringen, The Fast and the Furious-filmen*. Vi bruker også bindestrek der deler av sammensetningen er tall: *åtti-åra*.

Vi bruker apostrof ved genitiv:

- *når det gjelder representanten Holmås' innlegg [...]*

I tilfeller der representanten uttaler en forkortelse slik den skrives, transkriberer vi forkortelsen med tegnsetting: */ref/* blir *ref..*

## Uforståelige og ufullstendige ord

Når man ikke forstår et ord som blir sagt, skal man sette inn en uforståelig-tag: [Uhørbart]. Det er ulike grunner til at et ord kan være vanskelig å oppfatte; lyden er dårlig, representanten "snøvler", en forstår ikke ordets betydning m.m. Vi bruker [Uhørbart]-taggen for alle disse tilfellene.

Vi transkriberer rene nølelyder. Vokalisk nølelyd markeres *ee* og nasal nølelyd markeres *mm*. Vi markerer også ikke-språklige lyder med en ordlignende kvalitet (latter, *ehe*, *atsjo*) med fellestagger *qq*. Andre ikke-språklige lyder markeres ikke (bank med gavelen, bakgrunnsstøy, applaus o.s.v)

Dersom det dreier seg om nøling eller stamming i starten av et ord, eller inni et ord, markerer vi ikke dette på noe vis: */realitetsbe- ee handles/* transkriberes *realitetsbehandles*. Tydelig avbrutte ord markerer vi med tegnet *∅*. Tilfeller som: */vi går til stem- votering/* og */i regj- ee Stortinget/* blir transkribert henholdsvis *vi går til stem∅ votering* og *i regj∅ Stortinget*. Også reparering markeres med *∅*: */Afg-Afghanistan/, /stemmese- stemmeseddel/* og */realitetsbehun- behandles/* transkriberes *Afg∅ Afghanistan, stemmese∅ stemmeseddel* og *realitetsbehun∅ behandles*. Det er ikke alltid like åpenbart hva som er nøling og stamming og hva som bør regnes som en omstart. Dette blir en skjønnsmessig vurdering hver enkelt transkribør foretar.

Dersom ordet er uttalt noe sleivete, men det er tydelig hvilket ord som menes, skriver vi ut den normerte formen. Vi transkriberer for eksempel *post* når det blir uttalt */poft/*. Ellers skriver vi ut alle normerte ord, også sammensetninger og avledninger, selv der det er tydelig fra konteksten at det ikke er det ordet som er "ment": */statsjorden/* for det som antakelig skal være *statsråden* skrives *statsjorden*. På samme måte blir henholdsvis */næringsdrivet/* og */riksvisjonen/* transkribert *næringsdrivet* og *Riksvisjonen*, til tross for at det sannsynligvis er *næringslivet* og *Riksrevisjonen* som er ment her. Dersom det er vanskelig å oppfatte hele eller deler av ordet, bruker vi [Uhørbart]-taggen. Vi bruker ikke mye tid på slike avgjørelser, vi bruker skjønn.

Dersom to eller flere talere snakker i munnen på hverandre bruker vi taggen [Overlappende]. Vi bruker denne taggen relativt liberalt i transkripsjonen. Det vil si at vi bruker [Overlappende]-taggen i alle tilfeller hvor talere snakker i munnen på hverandre – selv om det er helt klart hvem som har turen, og/eller vi hører en taler mye tydeligere enn en annen. I transkripsjonen gjør vi all overlappende tale til ett token som vi tagger som [Overlappende].