Reorganiserte taledatabaser for norsk talegjenkjenning fra Nordisk språkteknologi

Språkbanken, juli 2020

Innledning

Språkbanken ved Nasjonalbiblioteket mottok i 2011 flere talemålsressurser for de skandinaviske språkene fra konkursboet etter firmaet Nordisk språkteknologi (NST), blant dem flere taledatabaser for diktering, ASR og talesyntese. Taledatabasene for talegjenkjenning norsk ble i 2013 reorganisert, og dette dokumentet beskriver den reorganiserte versjonen av disse databasene. Generell informasjon om taledatabasene finner man i den opprinnelige dokumentasjonen.

I tillegg til at vi tilbyr denne taledatabasen som en nedlastbar ressurs, kan man også søke i og spille av lydfilene på denne sida.

Den opprinnelige versjonen av databasen kan lastes ned her.

Innholdet i databasene

Taledatabasene består av manuskriptlest tale med tilhørende annotasjoner og metadata. Materialet er delt inn i en treningsdel, kalt *nor0463*, og en testdel, kalt *nor0464.* I tillegg inneholder ressursen en deldatabase for diktering, *ADB_OD_Nor.NOR*. Metadataene er skrevet inn i JSON-filer, én fil per opptakssesjon, mens lydfilene er delt inn i én fil per manuskriptlinje, som ofte tilsvarer en setning.

I den opprinnelige versjonen av materialet slik det ble overlevert fra NST, var filene organisert i en bestemt mappestruktur der mappenavnene var meningsbærende (jf. dokumentasjonen). Filnavnene var imidlertid ikke meningsbærende, og det fantes også tilfeller av filer med liktlydende navn i forskjellige mapper. Dette viste seg å være upraktisk, siden brukere måtte beholde den opprinnelige mappestrukturen for å kunne benytte seg av dataene. Språkbanken har derfor navngitt filene på nytt, slik at filnavnene er unike og meningsbærende uavhengig av mappestrukturen. De opprinnelige metadatafilene var i spl-format. Disse er konvertert til JSON-format. De konverterte metadatafilene er også anonymisert og tekstkodingen er UTF-8 istedenfor ANSI, som det opprinnelige materialet hadde. Se under for mer informasjon om metadatafilene og lydfilene.

Navngivningskonvensjonen for metadatafiler er som følger:
språkkode+stasjonsnummer+x+informantkode-opptaksdato-klokkeslett_opprinnelig filnavn.json
Filnavnet *no10x005-02071999-1529_r4630005.json* gir med andre ord følgende informasjon:

- språket er norsk
- stasjonsnummeret er 10[1]
- informantnummeret er 005
- opptaksdatoen er 2.7. 1999
- Opptaksklokkeslettet er 15.29
- det opprinnelige filnavnet var *r4630005.spl*

Navngivningskonvensjonen for lydfiler er som følger:
språkkode+stasjonsnummer+x+informantkode-opptaksdato-klokkeslett_opprinnelig filnavn_kanalnummer.json
Filnavnet *no10x005-02071999-1529_u0005001-1.wav,* som er lydfila til en av manuslinjene i metadatafilen over, indikerer at det opprinnelige filnavnet var *u0005001.wav* og at den inneholder lyden fra kanal 1 (mer om kanaler under). Lydfiler der begge kanaler er med, inneholder ingen spesifikasjon av kanal.

Mappestrukturen i den reorganiserte taledatabasen er som følger:

- *ADB_NOR_0463/* inneholder metadatafiler for treningssettet
- *ADB_NOR_0464/* inneholder metadatafiler for testsettet
- *ADB_OD_Nor.NOR/* inneholder metadatafiler for dikteringsdatabasen
- *lydfiler_16_1/no* inneholder undermapper med navn av typen *no10x005-02071999-1529*, altså *språkkode+stasjonsnummer+x+informantkode-opptaksdato-klokkeslett.* Disse inneholder alle kanal-1-lydfilene med spesifikasjonen gitt av mappenavnet
- *lydfiler_16_2/no* er organisert som *lydfiler_16_1/no*, men inneholder lyden fra kanal 2.
- *lydfiler_16_begge/no* er organisert som de to foregående mappene, men inneholder filer med lyd fra begge kanaler.

Metadatafilene

Metadatafilene er i JSON-format. Hver metadatafil inneholder informasjon om én opptakssesjon med en informant. Merk at én opptakssesjon svarer til flere lydfiler, siden lydfilene er delt opp i én lydfil per manuslinje. Felt i metadatafila som oppgir filnavn, mappestier og tegnkoding, gir informasjon om tingenes tilstand i det opprinnelige materialet fra NST. Som forklart over, er tegnkoding endra, og mappestrukturen er ikke lenger nødvendig for å identifisere korrekt fil.

[1] Merk at stasjonsnummeret i filnavnene ikke alltid samsvarer med verdien til *station* i metadatafilene. På tidspunktet når denne dokumentasjonen er skrevet, er det uklart hva denne diskrepansen kommer av.

Under kommer informasjon om de enkelte feltene:

- *info* inneholder metadata om informanten.
- *pid* inneholder identifikatoren for opptakssesjonen som blir brukt til å lage de nye filnavnene, altså *språkkode+stasjonsnummer+x+informantkode-opptaksdato-klokkeslett.*
- *session* og *system* inneholder metadata om opptakssesjonen, opprinnelig mappesti og diverse teknisk informasjon.
- *val_recordings* inneholder de enkelte de enkelte manuslinjene og metadata knytta til disse.
    - *text* inneholder manuslinjas tekst.
    - *file* inneholder det opprinnelige filnavnet til lydfila som gjengir manuslinja. (Nåværende filnavn finner man ved å koble sammen id-en i *pid*-feltet og det opprinnelige filnavnet, som beskrevet over.)
    - *DST, NOI, QUA, SND, SPC, UTT* er felt som skal inneholde informasjon om valideringa av opptakene. Disse feltene ser ut til å være ubrukte i de fleste tilfeller og inneholder kun default-verdier.
    - *t0, t1* og *t2* er tall knyttet til de enkelte manuskriptlinjene i de opprinnelige metadatafilene. Det er uklart hvilken funksjon disse hadde.
    - *type* klassifiserer manuskriptlinjene i forskjellige kategorier. Verdiene består ofte av en kategoriindikator + et tall, for eksempel *CD20*, hvor *CD* angir at manuskriptlinja er en sekvens med tall. Noen av typekodene er mer komplekse, for eksempel *ISp3*, hvor *IS* er kategoriindikatoren for setninger. Mens kategoriindikatorene delvis identifiserbare, er det ikke klart hvilken signifikans de øvrige tegnene i typekodene har. De følgende kategoriindikatorene ble identifisert da taledatabasen ble konvertert i 2013:
        - *FF:* teksten er et flerordstall, for eksempel “femtini tusen sju hundre”
        - *IS:* teksten er en setning
        - *CD:* teksten er en sekvens med tall
        - *pIW:* teksten er en staving av et ord
        - *prIW:* teksten er et personnavn (fornavn og etternavn)
        - *CIW:* teksten er et egennavn (sted eller person)
        - *phIW* og *IW:* teksten består av ett ord

Lydfilene

Lydfilene i den originale versjonen av taledatabasen var i wav-format med 16 kHz samplingfrekvens, 16 bit oppløsning og to kanaler (nær og fjern).[2] Disse filene finner man i mappa *lydfiler_16_begge/no* i den nye versjonen av databasen. I forbindelse med reorganiseringa i 2013 ble de to kanalene også splitta ut ved hjelp av programmet Sox. *lydfiler_16_1/no* inneholder wav-filer med den nære kanalen, mens *lydfiler_16_2/no* inneholder lyden fra den fjerne kanalen.

[2] I dokumentasjonen som ligger ved den originale versjonen av databasen, står det at filformatet skal være ukodede rådata, men filene som ligger i databasen, er i wav-format.

I dumpen i Språkbankens ressurskatalog er lydfilmappene splitta opp i flere *tar.gz*-filer for å unngå at størrelsen på de nedlastbare filene ikke blir for stor: *lydfiler_16_1_a.tar.gz, lydfiler_16_1_b.tar.gz*. Mappene er fordelt på samme måte for kanal 1, 2 og begge. For eksempel vil filene i *lydfiler_16_1_a.tar.gz lydfiler_16_2_a.tar.gz* and *lydfiler_16_begge_a.tar.gz* komme fra de samme opptakssesjonene.