Tiedoston puhetietokanta.xls (ja puhetietokanta.sav) sisältämä aineisto koskee puhetiedonhaun tutkimustarpeita varten koottua puhetietokantaa. Tietokanta koostuu 288 puhutusta uutisesta.

Tilastoyksikkö on yksittäinen tietokannan uutinen.

Aihe
Jokainen uutinen kuuluu yhteen aihekategoriaan. Uutisille annetut aihetunnukset liittyvät siihen, mihin tiedonhaun testikysymykseen ne vastaavat. Toisin sanoen tietokannan tekijä on valinnut tietokantaan joukon uutisia, jotka kaikki koskevat jotakin tiettyä aihetta, jotta näitä aiheita myöhemmin voisi käyttää tiedonhaun menetelmien testaukseen. Siispä esimerkiksi uutiset, joiden aihetunnus on 2, käsittelevät kaikki jollakin tavalla Latinalaisen Amerikan maiden velkataakkaa kun taas uutiset tunnuksella 3 koskevat Yhdysvaltojen polkumyyntisyytöksiä suomalaisia paperiyhtiöitä vastaan jne.

ID
ID on uutisen yksikäsitteinen tunnusnumero tietokannassa. Sillä ei ole muuta tehtävää kuin yksittäisen uutisen yksilöiminen.

Muuttujat

Uutisessa sanoja
Uutisen sisältämien sanojen määrä yhteensä. Laskettu siitä tekstistä, jonka pohjalta puhuttu uutinen on luettu ääneen.

Uutisessa merkkejä
Uutisen sisältämien merkkien määrä. Laskettu siitä tekstistä, jonka pohjalta puhuttu uutinen on luettu ääneen. Tyhjiä merkkejä ja välimerkkejä ei ole laskettu.

Kesto
Puhutun uutisen kesto sekunteina.

Tunnistuksen virheprosentti, v1--v4
Puheaineiston tunnistamiseen on käytetty äännetunnistinta. Äännetunnistin tunnistaa puheesta yksittäisiä äänteitä, eikä siis tuota puheesta esimerkiksi tekstiä. Tunnistuksen virheprosentti kertoo, kuinka suuri osuus kaikista tunnistimen antamista merkeistä ovat virheellisiä. Esimerkiksi 40% virheprosentti tarkoittaa, että 4/10 tunnistustuloksessa olevasta merkistä on tunnistettu väärin. Eri menetelmillä (v1--v4) saadaan erilaisia tunnistustarkkuuksia. Menetelmät perustuvat erilaisiin tapoihin käsitellä tunnistimen tuottamaa väliaikaista transkriptiota.

Sanamäärät sanojen pituuden mukaan
Uutisen sanamäärä (aiemmin) kertoo vain kuinka monta sanaa uutinen sisältää. Taulukossa on esitetty yhden uutisen sisältämät sanamäärät sanojen pituuden mukaan. Toisin sanoen taulukossa on esitetty erikseen kuinka monta 2, 3, 4, ..., jne. kirjaimen pituista sanaa jokainen uutinen sisältää.