

Suomenkielisen keskustelupuheen segmentoinnin ja nimikoinnin ongelmia

Mietta Lennes & Hanna Anttila

mietta.lennes@helsinki.fi

haanttil@cc.helsinki.fi

Puheen akustis-foneettista tutkimusta varten on tapana rajata puhesignaalista erilaisia kielellisiä tai foneettisia yksiköitä vastaavat segmentit, esim. lauseet tai äännökset, sanat, tavut, foneemit ja/tai äänneet eli foonit. Kullekin eri tasolla rajatulle segmentille annetaan myös tietynlainen symbolinen kuvaus tai tunniste, joka voi olla esimerkiksi ortografinen litteraatio tai foneettinen transkriptio. Tällaista menettelyä kutsutaan *nimikoinniksi* ja sitä käytetään yleensä puhetietokantaan liitettävien puhesignaalien indeksointiin, jolloin tietokannasta voidaan etsiä tarkasteltavaksi halutut osat kielellisillä tai foneettisilla kriteereillä ja kohdistaa analyysi niihin. Käytettävissä olevat hakukriteerit - ja siis myös tietokannan käyttöarvo - määräytyvät pitkälti sen perusteella, minkälaisia yksiköitä ja periaatteita nimikointivaiheessa on sovellettu. Nimikointityö vaatii tekijältään kokemusta ja huolellisuutta ja vie paljon aikaa. Toistaiseksi ei kuitenkaan ole saatavilla automaattisia työkaluja, joilla saavutettaisiin sama nimikointitarkkuus kuin ihmisen tekemänä.

Puheen segmentointi perustuu ajatukseen, että diskreetit kielen yksiköt (foneemit) edustuvat puheessa jonkinlaisina peräkkäisinä rakenteina (äänteinä), ja että näillä yksiköillä on kielessä hierarkkisia suhteita (lauseet koostuvat sanoista, sanat tavuista ja äänneistä jne.). Kun puhesignaalia tarkastellaan lähempää, huomataan, ettei äänneiden välillä ole selkeitä akustisesti tai kuulonvaraisesti määritettäviä rajakohtia. Segmentoitavien äänneiden lukumäärä ei tämän vuoksi ole yksiselitteinen. Jopa tehtävään koulutetut nimikoijat ovat usein erimielisiä äänneiden tarkoista rajakohdista ja käytettävistä foneettisista symboleista. Jos taas yritetään ensisijaisesti segmentoida puheesta kielen foneemeja vastaavia yksiköitä, tulee vastaan tilanteita, joissa oletettua yksikköä ei löydy puhesignaalista tai kahden foneemin rajakohtaa ei voida johdonmukaisesti määrittää. Spontaanin puheen taustalla oleva foneemirakenne ei myöskään aina ole ilmeinen, koska jotkut sanat eivät ole yksiselitteisesti tunnistettavissa puhesignaalia kuuntelemalla. Sellaisesta sanasta, joka ei ole vakiintunut kirjoitettuun kieleen ja joka esiintyy painottomana, on yllättävän vaikea määrittää foneemisia yksiköitä tai vaikkapa niiden kvantiteettia. Myös sana- tai tavarajan paikantaminen puhesignaalista on hankala tehtävä - syntyperäisen kielenpuhujan intuitiollakin.

Spontaanin ts. epämuodollisen puheen kohdalla nimikointiin liittyvät ongelmat ovat pahimmillaan, koska foneettinen vaihtelu on suurta ja koska arkipuhe on kielellisiltä rakenteiltaan erilaista kuin teksti. Esittelemme eräitä ongelmatapauksia, joita spontaanin suomenkielisen puheen nimikoija joutuu kohtaamaan, ja pohdimme niiden ratkaisumahdollisuuksia.