Loppuraportti
Monikielisten ääniuutisten automaattinen luominen tekstimuotoisista uutisista

Isoft.ai Oy
Yrttipellontie 1D
90230 Oulu

Sisältö

1. Tutkimuksen menetelmät ja aineistot

Tutkimuksessa käytettiin Helsingin Sanomien verkkolehden tekstimuotoisia uutisartikkeleita lukemalla uudet artikkelit HS Tuoreimmat RSS syötteen avulla.

Isoftin Ipodcast.ai -alustan avulla HS artikkeleista luotiin MP3 tiedostoformaatissa olevia ääniuutisia sekä suoritettiin tekstin automaattinen kääntäminen englanniksi.

Ääniuutisten luomisessa käytettiin lisäksi sekä Googlen että Microsoftin äänisynteesipalveluja. Amazonin vastaavaa palvelua ei voitu käyttää puuttuvan Suomen kielen tuen vuoksi.

Automaattisessa kääntämisessä käytettiin pääsääntöisesti Google Translate palvelua. Microsoftin vastaavaa palvelua käytettiin vain alkuvaiheessa.

Touko-lokakuun aikana analysoitiin 550 uutisartikkelia. Analysointiin käytettiin tarkoitusta varten kehitettyä web työkalua, jonka avulla ääniuutisia voitiin kuunnella ja verrata alkuperäisen uutisartikkelin tekstiin. Arvioinnista jätettiin pois artikkelit, joiden sisältö ei sovellu ääniuutisiin (kuten videoon perustuvat uutiset). Arvioinnit tallennettiin omaan raportointi tietokantaan, josta kerättiin loppuraportointiin tarvittavat tiedot

Analysointia suorittavat henkilöt kuuntelivat artikkelit ja kirjasivat virheet sekä huomiot äänen tasokkuudesta: Lisäksi analysoija tarkasti automaattisen käännöksen lopputuloksen ja kirjasi virheet automaattisessa kääntämisessä.

Analysoidut artikkelit sisälsivät kokonaisuudessaan noin 10 800 lausetta

Tutkimuksen vaiheet

2. Tutkimuksen tulokset

2.1. Ääniuutiset – havaitut virheet

Ääniuutisissa havaittiin kaikkiaan 140 vakavaa virhettä 547:ssä uutisartikkelissa

2.1.1. Virheiden esiintymistiheys

Virheiden esiintymistiheydet tutkimusaineistossa selvitettiin uutisartikkeli- ja artikkelien lause tasoilla.

Virheiden esiintymistiheys artikkeleissa oli noin 20%, joka viidennessä uutisartikkelissa oli vakava tekstistä puheeksi muunnosvirhe.

Virheitä	Artikkeleita	Esiintymistiheys/artikkeli
114	**547**	**20,84 %**

Virheiden esiintymistiheys lauseissa oli noin 1,29%.

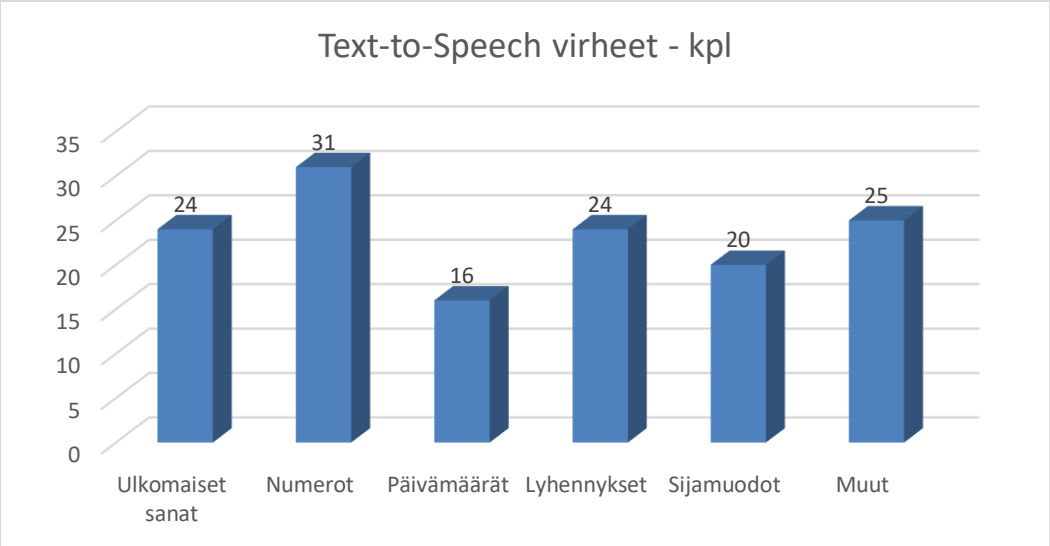
Virheitä	Lauseita	Esiintymistiheys/lause
140	**10815**	**1,29 %**

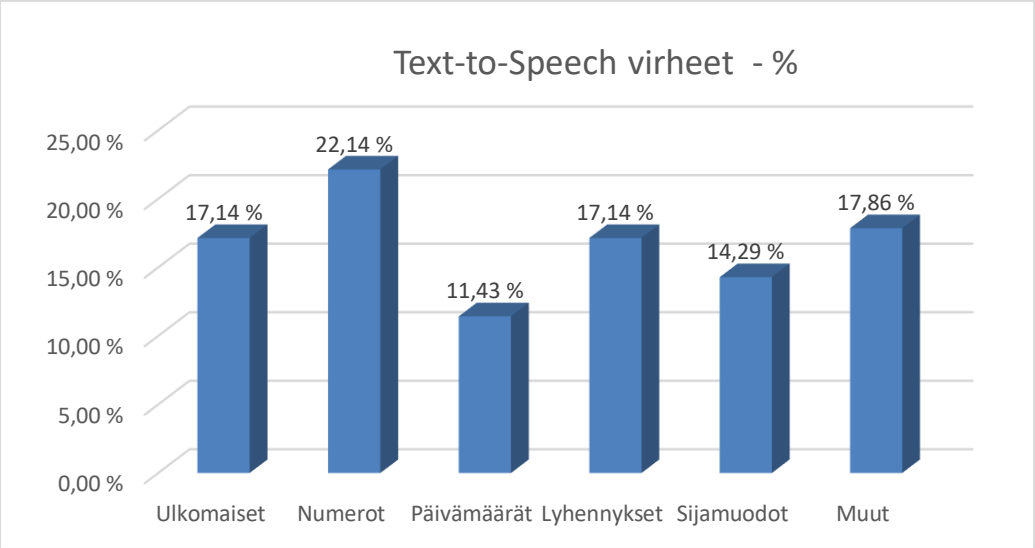
2.1.2. Virheiden tyypit

Vakavat virheet puhesynteesin luomisessa voitiin jakaa seuraaviin luokkiin:

- **Ulkomaiset sanat.**
  Ulkomaisten sanojen lausuminen virheellistä tai kömpelösti äännettyä (”rallienglantia”)
- **Numerot**
  Numerosarjat, desimaaliluvut, tekstistä ääneksi muunnos virheellinen
- **Päivämäärät**
  Päivien, kuukausien ja vuosien tekstistä ääneksi muunnos virheellinen
- **Lyhennelmät**
  Sanojen lyhennelmissä väärä merkitys tai lyhennys muuttuu yhtenäiseksi sanaksi
- **Sijamuodot**
  Suomen kielen sijamuoto virheellinen
- **Muut virheet**
  Kaikki muut virheet

Virheiden jakautuminen virheluokittain:





Kaiken kaikkiaan virheet jakautuivat melko tasaisesti eri luokkiin.

Eniten virheitä aiheuttivat numerot (22%), lyhennykset (17%) sekä ulkomaiset sanat (17%).

2.2. Automaattinen kääntäminen – havaitut virheet

Automaattisessa käännöksessä Suomesta Englanniksi havaittiin yhteensä 229 vakavaa virhettä. Vakavassa virheessä alkuperäinen merkitys muuttuu kokonaan toiseksi tai ei ymmärrettäväksi.

2.2.1. Virheiden esiintymistiheys

Virheiden esiintymistiheydet tutkimusaineistossa selvitettiin uutisartikkeli ja artikkelien lause tasoilla.

Virheitä	Lauseita	Esiintymistiheys/lause
229	**10815**	**2,12 %**

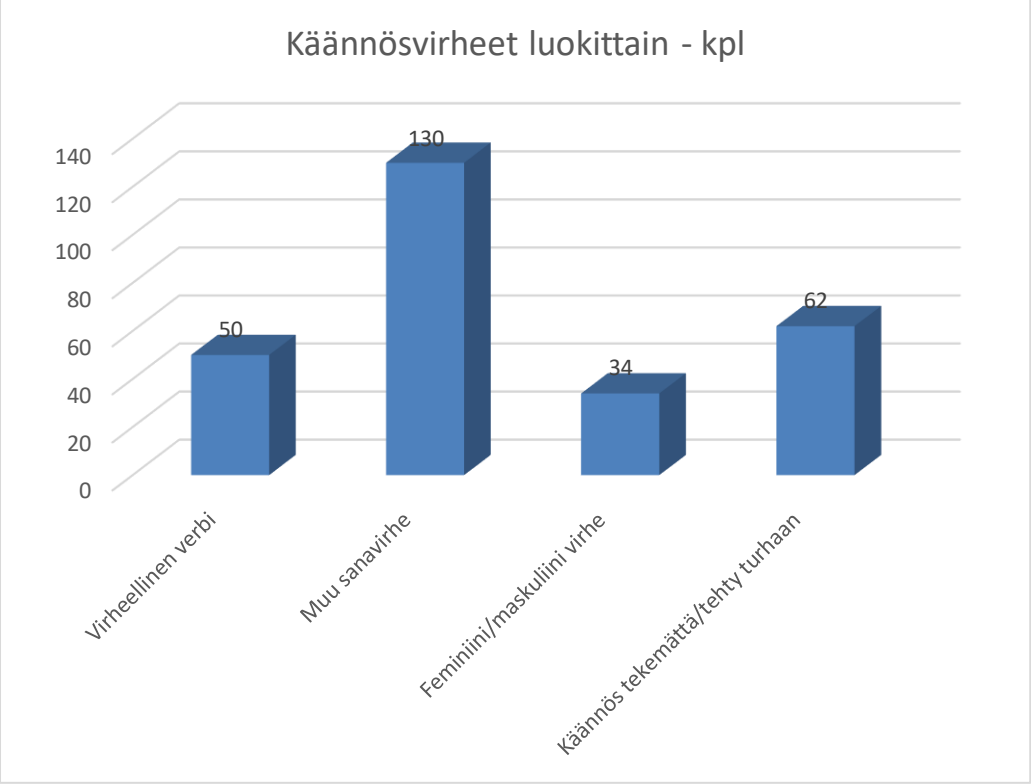
Virheellisiä artikkeleita	Artikkeleita	Esiintymistiheys/artikkeli
178	**547**	**32,54 %**

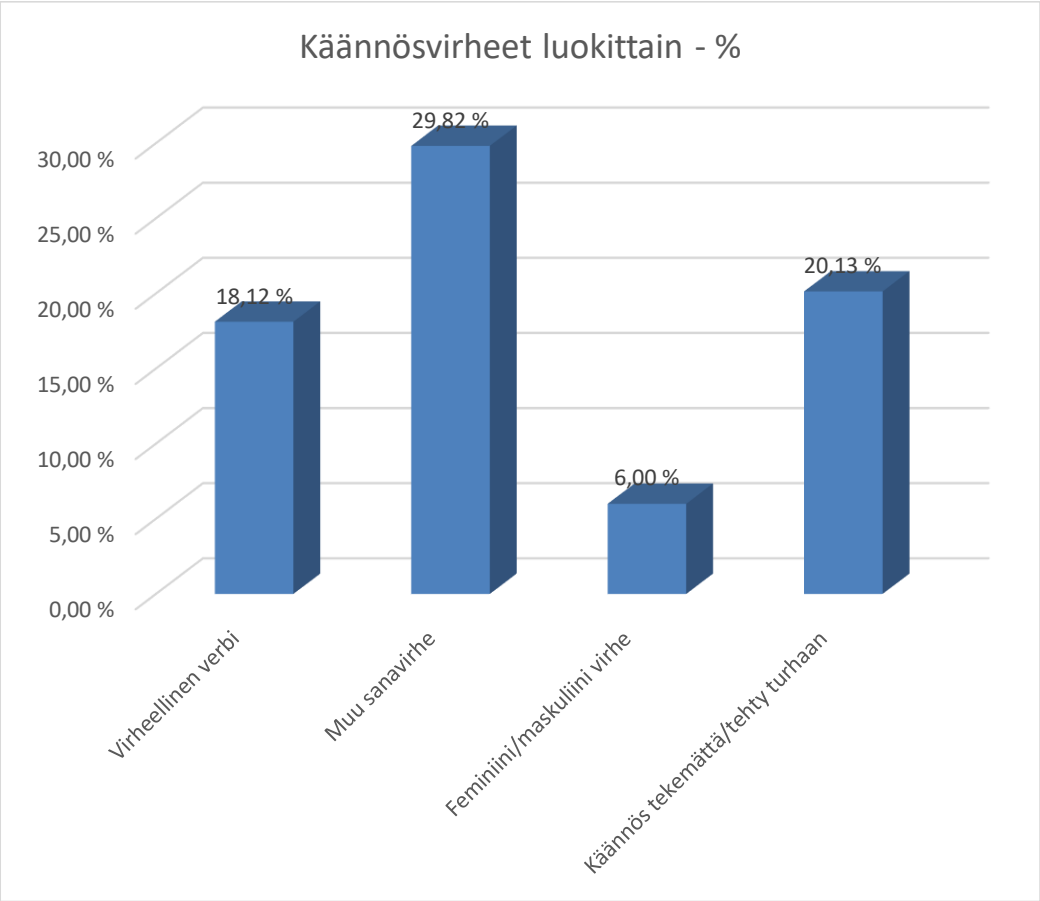
2.2.2. Virheiden tyypit

Virheet automaattisessa käännöksessä voitiin jakaa seuraaviin luokkiin:

- **Verbit.**
  Suomenkielinen verbi kääntyi muotoon, joka muutti lauseen merkityksen toiseksi tai teki lauseesta vaikeasti ymmärrettävän
- **Muut sanavirheet**
  Suomenkielinen sana (pois lukien verbit) kääntyi virheelliseen tai epäsopivaan muotoon, joka muutti lauseen merkityksen virheelliseksi tai vaikeasti ymmärrettäväksi
- **Feminiini/maskuliini -muunnosvirheet**
  Suomenkielinen feminiini/maskuliini muoto kääntyi virheellisesti väärään muotoon
- **Muut käännösvirheet**
  Suomenkielistä sanaa ei käännetty ollenkaan tai erillisnimi/lyhenne käännetty vaikka ei olisi pitänyt

Virheiden jakautuminen virheluokittain:





Eniten käännösvirheitä aiheuttivat sanavirheet (29,8%), käännös tekemättä tai turha käännös (20,13%) ja virheelliset verbit (18,12%).

3. Tulosten hyödyntäminen

3.1. Ääniuutiset

Suomenkielisten tekstimuotoisten uutisten muuntamisessa text-to-speech palvelujen avulla ääniuutisiksi havaittiin selkeitä ongelma-alueita joihin virheet keskittyvät.

Kaikki markkinoiden johtavat text-to-speech palvelut (Google, Microsoft, Amazon, IBM) tukevat Speech Synthesis Markup Language (SSML) standardin mukaista äänisynteesiin optimointi- ja korjaussyntaksia. SSML on W3C standardi, joka määrittelee tapoja ohjata puhesynteesin suorittamista (https://www.w3.org/TR/speech-synthesis11/).

Tutkimuksessa havaittiin, että SSML ominaisuuksia hyödyntämällä huomattava osa löydetyistä virheistä voidaan korjata.

Isoft on kehittänyt erilaisia ääniuutispalveluja useiden vuosien ajan. Tutkimuksen tuloksia hyödynnetään Isoftin uusien palvelujen tuotekehityksessä. Tavoitteena on tarjota korkealaatuinen ääniuutispalvelu suomalaisille media-alan yrityksille sekä muille yrityksille, joilla on tarpeita text-to-speech ratkaisujen alueella.

3.2. Automaattinen kääntäminen

Uutisartikkelien automaattisessa kääntämisessä suomesta englanniksi havaittiin lukuisia virheitä. Virheet pystyttiin luokittelemaan muutamaan pääongelma-alueeseen.
Käytetyn Googlen kääntäjän suurin ongelma on laajemman kontekstin ymmärtämisen puute. Käännökset onnistuvat kohtalaisen hyvin lauseen tasolla mutta laajemman kontekstin ymmärtämisessä on puutteita. Tämä aiheuttaa runsaasti virheitä.

Sekä Googlen että Microsoftin käännöspalveluja voidaan opettaa ymmärtämään paremmin suomen kieltä ja sen erikoisuuksia. Erilaisten sanastojen tai pidemmälle kehitettyjen uusien tekoälymallien avulla voidaan virheiden määrä saada pienemmäksi. On kuitenkin huomattava että tarvittava työmäärä voi olla huomattava.

Tutkimuksen tuloksia hyödynnetään Isoftin uusien palvelujen tuotekehityksessä uusien sanastojen ja tekoälymallien kehityksessä. Tavoitteena on tarjota korkealaatuinen käännöspalvelu suomalaisille media-alan yrityksille sekä muille yrityksille, joilla on tarpeita automaattisen kääntämisen ratkaisujen alueella.