Korpus verkkoteksteistä

- kielentunnistus

Korpus verkkoteksteistä

- Mikä korpus?
- Mikä ”verkkokorpus”?
- Miksi?
- Miten?
  - Hakukoneet
  - Esi-/jälkikäsittely
  - Www-sivujen lataus/siivous/analysointi
- Miten tekstin kieli tunnistetaan?

Kielentunnistus

- Tekniikat
  - N-grammit
  - HTTP-protokolla (Html-header), pelkkä URL?
  - Yleisten sanojen lista
  - Uniikit merkkiyhdistelmät (huono?)
- Kaksi/monikieliset tekstit
  - Esimerkiksi kirjanorja/uusnorja, baski/espanja
  - Miten käsitellään/erotetaan
  - Kielentunnistus kappaletasolla (jopa lausetasolla)

Lähteitä

- Web as Corpus:
  - The World Wide Web as Linguistic Corpus
  - Can we beat Google?
  - Corpus Linguistics and the Web
- Kielen tunnistus (yleinen ja www)
  - Natural language identification using corpus-based models
  - Language Identification in Web Pages
  - Web Page Language Identification Based on URLs
  - Language Identification for Multilingual Documents