HELSINKI UNIVERSITY OF TECHNOLOGY
Media Technology

Airo: Ontologiaperustainen indeksointi- ja hakutyökalu

Matias Frosterus, Olli Alm
Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK),
Laboratory of Media Technology
and
University of Helsinki, Department of Computer Science

http://www.seco.tkk.fi

UNIVERSITY OF HELSINKI

SeCo
semantic computing

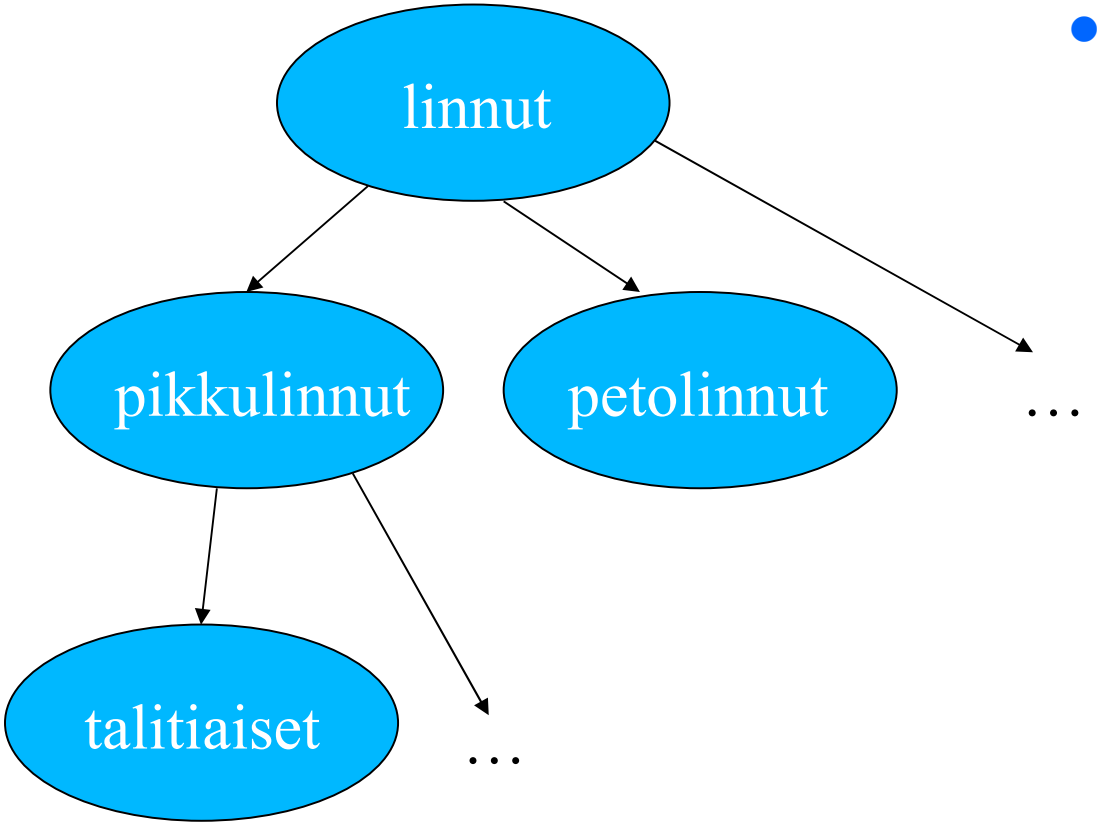
Tavoite

- Hakuominaisuuksien parantamien tekstihaulle ontologioita hyödyntämällä
  - sovelluskohteena esimerkiksi sanomalehtiarkistot
    - » miljoonia artikkeleita
    - » lisää joka päivä
      - vaatimuksena automaattisuus
- Tekstistä löydettävien ontologisten käsitteiden laajentaminen aineiston ontologista jatkohyödyntämistä silmällä pitäen

Menetelmä

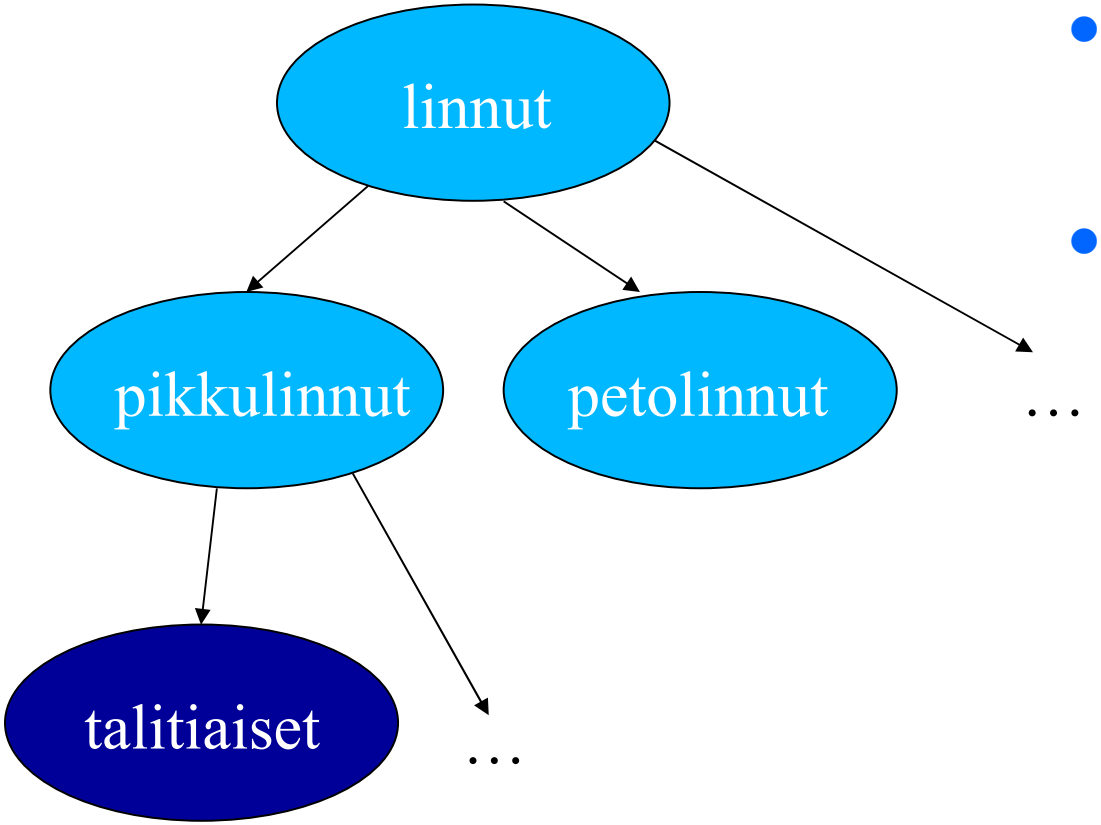
- Automaattinen käsitteiden tunnistaminen tekstistä
  - tapahtuu Pokalla
  - eri ontologioiden käsitteet talletetaan omiin indekseihinsä, jolloin haun aihepiiriä voi rajata valitsemalla ontologia, jonka käsitteitä haussa hyödynnetään
- Käsiteklusterointi

Käsiteklusterointi 1/4



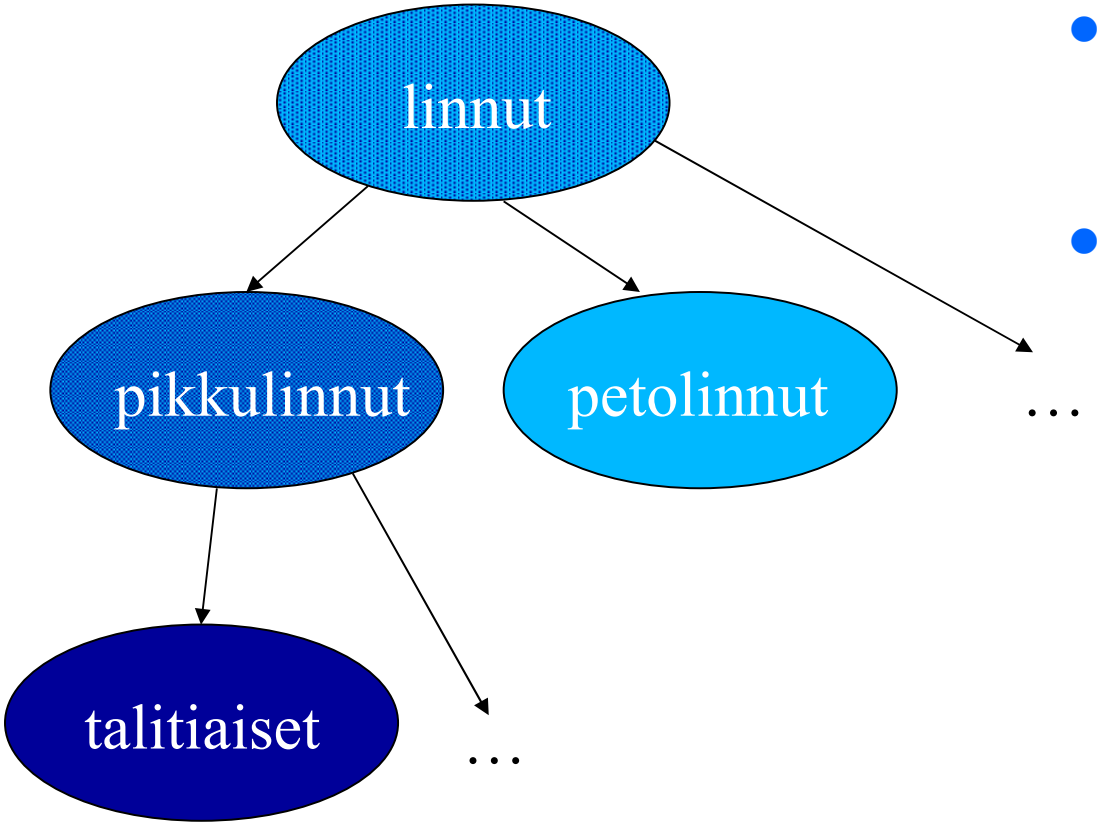
- Ontologinen hierarkia
  - nuolet kuvaavat aliluokkasuhteita

Käsiteklusterointi 2/4



- Ontologinen hierarkia
  - nuolet kuvaavat aliluokkasuhteita
- Tekstistä löytyy termi ”talitiainen”
  - täsmää ontologian käsitteeseen talitiaiset

Käsiteklusterointi 3/4



- Ontologinen hierarkia
  - nuolet kuvaavat aliluokkasuhteita
- Tekstistä löytyy termi ”talitiainen”
  - täsmää ontologian käsitteeseen talitiaiset
  - Nostaa myös lähikäsitteiden painoa
  - laajuus ja voimakkuus vapaasti konfiguroitavissa

Käsiteklusterointi 4/4 - Esimerkki

- ”Maija Meikäläinen on missivuotensa jälkeen luonut uraa mallina.”
  - Tekstistä löydetään mm. käsitteet missi ja malli
  - Termi ’malli’ vastaa kolmeen käsitteeseen
    - » ”pienoismalli”-malli
    - » ”valokuvamalli”-malli
    - » ”roolimalli”-malli
  - YSO:ssa missit-käsitteellä ja ”valokuvamalli”-mallit-käsitteellä on assosiatiivinen suhde
    - » Missi-termi voimistaa ”valokuvamalli”-tulkintaa eli kyseisen lauseen ”valokuvamalli”-malli-käsite saa lisää painoa
  - Hakukone tulkitsee lauseen relevantimmaksi ”valokuvamalli”-malli haulle kuin lauseen ”Uudet Volvo-mallit ovat saapuneet kauppoihin.”

Järjestelmän hyödyt

- Ontologiaa voidaan hyödyntää hakujen saannin parantamiseen
  - esim. Iranista ja öljystä kertova artikkeli vastaa hakuun ’öljyvaltiot’ vaikka sana ei edes esiintyisi
- Suosittelu
  - ”Nämä artikkelit vastasivat sanahakuun. Katso myös nämä, jotka sisältävät samoja käsitteitä.”
- Käsitteistys antaa pohjan aineiston jatkohyödyntämiselle ontologisia menetelmiä hyödyntäen

Airo

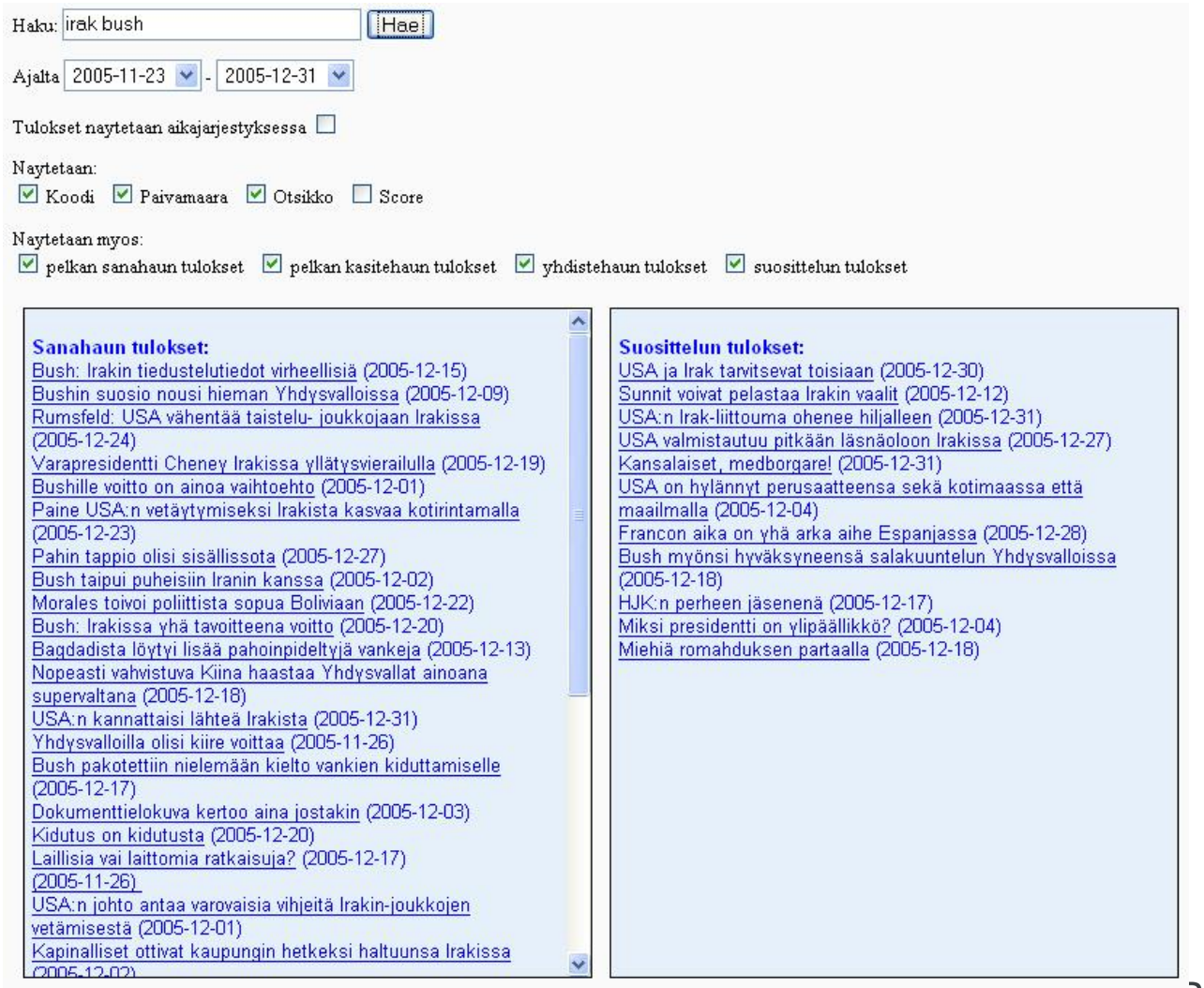
- Mahdollisuus sanahakuun ja käsitehakuun, sekä näiden yhdistelmään
- Voidaan esimerkiksi toteuttaa perinteinen sanahaku ja lisätä sen rinnalle suosittelu käsitteillä
- Yksinkertainen hahmokieli klusterointiohjeiden luomiselle

Airo

- Toteutuksessa käytettiin Yleistä suomalaista ontologiaa
- Haku toteutettiin Lucenella
  - käytössä mm. Wikipediassa
- Testattiin European Language Resources Associationin CLEF Test Suitella
  - Tulokset positiivisia

- Projektin aikana valmistui diplomityö otsikolla Tekstiaineiston ontologiaperustainen indeksointi ja haku

Airo



HELSINKI UNIVERSITY OF TECHNOLOGY
Media Technology

UNIVERSITY OF HELSINKI

SeCo semantic computing