FACULTY OF ARTS
Charles University

Valentina Dani

Tohtorikoulutettava – Matemaattisen kielitieteen tohtoriohjelma

Johdanto oppijansuomen korpuksiin

Sisältö

- Johdanto oppijansuomen korpuksiin
- Korpustekstien käsittely ja virheannotointi
- Yleiset virhetyypit

Johdanto oppijansuomen korpuksiin

(Ivaska 2014, Jantunen & Pirkola 2015)

- ICLFI: Kansainvälinen oppijansuomen korpus (International Corpus of Learner Finnish) (2007→)
  - sisältää yliopisto-opiskelijoiden kirjallisia tuotoksia
  - aineisto on kerätty yli 20 yliopistossa Suomen ulkopuolella
  - koko: noin miljoona sanetta
  - lisätietoja: https://www.oulu.fi/suomitoisenakielena/node/16078

- Edistyneiden suomenoppijoiden korpus LAS2 (The Corpus of Advanced Learner Finnish – LAS2) (2007→)
  - sisältää edistyneiden suomenoppijoiden kirjallisia tuotoksia
  - aineisto on kerätty Turun yliopistossa – korpus sisältää suomen ja sen sukukielten maisteriohjelman opiskelijoiden kirjoittamia tekstejä
  - koko: noin 630 000 sanetta
  - lisätietoja: https://jyx.jyu.fi/handle/123456789/45018

Muita oppijansuomen korpuksia

- YKI-korpus: Yleisten kielitutkintojen korpus
  - sisältää englannin, espanjan, italian, ranskan, ruotsin, saamen, saksan, venäjän ja suomen kirjoitettujen ja puhuttujen kielitaitotekstien aineistoa
  - suomen kielen tutkinnon suorittajat ovat pääasiassa henkilöitä, jotka tarvitsevat todistuksen Suomen kansalaisuuden hakemista varten (Jantunen & Pirkola 2015: 89)
- Cefling-korpus:
  - sisältää yläkouluikäisten suomen ja englannin oppijoiden kirjoitelmia
- Topling-korpus:
  - sisältää suomi toisena kielenä-oppijoiden sekä englannin ja ruotsin oppijoiden kirjallisia tuotoksia
  - aineisto on diakroninen: korpus sisältää alakoululaisten, yläkoululaisten ja lukiolaisten kirjoittamia tekstejä
- Muita oppijansuomen korpuksia: Dialuki, Long Second

Korpustekstien käsittely ja virheannotointi

- Korpusteksteihin on mahdollista lisätä useita tietoja, joiden avulla käyttäjä voi suorittaa erilaisia hakuja:
  - taustatiedot: tietoja kirjoittajasta ja tekstistä
  - lemmatisointi: tieto sanan perusmuodosta eli lemmasta
  - kieliopillinen ja syntaktinen annotointi: sanaluokan, sijamuodon sekä lauseenjäsenen koodaaminen
  - virheannotointia: virheannotointi mahdollistaa virheiden hakua aineistosta. Virhekoodattu korpus paljastaa epätyypilliset muodot ja mahdollistaa virheiden tehokkaan hakemisen virhetyypin tai tietyn kielenoppijaryhmän mukaan. Materiaalista voidaan löytää sekä odotuksenmukaisia että täysin ennakoimattomia virheitä. (Jantunen, Brunni, Lehto, & Skantsi, 2014: 65)

Korpustekstien käsittely ja virheannotointi

Mitä tein kesälomalla

Minusta kesäloma oli, kuten tavallista, liian lyhyt (kuitenkin se kesti kolme kuuta). Loman alussa kaksi kaveriani tulivat minulle Krumloviin. He jäivät meille koko viikonloppua ja koimme yhdessä kaupungissa luonnossakin. On vahinko, että meille ei ollut enemmän aikaa. Muuten minun täytyi tehdä bakalaarisen esseeni, koska haluan lopulta lopettaa opiskeluni humanistisessa tiedekunnassa (anteeksi, mutta en ole varmaa, miten tämä sanoa suomeksi). Valitettavasti minun täytyy tunnustaa, että en ollut oikein ahkeraa, niin että en ole tehnyt sitä loppuun. [TS0001f]

…

Korpustekstien käsittely ja virheannotointi

Mitä tein kesälomalla

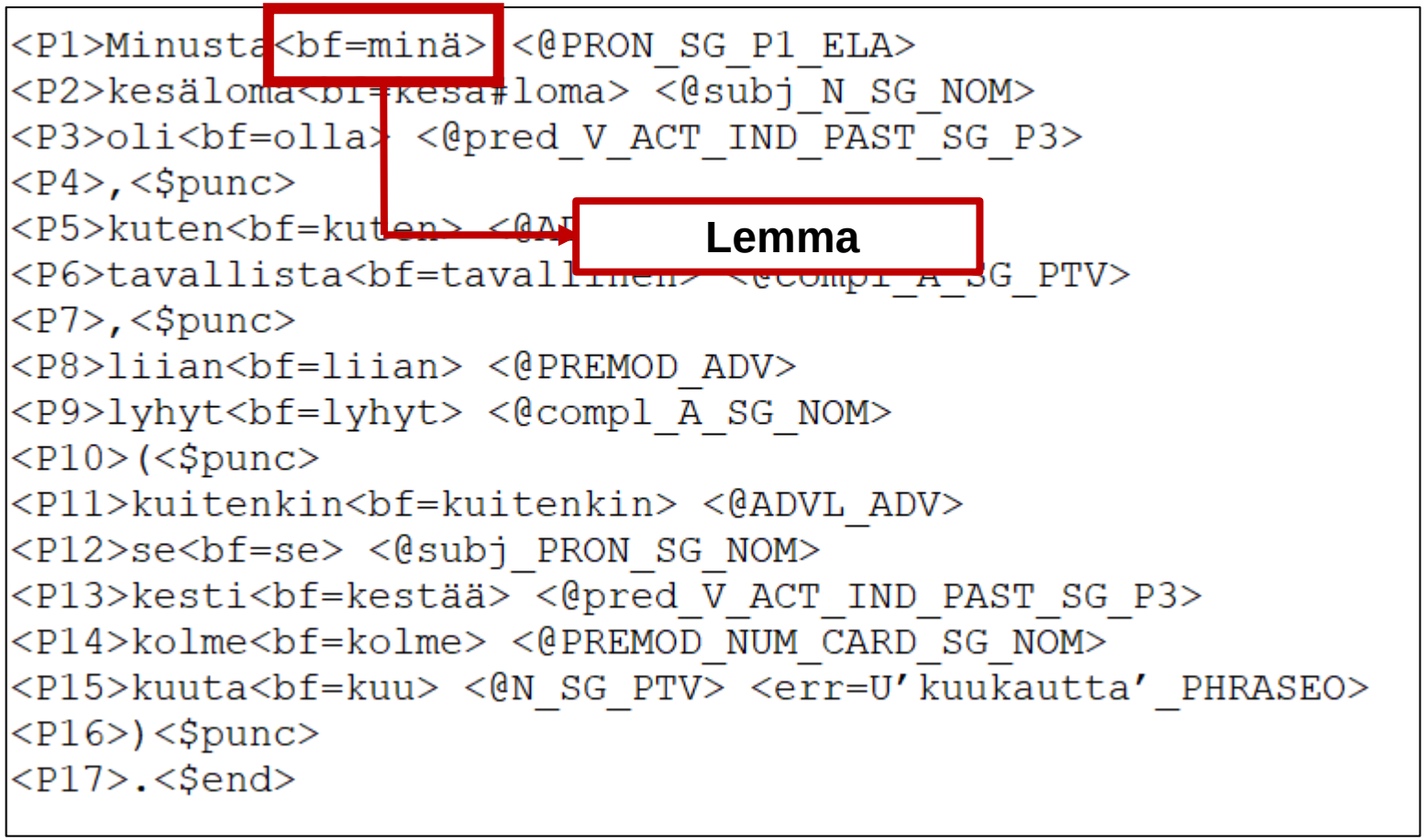
Minusta kesäloma oli, kuten tavallista, liian lyhyt (kuitenkin se kesti kolme kuuta). Loman alussa kaksi kaveriani tulivat minulle Krumloviin. He jäivät meille koko viikonloppua ja koimme yhdessä kaupungissa luonnossakin. On vahinko, että meille ei ollut enemmän aikaa. Muuten minun täytyi tehdä bakalaarisen esseeni, koska haluan lopulta lopettaa opiskeluni humanistisessa tiedekunnassa (anteeksi, mutta en ole varmaa, miten tämä sanoa suomeksi). Valitettavasti minun täytyy tunnustaa, että en ollut oikein ahkeraa, niin että en ole tehnyt sitä loppuun.

...

Korpustekstien käsittely ja virheannotointi

```
<P1>Minusta<bf=minä> <@PRON_SG_P1_ELA>
<P2>kesäloma<bf=kesä#loma> <@subj_N_SG_NOM>
<P3>oli<bf=olla> <@pred_V_ACT_IND_PAST_SG_P3>
<P4>,<$punc>
<P5>kuten<bf=kuten> <@ADVL_ADV>
<P6>tavallista<bf=tavallinen> <@compl_A_SG_PTV>
<P7>,<$punc>
<P8>liian<bf=liian> <@PREMOD_ADV>
<P9>lyhyt<bf=lyhyt> <@compl_A_SG_NOM>
<P10>(<$punc>
<P11>kuitenkin<bf=kuitenkin> <@ADVL_ADV>
<P12>se<bf=se> <@subj_PRON_SG_NOM>
<P13>kesti<bf=kestää> <@pred_V_ACT_IND_PAST_SG_P3>
<P14>kolme<bf=kolme> <@PREMOD_NUM_CARD_SG_NOM>
<P15>kuuta<bf=kuu> <@N_SG_PTV> <err=U'kuukautta'_PHRASEO>
<P16>)<$punc>
<P17>.<$end>
```

Korpustekstien käsittely ja virheannotointi




```
<P1>Minusta<bf=minä> <@PRON_SG_P1_ELA>
<P2>kesäloma<bf=kesä#loma> <@subj_N_SG_NOM>
<P3>oli<bf=olla> <@pred_V_ACT_IND_PAST_SG_P3>
<P4>,<$punc>
<P5>kuten<bf=kuten> <@AD
<P6>tavallista<bf=tavallinen> <@compl_A_SG_PTV>
<P7>,<$punc>
<P8>liian<bf=liian> <@PREMOD_ADV>
<P9>lyhyt<bf=lyhyt> <@compl_A_SG_NOM>
<P10>(<$punc>
<P11>kuitenkin<bf=kuitenkin> <@ADVL_ADV>
<P12>se<bf=se> <@subj_PRON_SG_NOM>
<P13>kesti<bf=kestää> <@pred_V_ACT_IND_PAST_SG_P3>
<P14>kolme<bf=kolme> <@PREMOD_NUM_CARD_SG_NOM>
<P15>kuuta<bf=kuu> <@N_SG_PTV> <err=U'kuukautta'_PHRASEO>
<P16>)<$punc>
<P17>.<$end>
```

Korpustekstien käsittely ja virheannotointi

```
<P1>Minusta<bf=minä> <@PRON_SG_P1_ELA>
<P2>kesäloma<bf=kesä#loma> <@subj_N_SG_NOM>
<P3>oli<bf=olla> <@pred_V_ACT_IND_PAST_SG_P3>
<P4>,<$punc>
<P5>kuten<bf=kuten> <@ADVL_ADV>
<P6>tavallista<bf=tavallinen> <@
<P7>,<$punc>
<P8>liian<bf=liian> <@PREMOD_ADV
<P9>lyhyt<bf=lyhyt> <@compl_A_SG_NOM>
<P10>(<$punc>
<P11>kuitenkin<bf=kuitenkin> <@ADVL_ADV>
<P12>se<bf=se> <@subj_PRON_SG_NOM>
<P13>kesti<bf=kestää> <@pred_V_ACT_IND_PAST_SG_P3>
<P14>kolme<bf=kolme> <@PREMOD_NUM_CARD_SG_NOM>
<P15>kuuta<bf=kuu> <@N_SG_PTV> <err=U'kuukautta'_PHRASEO>
<P16>)<$punc>
<P17>.<$end>
```

Kieliopillinen annotointi

Korpustekstien käsittely ja virheannotointi

```
<P1>Minusta<bf=minä> <@PRON_SG_P1_ELA>
<P2>kesäloma<bf=kesä#loma> <@subj_N_SG_NOM>
<P3>oli<bf=olla> <@pred_V_ACT_IND_PAST_SG_P3>
<P4>,<$punc>
<P5>kuten<bf=kuten> <@ADVL_ADV>
<P6>tavallista<bf=tavallinen> <@compl_A_SG_PTV>
<P7>,<$punc>
<P8>liian<bf=liian> <@PREMOD_ADV>
<P9>lyhyt<bf=lyhyt> <@compl_A_SG_NOM>
<P10>(<$punc>
<P11>kuitenkin<bf=kuitenkin> <@ADVL_ADV>
<P12>se<bf=se> <@subj_PRON_SG_NOM>
<P13>kesti<bf=kestää> <@pred_V_ACT_IND_PAST_SG_P3>
<P14>kolme<bf=kolme> <@PREMOD_NUM_CARD_SG_NOM>
<P15>kuuta<bf=kuu> <@N_SG_PTV> <err=U'kuukautta'_PHRASEO>
<P16>)<$punc>
<P17>.<$end>
```

Virheannotointi

Yleiset virhetyypit

- Kongruenssi:

Loman
alussa
kaksi
kaveriani
tulivat <AGR_VERBAL>
minulle
Krumloviin

- Predikatiivin sijavalinta:

anteeksi
mutta
en
ole
varmaa <PRED_PAR_NOM>
miten
tämä
sanoa
suomeksi

Yleiset virhetyypit

- Sijavalinta (muu lauseenjäsen):

He
jäivät
meille
koko
viikonloppua <OTHER_PAR_TRA>

- Verbivalinta:

Ja
koimme <V_LEX_"kokea">
yhdessä
kaupungissa
luonnossakin

Väitöstutkimus

- Korpuspohjainen tutkimus tšekinkielisten ja venäjänkielisten suomenoppijoiden oppijankielestä ICLFI-aineiston perusteella:
  - morfosyntaktiset virheet: virheet subjektin, objektin, predikatiivin ja adverbiaalin sijavalinnassa. Puuttuva tai ylimääräinen subjekti, objekti, predikatiivi, adverbiaali
  - kielto: virheet kieltomuotojen muodostamisessa sekä virheellinen sijavalinta ei-myönteisissä lauseissa
  - lausetyyppi: virheiden taajuus eri lausetyypeissä
  - verbivalinta: leksikaaliset virheet verbivalinnassa
- Yleisimmät virhetyypit:
  - adverbiaalin sijavalinta
  - verbivalinta
  - predikatiivin sijavalinta
  - objektin ja subjektin sijavalinta
  - kongruenssi

Esimerkkilauseet

- Adverbiaalin sijavalinta:
  - *”Mari tajuaa Takanašille ansiostaan, että hänen siskonsa tarvitsee häntä”* [TS0004]
- Verbivalinta:
  - *”he pitävät kaikista talven urheilusta. Heidän poikaan kanssa koulutavat jääkiekkoa järvellä”* [TS0018c]
- Predikatiivin sijavalinta:
  - *” Minä pidän kesästä koska kesällä aurinko paistaa ja päivät ovat lämpimät”* [TS0043a]
- Subjektin sijavalinta:
  - *” Makuuhuonen keskellä on sänky ja pöytällä on kukat”* [TS0015f]

Esimerkkilauseet

- Objektin sijavalinta:
  - *”Hän opiskelee Kulttuurituotanto, musiikkibusiness (esim. musiikkifestivaalien järjestäminen)”* [TS0022h]

- Kongruenssi:
  - *”Mutta siitä, mitä enemmistö ihmisistä tekevät ennen joulua, sanoisit melkein, että he valmistelevat sodaksi”* [TS0001d]

Kiitos huomiosta!

Lähde: Fingerpori

Lähteet

- Ivaska, I. (2014). The Corpus of Advanced Learner Finnish (LAS2): Database and toolkit to study academic learner Finnish. *Apples: journal of applied language studies, 8*(3), pp. 21-38. Retrieved February 12th, 2020, from http://apples.jyu.fi
- Jantunen, J. H., Brunni, S., Lehto, L.-M., & Skantsi, V. (2014). Oppijankieliaineistojen annotointi – esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut. (M. Mutta, P. Lintunen, I. Ivaska, & P. Peltonen, Eds.) AFinLA-e: soveltavan kielitieteen tutkimuksia(7), pp. 60-80. Retrieved December 4th, 2020, from https://journal.fi/afinla/article/view/48160
- Jantunen, o., Brunni, S., & University of Oulu, Department of Finnish Language (2013). International Corpus of Learner Finnish [text corpus]. *Kielipankki*. Retrieved from http://urn.fi/urn:nbn:fi:lb-20140730163
- Jantunen, J. H., & Pirkola, S. (2015). Oppijansuomen sähköiset tutkimusaineistot. Nykytilanne. *Virittäjä, 119*(1), pp. 88-103. Retrieved February 8th, 2020, from https://journal.fi/virittaja/article/view/46508