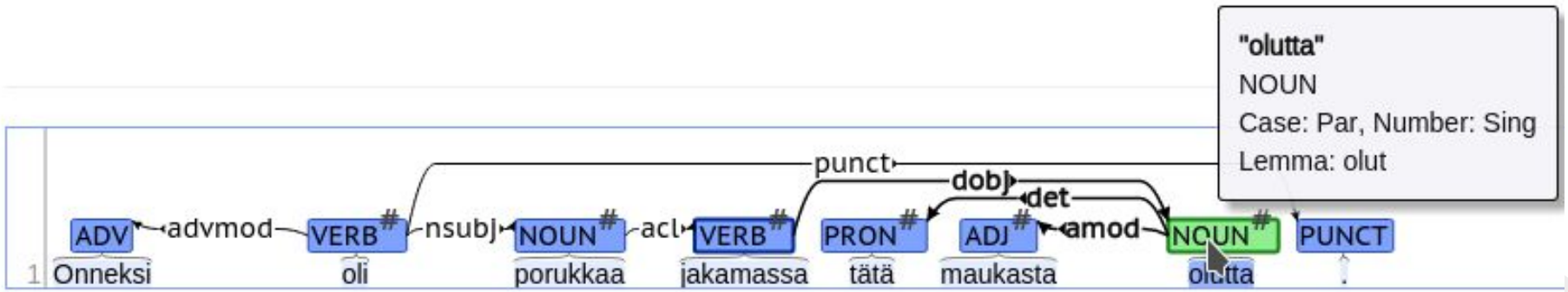
# ParseBanks for linguists

*Filip Ginter*
*Dept of IT - University of Turku, Finland*

*bionlp.utu.fi  www.evexdb.org  universaldependencies.org*

# Finnish Internet Parsebank

- 3.6B tokens of UD-parsed Finnish data
  - UD = Universal Dependencies
  - data off the Internet: CommonCrawl + own crawl
  - more data in the pipeline



# Why?

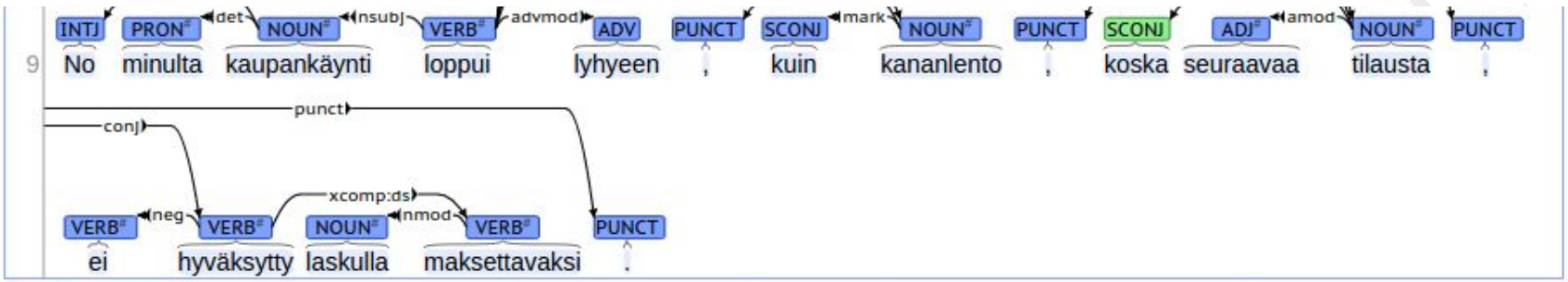
- Motivation for NLP quite obvious
  - Language models
  - Corpus for embeddings
  - Un-/semi-supervised methods
  - ...
- But how about linguistics?
  - What would a linguist need a parsebank for?
  - This talk: a little recap of what tools and resources we make available and whether/how the linguists use them + ideas for improvements (from the audience)

# #1 - Syntax search

- Major effort spent on an online dependency tree search tool which could scale up
  - Expectation: Linguists do their own searches and we play with word2vec all days
  - Reality: Linguists swamped with thousands upon thousands of hits, most of them parsing errors, helpless but not discouraged
  - With some post-processing, surprisingly good results can be obtained

# Syntax search (cont.)

- *"because NOUN"* - success
- *"partitive subjects in transitive clauses"* - nightmare with good end (paper upcoming)
- ...few more similar use cases in the works
  - quantifier distribution (moni vs monta vs montaa, ...)
  - specific fixed expressions (NOM TRA - sika mieheksi)
  - ...
- Syntax definitely of help in locating the examples, but not necessarily of interest on its own right




Original page Avaamme jalustan , paketti tiputettu kuljetuksen yhteydessä , rullan ruostumaton akseli vääntynyt , koska se on tiputettu 60kg ja seuraus rulla liian matalalla ja ei pyöri koska ottaa kiinni ! Irrotin akselin pois , murtui liimauksista . Taottu alasinta vasten suoraksi , takaisin paikoilleen . **Pysyy koska painoa päällä .** Virheellisesti saimme toisen vastaavanlaisen varjon joka oli ollut rikki . Tämä oli ehjä , ainoastaan yläosassa , kierrettävän muovi " sorvatun " alueen alla tääsäkin varjossa n . 8x8mm reikä ! Suoja moitteeton ja hyvä .

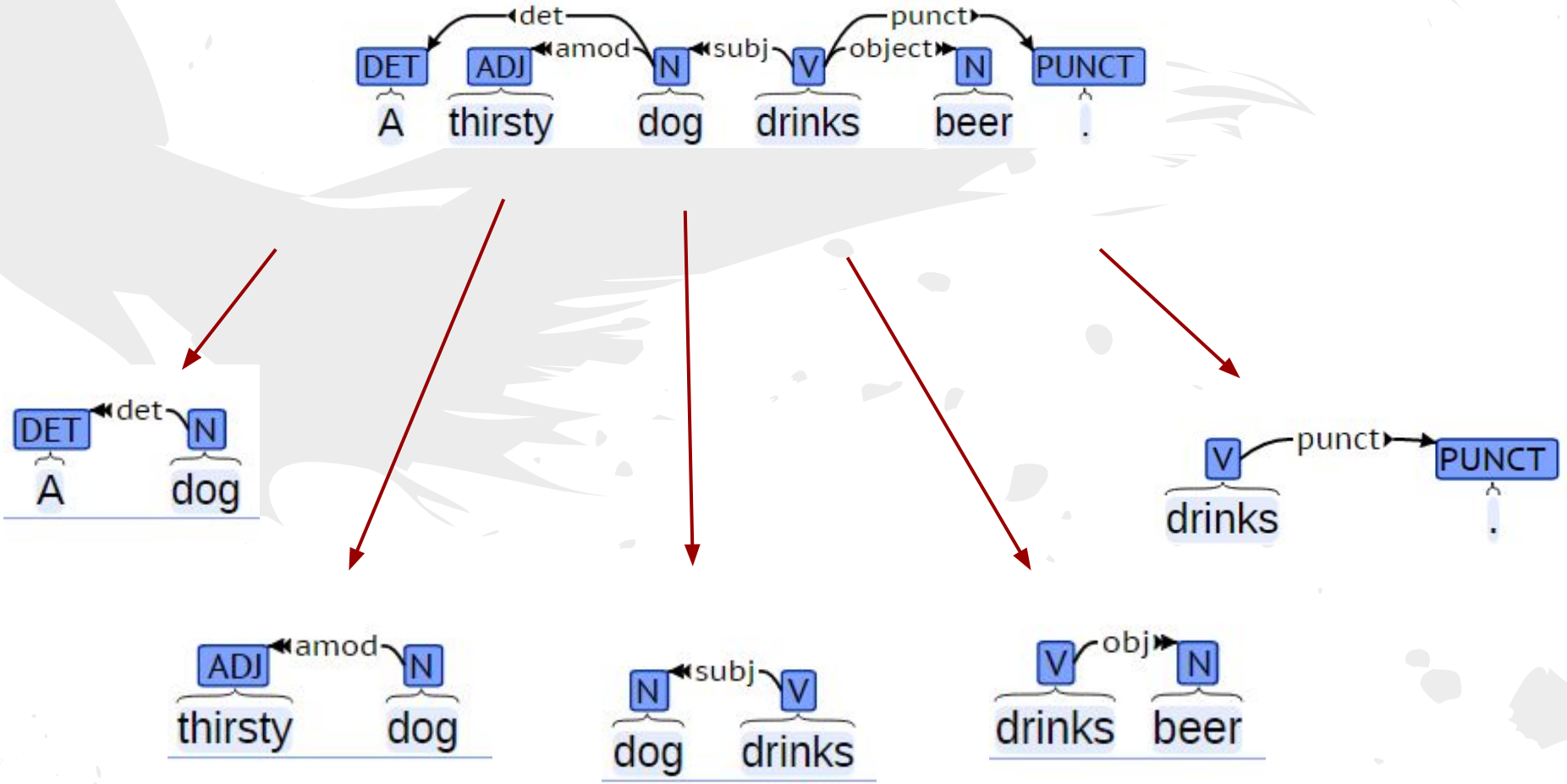

Original page Helppo . 05.04.2014 Hyvä . 05.04.2014 Jouduin noutamaan 400x500x620 sm kokoisen laatikon taksilla 4 km päästä , vaikka lähelläni on postin palvelupiste 150 metrin päässä . **Otin yhteyttä asiakaspalveluunne ja annoin 3 muuta vaihtoehtoa jonne paketin voisi toimittaa , koska kyseinen Siwa ei päivittäisen kulkureittini varrella , mutta tämä ei ollut mahdollista .** Koska muiden verkkokauppojen kautta toimitukset saan haluamallani tavalla , valitsen mieluummin sellaisen , jolla on tarjota minulle parhaiten sopivat toimituspalvelut . Kaikki muu on vaivatonta , mutta tämä ketjun viimeisin lenkki pettää eli toimitatte teille helpoimmalla tavalla tuotteen asiakkaalle saatuanne maksusuorituksen . Tuote näyttää hyvältä , mutta kattovalaisin odottaa vielä asennusta , joten toimivuudesta en tiedä .

# Syntax search (cont.)

- Most common requests (aka our TODO list)
  - Source document URL and context of the hit
  - Don’t show that messy tree, just highlight the words
  - Aggregation and statistics
  - Word order and linear distance restrictions

# Syntactic n-grams

- Goldberg et al. style syntactic n-grams
  - Available online, naturally
- Numerous papers replacing more traditional keyword features with delexicalized syntactic n-grams
  - Translation universals (re-)discovered by a linear classifier
  - Syntactic n-gram profiles of text using emoticons no different from other text. Emoticons not restricted to syntactically-poor contexts.
- From keywords to keystructures
  - Keystructures in different genres


det
amod
subj
punct
object
DET
ADJ
N
V
N
PUNCT
A
thirsty
dog
drinks
beer
.
DET
det
N
A
dog
ADJ
amod
N
thirsty
dog
N
subj
V
dog
drinks
V
obj
N
drinks
beer
V
punct
PUNCT
drinks
.

# NoSketchEngine

- Concordance search
- In active use by the lexicographers
- NoSketchEngine
  - A real problem - no word sketches - no real summary information on word usage
- Need correct lemmas for novel wordforms
  - THE thing for lexicographers, naturally
  - Relatively difficult to get right for Finnish

# word2vec / embeddings

- little demo on the web
  - Poor-man's synonym dictionary
  - In active use when designing datasets for eye-tracking studies in (psycho)linguistics
- Downloadable models
  - Used in ongoing research on human cognition at the [Censored] University
- Autoencoder-based prediction of selectional preferences
  - S+V goes in, O goes out
  - study in language processing in (psycho)linguistics

# twitter

- word2vec embeddings used to cluster tweets
- Digital Discourse Analysis of *stance* in twitter
- Paper upcoming

# Recap

- Interest in use of the parsebank's syntax but more needed in terms of tools before the linguists manage on their own
  - More expressive power in search
  - Aggregation and statistics on the results: "sketches"
- Diving the parsebank data into subcorpora by genre - better focus / context in studies

# Thanks

Jenna Kanerva - Aki Kyröläinen - Veronika Laippala - Juhani Luotolahti - Sampo Pyysalo - Tapio Salakoski

Kone Foundation

Emil Aaltonen Foundation