Practical syntactic and semantic analysis of Finnish on a large scale

Filip Ginter
University of Turku

Turku NLP group
bionlp.utu.fi

Turku NLP group

- bionlp.utu.fi
- Founded in 2001
- A number of projects in NLP of biomedical texts (scientific English)
- Since 2008 a concentrated effort on building open-source tools and resources for statistical Finnish NLP
- 140+ publications on various NLP-related topics

What is the Turku Natural Language Processing group doing for Finnish NLP?

Turku NLP group - relevant projects

- **Treebank** - Corpus with manually annotated trees
- **Parser** - Statistical parser trained on the Treebank
- **Parsebank** - Large corpus of syntactically parsed text
- **Propbank** - Manual annotation of predicate-argument relations in the Treebank

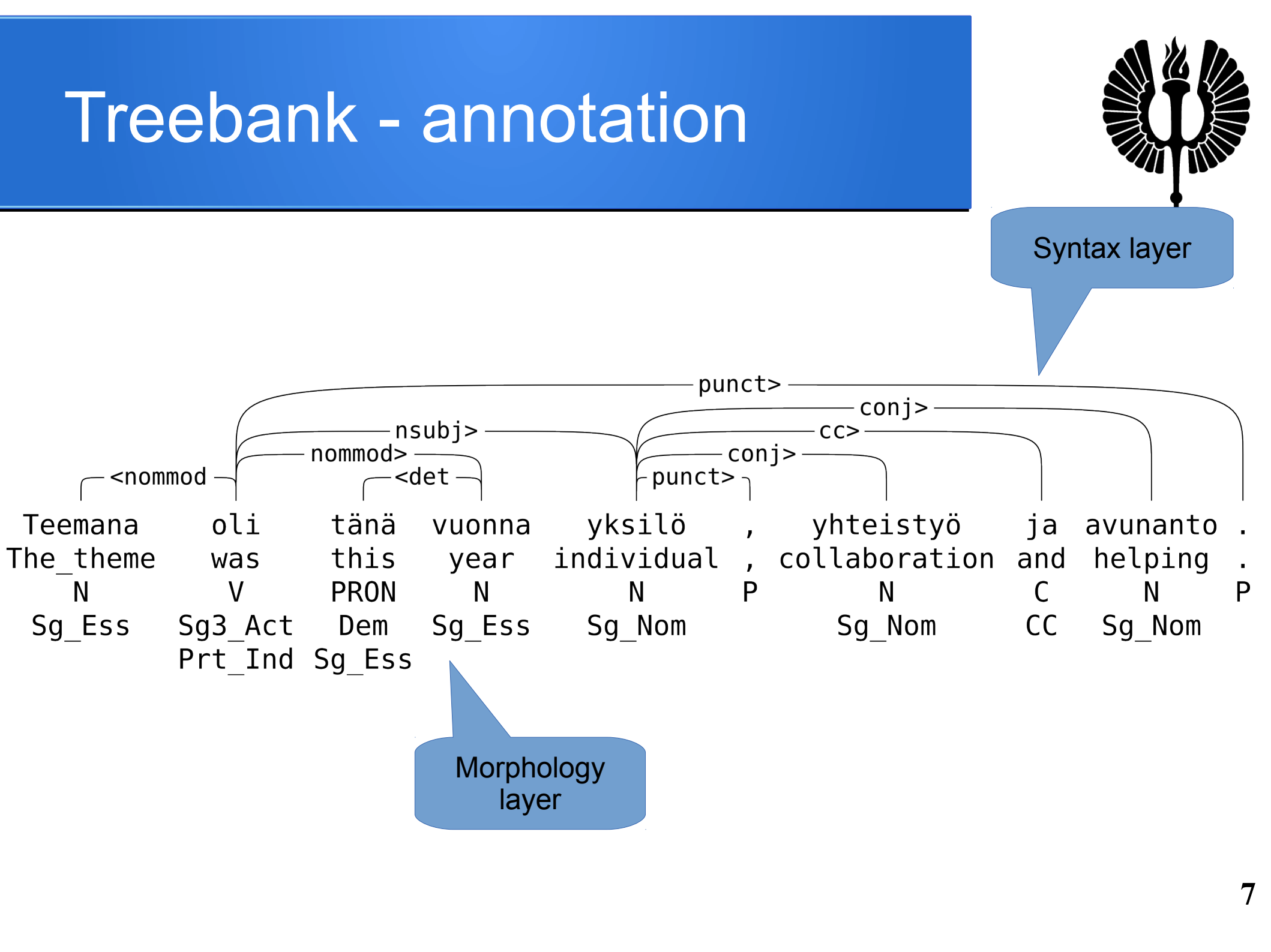
Turku Dependency Treebank (TDT)

Treebank

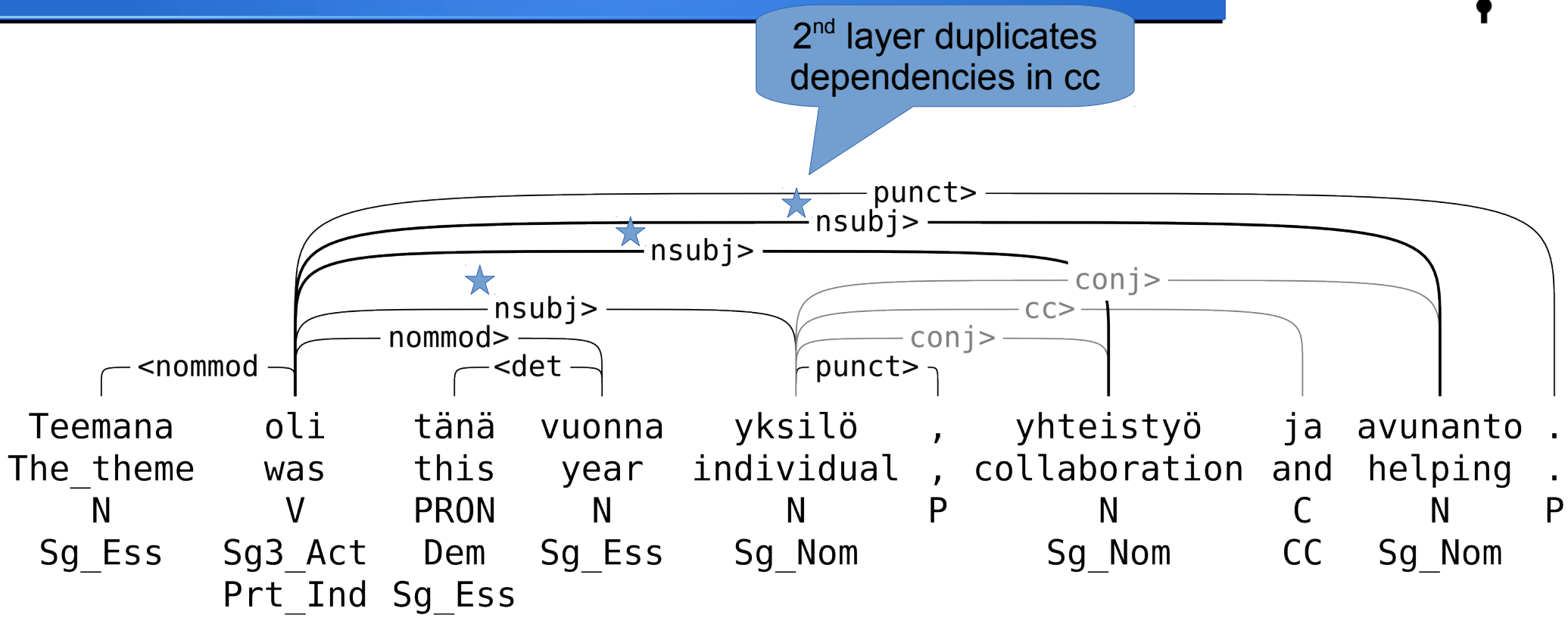
- Corpus with manually annotated syntax

  Completed project (2009-2013)
- Primarily intended as training data for statistical NLP
  - parsers, taggers

Treebank - annotation



Treebank - annotation



Treebank annotation

- Stanford Dependency Scheme (SD)
  - Scheme: choice of dependency types and analyzes for common structures
  - Application-friendly: content words as heads
  - NLP engineer-friendly (i.e. understandable to non-linguists)
  - Now becoming main-stream (was a good bet)
  - Works well for Finnish

Treebank - text

- 205,000 tokens / 15,000 sentences
- 10 text sources: Wikipedia, Taloussanomat, blogs, European Parliament, legal text, fiction, student magazines...
- Built specifically for statistical NLP:
  - Randomized choice of input material
  - Actual text as it's written

Treebank - availability

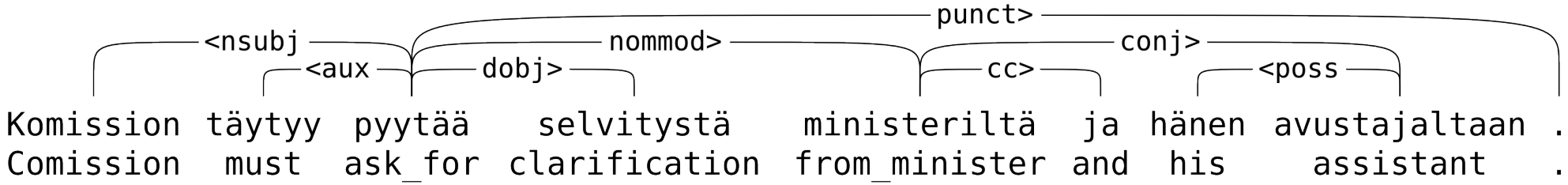
- http://bionlp.utu.fi/
- **Creative Commons (CC-BY-SA)**
  - Share Alike & Attribute
  - **No non-commercial clause**
- Also online browsing and querying

Treebank availability

- Soon to be included in Google's *Universal Dependency Treebank Project*
- *code.google.com/p/uni-dep-tb/*
- Treebanks for a number of languages
  - Currently five released, many more coming
  - Harmonized annotation in the SD scheme
  - Same set of morphological tags and dependency labels
- **Scheme unification brings major advantages**

...now that we have the Treebank...

Statistical Dependency Parser for Finnish



Parser #1

- **Graph-based parser**

  https://code.google.com/p/mate-tools/

- **POS tagging:** OMorFi+HunPOS

  https://code.google.com/p/omorfi/

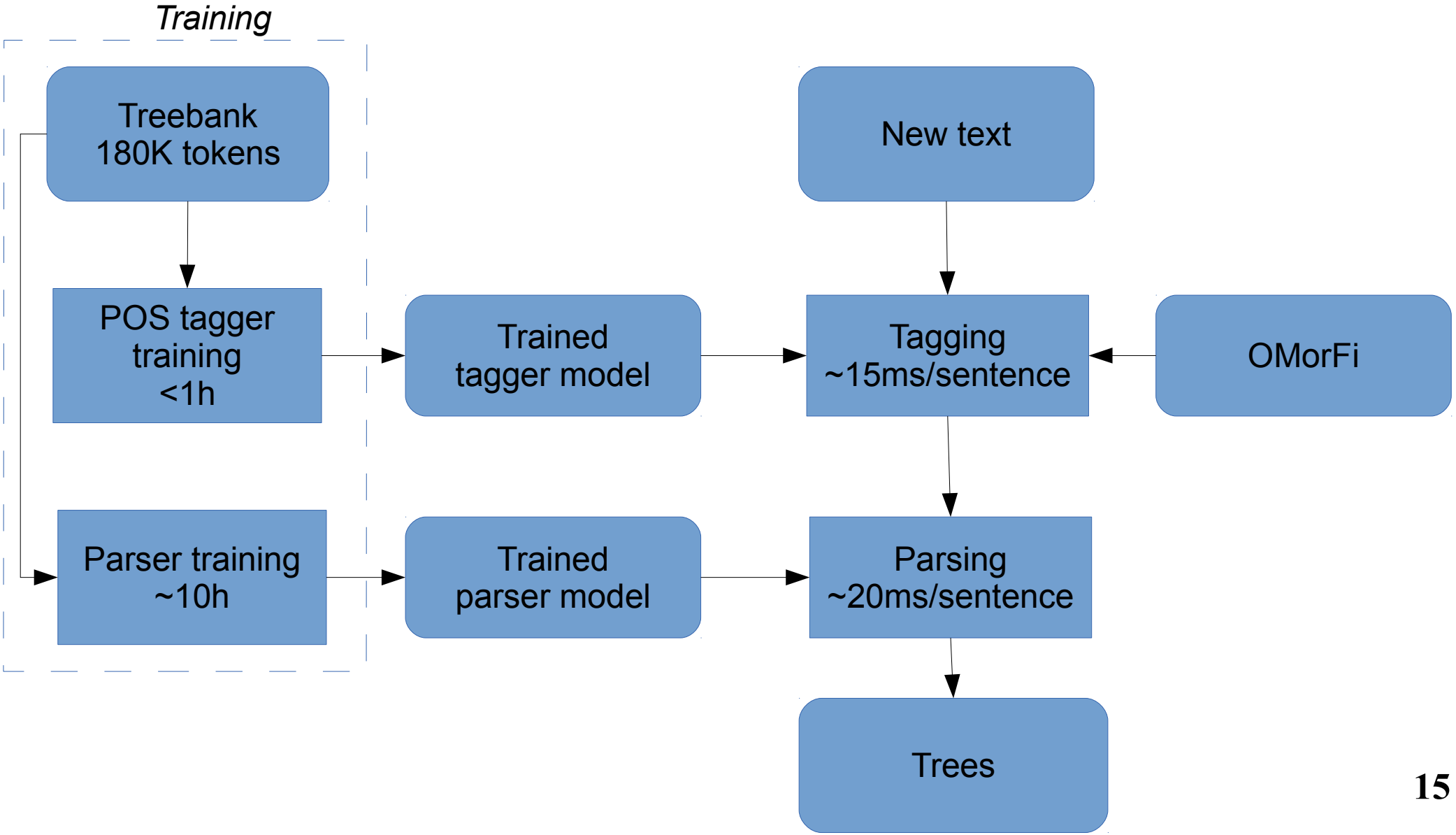
  https://code.google.com/p/hunpos/

- **Sent. splitting and tokenization:** OpenNLP

  http://opennlp.apache.org/

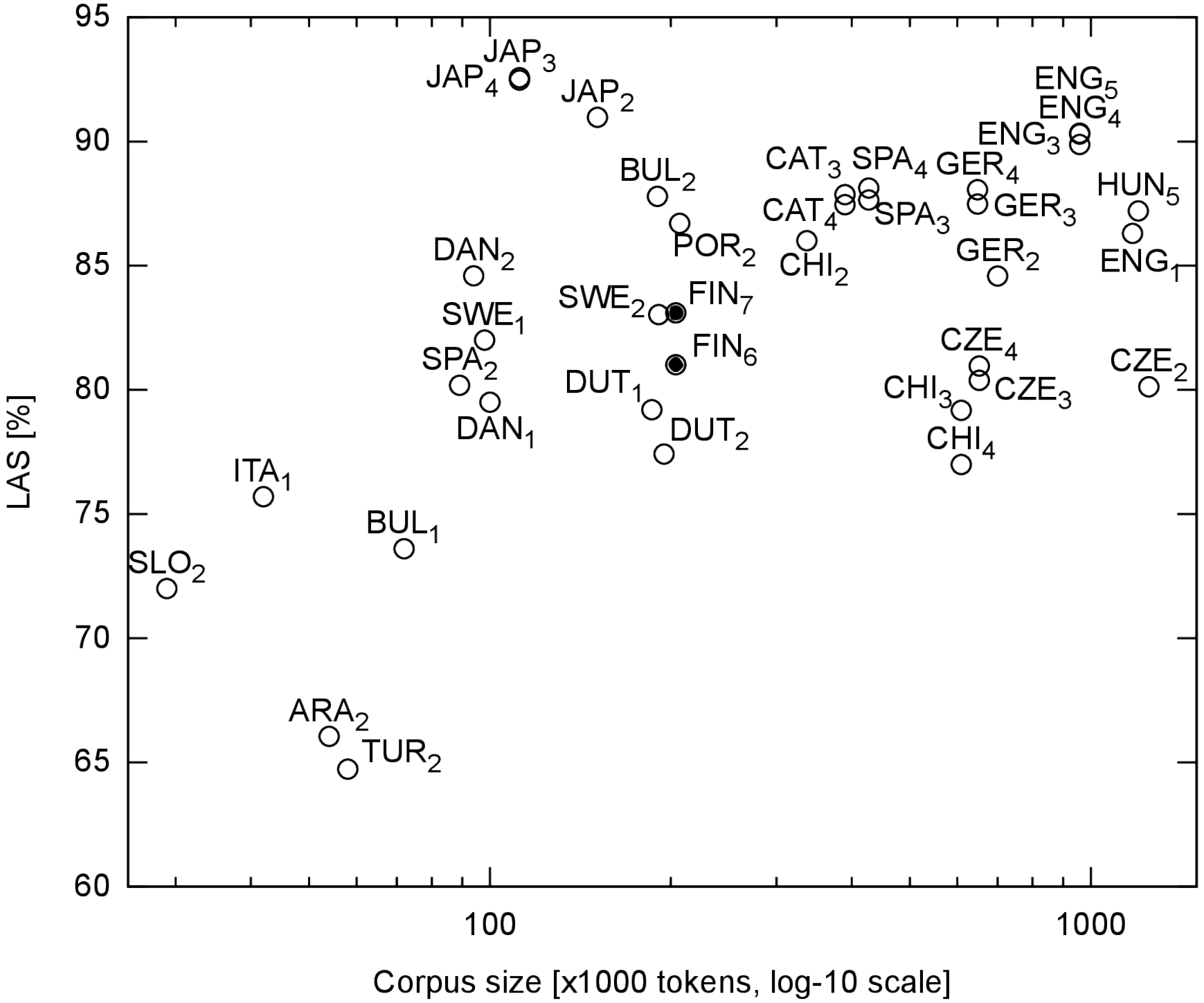
- All tools trained on the Turku Dependency Treebank

Parsing pipeline #1



Parser #2

- Recent development
- Transition-based parser
- Joint inference of morphology and syntax
- Best published results for dependency parsing of Czech, Finnish, German, Hungarian, and Russian



LAS = Labeled Attachment Score

Availability

- Current state: “research prototype”
  - Available for testing
- More polished public release later this year
- License:
  - OMorFi: LGPLv3
  - Parser, pre/post-processing code: GPLv2

=> <u>Fully open-source parsing stack for Finnish</u>

Parser - applications

- Finnish -> English Machine Translation
- Joint project with Convertus AB, Uppsala
  - Adaptation of the parser to the education domain (annotated extra in-domain data and re-trained the parser)
  - Syntactic analyzes used to improve MT

Parser - applications

- Parsebank of Finnish parliamentary and legal texts
- Joint project with Lingsoft, Inc
- Distributed by FIN-CLARIN as “FinnTreeBank ver 3”
  - (Not to be confused with FinnTreeBank ver 1&2 which are manually annotated treebanks developed at University of Helsinki)

...now that we have the parser...

Finnish Internet Parsebank

Internet Parsebank

- Kone Foundation project

  (2014-2016)
- Objective: “Gather and parse as much of Finnish text as you can get off the Internet”
- Hopefully dozens of billions of tokens of fully parsed text at the end of the project
  - Depends on how much we can crawl

Parsebank

- Why?
- Massive parsed dataset opens possibilities for advanced statistical NLP methods
  - Technology: Lexical acquisition, parser self-training, word usage statistics, information extraction, distributional semantics, etc.
  - Linguistic research: topical clustering, rare structure search, variation research, corpus linguistics, etc.

Parsebank data

- Data source #1: Common Crawl
  - http://commoncrawl.org
  - ~12B tokens of raw Finnish identified
  - A lot less of “clean” de-duplicated text
  - NOT restricted to .fi domain
- Data source #2:
  - Homebrew crawl targeting .fi and seeds from CommonCrawl

Internet Parsebank

- Over 1B (1,000,000,000) tokens parsed in trial
  - 80M+ sentences
- Few thousand CPU core hours
  - One day on the CSC cluster! :)
- Data not (yet) publicly available: will be made available under same open principles as all other tools and data we develop

Parsebank applications

- Automatic expansion of OMorFi lexicon

  (ongoing project)
- Gathering new words AND inferring their inflection patterns fully automatically
- Reliable statistics from the massive amount of data
- Better lexicon → better analyzer → better parser → better applications

Parsebank applications

- Methods of distributional semantics

  (ongoing project)

- 1B+ tokens becomes a sufficient data size for modern distributional semantics
  - Random Indexing
  - Recent Neural network-based models

    https://code.google.com/p/word2vec/

Distributional semantics

- Example: most similar words to a query
- Only using statistics from running text
- Kaunis → ihastuttava, hurmaava, lumoava, viehättävä,viehkeä, sievä, ihana,...
- Pizza → pitsa, lasagne, pippuripihvi, valkosipulietana, alkuruoka, kebabannos,...
- Word arithmetic: Pariisi - Ranska + Ruotsi
  - Tukholma, Oslo, Hampuri, Berliini, Praha, Göteborg,...
  - Still very much work-in-progress
  - Based on the English demo at http://thisplusthat.me

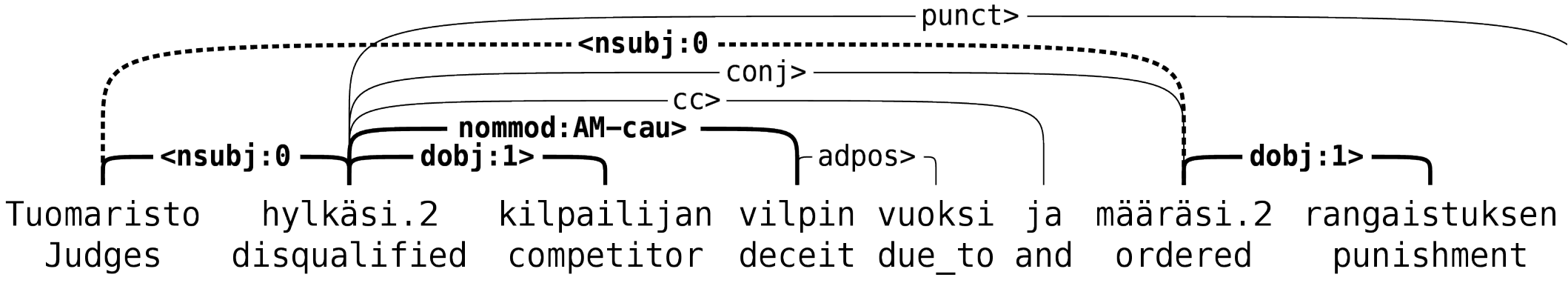
Propbank

- Finnish Proposition Bank
- Emil Aaltonen foundation project

  (2012-2014)
- Adds a semantic role labeling annotation on top of the Turku Dependency Treebank
- Modelled after the English PropBank annotation

Semantic role labeling

- Verb senses are disambiguated
- The role of every argument and modifier is annotated



Propbank - availability

- 35,000 verbs annotated, annotation nearly complete - still an ongoing project
- All data will be released under Creative Commons CC BY SA
  - Data release: likely first half of 2014
- Future plan:
  - Train a statistical semantic role labeling tool and analyze with it the Internet Parsebank

Recap

- Turku Dependency Treebank + modern statistical parsers → fully open parsing pipeline for Finnish
- Internet Parsebank → massive amounts of morphosyntactically analyzed data for modern statistical NLP methods
- Finnish Propbank → verb senses and verb argument roles
- **All data and tools fully open and free to use**

Selected references

- **Treebank and Parser #1:** Haverinen et al. (2013) Building the essential resources for Finnish: the Turku Dependency Treebank. Language Resources and Evaluation. Springer.
- **Parser #2:** Bohnet et al. (2013) Joint Morphological and Syntactic Analysis for Richly Inflected Languages. Transactions of the Association for Computational Linguistics. ACL.
- **FinnTreeBank ver 3:** Ginter et al. (2013) Building a Large Automatically Parsed Corpus of Finnish. Proceedings of NoDaLiDa'13
- **Propbank:** Haverinen et al. (2013) Towards a Dependency-based PropBank of General Finnish. Proceedings of NoDaLiDa'13

- More at: http://bionlp.utu.fi/publications.html

Thanks for your attention!

...and thanks to my many colleagues and collaborators involved in the projects:

Dr. Bernd Bohnet, Katri Haverinen,
Jenna Kanerva, Samuel Kohonen,
Dr. Veronika Laippala, Juhani Luotolahti,
Anna Missilä, Prof. Joakim Nivre,
Dr. Stina Ojala, Prof. Tapio Salakoski,
Simo Vihjanen, Timo Viljanen