TurkuNLP: Methods and Resources

Filip Ginter • Tapio Salakoski • Jari Björne • Kai Hakala • Katri Haverinen • Suwisa Kaewphan • Jenna Kanerva • Aki-Juhani Kyröläinen • Veronika Laippala • Juhani Luotolahti • Farrokh Mehryary • Hans Moen • Sampo Pyysalo • Aleksi Vesanto

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

Turku NLP group (est. 2001)

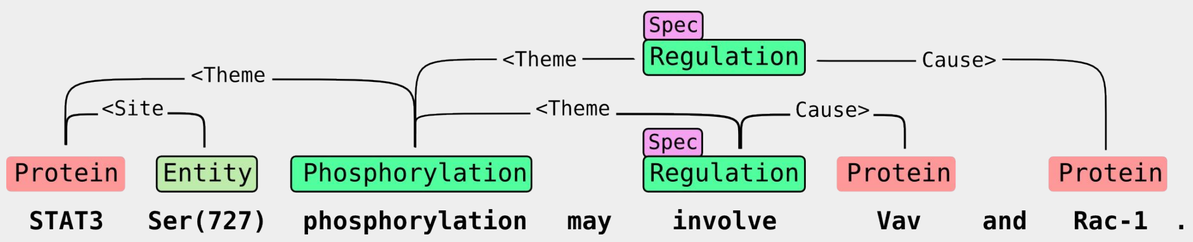
BioNLP

1. NLP in the biomedical domain
2. TEES system for biomedical event extraction
3. www.evexdb.org

NLP for Finnish

4. Turku Dependency Treebank (TDT) - Finnish Treebank
5. Finnish PropBank
6. Finnish Dependency Parser pipeline
7. Finnish Parsebank
8. Universal Dependencies

Extracting biological events



Events:

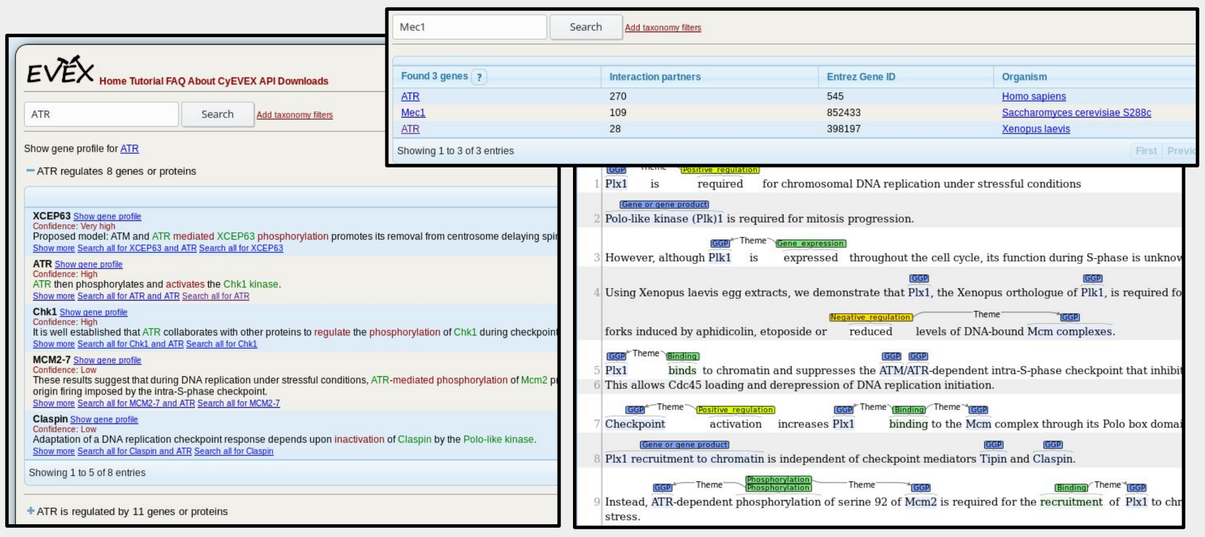
1. Vav → regulates → phosphorylation → STAT3
2. Rac-1 → regulates → phosphorylation → STAT3

TEES

- General-purpose event extraction system
- Numerous top ranks in shared tasks
- Widely adopted in BioNLP
- Fully open-source
- http://jbjorne.github.io/TEES/

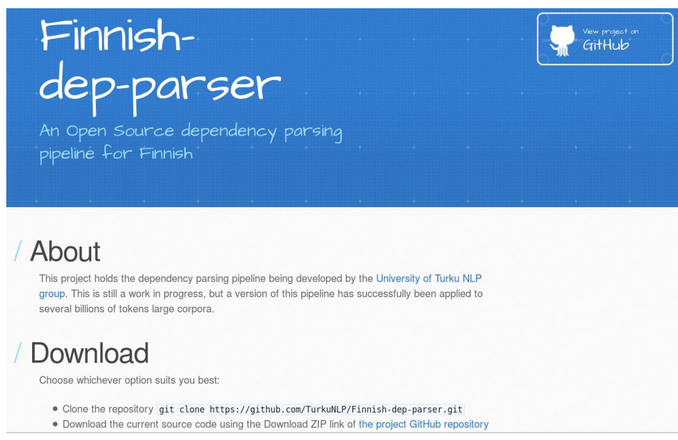
EVEX - www.evexdb.org

- Events extracted from the *entire* literature available for text mining (PubMed & PubMed Central literature databases)
  - **25M article titles and abstracts**
  - **1.4M full-text articles**
- EVEX: **40 million bio-molecular events** among more than **76 million gene/protein mentions**.
- Coverage recently expanded to over **190M *gene/protein*, 150M *chemical*, 100M *disease*, 70M *organism*** and **7M *cell line* mentions**.



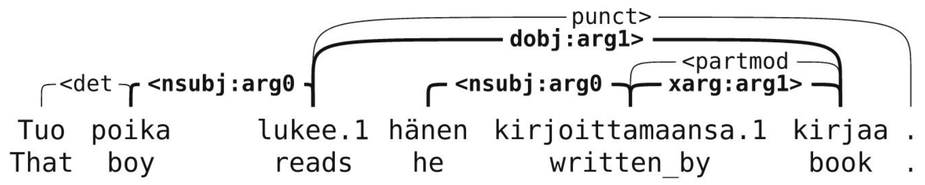
Turku Treebank and Parser

- 200K tokens fully manually annotated
- 10 sources / domains
- Now converted to Universal Dependencies
- Parser: OMorFi + MarMoT + Mate tools + glue code
- 82% LAS



Finnish PropBank

- All verb occurrences in the Treebank annotated
  - sense + arguments + roles
  - 30K verb occurrences



Finnish Parsebank

- CommonCrawl + own Internet crawl (still ongoing)
- 3.6B tokens, 275M sentences fully parsed
- ~5B tokens waiting for parsing
- Availability:
  - Download sentence shuffled upon request
  - SETS API + KORP (upcoming)

SETS: Dependency search in large ParseBanks

- Full syntactic search - subtrees, negations, morphology, universal quantification, linear distance
- Tested on 270M trees
- bionlp-www.utu.fi/dep_search
- Also accessible via web API:
  - Finnish Parsebank + Suomi24
  - All Universal Dependencies treebanks
- Powers Universal Dependencies content validation