Turku
Natural Language Processing
Infrastructures

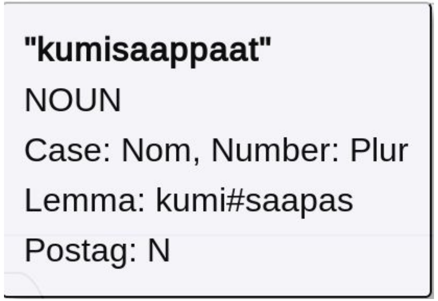
Aleksi Vesanto, TurkuNLP,
University of Turku

This presentation

- Turku-neural-parser-pipeline
- Finnish Internet Parsebank
- Depsearch
- Text reuse detection with BLAST

Turku-neural-parser-pipeline

- Word and sentence segmentation
- Part-of-speech and morphological tagging
- Lemmatization
- Syntactic parsing



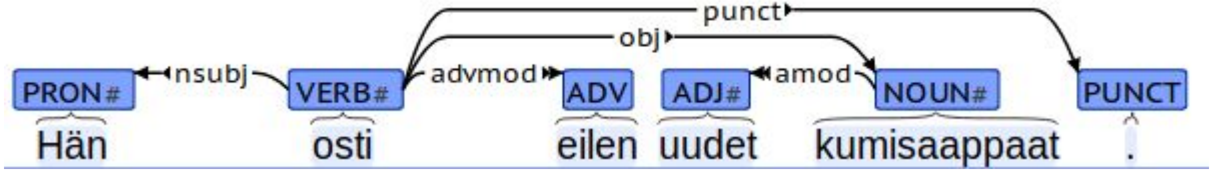
"kumisaappaat"

NOUN

Case: Nom, Number: Plur

Lemma: kumi#saapas

Postag: N



Turku-neural-parser-pipeline

- Different neural components combined into single pipeline → runnable with one command
- Pretrained models for 50+ languages
- 1st on lemmatization and 2nd on syntactic parsing and morphological tagging out of 26 teams

https://turkunlp.github.io/Turku-neural-parser-pipeline/

Finnish Internet Parsebank

- 8 billion token corpus of web crawled Finnish text
  - Only sentence-structured, document-level deduplicated text included, no menus, lists, etc.

- Fully morpho-syntactically analyzed

Finnish Internet Parsebank - Use cases

- Linguistic studies
  - Easy access to pre-analyzed text for corpus linguistics

- Language technology
  - Ready-made Finnish w2v models
  - E.g. http://bionlp-www.utu.fi/wv_demo/
  - Raw material for language technology research

Universal Web Parsebanks

- How about other languages?
- Similar web crawl repeated for 40+ languages
- Sentence shuffled versions available at

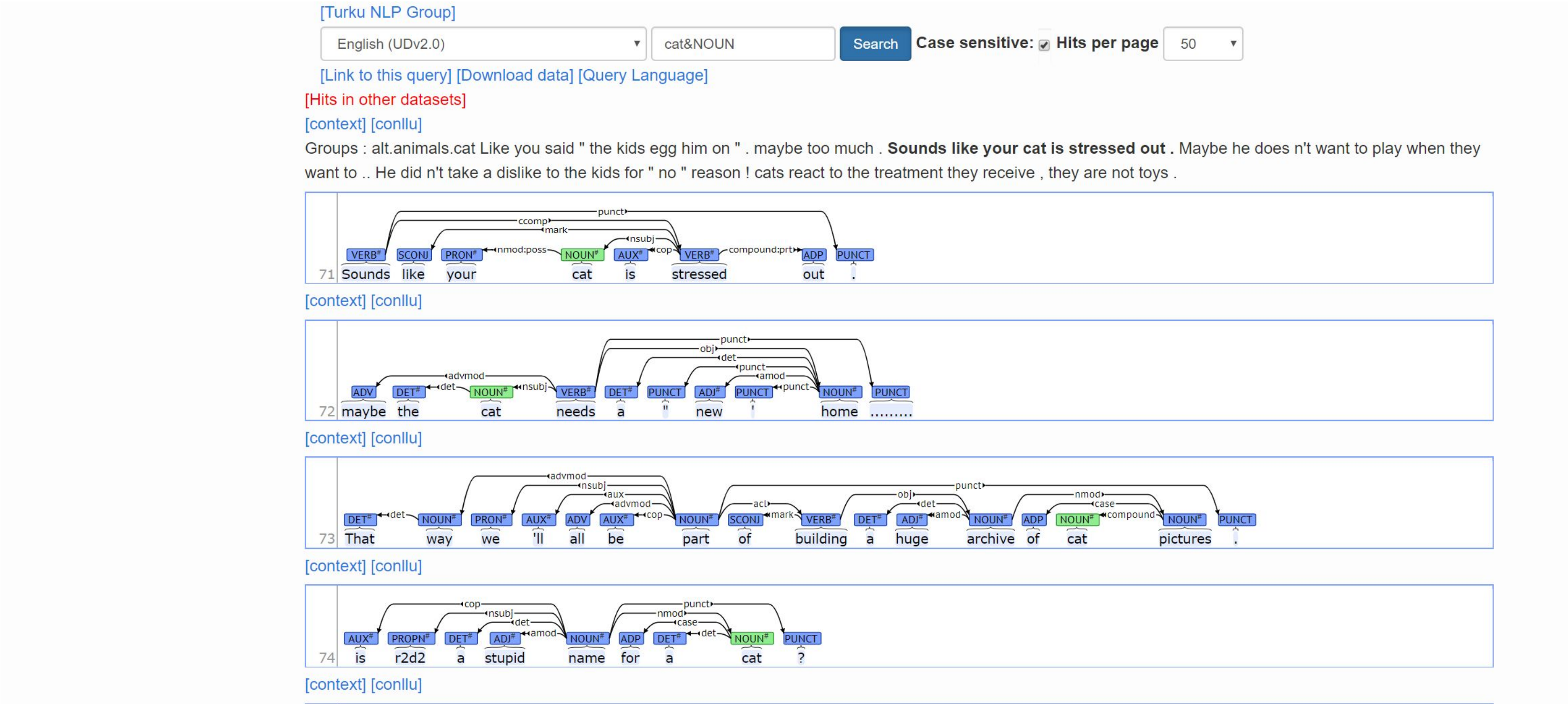
  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- Full document structure exists, but copyright laws do not allow freely distributing the data

Depsearch

- Search tool for dependency graphs
-
- Finnish Parsebank and the 40+ other languages searchable

http://bionlp-www.utu.fi/dep_search

Depsearch



[Turku NLP Group]

English (UDv2.0) ▾ | cat&NOUN | Search | **Case sensitive:** ☑ **Hits per page** 50 ▾

[Link to this query] [Download data] [Query Language]

[Hits in other datasets]

[context] [conllu]

Groups : alt.animals.cat Like you said " the kids egg him on " . maybe too much . **Sounds like your cat is stressed out .** Maybe he does n't want to play when they want to .. He did n't take a dislike to the kids for " no " reason ! cats react to the treatment they receive , they are not toys .

71 Sounds like your cat is stressed out .

[context] [conllu]

72 maybe the cat needs a " new ' home ....

[context] [conllu]

73 That way we 'll all be part of building a huge archive of cat pictures .

[context] [conllu]

74 is r2d2 a stupid name for a cat ?

[context] [conllu]

Text reuse detection with BLAST

- Finds reused passages in very noisy corpora
  - OCR scanned documents
  - The quality does not need to be that great

- News, advertisements, citations etc.

Text reuse detection with BLAST

- Uses NCBI BLAST as engine
- Program designed for comparing and aligning biomedical sequences, like proteins

- Finds overlapping sequences in a large sequence database (used on whole genomes)

Text reuse detection with BLAST

- NCBI BLAST reads proteins, not text
- Encode text data into proteins first
- 23 distinct amino acids to work with
- Find the 23 most used characters from the data, form character → amino acid mapping
- NCBI BLAST outputs a pairwise alignment for all sequences
  - Shows the regions that overlap, .i.e. are text reuse
- Results are then clustered
  - Each cluster contains all mentions of particular reused passage
  - e.g. article in a newspaper

Text reuse detection with BLAST - Use cases

- Newspapers
  - See how news spread, which became viral etc.
  - http://comhis.fi/clusters
- Books
  - See which books cited or plagiarised others

Example pair

Multa t\ä@tä fyNlkÄsiii kchtalostu ,ct , Abouil Asi,3 wic!lä ticiun't>t ,mitää>«, » vaalii luiftti iloista M,m<iä Tshiragauissa, ©elä fi:föf3>i'öi että uiUfatfpäim -uhkaisiloui i Hviarat, miinto fu^tiaani 'fatifefi- fuffotai» lÄuja THi roinin, puutarhassa ja, ipici'ilitsi hwi'tt<iiöii fmmiamcrk^iUi ja anoo» »imilyMla,

Mutta tästä synkästä kohtalosta ei Abbul Asib »ielä tiennyt mitään, vaan »ietti iloista elämää TshiraganiSsa. Sekä sis»Stä «ttä ulkoapäin uhkasivat «aarat. mutta sulttaani katseli lukkotaisteluja Tfhiiaaanin puutarhassa ja palkitsi voittajan lunnicnnerleillä ja ar° vonimityksillä.

Thanks for listening!

- Parser: https://turkunlp.github.io/Turku-neural-parser-pipeline/
- Parsebanks: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- Depsearch: http://bionlp-www.utu.fi/dep_search
- BLAST: https://github.com/avjves/textreuse-blast, http://comhis.fi/clusters