Querying a Dozen Corpora and a Thousand Years with Fintan

Christian Chiarcos*
chiarcos@em.uni-frankfurt.de

Christian Fäth*
faeth@informatik.uni-frankfurt.de

Maxim Ionov*
ionov@cs.uni-frankfurt.de

* Applied Computational Linguistics (AcOli), Goethe-Universität Frankfurt, Germany
* Institute for Digital Humanities (IDH), Universität zu Köln, Germany

Fintan – Flexible INtegrated Transformation and Annotation eNgineering

Transforming heterogeneous data in a unified way
(Fäth et al. 2020)

- Convert any kind of language resource to RDF graphs.
- Manipulate/link/transform graphs with SPARQL.
- Serialize as RDF or in conventional corpus formats

Modular: Pipelines broken into small, reusable pieces
Reusable: Same RDF vocabulary => same modules
Extensible: Add your own (SPARQL, Docker, Java, …)
Scalable: Stream processing & parallelization

https://github.com/Pret-a-LLOD/Fintan (wrapper repo)
https://github.com/acoli-repo/conll-rdf (CoNLL customization)

Use Case: Order of Nominal Dative and Accusative Arguments in Historical German

Polomeus schreib die tat dem kunig
Ptolemeus wrote the deed to the king.

Si zeig dem kunig den mandel und ir gewant
She showed her coat and garment to the king

Fintan Workflow

Preprocessing => input format(s) (outside Fintan)
Fintan Loader => RDF (for XML, CoNLL, …)
Fintan Updater => uniform data model
Fintan Updater => query pre-compilation
(F optional, e.g., labelled edges => object properties)
Fintan Formatter => export and/or query

Evaluation => external tools

Uniform Data Model

RDF vocabularies:
nif (words, sentences)
conll (dependencies, word annotations)
powla (phrases, trees, labelled edges)
tiger (phrase and edge annotations)
* currently points to TIGER-XML documentation

12 corpora, 1000 years => heterogeneous:
- dependency syntax (4 formats)
- phrase structure syntax (3 formats)
- flat span annotations (1 format)

Observations
- Fintan provides natural support for querying with SPARQL
- Queries wrapped in Fintan workflowcan be easily replicated
- Unlike query tools like ANNIS, SPARQL supports deduplication in aggregates
- We explore routes to integration

Results
- Smoothed over 100 year windows, the corpus statistics form a continuous plot
- Continuous trend to restrict ACC>DAT since the 14th c.

Example (Spans)
(Old High German TCodex, visualized with ANNIS)

Preprocessing (CoNLL)
(CoNLL conversion, cf. Chiarcos & Schenk 2018)

Loader/Updater (CoNLL-RDF)
IIOB-notation is resolved during loading … but nesting of spans is to be asserted with SPARQL

Querying (SPARQL)