AN OPEN-SOURCE CORPUS-QUERY PLATFORM FOR THE ANALYSIS OF VERY LARGE MULTIPLY ANNOTATED CORPORA

KORAP

MAIN AIMS OF KORAP

- always meet requirements of a scientific tool [1]
- core-sustainability for 15-20 years
- support for arbitrarily large corpora
- support for any number of potentially concurring annotation layers
- support for virtual corpora/collections definable on internal and external text properties [2]
- extensible also by external developments

MOtivation

- German Reference Corpus DeReKo [3] now contains more than 26 billion words
- growing by 1.5 billion words/year
- old query system COSMAS II only supports 8 billion words per archive and only 2 annotation layers

GENERAL ARCHITECTURE

- modular design with replaceable components
- support for multiple specialized backends

KALAMAR (FRONTEND)

- KMB views for matches
- table views of morphological annotations
- tree views of hierarchical annotations

KORAL (QUERY SERIALIZER)

- translates queries into a general query protocol, called KoraQuery [4]
- query helper for multiple tag sets
- creation of virtual collections
- embedded interactive documentation

KORAP (LUCENE BASED BACKEND)

- documents are stored as indexed field sets, including primary data, metadata, and annotations
- annotations are indexed as term vectors with additional information
- meta information can be used to narrow the search space by defining virtual-collections
- supports a large subset of KoraQuery by utilizing a set of index specific query mechanisms, widely extending the core functionality of Lucene:
  - full text search, token-based annotation search, span-based annotation search, distance search, positional search, nested queries, etc.

HORIZONTAL SCALABILITY

- time critical tasks are distributable among worker nodes
- if the system gets too slow, add another node
- computational problems: financial overhead
- cheaper hardware can be used
- improved failure tolerance

LET THE CODE COME TO THE DATA!

- terabytes of corpus data are too bulky to move
- and by license contracts not allowed to move

SOURCE CODE AND LICENSE

- published under BSD-2-license
  - http://github.com/KorAP
  - http://korap.ids-mannheim.de/gemr/

CURRENT STATE (AS OF 7/2015)

- IDS internal alpha-version running since 2/2014
- still to be published:
  - Kustvakt
  - Kanalito (distribution layer, part of Krill)
  - Karang
  - example pipeline for ingesting corpora with annotations
  - new work item at ISO TC37 SC4: Corpus Query Lingua Franca (CQLF) [5]

SUSTAINABILITY

- 2.5 permanent FTE for further development, maintenance, support
- cooperations desired on:
  - comparable corpora (situated at different places)
  - KoRAp development and extension in general

REFERENCES