Piotr Bański, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Eliza Margaretha, Andreas Witt

INSTITUT FÜR DEUTSCHE SPRACHE

KORAP

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

MAIN AIMS OF KORAP

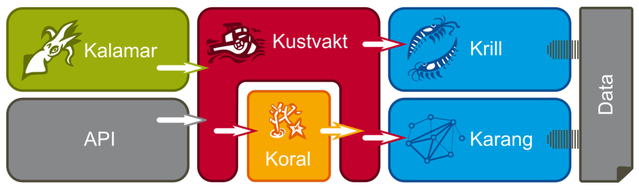
- always meet requirements of a scientific tool [1]
- core-sustainability for 15-20 years
- support for arbitrarily large corpora
- support for any number of potentially concurring annotation layers
- support for virtual corpora / collections definable on internal and external text properties [2]
- extensible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

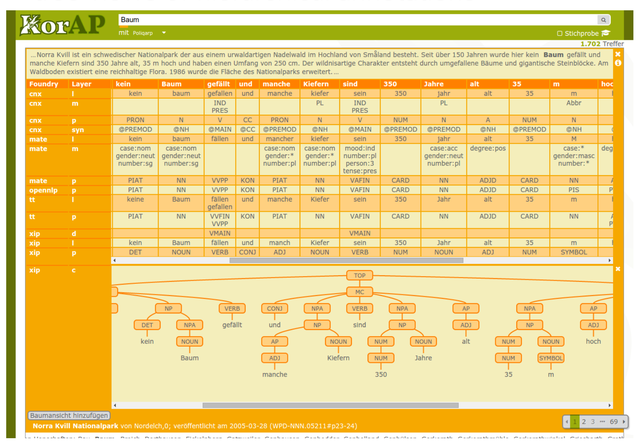
- modular design with replaceable components



- support for multiple specialized backends

KALAMAR (FRONTEND)

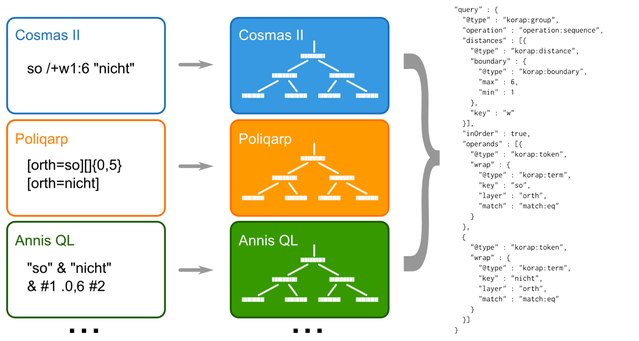
- KWIC views for matches
- table views of morphological annotations
- tree views of hierarchical annotations



- query helper for multiple tag sets
- creation of virtual collections
- embedded interactive documentation

KORAL (QUERY SERIALIZER)

- translates queries into a general query protocol, called KoralQuery [4]



- currently supported query languages:
  - COSMAS II QL
  - Poliqarp QL (CQP-dialect/extension)
  - ANNIS QL
  - CQL 1.2 (Clarin-FCS subset)

KARANG (NEO4J BASED BACKEND)

- complex annotation is represented as an arbitrary multigraph with properties
- vertices correspond to entities (words, …, sentences, …, texts)
- edges represent relations (dependency, domination, sequence, …)
- possible to run hierarchical cross-foundry queries spanning over several degrees of separation

KRILL (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 KoralQuery by utilizing a set of index specific query mechanisms, widely extending the core functionality of Lucene:
  - fulltext search, token-based annotation search, span-based annotation search, distance search, positional search, nested queries, etc.

KUSTVAKT (USER AND POLICY MANAGEMENT)

- takes queries and rewrites them to restrict the scope of a search to documents the user is allowed to access [4], [6]
- may also inject further properties the user has set
- example: injection of a collection constraint and the preferred annotation foundry for the "pos" layer:

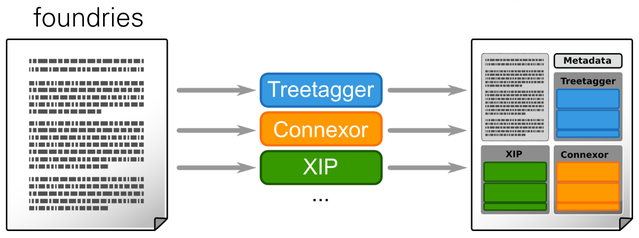
```
{
  "@context" : "http://korap.ids-mannheim.de/ns/
                koral/v0.3/context.jsonld",
  "query": {
    "@type": "korap:group",
    "operation": "operation:position",
    "frames": ["frame:contains"],
    "operands": [{
      "@type": "korap:span",
      "layer" : "c",
      "foundry" : "cnx",
      "key": "np"
    },{
      "@type": "korap:token",
      "wrap" : {
        "@type": "korap:term",
        "layer": "pos",
        "key" : "NE"
      }
    }]
  }
}
```

```
{
  "@context" : "http://korap.ids-mannheim.de/ns/
                koral/v0.3/context.jsonld",
  "query": {
    "@type": "korap:group",
    "operation": "operation:position",
    "frames": ["frame:contains"],
    "operands": [{
      "@type": "korap:span",
      "layer" : "c",
      "foundry" : "cnx",
      "key": "np"
    },{
      "@type": "korap:token",
      "wrap" : {
        "@type": "korap:term",
        "foundry": "mate",
        "layer": "pos",
        "key" : "NE"
      }
    }]
  },
  "collection": {
    "@type": "korap:doc",
    "key": "corpusID",
    "type": "type:string",
    "match": "match:eq",
    "key": "A00"
  }
}
```

- rewrites can be made transparent to the user (traceable, replicable)
- more efficient than filtering query hits
- backends and frontends can be developed without paying attention to authorization
- token-based authorization: OAuth2
- credentials via Shibboleth SSO or plain login

DATA MODEL

- separation of observed primary data and interpretations (annotations)
- arbitrary number of annotation layers organized in foundries



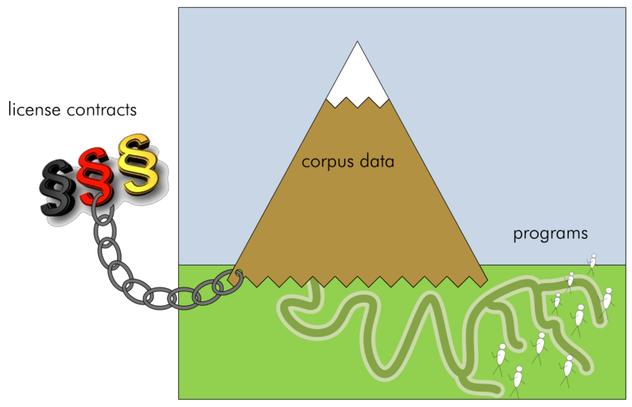
- add new annotation layers by extending KorAP!

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



- so let the code come to the data!

HOW CAN USER-CODE ACCESS THE DATA?

- via KorAP's REST-API
  - directly accessible for all kinds of clients
  - e.g from an R-script or another UI/frontend
- by contributing KorAP extensions, e.g. API extensions
- by adding completely new or alternative components

SOURCE CODE AND LICENSE

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

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

[1] P. Bański, P. M. Fischer, E. Frick, E. Ketzan, M. Kupietz, C. Schnober, O. Schonefeld, and A. Witt, "The new IDS corpus analysis platform: challenges and prospects", in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), ELRA, 2012.

[2] P. Bański, E. Frick, M. Hanl, M. Kupietz, C. Schnober, and A. Witt, "Robust corpus architecture: A new look at virtual collections and data access", in Corpus Linguistics 2013 Abstract Book, A. Hardie and R. Love, Eds., Lancaster: UCREL, 2013, pp. 23–25.

[3] M. Kupietz, C. Belica, H. Keibel, and A. Witt, "The german reference corpus DeReKo: a primordial sample for linguistic research", in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010.

[4] J. Bingel and N. Diewald, "KoralQuery – a general corpus query protocol", in Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, Vilnius, Lithuania, 2015.

[5] ISO, "Standard in development: BS ISO 24623-1 Language resource management – Corpus Query Lingua Franca (CQLF) Part 1: Metamodel", ISO, Geneva, Tech. Rep., 2014.

[6] P. Bański, N. Diewald, M. Hanl, M. Kupietz, and A. Witt, "Access control by query rewriting: the case of korap", in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ELRA, 2014.

Contact
Corpus Linguistics
Institut für Deutsche Sprache
PO Box 10 16 21
68016 Mannheim
Germany

Phone: +49 621.1581-409
Fax: +49 621.1581-200

www.ids-mannheim.de/kl/
corpuslinguistics@ids-mannheim.de

Address
Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim
Germany
Phone: +49 621.1581-0
Fax: +49 621.1581-20
info@ids-mannheim.de

www.ids-mannheim.de

Member of the Leibniz Association