INSTITUT FÜR
DEUTSCHE SPRACHE

Marc Kupietz, Harald Lüngen, Piotr Bański, Cyril Belica

MAXIMIZING THE POTENTIAL OF VERY LARGE CORPORA

50 Years of Big Language Data at IDS Mannheim

CMLC-2, 2014-05-31, Reykjavik

HISTORY OF CORPORA AT IDS MANNHEIM

1964 foundation of Institut für Deutsche Sprache
1967 first corpus construction project launched
1967 first corpus released (punchcarded): Mannheimer Korpus I
1972 Mannheimer Korpus II
1982 Bonner Zeitungskorpus
1999 label “Deutsches Rerenzkorpus”

HISTORY OF SEARCH AND ANALYSIS TOOLS

1982 first concordancer: REFER

- up-to 17 million words

1992 COSMAS online

- up-to 4 billion words
- supports creation of virtual corpora
  (based on content and metadata query results)

2003 COSMAS II in production

2014 next generation corpus analysis platform KorAP

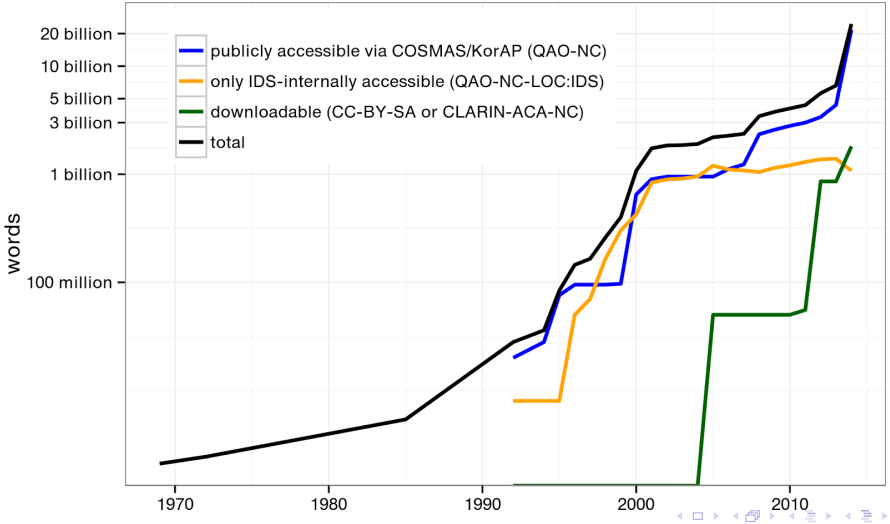
GERMAN REFERENCE CORPUS DEREKO

- empirical basis for linguistic research on contemporary German
- 24 billion words, continuously expanded
- all kinds of text types
- multiple POS, constituency, dependency annotations
- 32,000 users

DEREKO-DESIGN

- primordial sample design:
  - not intended to be balanced in any way
  - sample composition should be user-defined: “virtual corpora”
- maximizes the usefulness and re-usability of the data for a maximum number of applications
- expansion can concentrate on maximization of size and diversity

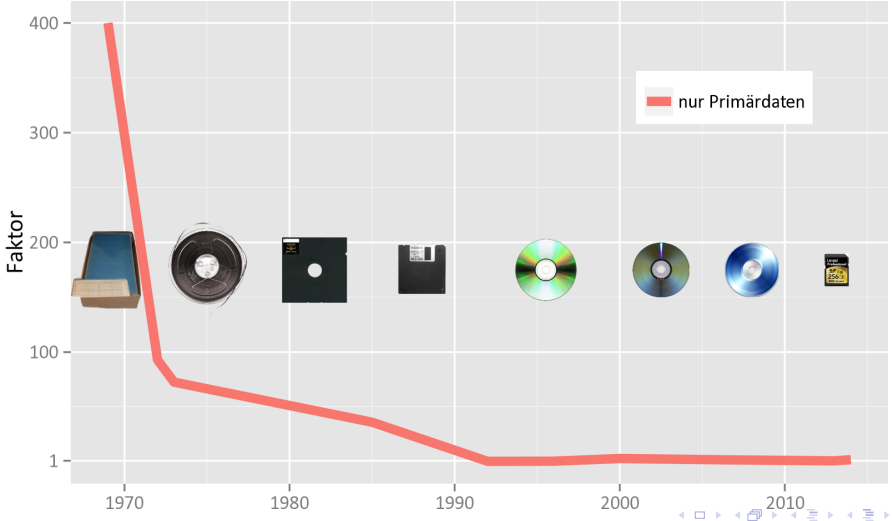
DEVELOPMENT OF DEREKO-SIZE AND AVAILABILITY SINCE 1969



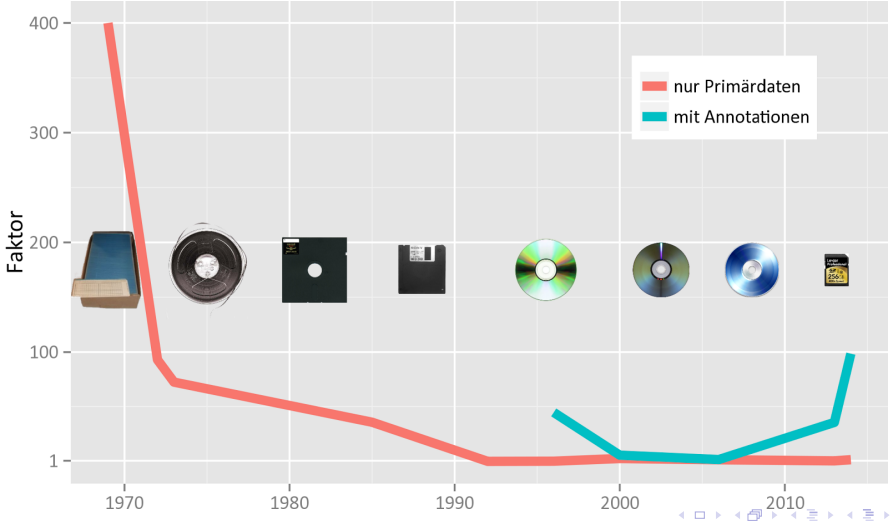
RECENT ACQUISITIONS

- Wikipedia articles and talk pages (1 billion tokens)
- parliamentary debates protocols (Project PolMine, 360 million tokens)
- fiction (6 million tokens)
- news database archive (17 billion tokens)
- current growth rate: 1.7 billion tokens/year

DEREKO SIZE IN RELATION TO CURRENT PORTABLE STORAGE MEDIA



DEREKO SIZE IN RELATION TO CURRENT PORTABLE STORAGE MEDIA



BALANCE OF INTERESTS

Important factors for licensing

HOW TO LICENSE A VERY LARGE CORPUS?

- for a large corpus, »download-first« licenses would cost hundreds of millions of euros
- way out:
  - technical solutions to maximize the usefulness of the data without interfering with interests of rights holders
  - *put the computation near the data, if the data is not allowed to move*
- strictly *non-consumptive* use
- *query and analysis only* licenses
- corpus analysis software that makes download redundant

COSMAS II

- EU-funded project started in 1994
- in production since 2003
- collocation analysis, search assistant, virtual corpora …
- supports up to 8 billion words with ~ 1.5 annotation layers
- not horizontally scalable
- indices should be held in RAM
- 20 year old code base -> high development costs

KORAP

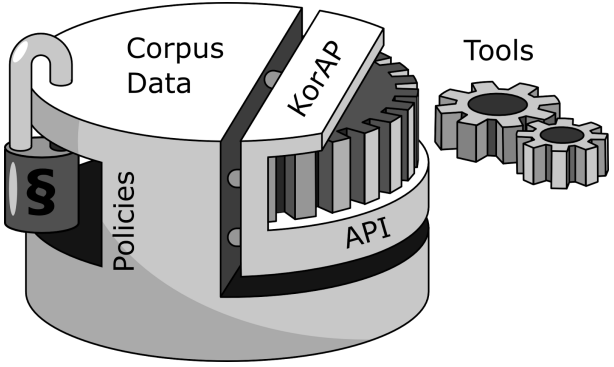
Next-generation corpus analysis platform

- Leibniz-funded project 2011-2015
- Foundry-based document data model
  - radical stand-off system: raw text separate from annotations
  - annotation layers grouped into clusters (“foundries”)
  - theoretically unlimited number of potentially conflicting descriptions
- Virtual collections: subject to access rights, on the basis of
  - text-internal properties
  - text-external properties
  - annotation properties
- Modular design, with currently two backends (Lucene/Solr, Neo4j)
- Contains reference implementation of ISO CQLF

SCALABILITY

- horizontal scalability:
  - Neo4j – natively,
  - Solr for the Lucene backend
- unlimited number of tokens
- unlimited number of foundries (=annotation clusters)
- unlimited number of layers

BRINGING THE COMPUTATION NEAR THE SECURED DATA



- Flexible access-control system for texts, annotations, and virtual collections
- Data exposed securely via an API
- Queries addressing protected portions of data are rewritten
- Planned: sandboxes for non-remote API access