Das Deutsche Textarchiv als Repositorium und Werkzeug

Frank Wiegand (BBAW)
Deutsches Textarchiv
www.deutschestextarchiv.de | wiegand@bbaw.de

Deutsches Textarchiv

- Referenzkorpus für die schriftliche neuhochdeutsche Sprache (1600–1900)
- 2288 Werke verfügbar (ca. 140 M Tokens)
- Formate: TEI-XML (DTA-Basisformat), Text, HTML, epub, TCF → CC by/nc
- Metadaten:
  - TEI-Header, CMDI, Dublin Core
  - OAI-PMH, BEACON

Persönliches Korpus

- Textauswahl nach:
  - Genre (Belletristik, Wissenschaft, Gebrauchsliteratur)
  - Entstehungszeit
  - Erfassungsmethode
  - bildgebende Einrichtung
  - Umfang
  - …

OAI-PMH-Repository

- Basisinformationen:

  http://fedora.dwds.de/oai-dta/?verb=Identify

- alle Sets:
  - http://fedora.deutschestextarchiv.de/oai-dta/?
    verb=ListSets
- alle Titel zu „Wissenschaft“

  http://fedora.dwds.de/oai-dta/?
  verb=ListIdentifiers&
  set=subject:wissenschaft&
  metadataPrefix=oai_dc
- http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Metadaten – CMDI

http://fedora.deutschestextarchiv.de/oai-dta/?verb=GetRecord&metadataPrefix=cmdi&identifier=oai:dta:adams_elektricitaet_1785

```
<Resources>
  <ResourceProxyList>
    <ResourceProxy id="dta-adams_elektricitaet_1785.xml">
      <ResourceType mimetype="application/xml">Resource</ResourceType>
      <ResourceRef>
        http://www.deutschestextarchiv.de/book/download_xml/adams_elektricitaet_1785
      </ResourceRef>
    </ResourceProxy>
    <ResourceProxy id="dta-adams_elektricitaet_1785.xhtml">
      <ResourceType mimetype="application/xhtml+xml">Resource</ResourceType>
      <ResourceRef>
        http://www.deutschestextarchiv.de/book/download_html/adams_elektricitaet_1785
      </ResourceRef>
    </ResourceProxy>
    <ResourceProxy id="dta-adams_elektricitaet_1785.text">
      <ResourceType mimetype="text/plain">Resource</ResourceType>
      <ResourceRef>
        http://www.deutschestextarchiv.de/book/download_txt/adams_elektricitaet_1785
      </ResourceRef>
    </ResourceProxy>
    <ResourceProxy id="dta-adams_elektricitaet_1785.landing_page">
      <ResourceType>LandingPage</ResourceType>
      <ResourceRef>
        http://www.deutschestextarchiv.de/adams_elektricitaet_1785
      </ResourceRef>
    </ResourceProxy>
```

Linguistische Annotation

- Ausgangspunkt: orthographie-übergreifende Suche
  - orthographische Normalisierung
  - Lemmatisierung
  - Part of Speech, Phonologie
  - Serialisierung
  - Thesaurus (GermaNet)
- Suchmaschine: DDC
- Schnittstellen: OpenSearch, DTA::CAB

Suchanfragen mit DDC

- DDC – Überblick
  - tokenbasiert, Sätze als Treffer
  - Filter: Wortart, Kontext, Metadaten
  - Formate: HTML, KWIC, Text, JSON, YAML, Atom, RSS, Histogramme, DDC Raw Data
- DTA::CAB integriert (Abfrage und Index)
- DDC ist Open Source:
  - http://sourceforge.net/projects/ddc-concordance/

DDC – mächtige Syntax

- lexembasierte Suche:

  `Teil` → Teil, Teile, Theile, Teils, Theyls, Thail, …

  (nicht: teils)
- exakte Wortform:

  `@Theil` → Theil

  (aber nicht: Teil, Teile, Theile, Teils, Theyls, Thail, teils …)
- Trunkierung: `*teil`, `teil*`, `*teil*`
- Phrasensuche: `"gutes Beispiel"`
- Abstandssuche: `"gutes #2 Beispiel"`

DDC – Syntax (2)

- mehrere Indexfelder pro Token:
  - `$l=kommandieren`

    kommandieren, kommandiert, commandiren, commandiret, …
  - `"$p=ADJ* *zeit"`

    schöne Lebenszeit, herrliche Hochzeit, grässliche Schulzeit, …

    (Basis: Stuttgart-Tübingen-Tagset STTS)
  - `$u=Wachſtube`

    Wachſtube, Wachſtuben

    (aber nicht: Wachstube, Wachstuben)
  - `$w=Wachstube`

    Wachstube, Wachstuben, Wachſtube, Wachſtuben

DDC – Syntax (3)

- Verknüpfungen:
  - `voll && ganz`
  - `voll || ganz`
  - `Perlen && Säue && !werfen`
- Expansionen:
  - `$w=still|pho`

    still, Stil, stiehl, Stiel, …
  - `Fisch|gn-asi`

    Aal, Karpfen, Karpffen, Sardelle
  - `Term|`*`http://example.com/my-expander`*

DDC – Syntax (4)

- Reguläre Ausdrücke (PCRE):
  - `/hoffnungs(voll|los)e/i`
- Filter:
  - `* with $con=/head/`
  - `* with $r=/aq/`
  - `* #has[author,/Goethe/]`
  - `* #date[1800]`

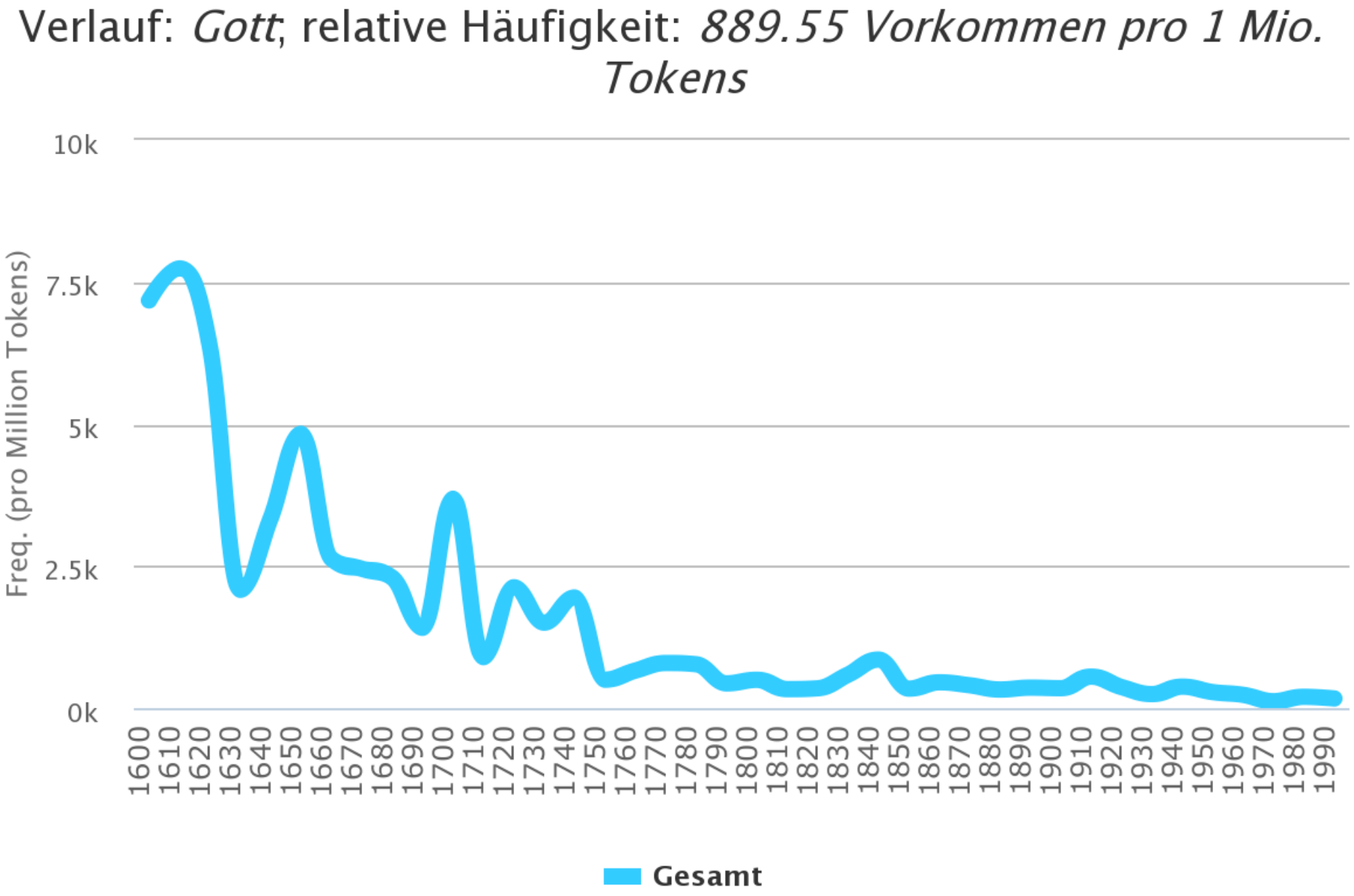
DDC – Abfragen

- Zum Einstieg:
  www.deutschestextarchiv.de
- Open-Search-Wrapper:

  http://kaskade.dwds.de/dtaos
- Webservice:

```
$ curl -d q=Gold -d fmt=json \
http://kaskade.dwds.de/dtaos/dta.perl
```

Verlaufskurven



DTA::CAB

- http://www.deutschestextarchiv.de/demo/cab
- Webservice (HTTP) für eigene Dokumente:
  - Serialisierung
  - Tokenisierung (→ moot/WASTE)
  - Analyse
- Formate:
  - XML (TEI, TCF), JSON, YAML, CSV, …
- Demo ist frei verfügbar (< 500 KB Daten)

DTA::CAB

- $ curl 'http://www.deutschestextarchiv.de/demo/cab/query?fmt=json&q=geweſen'


```
"moot" : {
  "word" : "gewesen",
  "lemma" : "sein",
  "tag" : "VAPP"
},
"hasmorph" : 1,
"lang" : [ "de" ],
"xlit" : {
  "latin1Text" : "gewesen",
  "isLatinExt" : 1,
  "isLatin1" : 1
}
```

Danke.

- www.deutschestextarchiv.de/api
- www.deutschestextarchiv.de/download

- wiegand@bbaw.de
- @textarchiv