Wie kann man die Suche nach historischen Korpora zentraler, einheitlich strukturierter realisieren?

Thomas Krause*, Anke Lüdeling*, Carolin Odebrecht*, Laurent Romary+, Peter Schirmbacher*, Dennis Zielke*
Humboldt-Universität zu Berlin*, Inria France+
http://www.laudatio-repository.org

Digital Humanities Berlin 28.02.2014
“Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen”

1. Beispiel für linguistische Forschungsdaten

- **RIDGES** (Register in German Diachronic German Science[13]) Herbology Projekt untersucht die Entstehung und Entwicklung der deutschen **Wissenschaftssprache**
- Integration des Projektes in die **Lehre**
  - Studierende erstellen historische Korpora
  - drei Jahrgänge arbeiten zusammen an einem Korpus
    - **Erweiterung der Textgrundlage**
    - **Veränderung der Korpusarchitektur** (multiple Segmentierung für eine Schritt für Schritt Normalisierung)
    - **Korrektur, Anpassung und Erweiterung der Annotationen**

2. Korpuserstellung und Veröffentlichung

- Herausgabe des Korpus auf **Projekthomepage**, inklusive Download aller Formate unter **CC-BY 3.0**[9]
- **freie, idiosynkratische Dokumentation** der Korpuserstellung und der Annotationsrichtlinien
- Zugänglichkeit der Daten über Suchtool ANNIS[7]

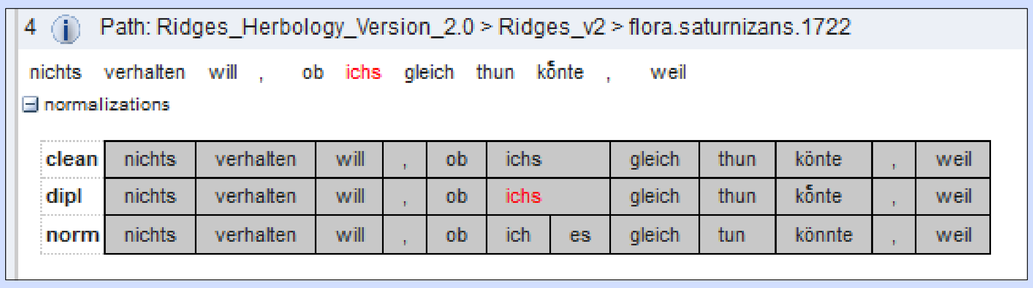
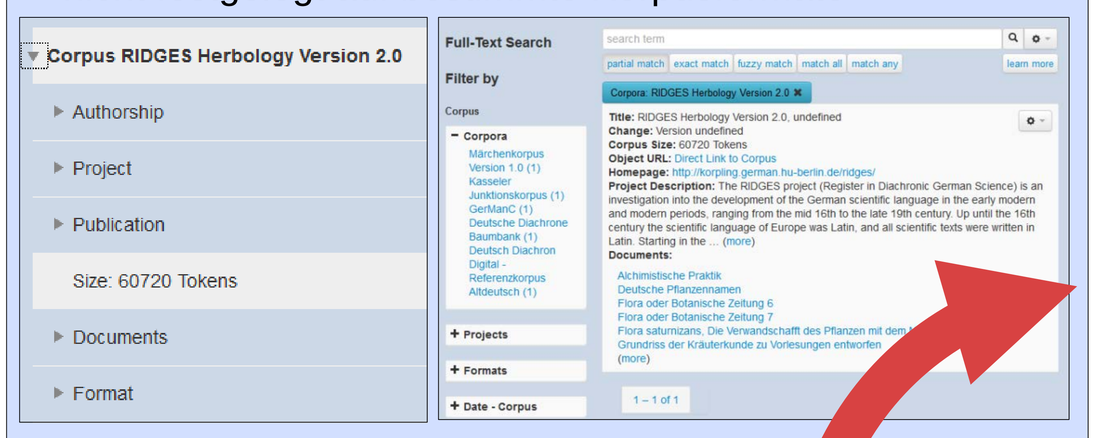


Abbildung: RIDGES Herbology Corpus V2.0 in ANNIS Suchinterface – Multiple Tokenisierung und Normalisierung[3]

3. Zentrale, einheitlich strukturierte Anzeige, Zugriff und Wiederverwendung von historischen Korpora durch LAUDATIO

Speicherung / Zugriff auf bestehendes Korpus

- Facettensuche
- strukturierte Anzeige der Dokumentation
- nicht festgelegt auf bestimmte Korpusformate

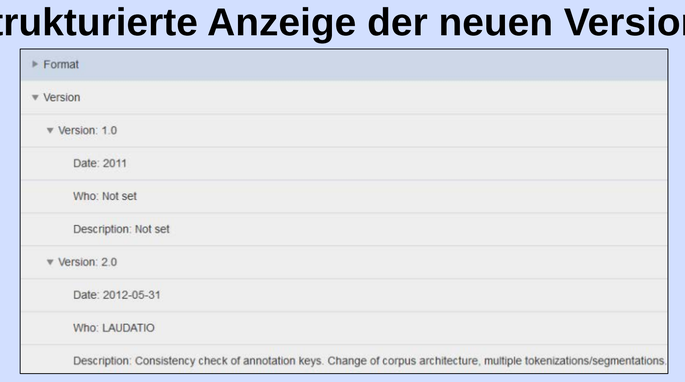


Metadatenschema (TEI-ODD[1,2])

- Dokumentation des Korpus mit Metadaten in TEI XML[5,6]
- Veröffentlichungsgeschichte des Korpus

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader type="CorpusHeader">
    <fileDesc>
      <titleStmt>
        <title type="Corpus">RIDGES Herbology Version 3.0</title>
        <editor role="CorpusEditor" n="1" />
        <author role="Annotator" n="1" />
      </titleStmt>
      <extent type="Tokens">122698</extent>
      <publicationStmt>
        <authority>Humboldt-Universität zu Berlin</authority>
        <idno>RIDGES Herbology Project.</idno>
        <availability status="free">
          <p>Open Source Project. Open Source Project. All corpus data generated by the RIDGES project
            is licensed under a Creative Commons Attribution 3.0 Unported License.</p>
        </availability>
        <date when="2011" type="CorpusRelease">First corpus release.</date>
        <date when="2012-05-31" type="CorpusRelease">Second corpus release.
Correction of the first version.</date>
        <date when="2013-06-15" type="CorpusRelease">Third corpus release.
Extension of the corpus.</date>
      </publicationStmt>
```

Strukturierte Anzeige der neuen Version



Download
Weiterverarbeitung
Konfiguration
Upload

Notwendige Anpassung des Interface via JSON

- motiviert durch Erweiterungen des Metadatenschemas

```
"3":{"value":"Publication",
        "1":{ "value":"Authority: $teiHeader->fileDesc->publicationStmt->authority"
        },
        "2":{ "value":"Project Name: $teiHeader->fileDesc->publicationStmt->idno"
        },
        "3":{ "value":"Availability Status: $teiHeader->fileDesc->publicationStmt-
         >availability['status']"
        },
        "5":{ "value":"$teiHeader->fileDesc->publicationStmt->availability->p"
        },
        "6":{ "value":"Corpus Release: $teiHeader->fileDesc->publicationStmt->date['when']"
        }
      },
```

Veröffentlichung der neuen Version

- Upload neue Version des Korpus

4. Technik[12]

- basiert auf der Repositorysoftware Fedora Commons[4]
- automatische Validierung durch RelaxNG-Schema[2]
- automatische Registrierung und Verwaltung der PIDs für jede Korpusversion[11]
- Indexierte und facettierte Suche mithilfe der Suchservertechnologie ElasticSearch[10]
- JSON-Editor zur Konfiguration der TEI XML Header in der GUI
- Schnittstelle zum Such- und Visualisierungstool ANNIS

5. Ausblick

- Erweiterung auf historische Textkorpora aus verschiedenen Geisteswissenschaften
- Entwicklung einer Schnittstelle zu Konverter-Framework SaltNPepper[8]
- Vernetzung der Inhalte mit anderen Datenbanken (Linked Open Data & Interoperabilität mit Metadatenformaten)

Referenzen [1] Burnard, L., Bauman, S. (Hg.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.* [2] Burnard, L., Rahtz, S. (2004) *RelaxNG with Son of ODD. Extreme Markup Languages.* [3] Krause, T., Lüdeling, A., Odebrecht, C., Zeldes, A. (2012) Multiple Tokenization in a Diachronic Corpus. *Exploring Ancient Languages through Corpora Conference (EALC). 14.-16.6.2012, Oslo.* [4] Lagoze, C., Payette, S., Shin, E., Wilper, C. (2006) Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries, 6(2)*. S.124-138. [5] Odebrecht, C. (erscheint) Modeling Linguistic Research Data for a Repository for Historical Corpora. *Digital Humanities 2014 Conference.* 8.7.-12.7.2014, Lausanne. [6] Odebrecht, C., Krause, T. (2013) Metadata in an Infrastructure for Historical Corpora. *SFB 732 Incremental Specification in Context. Kolloquium.* 20.6.2013, Stuttgart. [7] Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009) ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009.* Liverpool, UK. [8] Zipser, F., Romary, L. (2010) A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta. | [9] Creative Commons-License https://creativecommons.org/licenses/by/3.0/de/ [10] Elastic Search http://www.elasticsearch.org [11] Handle-PIDs verwaltet durch die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen http://epic.gwdg.de/wiki/index.php?title=EPIC:API:v2:contribution [12] Technische Dokumentation des LAUDATIO Repository http://rtd.cms.hu-berlin.de/docs/laudatio-repository-documentation/ [13] RIDGES Herbology http://korpling.german.hu-berlin.de/ridges, Korpus durchsuchbar unter https://korpling.german.hu-berlin.de/annis3/