Experiences with BaseX and XQuery for diachronic German texts

Cerstin Mahlow
cerstin.mahlow@unibas.ch
Department of German, University of Basel

Prague, February 8, 2013

Challenges for Retrieval in Digital Humanities

- Growing interest in **diachronic perspectives**
- Language phenomena to be dealt with:
  - spelling variation
  - change in inflectional paradigms
  - change in syntactical patterns
  - word order
  - vocabulary
  - change in meaning of words
- Document phenomena to be dealt with:
  - digitization errors
  - variety of document structures
  - variety of pre-annotation
  - large collections of documents
- **Searching for concepts**

Context – the OLdPhras project

- "German Proverbs and idioms in language change. Online-dictionary for diachronic phraseology"
- Funded by Swiss National Science Foundation
- ➥ Describe phraseological change for German from 1650 until present and provide evidence of actual use from corpora
  - **Extract and annotate evidence of phrasemes**

Examples

- **der Apfel fällt nicht weit vom Stamm**
  (like father, like son)
  - *So fällt der Apfel vom dem Stamm nicht weit*
  - *so weit fällt kein Apfel vom Stamm*
  - *daß der Apfel nahe zum Stamme fällt*
  - *die birn nit wey vom baum falt*
- **jemandem Rede und Antwort stehen**
  (to answer all questions)
  - *wußte er ihr nicht Rede und Antwort zu geben*
  - *wo er Red und Antwort geben möge*
  - *in der That stand er auch in Gedanken dem Untersuchungsrichter Red' und Antwort*

Searching Phrasemes

- Typical citation forms of German phrasemes:
  - «*jmndm. Rede und Antwort stehen*» (to answer all questions)
  - «*ins Fettnäpfchen treten*» (to blunder)
  - «*kurz angebunden sein*» (to be brusque)
  - «*gegen den Strom schwimmen*» (to swim against the current)
  - ➥ Infinite verb form in the last position with only some explicitly marked valency slots
- Phrasemes in texts:
  - are inflected,
  - may have various syntactic roles,
  - vary with respect to word order of the lexical units
- Searching for phrasemes means:
  - Looking for syntactic patterns/constructions
  - Looking for co-occurences of words ➥ querying lemmas
  - Allowing for variation/modification

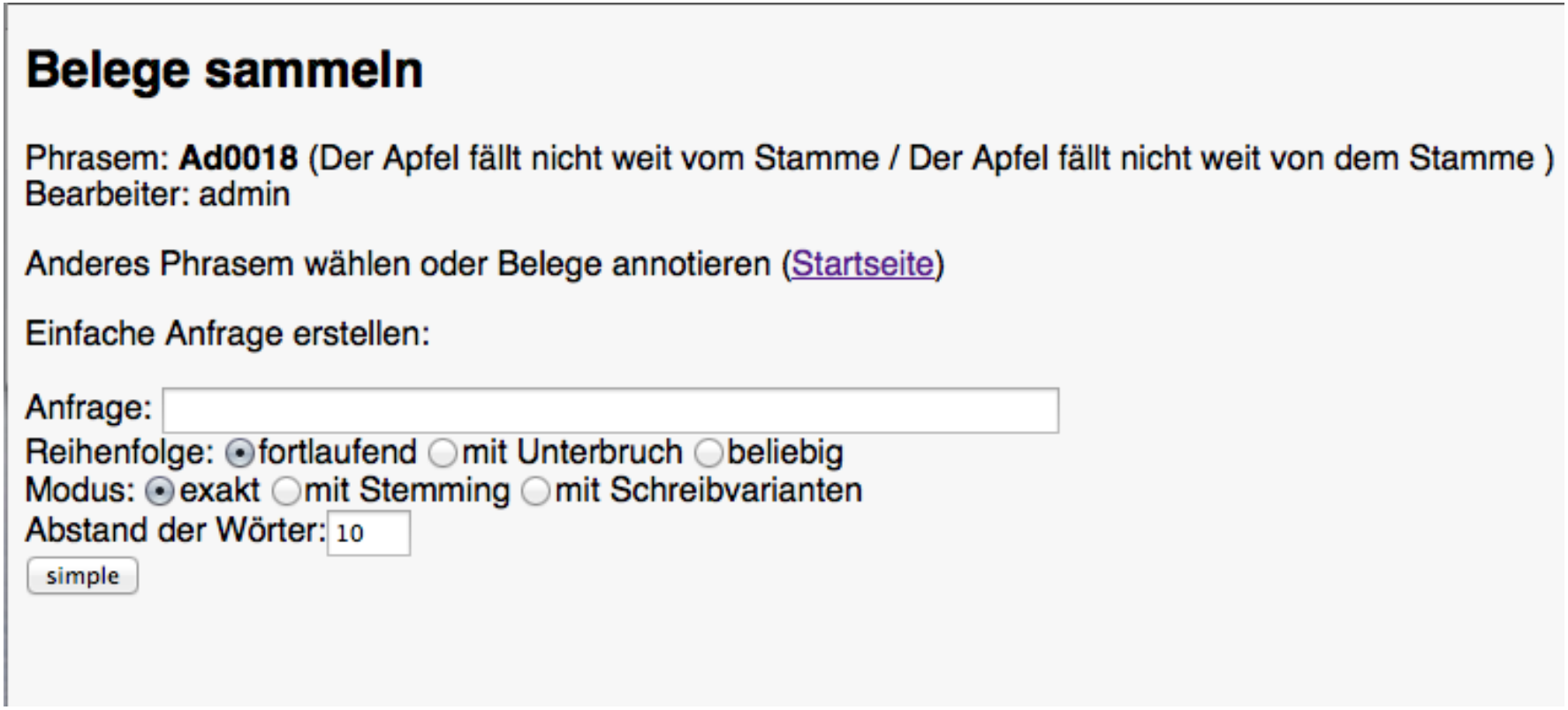
Working with Corpora

- Text collections:
  - **Deutsches Textarchiv (DTA)** (1780 to 1900, 75 M word forms)
  - **TextGrid Digital Library (DB)** (“beginning of publishing to 1930”, 87 M word forms)
  - **GerManC** (1650 to 1800, 2000-word samples)
- Use BaseX because of:
  - heterogeneous TEI annotation
  - lack of linguistic annotation
  - size of data
  - use via the web
  - speed

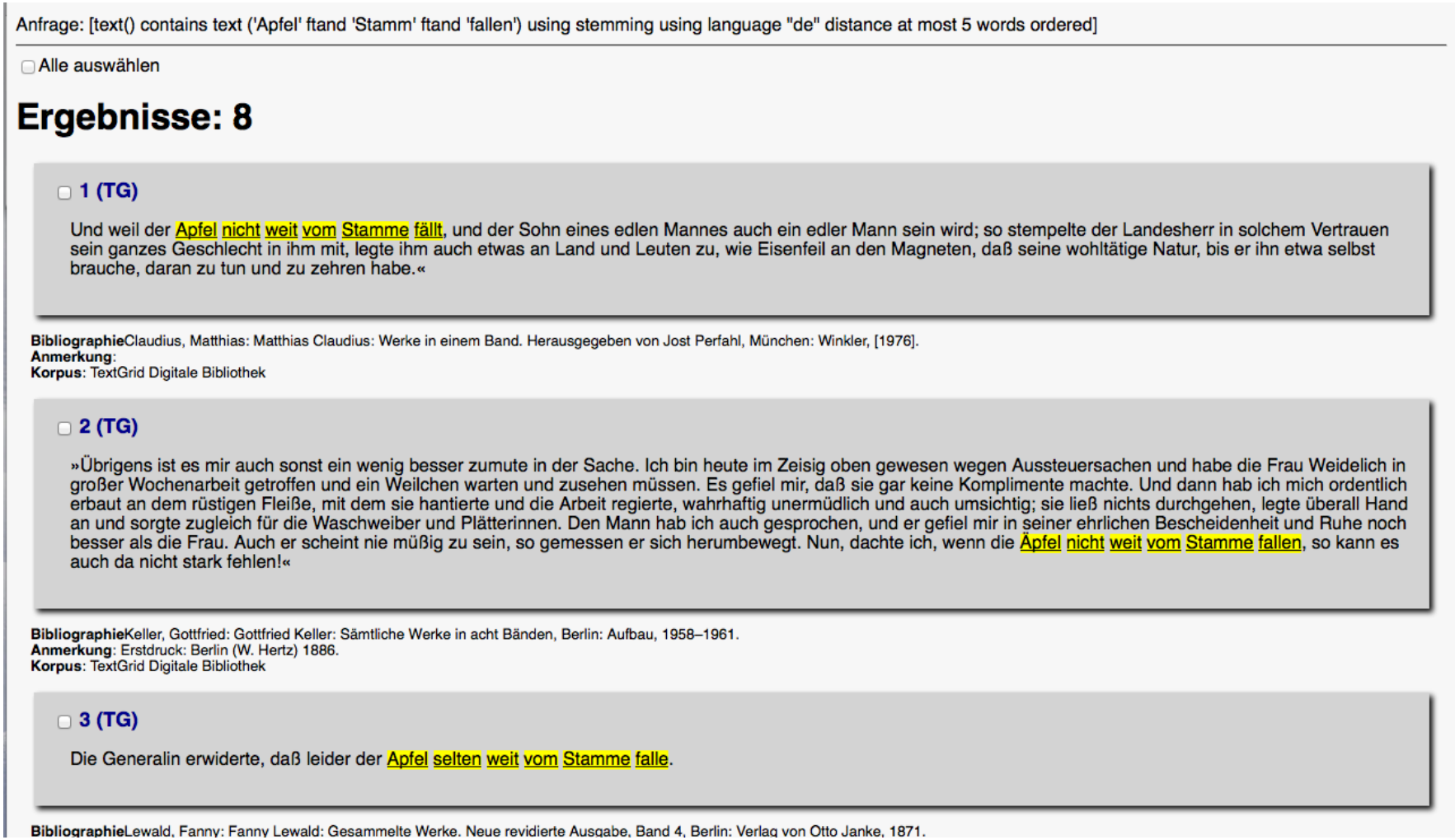
Challenges

- Recall/Precision
  - High recall (do not miss an evidence)
  - High precision (reduce manual effort)
- ➥ “Good” queries
- Size of data (too little and too much)
- **Concurrent user-actions**
- Overlapping/nested mark up of evidence, aka **mixed content**
- Main focus on content, not on XML-structure
- Add information to the corpus instead of extracting matches (with context)
- Users are linguists

Querying the Text Collection



Inspecting Results



Anfrage: [text() contains text ('Apfel' ftand 'Stamm' ftand 'fallen') using stemming using language "de" distance at most 5 words ordered]

Alle auswählen

Ergebnisse: 8

1 (TG)

Und weil der Apfel nicht weit vom Stamme fällt, und der Sohn eines edlen Mannes auch ein edler Mann sein wird; so stempelte der Landesherr in solchem Vertrauen sein ganzes Geschlecht in ihm mit, legte ihm auch etwas an Land und Leuten zu, wie Eisenfeil an den Magneten, daß seine wohltätige Natur, bis er ihn etwa selbst brauche, daran zu tun und zu zehren habe.«

BibliographieClaudius, Matthias: Matthias Claudius: Werke in einem Band. Herausgegeben von Jost Perfahl, München: Winkler, [1976].
Anmerkung:
Korpus: TextGrid Digitale Bibliothek

2 (TG)

»Übrigens ist es mir auch sonst ein wenig besser zumute in der Sache. Ich bin heute im Zeisig oben gewesen wegen Aussteuersachen und habe die Frau Weidelich in großer Wochenarbeit getroffen und ein Weilchen warten und zusehen müssen. Es gefiel mir, daß sie gar keine Komplimente machte. Und dann hab ich mich ordentlich erbaut an dem rüstigen Fleiße, mit dem sie hantierte und die Arbeit regierte, wahrhaftig unermüdlich und auch umsichtig; sie ließ nichts durchgehen, legte überall Hand an und sorgte zugleich für die Waschweiber und Plätterinnen. Den Mann hab ich auch gesprochen, und er gefiel mir in seiner ehrlichen Bescheidenheit und Ruhe noch besser als die Frau. Auch er scheint nie müßig zu sein, so gemessen er sich herumbewegt. Nun, dachte ich, wenn die Äpfel nicht weit vom Stamme fallen, so kann es auch da nicht stark fehlen!«

BibliographieKeller, Gottfried: Gottfried Keller: Sämtliche Werke in acht Bänden, Berlin: Aufbau, 1958–1961.
Anmerkung: Erstdruck: Berlin (W. Hertz) 1886.
Korpus: TextGrid Digitale Bibliothek

3 (TG)

Die Generalin erwiderte, daß leider der Apfel selten weit vom Stamme falle.

BibliographieLewald, Fanny: Fanny Lewald: Gesammelte Werke. Neue revidierte Ausgabe, Band 4, Berlin: Verlag von Otto Janke, 1871.

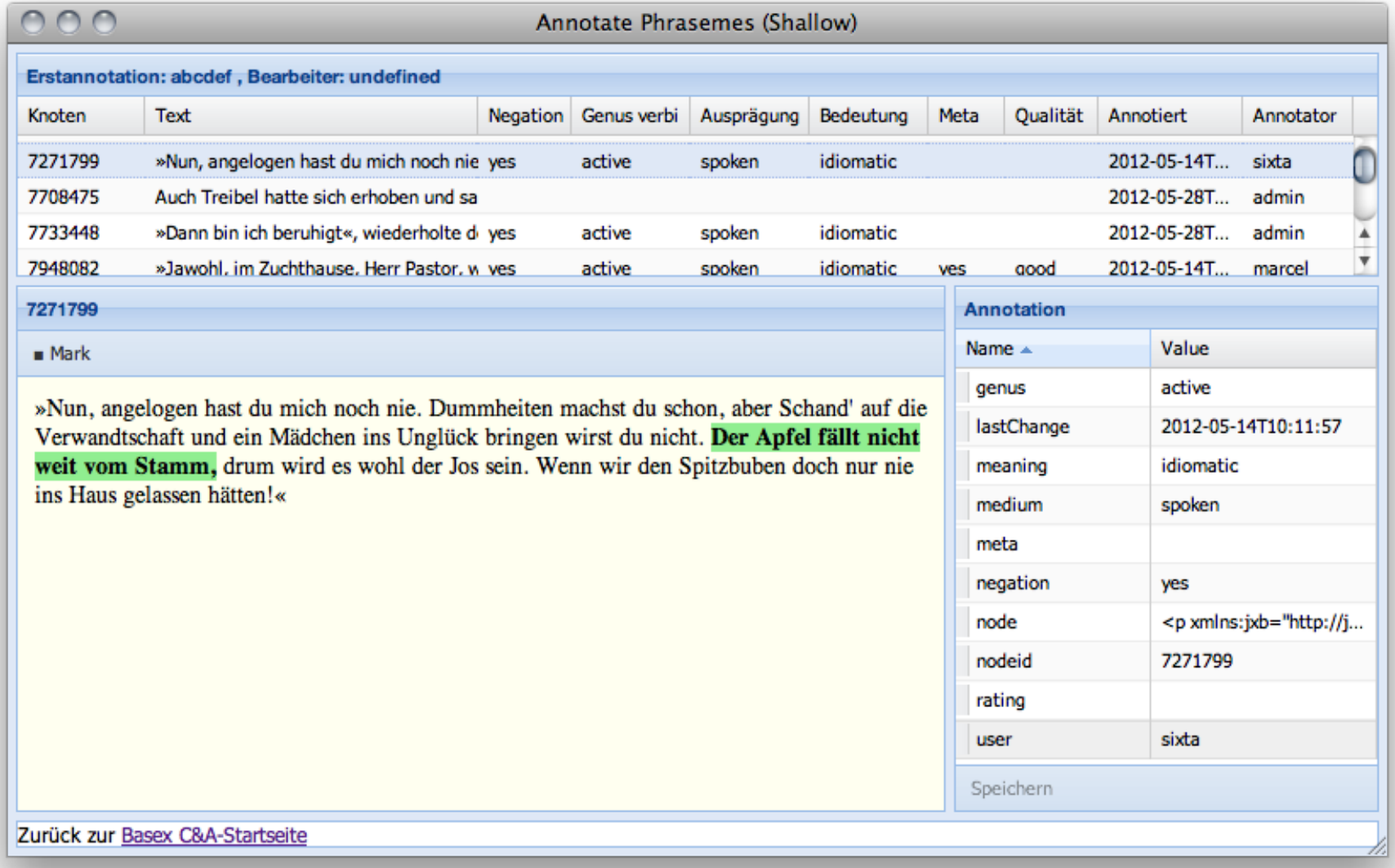
Collecting Results

```
<collection>
  <entry time="2012-03-04T17:43:29">
    <node>14452125</node>
    <query>[text() contains text ('Bank' ftand 'fallen') using stemming using language "de"
distance at most 6 words ordered]</query>
    <person>marcel</person>
    <phraseme>Ad0032</phraseme>
    <selected>no</selected>
  </entry>
  <entry time="2012-03-04T17:43:29">
    <node>23118042</node>
    <query>[text() contains text ('Bank' ftand 'fallen') using stemming using language "de"
distance at most 6 words ordered]</query>
    <person>marcel</person>
    <phraseme>Ad0032</phraseme>
    <selected>no</selected>
  </entry>
  <entry time="2012-03-04T17:43:29">
    <node>26984924</node>
    <query>[text() contains text ('Bank' ftand 'fallen') using stemming using language "de"
distance at most 6 words ordered]</query>
    <person>marcel</person>
    <phraseme>Ad0032</phraseme>
    <selected>no</selected>
  </entry>
  <entry time="2012-03-04T17:43:29">
    <node>31298848</node>
    <query>[text() contains text ('Bank' ftand 'fallen') using stemming using language "de"
distance at most 6 words ordered]</query>
    <person>marcel</person>
```

Current Results

- 469 phrasemes searched, resulting in
  - 42’842 positive hits
  - 94’434 negative hits
  - 1 to 47 queries per phraseme

Current Annotation Interface



Approach for Annotation

- Using ExtJS
- Process hits one by one
  - inspect node and bibliographic data and decide about editing
  - mark evidence (= **adjust mark u**p)
  - add attributes
    - automatically for query term, ID, time-stamp, etc.
    - manually for idiomaticity, negation, mediality, etc.
  - update node

Issues

- Conflicting multi-user actions
  (might be due to our web server, probably improved using BaseX 7.6)
- Framework for annotation task
- Marking mixed content and storing marked mixed content
- Encoding for fuzzy search (especially wrt DTA and GerManC)
  - long s (U+017F) should be treated as equivalent to regular *s*,
  - handling of combining characters,
    e.g., a with superscripted e (U+0061 U+0364) should match *a* and *ä*