3. CLARIN-D Fach-AG Workshop (30. Juni 2015)
„Fächerübergreifende Perspektiven durch digitale Forschungsinfrastrukturen“

Korpusanalyse mit dem ‚polmineR‘

Prof. Dr. Andreas Blätte
Professur für Public Policy und Landespolitik
UNIVERSITÄT DUISBURG-ESSEN

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

NRW School of Governance
www.nrwschool.de

PolMine

Yet another tool for corpus analysis?!

Eine Frage der Forschungslogik ...

- Quantifizierung / sprachliche Muster
  und Entwicklung von Interpretationen
- analytische Induktion („abduktive“ Forschungslogik)
- „Dialog mit den Daten“ *=> interaktive Datenanalyse*

Eine Frage der Daten ...

- Metadatenstruktur des PolMine-Plenarprotokollkorpus
  (Zeitstempel, Sprecher, Parteizugehörigkeit, etc.):

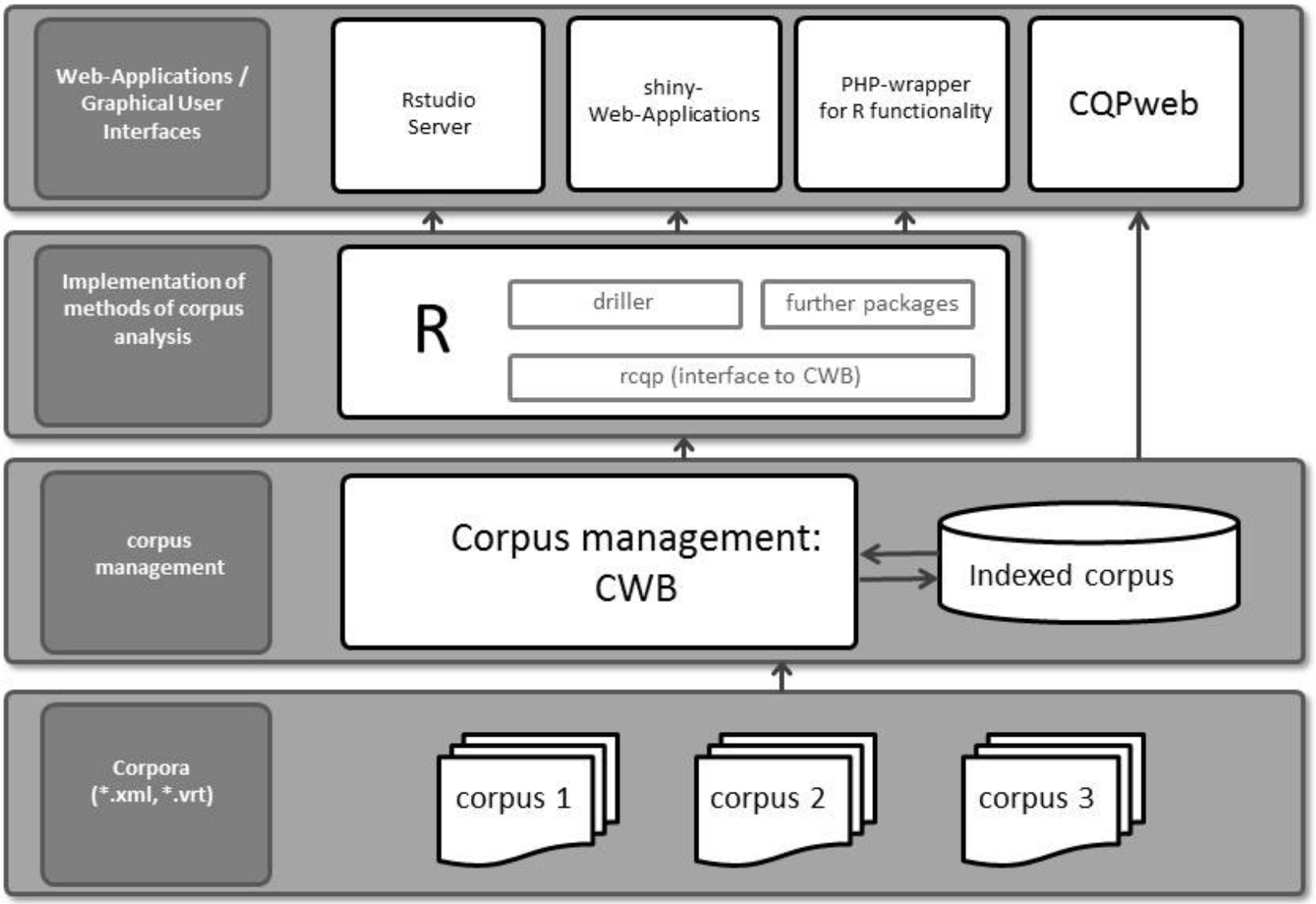
  => flexible Bildung von Subkorpora
- Umfang der Daten (70 – 250 Millionen Token)

  => Performanz der Analyseumgebung

Zielsetzungen

- Interaktvität *(Analyseumgebung in R)*
- Performanz *(Nutzung der CWB als backend)*
- Flexibilität *(Bandbreite der R-Pakete)*
- Quelloffenheit *(Publikation über GitHub)*
- Dokumentation *(R-Zwang zur Dokumentation)*
- Portabilität *(nur Linux / Mac OS)*
- Usability *(ein Tool für Verwegene)*

Schichtenarchitektur des PolMine-Toolkit



Leit- und Designideen

- Eine Paketfamilie (vgl. http://www.github.com/PolMine)
  - polmineR als Basispaket, dazu Plugins (z.B. polmineR.graph)
  - ctk (corpus toolkit) als Gegenpart zum polmineR für Korpusaufbereitung
- Objektorientierung
  - Implementierung mit S4-System (S4-Klassen/Methoden)
  - Nutzung gängiger generischer Funktionen (show, summary, length, barplot etc.)
- Parallelisierung
- Selbstbeschränkung
  - Interfaces / Exportfunktionen zu anderen Paketen / Programmbibliotheken (tm, topicmodels, ctm, mallet etc.)

Workflow – Beispiel 1

- Anlegen einer „Partition“
  merkel <- partition(„PLPRBTXT“, def=list(text_speaker=„Angela Merkel“))
- Frequenzanalyse
  tf(merkel, query=„Finanzkrise“)
  tf(merkel, query=,“Griechenland“ [] „Finanzkrise“‘)
- Kollokationsanalyse
  finKoll <- context(merkel, „Finanzkrise“)
  head(finKoll, n=25)
  as.data.frame(finKoll)
- Konkordanz- bzw. KWIC-Analyse
  kwic(finKoll)
  kwic(merkel, „Finanzkrise“)
  kwic(„PLPRBTXT“, „Finanzkrise“)

Workflow – Beispiel 2

- Anlegen eines „partitionBundle“

```
y2001 <- partitionBundle(
      „ARENEN“, def=list(text_year=„2001“),
      var=list(text_filename=NULL), tf=„word“
    )
```

- Umwandlung in Matrix

```
dtm <- as.DocumentTermMatrix(y2001, pAttribute=„word“)
```

- Reduktion der Matrix

```
dtm2 <- trim(dtm, ...)
```

- Berechnung eines Topic Models

```
tmodel <- LDA(dtm, ...)
```

polmineR-Kernfunktionalität

Methode	Beschreibung
partition()	Anlegen einer Partition
partitionBundle()	Anlegen eines Bündels von Partitionen
tf()	Termfrequenzen
dispersion()	Häufigkeitsverteilung
enrich()	Anreichern eines Objekts
trim()	Beschneiden eines Objekts
context()	Kontext- und Kollokationsanalyse
keyness()	Schlagwortberechnung / Termexktration
collocations()	Berechnung aller Kollokationen in Partition
kwic()	Konkordanzen / kwic
browse()	Ausgabe in Browser
mail()	Export per Email

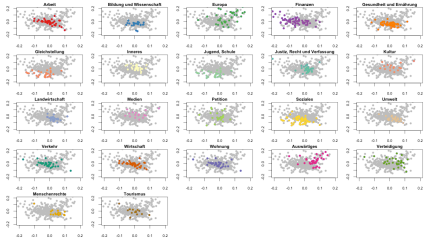
polmineR-Plugins

Paketname	Beschreibung
polmineR.shiny	shiny-Applikationen zu Kernfunktionen
polmineR.graph	Analyse von Kollokationsgraphen
polmineR.plpr	Datenspezifisches Paket für Plenarprotokollkorpora
polmineR.export [x]	Exportfunktionen / Rekonstruktion von XML
polmineR.press [x]	Datenspezifisches Paket für Zeitungskorpora
polmineR.sampleCorpus	Beispieldaten (Auszug Plenarprotokollkorpus)

[x] nicht öffentlich

Anwendungsfälle

- Kohärenz von Fraktionen im Deutschen Bundestag
  (visuelle Analyse von Korpusähnlichkeiten)



- Thematisierungs-/Diskursverhalten von Abgeordneten mit Migrationshintergrund
  (datengeleitete Diktionärsentwicklung / Termextraktion | Korpusähnlichkeiten)

- Sind Abgeordnete mit Migrationshintergrund typische MdBs für ihre jeweilige Bundestagsfraktion
  (Dimensionsreduzierung | Diskriminanzanalyse)

- Trends in der Kontextvariation der „Vielfalt“
  (Frequenzanlayse | Topic Models)

Perspektiven / To Do

- Kohärenz / Debugging
- Verbesserung Dokumentation
  (Vignette, Beispiel-Code, Video-Tutorials)
- Nutzerfreundlichkeit
  (Browserausgabe, shiny-Apps)
- Performanz durch Rcpp-Implementierung
- Wartung rcqp als API zur CWB
- Zugänglichkeit / Server-Installation von RStudio