Deutsche politische Reden – Korpus und Visualisierung

Adrien Barbaresi
ICAR / ENS Lyon

Vorstellung – Korpora und Werkzeuge meiner Doktorarbeit verfügbar machen

- Thema : **eine korpusbasierte Annäherung und Visualisierung der Textverständlichkeit**.
  Betreuung : Pr. Benoît Habert (ENS Lyon) und Pr. Henning Lobin (Universität Gießen)
  → Ziel ist es, einen Überblick zu verschaffen. Erst annotieren, dann klassifizieren und einstufen.
- **Von einem Black Box zu einem White Box-Modell**
  → Die Arbeit an Korpora, deren Übertragung möglich ist, soll vorgezogen werden.
  → Korpora und Werkzeuge sollen am Ende verfügbar gemacht werden.

Diese Ressource ist online verfügbar : **http://purl.org/corpus/german-speeches**

Inhalt und Zusammensetzung der Subkorpora

⇒ von offiziellen Quellen heruntergeladen : Webseiten von dem Bundespräsidialamt und Bundeskanzleramt.

Bundespräsidentenkorpus
1 442 Texte, 2 392 074 Tokens.
01.07.1984 – 17.02.2012

Präsident	Texte	Tokens
Johannes Rau	568	961 538
Horst Köhler	527	774 563
Christian Wulff	202	285 893
Roman Herzog	131	322 468
Richard von Weizsäcker	14	47 612

Bundesregierungkorpus
1 836 Texte, 3 893 766 Tokens.
11.12.1998 – 06.12.2011

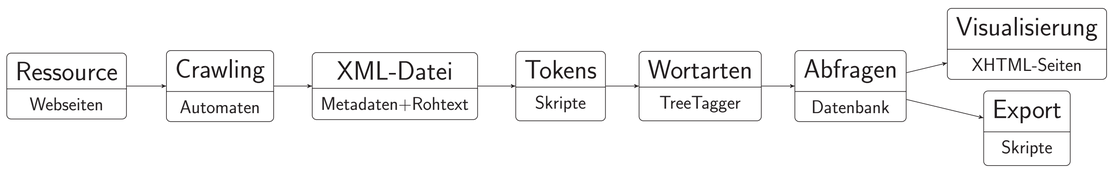
Politiker(in)	Texte	Tokens
Angela Merkel	610	1 641 481
Gerhard Schröder	420	984 373
Bernd Neumann	248	281 694
Christina Weiss	206	299 205
keine Angabe	*92*	*204 066*
Michael Naumann	61	120 558
Julian Nida-Rümelin	48	92 683
Thomas de Maizière	43	88 865
Hans Martin Bury	42	73 520
Joschka Fischer	32	55 511
Rolf Schwanitz	23	28 078
Frank-Walter Steinmeier	10	23 444

vorläufige Werte, Stand : 28.02.2012

Auswahl und Probleme

- Interviews, Reden auf anderen Sprachen und doppelte Reden → ggf. ausgelassen
- Schlecht klassifizierte Reden (andere bzw. fremde Persönlichkeiten)
- Keine typographischen Normen

Arbeitsfluss und erwähnenswerte Komponenten



Bash Master-Skript, das durch die meisten Etappen führt (baukastenartige Architektur)
Perl *focused crawling*, Tokenisierungsskript von Stefanie Dipper + Korrekturen, Reguläre Ausdrücke (Vorbereitung, Metadatenbestimmung), Stichwörterbestimmung, Erstellung der XML-Dateien
SQLite Kommunikation durch das Perl DBI Modul, « internes » Format, schnellere Abfragen

Export im XML Format und Veröffentlichung

- Ein möglichst TEI-kompatibles inline XML Format
  - u.a. `s` und `w` Tags und `xml:id` Attribut.
  - Metadaten : Titel, Redner(in), Datum, (Ort), Quelle, Exzerpt und Anrede (automatisch bestimmt und daher nicht fehlerfrei).
  - Verfügar mit oder ohne maschinelle Tokenisierung und Lemmatisierung.
- Wiederveröffentlichung → Grauzone ?
  - § 48 UrhG : öffentliche Reden können frei wiederveröffentlicht werden.
  - Creative-Commons Lizenz BY-SA 3.0 (Namensnennung und Weitergabe unter gleichen Bedingungen).

Darstellung

Technologie

⇒ **Valide XHTML/CSS Dokumente**, gleiche Wiedergabe bei den meisten Browsern.
⇒ JavaScript : Reiter-Navigation, Wörtermarkierungen, eilige Ergänzung der Webseiten beim Laden.

Gliederung der automatisch generierten Seiten

1. **Übersicht** mit allen Texten und deren **Stichwörtern** (hauseigener, noch experimenteller Algorithmus).
2. **Listen** von Schlüsselwörtern, **für die die Ergebnisse im Voraus erstellt worden sind.**
3. Die Entwicklung der erwähnten Wörter wird anhand von **Balkendiagrammen** dargestellt.
   Verteilung **pro Jahr, pro Redner, pro Text und schließlich Darstellung des Kontextes möglich.**
   Die Größe einer Nummer weist auf die Häufigkeit des gesuchten Ausdruckes hin.
4. Links zu den **Urtexten**, ggf. werden **gesuchte Wörter markiert**.

Beispiel : das Wort Krieg im Bundespräsidentenkorpus

Künftige Arbeit

- Regelmäßige Erweiterung im Laufe der Zeit.
- Weitere Annäherung an das XML-TEI Format.
- Visualisierung beliebiger Ausdrücke (Funktion bisher aus Infrastrukturgründen nicht vorhanden).
- Integration in andere Korpora (z. B. im Korpus elektronischer Texte von Gertrud Faaß und Ulrich Heid, Universität Hildesheim, ebenso am DGfS-CL 2012 vorgestellt).
- Arbeit an der Erkennung von wichtigen Schlüsselwörter.

École Normale Supérieure de Lyon – 15, parvis René Descartes F-69342 Lyon BP 7000 cedex 07 http://perso.ens-lyon.fr/adrien.barbaresi/ adrien.barbaresi@ens-lyon.fr