Word Continuum: WoCo Explorer

Alex Flückiger, Alon Cohen & Brian Strebel

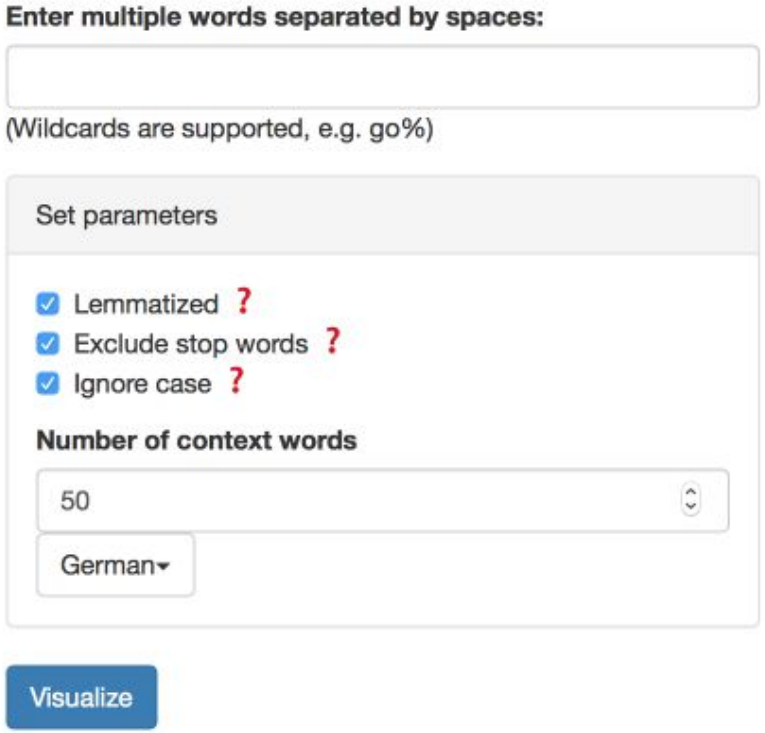
Projektbeschrieb

- Visualisierung von Bedeutungsunterschieden
  - Gebräuchliste Verwendung von Wörtern in einem bestimmten Kontext
  - Darstellung der kontextuellen Verwendung inhaltsähnlicher Wörter (z.B. Flüchtling/Migrant/Ausländer)
  - Semantische Relationen visualisiert als Sankey Diagramm

- Definition Kontext
  - Kontextfenster auf Satzebene
  - Relationen zwischen gesuchten Wörtern und weiteren Tokens im Satz
  - Lemmatisierung und Filterung von Stoppwörtern optional

Benutzeroberfläche

- Optionen für den Benutzer
  - Wildcards im Suchbegriff
  - Lemmatisierter Input
  - Output
    → semantisch: Lemma
    → syntaktisch: Lemma / Type
  - Filterung von Stoppwörtern
  - Gross- bzw. Kleinschreibung ignorieren
  - Anzahl Kontextwörter
  - Sprachauswahl aktuell DE/EN/FR
    → Kann beliebig erweitert werden



Mögliche Anwendungen

- Ursprünglich geplant:
  - Semantische Einbettung

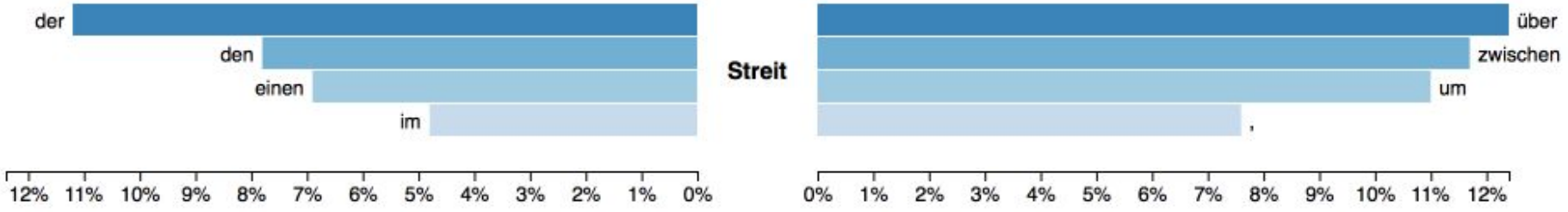
- Umgesetzt:
  - Syntaktische und Semantische Einbettung
    - Korpuslinguistik oder Hilfe für nicht Muttersprachler
    - Absolute und relative Häufigkeit einsehbar
    - Bedeutungsähnliche Wörter oder Antonyme
    - Syntaktische Information dank Neighbour-Chart
    - Kollokationen, Präpositionen
    - Untersuchung soziolinguistischer Wortverwendungen
      - Z.B. männliche und weibliche Formen desselben Substantivs

Frontend

- Ressourcen
  - Bootstrap (v3): Site-Design
  - JQuery (v3): Event-Handling & Backend-Verbindung
  - D3 (v4): SVG Grafiken
    - Sankey (vordefinierte API)
    - Neighbouring Chart (Eigenbau)
- Validität
  - HTML5 & SVG 1.1

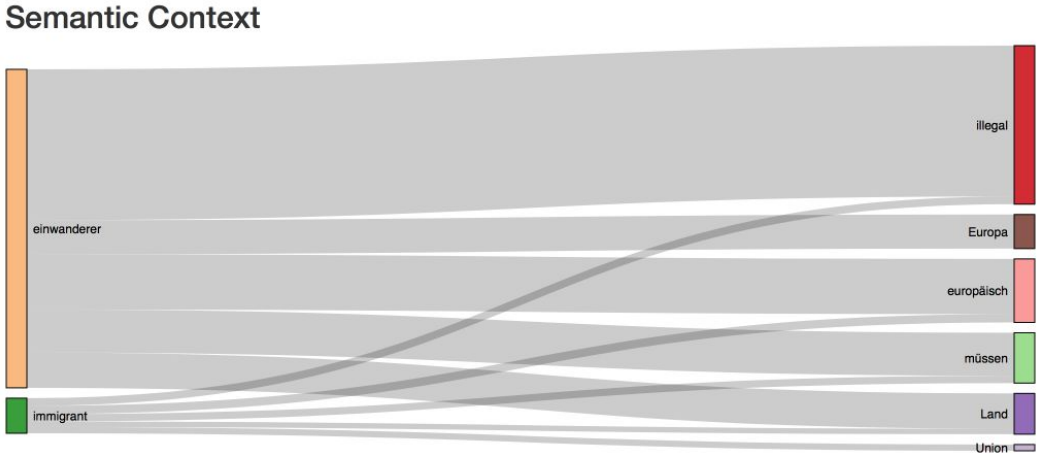
Syntactic Neighbours

- Direkte Nachbarn per Term (Grouped Chart)
  - vier häufigste Nachbar-Tokens links & rechts
  - relative Häufigkeit gegenüber Such-Term (geordnet)
  - zeigt häufige Kollokationen & Präpositionen
  - keine Filterung von Stoppwörtern & Interpunktion
  - Dynamische Höhe & Label-Abstände



Sankey Diagramm

- Distributionale Semantik
  - ‘You shall know a word by the company it keeps’ (Firth, J. R. 1957:11)
  - Distributionale Hypothese: Wörter, die häufig im selben Kontext auftauchen tendieren dazu ähnliche Bedeutungen zu haben
  - Darstellung der häufigsten Wörter die mit den Suchbegriffen im selben Satz erscheinen



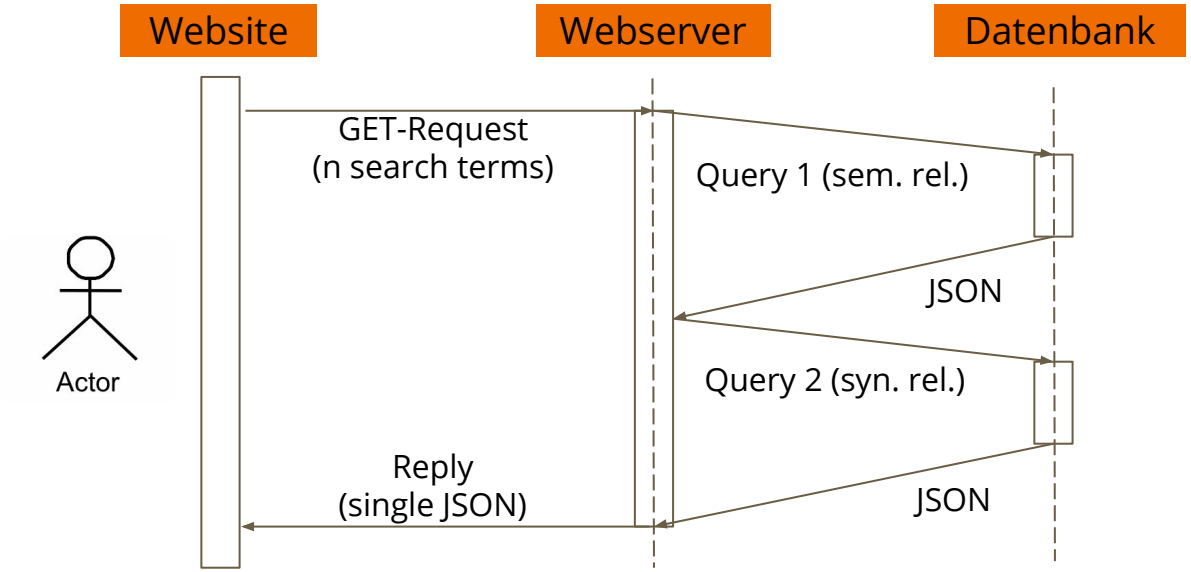
Javascript

```
1   // Sankey Diagram
2   function draw_sankey_diagram(data){
3
4       var dynamic_height = 50 * data.nodes.length;
5
6       var svg = d3.select(".chart_sankey"),
7           width = +svg.attr("width"),
8           height = dynamic_height;
9
10      svg.attr("height", dynamic_height)
11
```

HTML

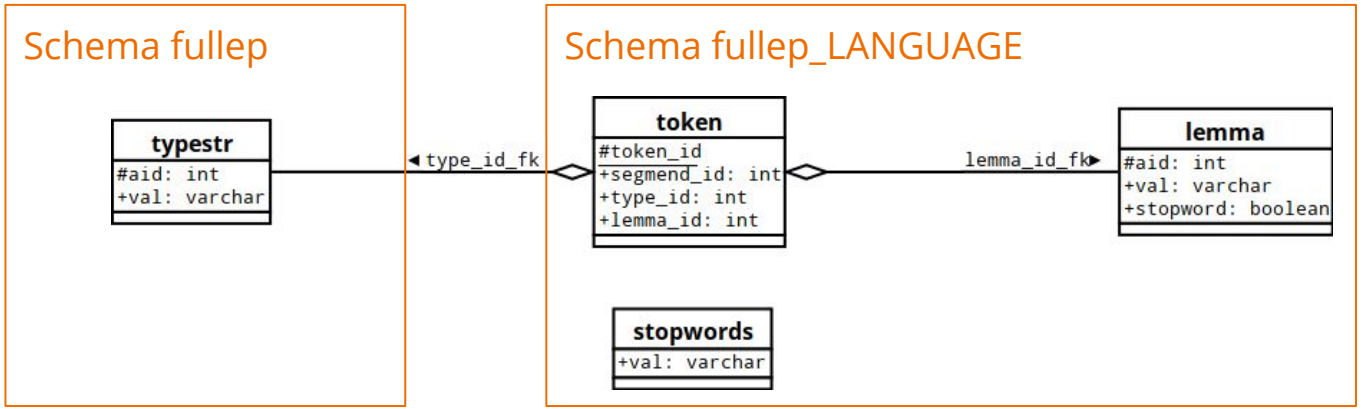
```
90              <div class="col-sm-12">
91                  <svg class="chart_sankey" width="1000"></svg>
92                  <script src="chart_sankey/sankey_code.js"></script>
93              </div>
```

Anfrage: Sequenzdiagramm



Projektdatenbank: Modell & Daten

- Sprachspezifische DB-Schema (DE, EN, FR)



- Daten: Fullep9-DB & Stoppwörter aus verschieden Quellen

Indizes & Queries

- B-tree indexierte Attribute
  - foreign keys
  - Lemma- & Type-Strings
  - funktionaler Index auf lower()
- Queries
  - Dynamisch zusammengesetzt & parametrisiert mit CASE WHEN
  - Extensive Nutzung von CTE, Window Functions & Subqueries
  - JSON-Objekt direkt in PostgreSQL
- Gute Performanz

Demonstration

- WoCo Explorer

Fragen?