Herausforderungen in der Nutzung vorhandener Tools für arabische Daten

Tillmann Feige und Alicia González

Vorgehen

1 Hintergründe & Workflow

2 Die Annotation

2.1 Anforderungen

3 Visualisierung

3.1 Anforderungen

4 Nachhaltigkeit

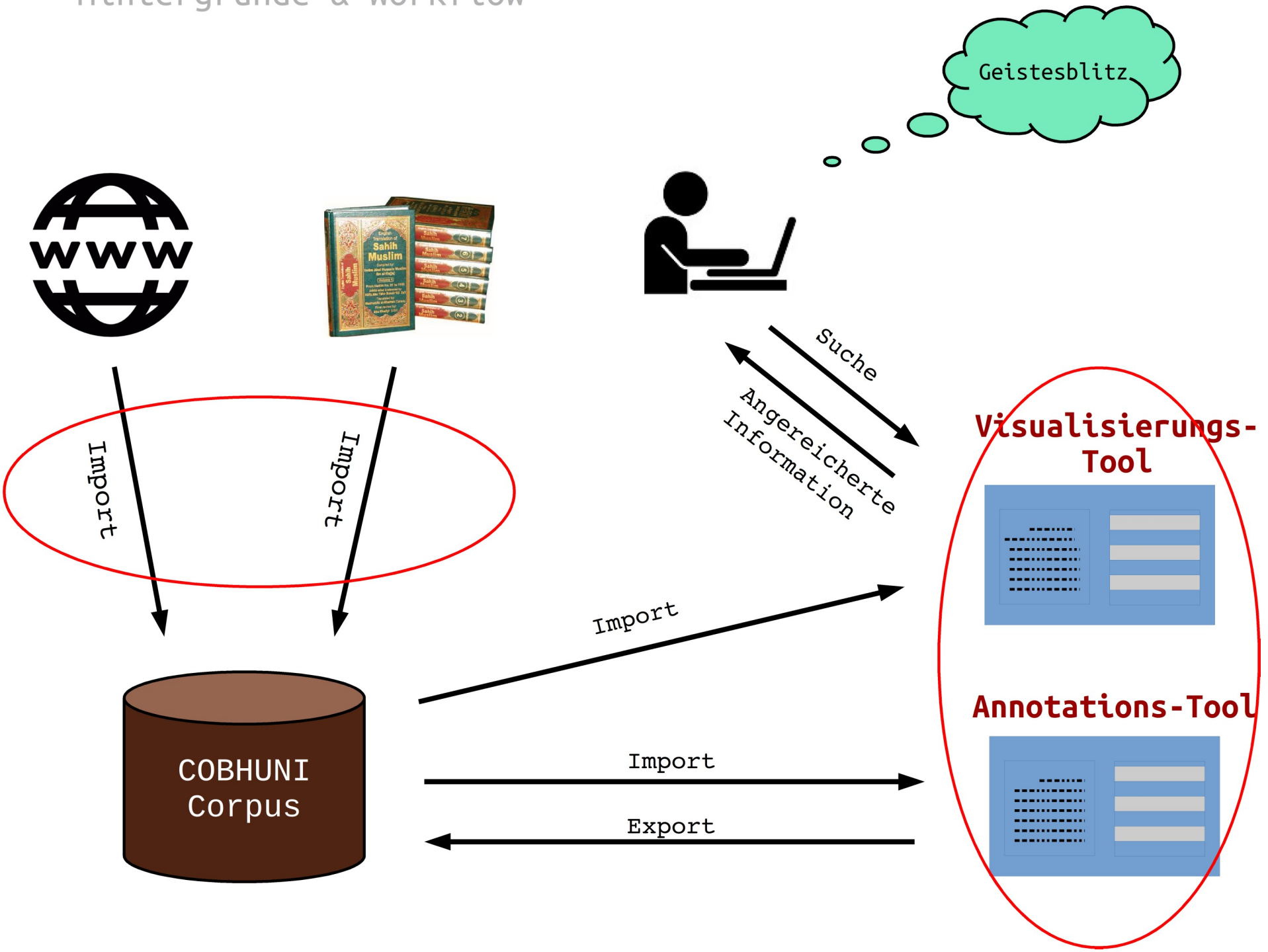
Das Setting wird durch das Projekt COBHUNI vorgegeben, in dem dem Vorstellungen des vorgeburtlichen Lebens in der islamischen Welt diachronisch untersucht werden.

ثُمَّ جَعَلْنَٰهُ نُطْفَةً فِى قَرَارٍ مَّكِينٍ ﴿١٣﴾
ثُمَّ خَلَقْنَا ٱلنُّطْفَةَ عَلَقَةً فَخَلَقْنَا
ٱلْعَلَقَةَ مُضْغَةً فَخَلَقْنَا ٱلْمُضْغَةَ
عِظَٰمًا فَكَسَوْنَا ٱلْعِظَٰمَ لَحْمًا ثُمَّ
أَنشَأْنَٰهُ خَلْقًا ءَاخَرَ فَتَبَارَكَ ٱللَّهُ أَحْسَنُ
ٱلْخَٰلِقِينَ ﴿١٤﴾

Unterstützung durch computerlinguistische Methoden:

- Arabische Texte müssen annotiert werden
  - Semantisch
  - Morphologisch (Lemma, POS)
- Suche in den verschiedenen Layern und Visualisierung der annotierten Texte

Hintergründe & Workflow



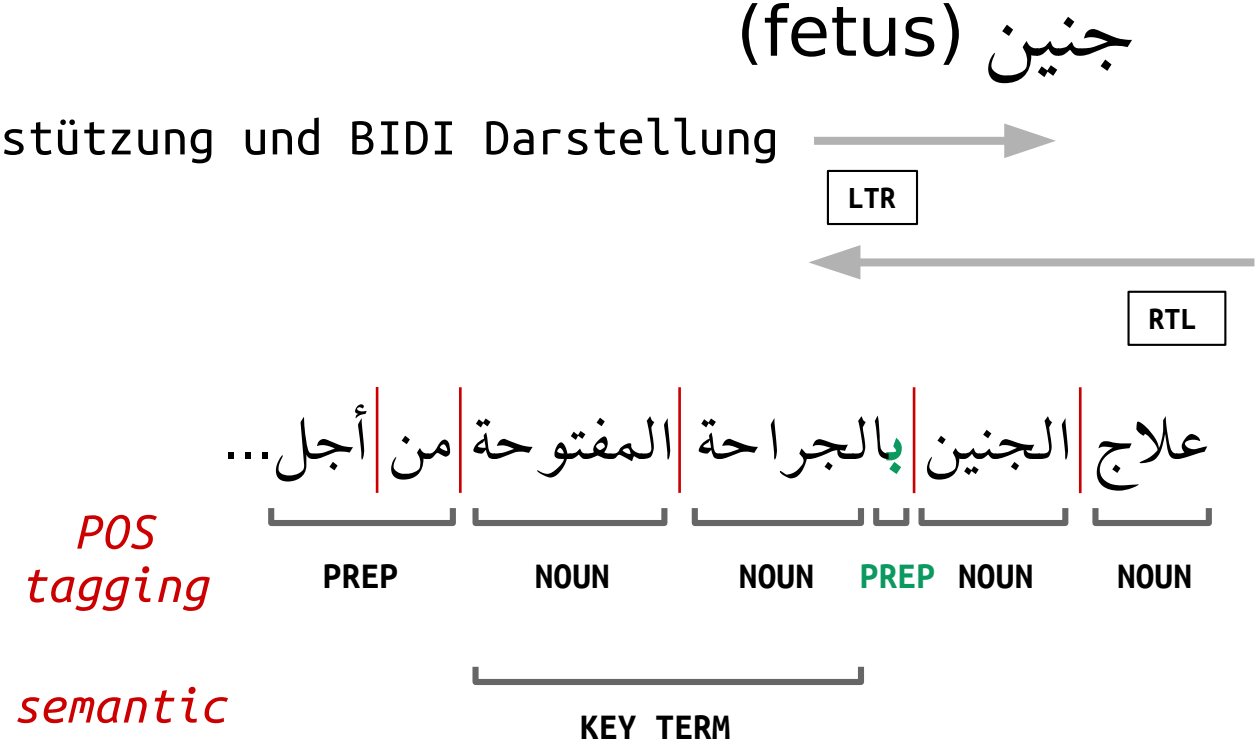
- Bisher:
  - Nur semantische und Named Entity-Annotation
  - Daher auch manuelle Annotation
- Geplant:
  - Simple Tagsets (flache Annotation) für POS und Lemma
  - Semi-automatische Annotation

Die Annotation: Das Semantische Tagset

- 1 MOTIVES
  - 1.1 Before the unborn
    - Semen and similarity / heredity
    - Semen as colors
    - Semen and coitus interruptus or contraceptives
    - Semen and wet dream
    - Sex act itself & its timing
  - 1.2 The unborn
    - Conception / fertilization
    - General / larger debate about predestination
    - Embryology: 40 days
    - Embryology: Ensoulment
    - Embryology: Angel visits Embryo
    - Embryology: expressed in a series of numbers
    - Embryology: Macrocosm – microcosm
    - Embryology: Embryo and link to resurrection & afterlife
    - Embryology: Link to (modern) science
    - Pregnancy: duration: Definition
  - 1.3 After the Unborn
    - Miscarriage / abortion and legal status of slave mother
    - Miscarriage / abortion and legal status of free mother
    - Miscarriage / abortion and legal status of the siqt
    - Abortion compared to killing a new-born
    - Menstruation
    - Breast-feeding
    - Legal status questions concerning the child after birth
- 2 METAMOTIVES
  - Philological exegesis
  - Hadith criticism
  - Latin script
- 3 NAMED ENTITIES → Eigenname

Technisch:

- Vollständige UTF-8 Unterstützung und BIDI Darstellung
- NLP Tools für Arabisch
- Konfigurierbare Tagsets
- Multi- und Subtoken
- Multilayer
- Overlaps
- Flexibler Im- und Export



Gewünscht:

- Browser-basiert
- Einfache Bedienbarkeit

Nicht festgelegt:

- Datenformat

Annotation: Die Anforderungen

Name	UTF8/ BIDI	Configurable tagset	Multi-token	Sub-token	Multi-Layer	Overlap	Flexible import	Browser	Good UX
Catma	(✓)	✓	✓	✓	✓	✓	✗	✓	(✓)
MAE	(✓)	✓	✓	✓	✓	✓	✗	✗	(✓)
WebAnno	(✓)	✓	✓	✓	✓	✓	✓	✓	✓
Atomic	(✓)	?	?	?	?	?	✓	✗	✗
GATE	(✓)	✓	?	?	?	?	(✓)?	(✓)?	✗

Annotation: Die Anforderungen

Name	UTF8/ BIDI	Configurable tagset	Multi-token	Sub-token	Multi-Layer	Overlap	Flexible import	Browser	Good UX
Catma	(✓)	✓	✓	✓	✓	✓	✗	✓	(✓)
MAE	(✓)	✓	✓	✓	✓	✓	✗	✗	(✓)
WebAnno	(✓)	✓	✓	✓	✓	✓	✓	✓	✓
Atomic	(✓)	?	?	?	?	?	✓	✗	✗
GATE	(✓)	✓	?	?	?	?	(✓)?	(✓)?	✗

Annotation: Die Anforderungen

Name	UTF8/ BIDI	Configurable tagset	Multi-token	Sub-token	Multi-Layer	Overlap	Flexible import	Browser	Good UX
Catma	(✓)	✓	✓	✓	✓	✓	✗	✓	(✓)
MAE	(✓)	✓	✓	✓	✓	✓	✗	✗	(✓)
WebAnno	✓	✓	✓	✓	✓	✓	✓	✓	✓
Atomic	(✓)	?	?	?	?	?	✓	✗	✗
GATE	(✓)	✓	?	?	?	?	(✓)?	(✓)?	✗

Visualisierung des Korpus:

- Ist Repräsentation des technischen Teils des Projekts
- Wird von Anwendern genutzt, später auch öffentlich verfügbar gemacht

Technisch:

- Vollständige UTF-8 Unterstützung und BIDI Darstellung
- Darstellung von:
  - Multi- und Subtoken
  - Multilayer
  - Overlaps
- Browser-basiert
  - Permalinks
- Suche in allen Layern und Metadaten

Gewünscht:

- Einfache Bedienbarkeit
- Statistische Analysemöglichkeiten

ANNIS

- **Erfüllt die Anforderungen**

Aspekte der Nachhaltigkeit bei COBHUNI:

- Framework
  - Apache UIMA (WebAnnno): (✓)
    - **Tokenizer (Stanford NLP): ✗**
- Multilingualität
  - Probleme bei Arabisch: Bi-Direktionalität
    - Nicht vollständig umgesetzt

→ Selbst bei Arabisch: Teil-Eigenentwicklung notwendig

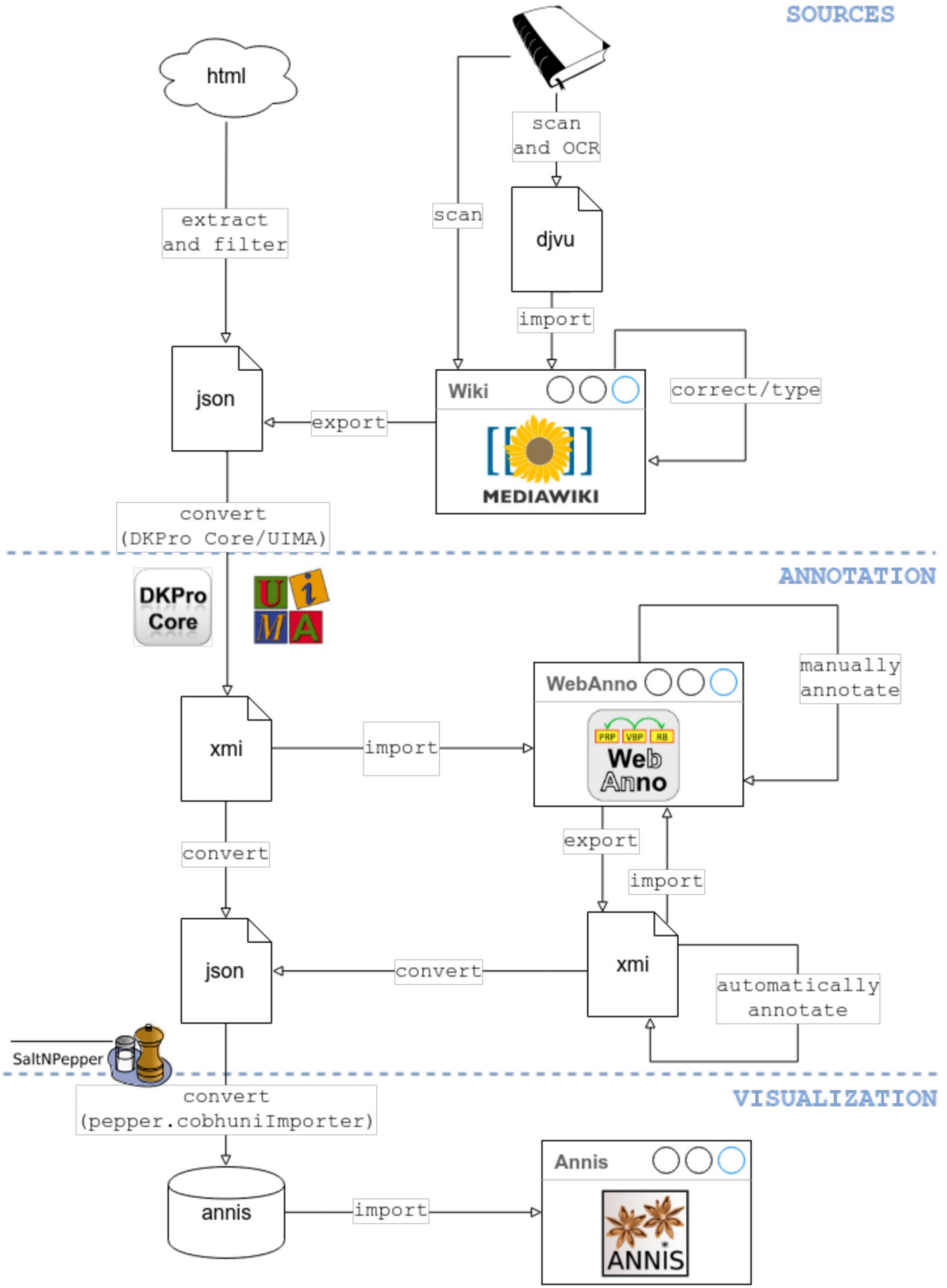
Aspekte der Nachhaltigkeit bei COBHUNI:

- Werkzeuge:
  - Darstellung als größtes Problem (BIDI)
  - Durch Anforderungen reduziert sich Auswahl erheblich
  - Aber für Arabisch und unseren Anwendungszweck gibt es Tools

Aspekte der Nachhaltigkeit bei COBHUNI:

- Daten:
  - Wir nutzen intern verschiedene Datenformate

Technical workflow



Aspekte der Nachhaltigkeit bei COBHUNI:

- Daten:
  - Kein TEI, da verschiedene Hierarchielevel
  - json mit stand-off Annotation ist simpler für interne Zwecke
  - Export in TEI ist vorerst nicht vorgesehen

Fazit:

- COBHUNI & Arabisch:
  - UD-POS Tags funktionieren
  - Arabic Stanford Parser funktioniert nicht für unsere Zwecke (Klassisches Arabisch)
  - Es gibt Tools, man benötigt aber Unterstützung der Entwickler
  - Wir umgehen komplette Eigenentwicklung, aber müssen teilweise nachbessern.

Danke!