

Textgrundlage

- Single-Sources & Standoff-Markup für Thesaurus Linguae Graecae
- bildet stabile und annotierbare Textgrundlage des Projekts
- 1823 Autoren, 7209 Werke, ca. 75 Mio. Wörter, ca. 1 Mio. Wortformen

Begriffsbildung

- theoretische Ausdifferenzierung des Paraphrasenbegriffs
- operationalisierbar und mit technischen Ansätzen kompatibel
- Kategorisierungssystem zur Charakterisierung der Einzelreferenz

Referenzierungssystem

- Adaption von URNs der Canonical Text Services (CTS)
- zur wortgenauen Referenzierung des Textkorpus
- urn:cts:ppd:tlg0059.tlg030:000:327_a.1-327_a.3@[8]

Goldstandard und Referenzannotierer

- Sammlung bekannter Referenzen auf Platon
- zur Evaluation der Algorithmen und zur Begriffsbildung
- 216 mit graphischem Werkzeug detailliert erfasste Referenzen

Linguistische Annotation

- Kombination von Morpheus, Mate-Tools und Sonderregeln
- als wichtige Datenquelle für automatische Verfahren
- eindeutige Lemmatisierung mit einer Genauigkeit von über 90%

Helleninet

- Altgriechisches Wortnetz auf Basis vielfältiger digitaler Ressourcen
- zur maschinellen Identifikation semantischer Wortrelationen
- Netz mit derzeit 15.145 automatisiert generierten Relationen

Worteinbettung

- Einbettung des Korpus in einen Vektorraum mittels word2vec
- zur maschinellen Erkennung bedeutungsähnlicher Wörter
- umfangreiche Optimierung mit erarbeitetem Fragenkatalog

Semi-automatische Paraphrasensuche

- Kombination verschiedener Ansätze:
 - auf Word Mover's Distance basierender Ansatz
 - Suche unter Verwendung komplexer n-Gramme
 - explorative Suche mittels „Rütteln“
- zur Erweiterung des Thesaurus zur Rezeption Platons
- qualitative & statistische Evaluationen belegen hohe Effektivität

<https://digital-plato.org>