Beschreibung der *Use Cases*

(R 5.2.1)

Version 13.02.2015
Cluster 5
Verantwortlicher Partner Universität Würzburg

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:

Anna Aurast, IEG
Frederik Baumgardt, INFAI
Marcus Held, IEG
Stefan Pernes, UWÜ
Steffen Pielström, UWÜ
Michael Piotrowski, IEG
Christof Schöch, UWÜ

Revisionsverlauf:

Datum	**Autor**	**Kommentare**
15.09.2014	alle	Erste Version
13.02.2015	Pielström	Überarbeitung des Layouts

Inhaltsverzeichnis:

1. Use Case: Narrative Techniken und Untergattungen im deutschen Roman

Beteiligte Partner: UWÜ und TUD

In diesem *Use Case* soll exemplarisch demonstriert werden, wie eine große Sammlung literarischer Texte genutzt werden kann, um mit Hilfe quantitativer Verfahren die historische Entwicklung narrativer Techniken - und in weiterer Folge auch die Entwicklung darauf aufbauender literarischer Kategorien - zu analysieren. Dabei ist es das Ziel, ein Set von *Best Practices*, Beispiel-*Workflows* und allgemeinverständlichen Tutorials zu erstellen, die es TextwissenschaftlerInnen ermöglichen, innovative, bereits vorhandene Werkzeuge flexibel auf ihre eigenen Daten anzuwenden.

1.1. Textgrundlage

Ausgangspunkt für den *Use Case* ist eine Textsammlung, die sich aus unterschiedlichen Quellen speist: Als erstes Korpus ist hier eine Teilmenge der Digitalen Bibliothek von *TextGrid*[1] zu nennen. Sie umfasst 400 deutschsprachige Romane aus dem Zeitraum von 1500 bis 1930, die im Volltext nach *TEI* kodiert vorliegen und mit einigen Metadaten versehen sind. Als Ergänzung zu diesem mehrfach aufbereiteten Korpus, wird eine Menge von 1.600 ebenfalls deutschsprachigen Romanen aus der *Gutenberg* Sammlung mit eingebunden, die jedoch nur im unkodierten Volltext vorliegen und eine schwankende Qualität aufweisen. Um die Übertragbarkeit und Robustheit der Lösungen weiter zu prüfen, werden die Werkzeuge auch auf eine *TEI*-kodierte Sammlung von 200 französischen Kriminalromanen des 19. und 20. Jahrhunderts angewandt.

1.2. Technische Grundlage

Mit dem auf *Apache UIMA* basierenden *DKPro* (*Darmstadt Knowledge Processing Software Repository*) stellt die TU Darmstadt eine bewährte und umfangreiche Sammlung leistungsfähiger Textverarbeitungs-, Annotations- und Analysewerkzeuge zur Verfügung, die sich insbesondere durch ihren modularen Aufbau und ihre Skalierbarkeit auszeichnet.
Um Anschlussfähigkeit an die unter TextwissenschaftlerInnen weit verbreitete Programmiersprache *Python* zu gewährleisten, wird eine Schnittstelle zu der auf *Java* basierenden *UIMA* Plattform eingerichtet. Dadurch soll es möglich werden, Ergebnisse aus *DKPro* direkt in *Python* weiterzuverarbeiten und gegebenenfalls auf die Zwischenergebnisse einzelner *DKPro*-Komponenten zuzugreifen.

1 http://www.textgrid.de/Digitale-Bibliothek

1.3. Rezepte für Analyseschritte

Das Desiderat des *Use Cases* ist es, eine Sammlung von *Best Practices* und Beispiel-*Workflows* zur Verfügung zu stellen. Anhand dieser "Rezepte" soll ein möglichst zugänglicher Einstiegspunkt geschaffen werden, um textwissenschaftliche Fragestellungen mit Hilfe computergestützter Methoden zu bearbeiten. Da es sich dabei praktisch immer um einen Ablauf handelt, der aus mehreren Arbeitsschritten besteht, sollen die einzelnen Teilprobleme im Folgenden kurz beschrieben werden.

Part-of-Speech Tagging: Ein wichtiger Verarbeitungsschritt, der auf die folgenden, komplexeren Verfahren vorbereitet, ist das *POS-Tagging*, d.h. das automatisierte Erkennen der grammatikalischen Funktion eines Wortes im Satzgefüge. Für das Deutsche gilt der *TreeTagger* gemeinhin als das leistungsfähigste Software-Paket, es werden jedoch auch alternative Rezepte getestet, die etwa mit dem *Stanford Parser*, oder dem speziell für Sprachen mit komplexer Morphologie entwickelten *hunpos* arbeiten. Die Übertragung des Verfahrens auf das Französische verspricht hier weitere Erfahrungswerte bei der Selektion und Anwendung unterschiedlicher Parser. Die beiden erwähnten Alternativen, *Stanford Parser* und *hunpos*, sind bereits im *DKPro Framework* verfügbar.

Named Entity Recognition: Ein weiterer vorbereitender Arbeitsschritt ist die automatische Erkennung von Namen, Orten und anderen Entitäten. Auch hier gibt es unterschiedliche, teilweise komplementäre Zugänge: Einerseits kann *NER* ungesteuert und ohne Rückgriff auf vorher definiertes Wissen durchgeführt werden, andererseits ist es auch möglich, mit *UIMA Ruta* einen regelbasierten Ansatz - ähnlich einer Mustererkennung - zu verfolgen.
Um die Güte der gewonnenen Ergebnisse zu überprüfen, kann sowohl auf ein bereits bestehendes, manuell annotiertes Korpus zurückgegriffen, als auch anhand von *WebAnno* kollaborativ ein "Gold Standard" erzeugt werden. Im Rahmen der *NER* soll auch überprüft werden, ob und inwieweit sich bestehende Modelle, die typischerweise auf Nachrichtentexten basieren, auch für literarische Texte eignen.

Sentiment Analysis: Gegenstand dieses Teilproblems ist es, die Polarität von Aussagen und in weiterer Folge, die Einstellungen zu bestimmten Themen festzustellen. Dabei kann die Sentiment Analysis Aufschluss geben über Bewertungen und Beurteilungen, die vermeindliche affektive Verfasstheit des Autors bzw. des Sprechers, sowie mögliche Kommunikationsabsichten. Auf literarische Texte angewandt, soll es damit möglich werden, auch Einstellungen des Erzählers zu den Figuren, sowie der Figuren untereinander zu bestimmen. Verfahren, die dabei zur Verfügung stehen, sind im wesentlichen die Klassifikation anhand von *Machine Learning*-Methoden und die Mustererkennung anhand von emotional besetzten Schlüsselwörtern.

Erkennung direkter Rede: Sowohl der Anteil von Dialogen an einem Text, als auch die stilistischen Nuancen einzelner sprechender Protagonisten können einen interessanten Aspekt im Erzählstil eines Autors darstellen. Darum ist die automatische Unterscheidung von direkter- und indirekter Rede in literarischen Texten ein wichtiger Schritt hin zur angestrebten quantitativen Beschreibung narrativer Techniken. Die Zuordnung von Textabschnitten direkter Rede zu einem bestimmten Protagonisten eröffnet weitere Möglichkeiten. Angestrebt ist darum ein Rezept zum automatischen Eintragen von Metadaten hinsichtlich direkter/indirekter Rede, und der Zuordnung derselben zu den Figuren und zum Erzähler. Um diese Aufgabe zu bewerkstelligen

sollen vorbereitende Verarbeitungsschritte in *DKPro* mit einer anschließenden regelbasierten Verarbeitung mit *UIMA Ruta* kombiniert werden.

Erkennung deskriptiver vs. narrativer Passagen: Die Unterscheidung verschiedener Texttypen (u.a. deskriptiv, narrativ, argumentativ) ist eine grundlegende Fähigkeit kompetenter Leser literarischer und anderer Texte. Deskriptive Passagen beschreiben oder charakterisieren Figuren und Objekte in der fiktionalen Welt; narrative Passagen erzählen Ereignisse und Handlungen in ihrem zeitlichen und kausalen Verlauf; argumentative Passagen schließlich beinhalten logisch verknüpfte Ideen und Tatsachen. Die Anteile solcher Passagen an einem
literarischen Text stehen in einem engen Zusammenhang mit der Gattungszugehörigkeit eines Textes und wirken sich auf seine stilistischen Merkmale aus. Daher ist es von grundsätzlichem Interesse, diese Unterscheidung algorithmisch zu modellieren und in Texten automatisch erkennen zu können. Grundlage für die Erkennung dieser Texttypen ist POS-Tagging (siehe oben), die konkreten Ansätze können dann ähnlich wie im Falle der Redewiedergabe sowohl aus dem Bereich des *Machine Learning* (SVM) als auch der regelbasierten Techniken sein.

Plotähnlichkeiten erkennen: Aufbauend auf genauen syntaktischen Informationen (POS-Tags sowie spezifisches *Parsing* nach Regeln der Dependenzgrammatik) ist es möglich, semantische Inhalte automatisch zu rekonstruieren. Wie auch bei der verwandten Aufgabe der Koreferenz-Resolution kann dabei im Wesentlichen ausgegangen werden von Informationen, die in Nominalphrasen (*entity-centric*) oder in Verbalphrasen (*event-centric*) enthalten sind. Ergebnis solcher Verfahren sind Datenstrukturen, die die Art der Ereignisse und die Rollen ihrer Teilnehmer enthalten (auch: *Frame* oder *Event Schema*). Solche ungesteuerten Ansätze der semantischen Erschließung von Textmaterial sind vielversprechend - insbesondere in ihrer Reichweite - sie sind jedoch auch limitiert in ihrer Genauigkeit. Verglichen mit einem Gold Standard bleiben einige Frames ambivalent, darüber hinaus ist es nicht einfach, solche Strukturen - auch über die Satzgrenze hinaus - korrekt zu desambiguieren. Deswegen ist es hier sinnvoll, auf Weltwissen zurückzugreifen und andere Verfahren in Kombination einzusetzen. So wären auch hier *Machine Learning* Ansätze (Klassifikation) oder regelbasierte Techniken denkbar.

Clustering in Gattungen: Es handelt sich dabei um ein Desiderat, das sich aus der Modellierung und Untersuchung der zuvor behandelten narrativen Techniken speist. Der Gattungsbegriff steht in diesem *Use Case* stellvertretend für theoriegeleitete literarische Kategorien, die anhand quantitativer Methoden nachvollzogen werden können und dadurch möglicherweise in einem neuen, empirischen, Licht erscheinen. Für die Modellierung von Untergattungen werden im Rückgriff auf die vorangegangenen Arbeitsschritte in einem iterativen Prozess *Feature*-Vektoren aufgebaut, die es erlauben, Rückschlüsse darüber zu ziehen und zu beobachten, in welcher Form sich derartige Merkmale hin zu Strukturen auf einer gattungsstilistischen Meso-Ebene aggregieren.

1.4. Ziele und Dissemination

Zu den hier beschriebenen Arbeitsschritten werden Rezepte entwickelt, die sowohl als vorläufige Versionen im Rahmen der *SVN*-integrierten Wiki-Dokumentation einsehbar sind, als auch im Zusammenhang mit Milestones in kompletten Paketen dis-

tribuiert werden sollen. Der Umgang mit diesen Rezepten wird in verständlichen Tutorials aufbereitet, mit dem Ziel, LiteraturwissenschaftlerInnen zu ermächtigen anhand von quantitativen Methoden ihr eigenes Material zu bearbeiten und ihre eigenen Fragestellungen zu beantworten. Dazu gehört auch die Vermittlung allgemeiner technischer Grundlagen, Informationen über das darunterliegende Framework (*UIMA*), sowie zur Datenaufbereitung und dem Trainieren von *NLP* Werkzeugen hinsichtlich spezifischer Fragestellungen. Die Rezepte, die zugehörige Dokumentation, sowie die noch stärker didaktisch aufbereiteten Lehr- und Lernmittel werden schließlich auf den DARIAH-DE und DARIAH-EU Portalen der Öffentlichkeit zur Verfügung gestellt.

2. Use Case: Bibliografien

Beteiligte Partner: IEG und Minf-BA

Der Use Case 2 „Biografien“ behandelt die automatische Entdeckung von Korrelationen zwischen Personen, Orten, Daten und Ereignissen und ist inhaltlich eng verbunden mit dem am IEG angesiedelten geschichtswissenschaftlichen Forschungsprojekt *Cosmobilities – Grenzüberschreitende Lebensläufe in den europäischen Nationalbiografien des 19. Jahrhunderts*[2]. In dieser seit dem 1. Mai 2014 von der Fritz-Thyssen-Stiftung geförderten Machbarkeitsstudie soll überprüft werden, inwiefern die Analyse grenzüberschreitender Lebensläufe in den europäischen Nationalbiografien unter Einbeziehung digitaler Ressourcen sowie weiteren biografischen Materials der historischen Forschung für das 19. Jahrhundert neue Impulse verleihen kann. Die gemeinsame Projektleitung liegt bei Prof. Dr. Johannes Paulmann (IEG) und Prof. Dr. Margit Szöllösi-Janze (LMU); die Machbarkeitsstudie wird von Sarah Panter am IEG, durchgeführt. Im Gegensatz zu vielen digitalen Projekten in der Geschichtswissenschaft, die sich vor allem auf die Datenerschließung fokussieren, kehrt die Studie die Herangehensweise jedoch um, denn es soll geprüft werden, ob sich aus der spezifischen Perspektive der *Cosmobilities* ein Frageraster sowie Kategorien für die Analyse serieller, digital verfügbarer Massendaten entwickeln lassen. Diese Vorgehensweise stellt damit ein hervorragendes Beispielszenario für die Anwendung von *Big Data*-Verfahren in einer Geisteswissenschaftlichen Fragestellung dar, und erlaubt zugleich, die Aktivitäten von Cluster 5 in die laufende Forschungstätigkeit am IEG einzubetten.

2.1. Ziele

Der Use Case erforscht, ob durch die Analyse der strukturierten Daten und der unstrukturierten Texte Verbindungen und Verknüpfungen zwischen individuellen historischen Lebensläufen und Internationalitätskriterien hergestellt werden können, die bislang für die geschichtswissenschaftliche Forschung noch im Verborgenen liegen. Das erste Ziel des Use Case ist also, die für das Projekt Cosmobilities relevanten Personengruppen in der Datenbasis zu identifizieren. Dies geschieht zunächst, indem die Korrelation verschiedener Merkmale zu ‚Mobilität‘ untersucht wird, etwa Geburts-, Wirkungs- und Sterbeorte, Berufe und Tätigkeiten, verwandtschaftliche Beziehungen sowie zeitliche Relationen. Dazu werden die biografischen Datenquellen in die Generische Suche integriert und die Suchmöglichkeiten so erweitert, dass Abfragen nach biografisch relevanten Aspekten möglich sind.

2 http://www.ieg-mainz.de/Forschungsprojekte------_site.site..ls_dir._nav.17_f.69_likecms.html

Das zu untersuchende Datenmaterial stellt durch seine Kombination von strukturierten und unstrukturierten Informationen ein interessantes Forschungsobjekt für die Entwicklung maschineller Auswertungsverfahren dar, das die klassische Data-Mining-Methoden und computerlinguistische Methoden zusammen bringt. Darüber hinaus können die Daten als Grundlage für die Untersuchung und Visualisierung historischer sozialer Netzwerke – auch in räumlicher und zeitlicher Dimension – dienen. Der Use Case wird außerdem quantitative Grundlagen für die Konzeption und Anwendung von kontrollierten Vokabularen für historische Orte, Personen und Ereignisse liefern (die Arbeiten zu Vokabularen wurden in Phase I begonnen und werden in Cluster 4, AP 4.3 fortgeführt).
Ein weiteres Ziel des Use Case ist es, mit Blick auf eine Verbesserung der digitalen Aufbereitung und Darstellung historisch-biografischer Informationen zu untersuchen, welche inhaltlichen und formalen Kategorien für die Weiterentwicklung auf dem Weg zum semantischen Web im Bereich der biografischen Forschung tatsächlich wissenschaftlich erforderlich und nutzbringend für historische Fragestellungen sind.

2.2. Datengrundlage

Als Datenbasis für den Use Case werden zunächst die biografischen Artikel in der deutsch- und englischsprachigen Wikipedia und in Wikidata herangezogen. Im weiteren Verlauf ist geplant, die Datenbasis um die Allgemeine Deutsche Biografie (ADB)[3] und evtl. weitere europäische Nationalbiografien (etwa das Oxford Dictionary of National Biography[4]) oder andere relevante Quellen zu ergänzen. Die ADB etwa enthält Biografien von ca. 26.500 Personen, die im deutschsprachigen Raum gewirkt haben und vor dem Jahr 1900 verstorben sind. Neben den wesentlichen Angaben zu Namen, Beruf, Geburt und Tod beinhalten die einzelnen Biogramme eine narrative Beschreibung des Lebenslaufs. Die ADB liegt als elektronischer Volltext vor. In der elektronischen Version ist jeder Artikel zusätzlich mit einer GND-Nummer versehen, durch die die beschriebene Person eindeutig identifiziert werden kann, sowie mit Referenzen zu weiteren Ressourcen. Die Daten aus den Nationalbiografien könnten somit als Korrektiv zu den im Web frei verfügbaren und strukturierten Daten aus Wikidata dienen, durch das fehlende Informationen oder vorhandene Schnittmengen identifiziert werden können.

2.3. Technische Grundlage

Die technische Grundlage des Use Case bildet zunächst die Generische Suche (http://dev3.dariah.eu/search/). Daraus hervorgegangen und für die besonderen Anforderungen des Use Case entwickelt wurde das CosmoSearch-Tool (http://dev3.dariah.eu/cosmotool/cosmosearch), das sich bereits in der Testphase befindet.

3 http://www.deutsche-biographie.de/
4 http://www.oxforddnb.com/

Mit dem CosmoSearch-Tool wird das Ziel verfolgt, eine übergreifende Suchmöglichkeit zu schaffen, welche die Eigenschaften einer Breiten- und Tiefensuche so vereint, dass eine dynamische Anpassung der Suche im Hinblick auf die Filterung und Facettierung von Suchanfragen sowie auch die Visualisierung von Suchergebnissen möglich wird. Die übergreifende Suche in eng assoziierten Datenquellen erlaubt eine detaillierte Auseinandersetzung mit den betrachteten Daten (Tiefensuche). Die Granularität der Betrachtung und Facettierung nimmt mit einer wachsenden Zahl von einbezogenen Datensätzen ggf. mangels reichhaltiger Verbindungen ab und nimmt schließlich die Form einer Breitensuche ein. Die Funktionalität der generischen Suche kann anhand der beiden Phasen der Anfrageverarbeitung sowie der Datenanalyse und -indexierung eingeordnet werden, welche grundsätzlich unabhängig voneinander operieren. Der Index wird hierbei als wesentliche Schnittstelle der Datenanalyse und -indexierung zugeordnet, da diese für die Erzeugung und Aktualisierung der Einträge in den Indices verantwortlich ist. Die Anfrageverarbeitung hingegen wertet den Index in Abhängigkeit von formulierten Anfragen, den für die konkrete Anfrage ausgewählten Datensätze zur Verfügung gestellten Informationen aus und liefert die Ergebnisse zurück an den Suchenden.
Für die Beantwortung von Suchanfragen nutzt CosmoSearch zunächst Daten aus Wikidata und Wikipedia. In einem ersten Schritt wurden bereits alle Berufsgruppen indexiert, um die Bestimmung von relevanten Personengruppen für die geschichtswissenschaftliche Fragestellung zu unterstützen.

2.4. Vorgehensweise

1. Arbeitsschritt: Durchsuchung von Wikipedia und Wikidata nach ausgewählten Berufsgruppen aus den vier übergeordneten Bereichen Wirtschaft, Politik, Wissenschaft und Kunst und einer Kontrollgruppe, die die Berufsgruppen Missionare, Diplomaten, Kolonialbeamte und Forschungsreisende umfasst. Bei der Kontrollgruppe wurde vorausgesetzt, dass die Lebensläufe ihrer Vertreter qua Karrierewege einen hohen Internationalitätsgrad aufweisen. Als zeitliche Einschränkung wurden die Geburtsdaten der zu untersuchenden Personen für den Zeitraum zwischen 1730 und 1900 festgelegt. Damit wird das ‚lange' 19. Jahrhundert in vollem Umfang berücksichtigt.
2. Arbeitsschritt: Feststellung des Internationalitätsgehalts der vier Gruppen und der Kontrollgruppe durch eine jeweils vertikale und horizontale Korrelation (d. h., einmal für jede übergeordnete Gruppe und einmal für das komplette Korpus) der Informationen bzw. Daten aus der deutsch- und englischsprachigen Wikipedia und Wikidata mit weiteren Suchkriterien wie Kosmopolitismus, Weltläufer, Internationalismus, Migration, Exil, Reise etc.
3. Arbeitsschritt: Nachdem die für das Projekt infrage kommenden übergeordneten Untersuchungsgruppen sowie deren Internationalitätsgehalt ermittelt wurden, sollen diese in einem weiteren Schritt miteinander in Verbindung gesetzt werden. Hierzu soll das ausgewählte Korpus mit Merkmalen wie Orte, Religion, Ethnizität, Geschlecht, Stand, Ausbildung etc. korreliert werden.
4. Arbeitsschritt: Die in den Arbeitsschritten 2 und 3 ermittelten Internationalitätsmerkmale und Internationalitätsindikatoren sollen in einem nächsten Schritt mit historischen Konjunkturen in Korrelation gebracht werden. Als mögliche historische Konjunkturen werden hierbei Ereignisse wie die Französische Revolution, der Wiener Kongress von 1815, die Revolution von 1848 u. Ä. berücksichtigt.

Die qualitative Auswertung der Ergebnisse aller Arbeitsschritte erfolgt am IEG.

Sind die Ergebnisse des ersten Suchlaufs ausgewertet, wird ein zweiter Suchlauf gestartet, nachdem die Datenbasis mit weiteren Daten aus ADB und ggf. anderen europäischen Nationalbiografien erweitert wurde.

3. Use Case: Identifikation von griechischem und lateinischem Text in einer Sammlung von 2 Millionen Texten

Beteiligte Partner: UWÜ und TUD

3.1. Vorraussetzungen, Methodik & Zielsetzung

Im dritten Use Case untersuchen wir das Werk des französischen Priesters und Herausgebers Jacques Paul Migne, der Mitte des 19. Jahrhunderts Editionen der **Texte frühchristlicher Kirchenväter** veröffentlichte. Die Dokumente entstammen dem Zeitraum vom Ursprung des Christentums bis zum Fall des Byzantinischen Reiches im 15. Jahrhundert und sind bekanntermaßen reichhaltig an Zitationen aus biblischen und antiken Texten, sowie Zitationen innerhalb des Korpus. Dieses ist geteilt in einen griechischen und einen lateinischen Teil, in Widerspiegelung der historischen Entwicklung des Christentums. Basierend auf den umfangreichen Forschungen der Patrologie (Studie der Kirchenväter) lassen sich viele der Texte gesichert in Raum und Zeit anordnen, und größtenteils sind die Autoren und deren Biographien bekannt. Daraus leiten sich viele potentielle Anwendungen auf ein digitalisiertes und annotiertes Migne Korpus ab.
Die Digitalisierung der Texte der Patrologia Latina und Patrologia Gracae findet im Rahmen des **OpenMigne** Projekts am Humboldt-Lehrstuhl für Digital Humanities der Universität Leipzig statt. Basierend auf dieser Arbeit soll der Use Case "Identifikation von griechischem und lateinischem Text in einer Sammlung von 2 Millionen Texten" eine erste Erschließung der intertextuellen Struktur des Korpus bieten. Hierzu werden einführend mit bekannten Verfahren Zitationsgraphen des Kernteils (Patrologia Latina) der zugrunde liegenden Dokumentensammlung berechnet und relativ zueinander evaluiert. In weiteren Schritten wird das Verfahren um anspruchsvollere Teilkorpora und neu zu entwickelnde Methoden ergänzt.
Ziel des Use Cases ist es, ausgehend von Zitationen in direkter Rede auf gleichsprachigen Urpsrungstexten, über Paraphrasen und wörtlichen Übersetzungen, zu Allusionen und Zitationen über sprachliche Grenzen hinweg, die Methodik schrittweise zu entwickeln und somit durch die **Erschliessung möglichst korrekter und vollständiger diachroner Zitationsspuren** weitere Forschung bezüglich der Evolution von Sprache und Kultur im christlich geprägten Kulturraum zu ermöglichen. Desweiteren ist es ein Anliegen, die entwickelten Verfahren für eine Anwendung über den Use Case hinaus aufzubereiten.

3.2. Glossar

Zitation: Zitationen sind Referenzierungen von Textstellen im gleichen oder anderen Dokument. Der Bezug zum Ursprungstext kann sich mit variierender Direktheit und Kenntlichkeit manifestieren, von verbatim Wiedergabe bis zu vagen Andeutungen. Darüber hinaus kann mit geringer werdender Direktheit der Zitation auch der Umriss des Ursprungtextes an Schärfe verlieren, z.B. bei der Zitation von Diskursen.

Zitationsgraph, Zitationsspuren: Durch die Vernetzung der Texte mittels Zitationen spannt sich ein Netzwerk von Referenzen durch den gesamten Korpus, das wir einen Zitationsgraphen nennen. Einzelne, chronologisch gerichtete Kantenverläufe innerhalb des Zitationsgraphen werden Zitationsspuren genannt.

Korpus, Teilkorpora: Zeitlich, sprachlich und strukturell bilden sich logische Einheiten in der betrachteten Textmenge. OpenMigne besteht aus einem griechischen und einem lateinischen Teil, die Bibel ist in mehreren Urpsrungssprachen und diversen Übersetzung referenziert, das antike Perseus Korpus beinhaltet lateinische und griechische Texte verschiedener Epochen.