Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit Nr. 118

Visuelle Analyse von Präpositionen in deutschen Texten

Melanie Zaiß

Studiengang:	Informatik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dr. Steffen Koch Dipl.-Phys. Qi Han
Beginn am:	7. April 2014
Beendet am:	7. Oktober 2014
CR-Nummer:	H.5.2, I.2.7

Kurzfassung

Um Systeme zu entwickeln, die Sprache verstehen und produzieren können, wie beispielsweise die maschinelle Übersetzung von Texten, beschäftigt sich die Computerlinguistik mit der Feststellung sprachlicher Gesetzesmäßigkeiten. Allerdings sind eine ganze Reihe von linguistischen Zusammenhängen noch nicht erschöpfend erforscht. Das Verstehen der unterschiedlichen Verwendung von Präpositionen in der deutschen Sprache ist ein Beispiel hierfür. Ziel dieser Arbeit ist es daher, einen interaktiven Analyseansatz für die Exploration von Präpositionen zu entwickeln.

Abstract

In order to develop systems that understand and produce language, such as the machine translation of texts, the computational linguistics are dealing with the determination of linguistic regularities. However, a number of linguistic correlations are not explored exhaustively. Understanding the different use of prepositions in the German language is an example. The aim of this thesis is therefore to develop an interactive analysis approach for the exploration of prepositions.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Listings

1. Einleitung

1.1. Motivation

Die natürliche Sprache ist das wichtigste Medium der Menschen zur Kommunikation und Informationsübergabe. Daher ist die Analyse der Sprache vor allem in der heutigen Informationsgesellschaft von großer Bedeutung. Die Computerlinguistik untersucht die Sprache aus einem besonderen Blickwinkel. Ihr geht es darum, sprachliche Gesetzmäßigkeiten explizit feststellen zu können, um auf dieser Basis Systeme zu entwickeln, die Sprache verstehen und produzieren können [LNG]. Der wissenschaftliche und technische Fortschritt der letzten Jahre in diesem Gebiet hat sich auch in der computergestützten Verarbeitung natürlichsprachlicher Texte gezeigt. Diese computergestützte Sprachverarbeitung findet unter anderem bei der maschinellen Übersetzung von Texten Anwendung. Um die maschinelle Textübersetzung beispielsweise verbessern zu können, müssen bestimmte Wortgruppen noch weitergehend erforscht werden. Die Computerlinguistik trifft in der deutschen Sprache oft auf ein Ambiguitätsproblem. Ein Beispiel hierfür ist die Wortgruppe der Präpositionen, denn Präpositionen sind oftmals mehrdeutig. Die Präposition *aus* wird beispielsweise verwendet um sowohl lokale, kausale, als auch modale Zusammenhänge zu beschreiben, wie bei den folgenden Beispielen zu sehen:

... *aus* Hannover ...
... *aus* beruflichen Gründen ...
... *aus* Papier ...

Das zeigt, dass eine Präposition je nach Verwendungszweck eine andere Bedeutung erhält. Um den Verwendungszweck einer Präposition automatisch zu erkennen, ist die Betrachtung der Kontextwörter nötig, d.h. Wörter, die im gegebenen Text in nächster Nähe zur entsprechenden Präposition stehen. Mit diesen Kontextwörtern bzw. deren Häufigkeit können Präpositionen auf Unterschiede und Ähnlichkeiten untersucht werden.

1.2. Zielsetzung

Diese Bachelorarbeit hat zum Ziel, einen interaktiven Ansatz für die Exploration und Analyse von Präpositionskontexten zu entwickeln. Aus diesem hochdimensionalen Kontextraum soll eine für den Nutzer verständliche Visualisierung erstellt werden. Dieser Ansatz soll verschiedene Interaktionsmethoden enthalten, die Annahmen bezüglich Präpositionsähnlichkeiten bzw. -unterschieden ermöglichen. Eine Evaluation des Ansatzes soll daraufhin mit Hilfe von Experten in diesem Bereich durchgeführt werden.

1.3. Gliederung der Arbeit

Die Arbeit ist in folgender Weise gegliedert:

Kapitel 2 – Grundlagen: erläutert wichtige Grundlagen zum Verständnis der Arbeit.

Kapitel 3 – Verwandte Arbeiten stellt ähnliche bisher durchgeführte Arbeiten vor.

Kapitel 4 – Konzept beschreibt das Konzept für die Entwicklung des Ansatzes.

Kapitel 5 – Implementierung geht auf die Implementierung bestimmter Komponenten ein.

Kapitel 6 – Experten-Feedback präsentiert Vorgehensweise sowie Erkenntnisse der Evaluation mit den Experten.

Kapitel 7 – Zusammenfassung und Ausblick fasst die Ergebnisse der Arbeit zusammen und stellt mögliche Anknüpfungspunkte vor.

2. Grundlagen

Das folgende Kapitel beschreibt die nötigen Grundlagen zum Verständnis dieser Arbeit. Zu Beginn wird der Begriff der Informationsvisualisierung näher erläutert.

2.1. Informationsvisualisierung

„Der Zweck der Informationsvisualisierung ist es, die kognitive Leistungfähigkeit zu erweitern und nicht nur interessante Bilder zu schaffen. Informationsvisualisierungen sollten für den Kopf das sein, was Autos für die Füße sind." [SJ09]

Durch die Informationsflut in der heutigen Zeit haben Forschungsbereiche, wie die Informationsvisualisierung, an Bedeutung gewonnen. Die Visualisierung soll bei der Bewältigung der vielen Information helfen, denn ca. 80% der Informationsaufnahme erfolgt beim Menschen über den Sehsinn [War00]. Im folgenden Abschnitt werden die Begriffe Information und Visualisierung erstmal genauer definiert [ULP], bevor auf die Begrifflichkeit der Informationsvisualisierung näher eingegangen wird.

Definition 2.1.1 (Information)
„Datenmenge, der eine bestimmte Bedeutung zugeordnet werden kann."

Definition 2.1.2 (Visualisierung)
„Vorgang etwas in sichtbare Beziehungen zu setzen oder in sichtbare Form zu überführen."

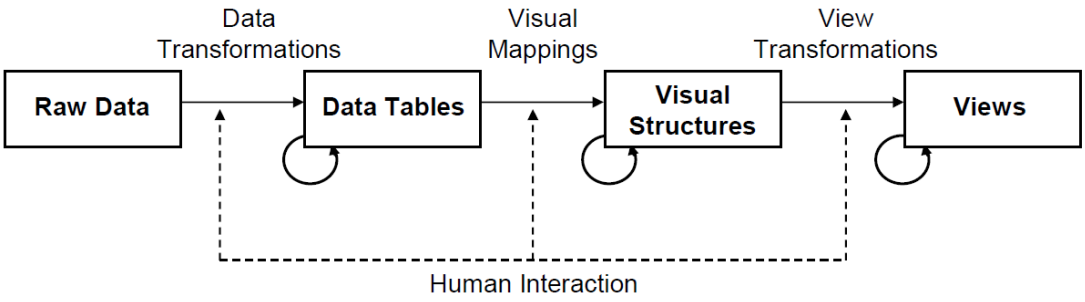

Abbildung 2.1.: Veranschaulichung des abstrakten Prozesses der Informationsvisualisierung. Angelehnt an das $Information\ Visualization\ Reference\ Model$ von Card et al. [CMS99]

Das oben genannte Zitat legt das Prinzip der Informationsvisualisierung bereits sehr nahe. Die Visualisierung soll hier einen guten Überblick über die in den Daten enthaltenen Informationen für den Betrachter bieten.

In der Abbildung 2.1 wird der abstrakte Prozess der Informationsvisualisierung veranschaulicht. Im ersten Schritt müssen die gegebenen Daten vorverarbeitet und transformiert, d.h. umgewandelt werden, um verschiedene Darstellungen daraus ableiten zu können und eine erweiterte Exploration der Daten zu ermöglichen. Die typische Aufgaben die der Vorverarbeitung angehören sind beispielsweise die Bereinigung oder Normalisierung von Daten. Die aus den Daten erzeugte Visualisierung muss vor allem das Ableiten von Erkenntnissen für den Betrachter ermöglichen. Führte die Visualisierung jedoch zu keinen überzeugenden Ergebnissen, so sollte der gesamte Prozess noch einmal überarbeitet werden. Der Grund für weniger überzeugende Visualisierungen liegt oft an einer ungeeigneten Transformation der Daten im ersten Schritt des Prozesses [KKEM10].

2.1.1. Textvisualisierung

Das Abbilden von textuellen Dokumenten, d.h. eine aussagekräftige Visualisierung von einzelnen Textdokumenten bis hin zu ganzen Bibliotheken, wäre vor allem für Wissenschaftler von großem Vorteil. Die Textvisualisierung kann als Zusammenfassung dienen oder ein Ausgangspunkt sein für ein anknüpfendes *Close Reading*, welches für die sorgfältige Interpretation von Textpassagen steht. Die verschiedenen Visualisierungstechniken können auch verwendet werden, um mehrere Texte zu vergleichen, wie beispielsweise Bücher von verschiedenen Autoren oder Reden von Politikern.

Würde man allerdings schnell mal ein Buch visualisieren wollen, so stößt man auf bestimmte Probleme. Zu beachten ist, dass es sich oft um Texte handelt, die mehrere tausende Wörter beinhalten. Für eine brauchbare Visualisierung müssen die Texte also erst einmal analysiert und zusammengefasst werden. Das zentrale Problem hier ist es eine effektive Analyseeinheit zu finden. Eine solche Einheit kann aus Buchstaben, aus Wörtern oder sonstigem bestehen. Der Stand der computergestützten Sprachverarbeitung bietet bereits eine Vielfalt einsetzbarer Einheiten. Bei der Auswahl muss ein Kompromiss zwischen Zuverlässigkeit und Gültigkeit gefunden werden. Auf der einen Seite können Computer sehr zuverlässig einzelne Wörter aus Texten herausfiltern und würden so den semantischen Teil dem Menschen überlassen, d.h. das Zusammensetzen bzw. deuten der Wörter. Andererseits gibt es auch bereits Programme, die Texte auf semantischer Ebene analysieren können, jedoch sind die Fehlerraten hier noch zu hoch und führen oft zu Fehlinterpretationen in der späteren Nutzung. Bei einer Visualisierung von Texten kann auch die Lesbarkeit immer wieder ein schwer zu handelndes Problem darstellen.

Es gibt verschiedene Konzepte zur Textvisualisierung, die die genannten Probleme zu kompensieren versuchen. In der Abbildung 2.2 wurde die sogenannte *Phrase Net*-Technik verwendet. Das *Alte* und das *Neue Testament* werden hier verglichen, indem jeweils relevante und am häufigsten vorkommenden Wörter dargestellt werden. Die Schriftgröße der Wörter und die Dicke der Pfeile stehen für die Vorkommnisse der einzelnen Wörter bzw. der Wörter im selben Kontext [HWV09]. Im Rahmen dieser Bachelorarbeit soll ebenfalls, durch eine geeignete Technik, aus den Daten mit den Präpositionskontexten eine Textvisualisierung entwickelt werden.

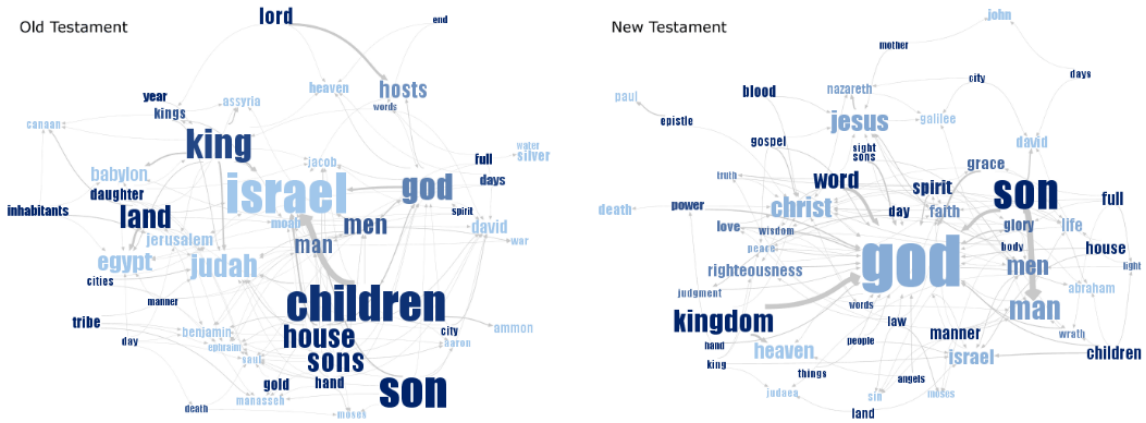

Abbildung 2.2.: Vergleich des Alten und Neuen Testaments durch die Textvisualisierung $PhraseNet$ [HWV09]

2.2. Local Mutual Information (LMI)

Das *Local Mutual Information*-Maß kommt aus der Informationstheorie. Im Bereich der Computerlinguistik wird es beispielsweise verwendet um ein Maß für die Stärke des Zusammenhangs eines Wortpaares zu erhalten. Die, für diese Arbeit, gegebenen Daten enthalten LMI-Werte für Wortpaare bestehend aus jeweils einer Präposition und einem Kontextwort. Das LMI-Maß wird mit Hilfe einer beobachteten Häufigkeit H_1 und einer erwarteten Häufigkeit H_2 berechnet:

$$LMI = \mathbf{H}_1 \log\left(\frac{\mathbf{H}_1}{\mathbf{H}_2}\right)$$

Die beobachtete Häufigkeit ist die tatsächliche Vorkommenshäufigkeit eines Wortpaares in Texten, während die erwartete Häufigkeit nur eine Voraussage des gemeinsamen Vorkommens eines Wortpaares ist [COL].

2.3. Kosinus-Ähnlichkeitsmaß

Die Kosinus-Ähnlichkeit zwischen zwei Vektoren ist ein Maß um die Ähnlichkeit zwischen diesen zu ermitteln. Hierfür wird der Kosinus des eingeschlossenen Winkels, der zwei Vektoren A und B, wie folgt berechnet:

$$\cos(\theta) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i\mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}}$$

Die Ähnlichkeit zwischen den Vektoren wird also, wie in Abbildung 2.3 genauer zu sehen, anhand der Richtung der Vektoren bestimmt. Vektoren, die genau entgegengerichtet sind, erhalten als Messwert -1 und im Fall, dass sie genau gleichgerichtet sind den Wert 1. Eine 0 als Ergebnis ergibt sich, wenn

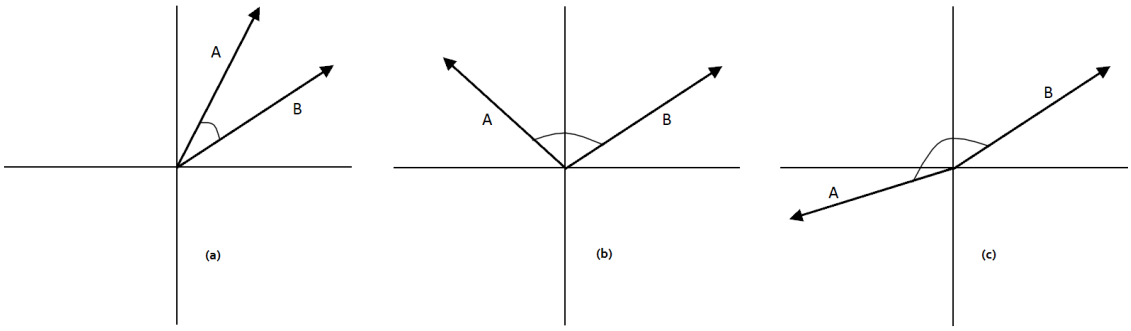

Abbildung 2.3.: Kosinus-Ähnlichkeit: (a) Winkel zwischen den Vektoren A und B ein wenig größer als 0°, d.h. Kosinus-Ähnlichkeit ist nahe der 1 bzw. 100%; (b) Winkel zwischen den Vektoren A und B nahe zu orthogonal, d.h. Kosinus-Ähnlichkeit ist fast 0 bzw. 0%; (c) Winkel zwischen den Vektoren A und B nahe den 180°, d.h. Kosinus-Ähnlichkeit ist fast -1 bzw. -100%

die Vektoren orthogonal zueinander sind und steht somit für eine Kosinus-Ähnlichkeit von 0%. Die anderen Zwischenwerte dagegen zeigen inwiefern sich die Vektoren ähneln bzw. nicht ähneln [MRS08]. Dieses Ähnlichkeitsmaß wird in der vorliegenden Arbeit verwendet um die Ähnlichkeiten zwischen den Präpositionen zu ermitteln.

2.4. t-Distributed Stochastic Neighbor Embedding (t-SNE)

Laurens van der Maaten und Geoffrey Hinton beschreiben in ihrer Arbeit "Visualzing Data using t-SNE", [MH08] ein Verfahren zur Dimensionsreduktion hochdimensionaler Daten. Eine Dimensionsreduktion ist für den Ansatz dieser Arbeit notwendig, um aus dem hochdimensionalen Kontextraum eine zweidimensionale Visualisierung erhalten zu können. Das sogenannte *t-Distributed Stochastic Neighbor Embedding* (t-SNE) Verfahren visualisiert hochdimensionale Daten, indem für jedes Datenobjekt aus dem hochdimensionalen Datensatz ein Punkt im zwei- oder dreidimensionalen Raum ermittelt wird. Die daraus resultierende Punktwolke soll Punkte nahe beieinander abbilden, wenn die entsprechenden Datenobjekte sich ähnlich sind und unähnliche Datenobjekte als weit voneinander entfernte Punkte darstellen. Das t-SNE unterscheidet sich von anderen Verfahren zur Dimensionsreduktion vor allem dadurch, dass es in der niederdimensionalen Projektion sowohl lokale als auch globale Strukturen der hochdimensionalen Daten so gut wie möglich zu erhalten versucht.

Als erstes berechnet das Verfahren, durch die Umwandlung der hochdimensionalen euklidischen Abstände zwischen Datenobjekten in Wahrscheinlichkeiten, die Ähnlichkeiten der Datenobjekte. Die Ähnlichkeit eines Datenobjekts x_i, aus dem hochdimensionalen Datensatz X mit N Datenobjekten $\{x_1, x_2, ..., x_n\}$, zu einem anderem Datenobjekt x_j, ist die bedingte Wahrscheinlichkeit $p_{j|i}$, dass x_i als Nachbar x_j wählt. Für im hochdimensionalen Raum nahe liegende Datenobjekte fällt die Wahrschein-

lichkeit $p_{j|i}$ hoch aus. Die Formel von $p_{j|i}$ berechnet die Wahrscheinlichkeit unter Verwendung der Varianz der Gaußfunktion σ_i zentriert auf x_i:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}$$

Da nur Ähnlichkeiten zwischen verschiedenen Datenobjekten relevant sind, werden die Wahrscheinlichkeiten $p_{i|i}$ auf 0 gesetzt. Unter Verwendung der vorherigen Formel kann die paarweise Ähnlichkeit im hochdimensionalen Raum wie folgt berechnet werden:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Die Datenpunkte des niederdimensionalen Datensatzes $Y = \{y_1, y_2, ..., y_n\}$ sollen, die sich aus p_{ij} resultierenden Ähnlichkeiten soweit wie möglich wiedergeben. Unter Verwendung der studentschen t-Verteilung mit einem Freiheitsgrad, welches die Abbildung von größeren Distanzen aus dem hochdimensionalen Raum im niederdimensionalen Raum verbessert, wird q_{ij}, d.h. die paarweise Ähnlichkeit von niederdimensionalen Punkten, berechnet:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

Die Kullback-Leibler Divergenz ist ein Maß um die Differenz zwischen den zwei Wahrscheinlichkeitsverteilungen P und Q zu ermitteln:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Sie dient im t-SNE als Kostenfunktion C und wird unter Verwendung des Gradientenabstiegs minimiert:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_i)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1})$$

2.5. Technische Grundlagen

In diesem Abschnitt sollen die für diese Arbeit notwendigen technischen Grundlagen beschrieben werden. Dazu gehören die, für die Umsetzung des Ansatzes, verwendeten Werkzeuge wie *Prefuse*, die *Google Guava Library* und die gegebenen Daten.

2.5.1. Prefuse

Die interaktiven Visualisierung für die vorliegende Arbeit wurde unter anderem durch das Software-Werkzeug *Prefuse*[1] realisiert, welches speziell für das Erstellen von Informationsvisualisierungen

[1] http://prefuse.org/

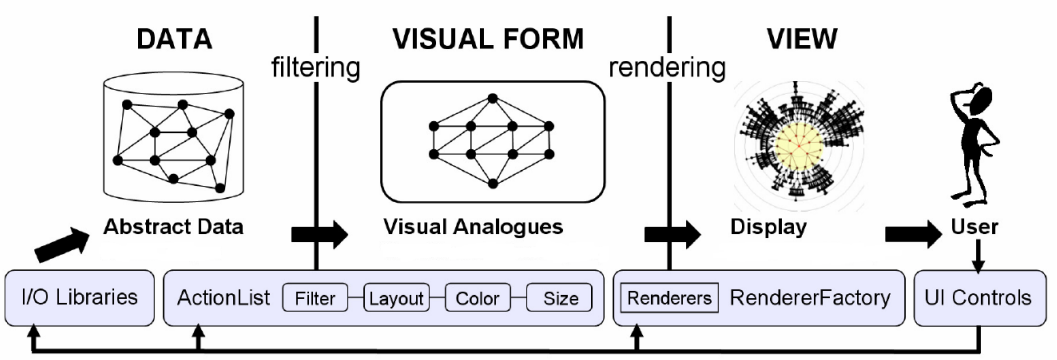

Abbildung 2.4.: Prozess für die Erstellung interaktiver Visualisierungen mit Prefuse [Hee04]

entwickelt wurde. Es ist komplett in Java geschrieben und basiert auf der *Java 2D* Grafikbibliothek. Die Internetpäsenz dieses Software-Werkzeugs stellt das Werkzeug selbst, dessen Quellcode und einige interaktive Beispiele zur Verfügung. Das in der Abbildung 2.3 dargestellte Diagramm zeigt, welche Schnittstellen für eine Visualisierung bereit stehen und wie der Ablauf des Visualisierungsprozesses unter Verwendung von *Prefuse* genau aussieht.

Dieser Prozess beginnt mit dem Datenimport, welcher durch *Prefuse* mit bestimmten Schnittstellen unterstützt wird, die die Verbindung zu SQL-Datenbanken und die Übernahme aus Textdateien in den verschiedensten Formaten wie beispielsweise `.csv`-Dateien, ermöglichen. Die abstrakten Datenobjekte werden zu Tupel, in einer von *Prefuse* bereitgestellten Tabelle, transformiert. Aus diesen Daten werden jene Datenobjekte herausgefiltert, die für die Visualisierung relevant sind. Daraus entstehen die *VisualItems*, welche die Datenobjekte als visuell interaktive Objekte repräsentieren. Diese *VisualItems* speichern visuelle Eigenschaften, welche durch die *Action*-Module festgelegt werden und bieten viele Verarbeitungsmöglichkeiten für die visuellen Daten. Dazu gehören Eigenschaften der *VisualItems* bezüglich der räumlichen Anordnung (*Layout*), Farbe, Größe und Schriftzuordnung [Bä07].

Das endgültige Erscheinungsbild eines *VisualItem* wird allerdings von einem *Renderer*, der die Anweisungen für die Erstellung der des Elements enthält, bestimmt. Mit Unterstützung des *RendererFactory* wird entschieden, welcher *Renderer* letztlich für das gegebene *VisualItem* eingesetzt werden darf.

Die *Action*-Schnittstelle wurde so konzipiert, dass es Entwicklern möglich ist individuelle Aktionen zu erstellen, um ihre Ziele für die erwünschte Visualisierung zu erreichen. Ein *ActivityManager* führt die *Actions* mit Hilfe von *ActionLists* aus und ermöglicht z.B. Animation und zeitbasierte Verarbeitung von der Visualisierung der *VisualItems*.

Die Klasse *Display* ist eine Unterklasse von *javax.swing.JComponent* und ermöglicht die Visualisierung der *VisualItems* auf dem Bildschirm. Benutzerinteraktionen durch Maus und Tastatur mit der *Display*-Komponenten kann der Entwickler auf Basis der *ControlListener*-Schnittstellen entwerfen [Hee04].

2.5.2. Google Guava Library

Google Guava[2] ist eine Open-Source Sammlung an Bibliotheken für Java, die von Google entwickelt und gewartet werden. Das *Guava* Projekt bietet eine breite Palette an neuen Funktionalitäten für Software-Entwickler bezüglich *Collections*, Stringmanipulationen oder der *I/O*-Unterstützung, die auch Verwendung fanden bei der Realisierung der Visualisierung für diese Arbeit. Unter anderem bietet es auch sogenannte *Basic Utilities*, die das Programmieren mit Java erleichtern sollen.

2.5.3. Relevante Daten

Die in dieser Arbeit verwendeten Daten wurden vom Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart zur Verfügung gestellt. Sie liegen als `.txt`-Dateien vor, die Information über Kontextwörter und deren Auftreten mit Präpositionen einschließen bzw. die eine ambige Zuordnung der Präpositionen zu semantischen Klassen enthalten.

Eine Datei enthält Präpositionen-Verb-Tupel, d.h. 4 Spalten mit Tabulatoren getrennt. In der ersten Spalte steht jeweils die Präposition, in der zweiten Spalte das Verb, in der dritten Spalte die Häufigkeit (wie oft steht das Verb in den Texten in nächster Nähe zur entsprechenden Präposition) und in der dritten Spalte der LMI-Wert. Eine weitere Datei enthält Präpositionen-Nomen-Tupel. Das Format ist dasselbe wie bei der erst erwähnten Datei, nur dass in der zweiten Spalte Nomen statt Verben stehen.

Die Datei mit den semantischen Klassen der Präpositionen, die sogenannten *Gold Standards*, enthält zwei Spalten mit Tabulatoren getrennt. In der ersten Spalte steht jeweils die Präposition und in der zweiten Spalte die semantische Gold-Standard-Klasse. Da die Zuordnung ambig ist, tauchen viele Präpositionen in mehreren Klassen auf.

[2] https://code.google.com/p/guava-libraries/

3. Verwandte Arbeiten

In diesem Kapitel werden bereits veröffentlichte Arbeiten, welche sich mit einer Visualisierung von hochdimensionalen Daten beschäftigt haben, vorgestellt. Die daraus gewonnenen Erkenntnisse hatten einen Einfluss auf die grundlegende Konzipierung und Durchführung dieser Arbeit. In den folgenden Abschnitten werden die Ideen und Vorgehensweisen spezieller Arbeiten aus diesem Bereich kurz beschrieben.

3.1. Visualisierung hochdimensionaler Daten

Die Arbeit "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents" von Wise et al. [WTP+95] hatte das Ziel eine Anwendung zu entwickeln, die das Analysieren einer großen Menge an Textdokumenten vereinfacht. Die Textdokumente sollten so visualisiert werden, dass bestimmte Merkmale leicht zu erkennen sind und somit das vereinzelte Durchlesen jedes Dokuments nicht mehr notwendig ist, um Gemeinsamkeiten feststellen zu können. Allerdings stellen

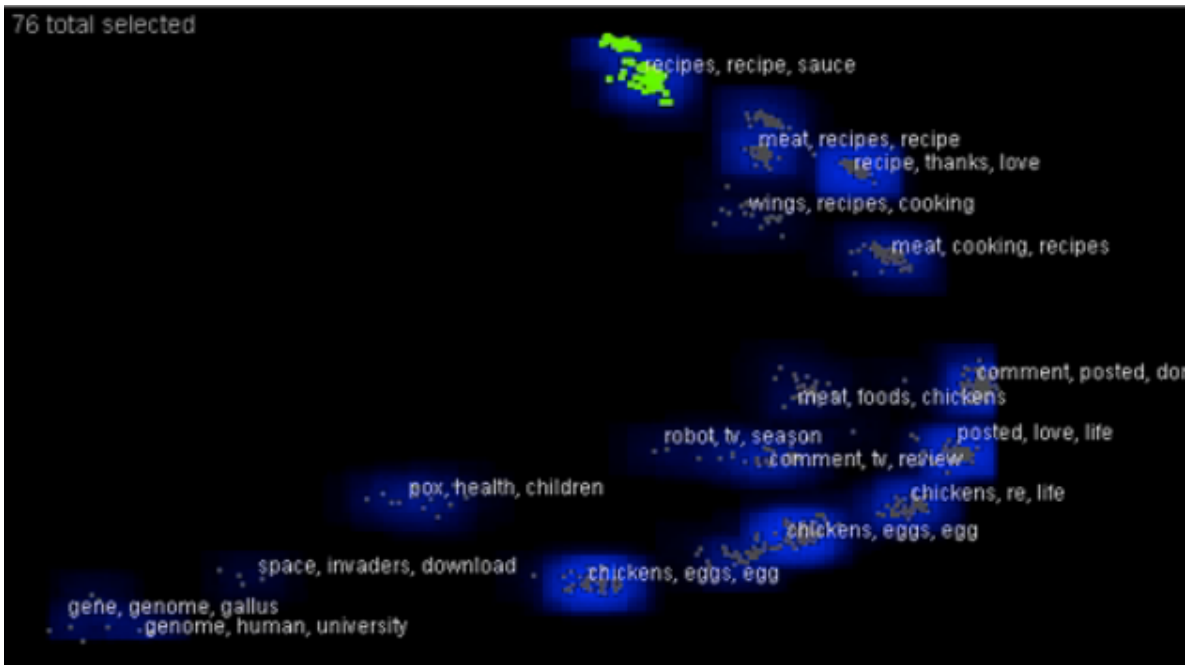

Abbildung 3.1.: Galaxy-Ansicht der Textdokumente als (Sternen)-Punkte [SPI]

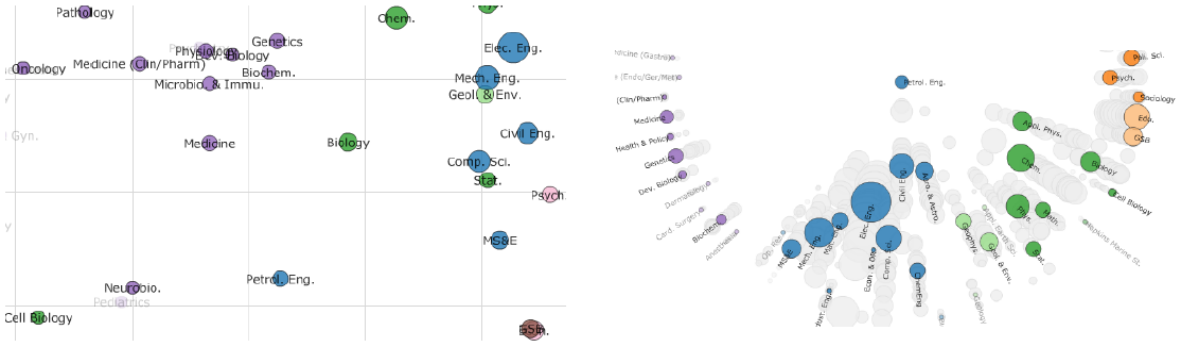

Abbildung 3.2.: *Interpretation*-Ansicht (rechtes Bild) und *Trust*-Ansicht (linkes Bild) [CRMH12]

diese Textdokumente mit ihrem Inhalt hochdimensionale Datenobjekte dar und können nicht ohne weitere Bearbeitung brauchbar visualisiert werden, da Visualisierungen mit zu vielen Dimensionen für den Menschen nur schwer nachvollziehbar sind. Merkmale der Dokumente, wie beispielsweise die Häufigkeit eines Wortes, wurden durch eine computergestützte Textanalyse festgestellt. Somit konnten die Dokumente jeweils als hochdimensionale Vektoren repräsentiert werden. Diese Vektordarstellung ermöglichte, dass die Dokumente direkt verglichen, gefiltert und umgewandelt werden konnten. Die Idee war die Dokumente in einem zweidimensionalen Raum darzustellen, um eine für den Betrachter verständliche Visualisierung zu erhalten. Die hochdimensionalen Vektoren wurden durch ein Verfahren des *Multidimensional Scaling* (MDS) auf zweidimensionale Punkte reduziert, dieses Verfahren soll soweit möglich die Struktur des hochdimensionalen Raums beibehalten. Die daraus resultierende zweidimensionale Visualisierung, wie in Abbildung 3.1 zu sehen, wurde als eine sogenannte Galaxy-Ansicht dargestellt, in der die (Sternen-) Punkte die Dokumente verkörpern. Die Beschriftungen in weißer Schrift zeigen die dominierenden Themen der Dokumente in der jeweiligen in blau gekennzeichneten Gruppen an. Diese Darstellung bietet bereits einen guten Einblick wie die Inhalte mehrerer Dokumente verbunden sind. Diese Visualisierung soll vor allem die Ähnlichkeit zwischen den Dokumenten aufzeigen, d.h. je ähnlicher die Dokumente sich in ihrem Kontext und Inhalt sind, desto näher werden sie im zweidimensionalen Raum abgebildet. Die Betrachter haben dadurch schon einen ersten guten Eindruck von Mustern und Tendenzen, die sich aus der gegebenen Mengen an Dokumenten ergeben. Ein großer Unterschied dieser Arbeit zur vorliegenden Bachelorarbeit ist, dass Textdokumente als die zu analysierenden Datenobjekte betrachtet werden statt die Präpositionskontexte.

Des Weiteren findet sich die Arbeit "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis" von Chuang et al. [CRMH12] in welcher nützliche Erkenntnisse, über die Visualisierung hochdimensionaler Daten, für die vorliegende Arbeit beschrieben werden. Die in der erwähnten Arbeit beschriebenen Erkenntnisse stammen von den Erfahrungen die bei der Entwicklung des *Dissertation Browsers*[1] für die Stanford University gesammelt wurden. Das *Dissertation Browser* Tool ermöglicht einen Vergleich von über 9000 Doktorarbeiten nach Themenähnlichkeiten. Für die zweidimensionale Projektion dieser Daten wurde ähnlich wie in der vorherigen Arbeit vorgegangen

[1] http://www-nlp.stanford.edu/projects/dissertations/browser.html/

und somit ein Verfahren zur Dimensionsreduktion verwendet. In dieser Darstellung werden, wie in Abbildung 3.2 zu sehen, nur die verschiedenen wissenschaftliche Bereiche der Universität abgebildet. Die Abstände zwischen den Bereichen bzw. Punkten ist von den Ähnlichkeiten zwischen den Doktorarbeiten abhängig. Da die erzeugte zweidimensionale Projektion allerdings ein leicht verzerrtes Abbild der wirklichen Ähnlichkeiten der Bereiche ist, gibt es noch die sogenannte *Trust*-Ansicht. Diese Ansicht (auf der rechten Seite der Abbildung 3.2) zeigt explizit den Abstand von einem fokussierten Bereich zu jedem anderem Bereich an. Die tatsächlichen Werte werden aus der entsprechende Zeile in der Ähnlichkeitsmatrix entnommen. Die Ähnlichkeiten werden hier als radiale Abstände von den fokussierten Bereich aus zu jedem anderen Bereich verwendet. Die übrigen Bereiche der Universität, die keine Ähnlichkeit bezüglich der Doktorarbeiten vorzuweisen haben, werden rund um den Kreis dargestellt. Die Abbildung 3.2 verdeutlicht mit dem Fall des Bereichs *Petroleum Engineering* weshalb die *Trust*-Ansicht wichtig ist für eine richtige Interpretation einer solchen zweidimensionalen Projektion hochdimensionaler Daten. Laut der Abbildung auf der linken Seite der Abbildung 3.2 erscheint *Petroleum Engineering* nahe den Bereichen Neurobiologie, Medizin und Biologie zu sein. Die rechte Visualisierung zeigt allerdings die unverzerrte Entfernungen von *Petroleum Engineering* zu den anderen Bereichen. Die Verbindung zur Biologie verschwindet, d.h. die Nähe zu Biologie war nur eine Fehldarstellung, die durch die Dimensionsreduktion erzeugt wurde. Die visuelle Darstellung des räumlichen Abstands in der ersten Ansicht ist interpretierbar, aber allein nicht vertrauenswürdig. Da auch bei der zweidimensionalen Visualisierung der Präpositionskontexte für diese Bachelorarbeit keine hundertprozentige Vertrauenswürdigkeit erwartet werden kann, wurde ebenfalls eine sogenannte *Trust*-Ansicht entwickelt. Diese Ansicht soll immer für jede einzelne Präposition die "echten" Abstände zu den anderen anzeigen lassen können.

In der Arbeit "Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches" von Boyack et al. [BND+11] findet eine Dimensionsreduktion von biomedizinischen Dokumenten statt. Hierfür werden neun verschiedene Ähnlichkeitsmaße bzw. Techniken zur Feststellung der Ähnlichkeiten zwischen den Dokumenten vorgestellt und verglichen. Ziel ist ein ein sogennantens Clustering, also eine Gruppierung der Dokumente basierend auf den neun Ähnlichkeitsmatrizen, die sich aus den verschiedenen Techniken ergeben. Unter anderem wurde die Kosinus-Ähnlichkeit unter Verwendung des *tf-idf*-Maßes vorgestellt. Das *tf-idf* ist ein Produkt gebildet aus der Vorkommenshäufigkeit *tf (term frequency)* und der inversen Dokumenthäufigkeit *idf (inverse document frequency)* und wird zur Gewichtung der Relevanz von Wörtern in Dokumenten eines Dokumentkorpus eingesetzt. Diese Technik wurde auch für die zweidimensionale Visualisierung der Präpositionskontexte in der vorliegenden Bachelorarbeit testweise verwendet.

Heulot et al. veröffentlichten ebenfalls eine Arbeit mit der Thematik Dimensionsreduktion von hochdimensionalen Daten. Ihre Arbeit "ProxiLens: Interactive Exploration of High-Dimensional Data using Projections" [HAF13] stellt eine interaktive Technik namens ProxiLens vor, welche eine kontinuierliche Navigation durch hochdimensionale Daten mit Hilfe einer zweidimensionalen Projektion ermöglicht. Diese interaktive Technik basiert auf einer semantischen Linse, die die fokussierten Punkte hervorhebt und falsche Nachbarn an den Rand der Linse drängt. Als falsche Nachbarn werden die bezeichnet, die in der zweidimensionalen Projektion benachbart sind, aber im hochdimensionalen Raum weit voneinander entfernt liegen. In Abbildung 3.3 ist eine Implementierung der ProxiLens-Technik zu sehen. Die hochdimensionalen Distanzen zwischen den Datenobjekten werden hier durch eine blaue Farbskala mit variierender Farbintensität angezeigt. Je heller die Farben der Skala, desto kürzer sind die hochdimensionalen Abstände zum fokussierten Punkt in der Mitte der Linse. Die

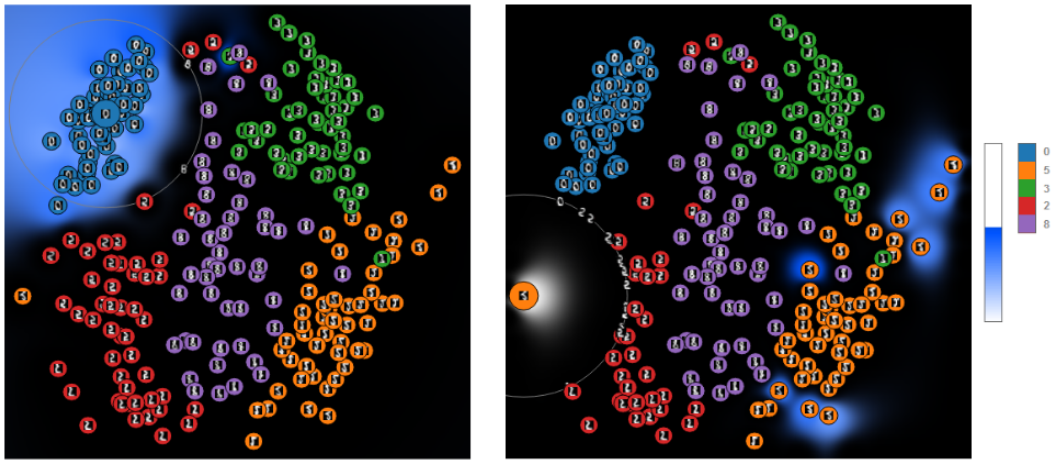

Abbildung 3.3.: ProxiLens-Ansicht in der die fokussierten Punkte hervorgehoben werden und falsche Nachbarn an den Rand der Linse gedrängt werden [HAF13]

unterschiedlichen Farben der visualisierten Objekte stehen für die verschiedene Klassen. Aus dem linken Bild der Abbildung 3.3 ist beispielsweise zu entnehmen, dass die Datenobjekte der Klasse 0 ein sehr dichtes Cluster im hochdimensionalen Raum bilden. Auf dem rechten Bild der Abbildung 3.3 steht ein Datenobjekt aus der Klasse 5 im Fokus. Es ist zu sehen, dass die visualisierten Objekte der Klasse 2 an den Rand der Linse platziert werden und Objekte aus der Klasse 5, die weit vom fokussierten Objekt entfernt sind, durch einen blauen Hintergrund hervorgehoben werden. Daraus ist zu schließen, dass die tatsächlichen Abstände des fokussierten Objekts zu den anderen stark von der zweidimensionalen Projektion abweichen. Die semantische Linse von ProxiLens ist verlgeichbar mit der *Trust*-Ansicht von Chuang et al. [CRMH12], welche ebenfalls eine "Überprüfung" des Wahrheitsgehalts der zweidimensionalen Projektion ermöglicht.

Für die Dimensionsreduktion von hochdimensionalen Datenobjekten stehen einem eine Vielzahl an Verfahren zur Verfügung. Diese ganzen Verfahren sind zwar alle in der Lage einen hochdimensionalen Datensatz in einen zwei- oder dreidimensionalen Datensatz zu formen, unterscheiden sich allerdings auch in vielen Aspekten. Diese Aspekte sind letzten Endes auch ausschlaggebend für die Wahl eines Verfahrens. Die Arbeit von García-Fernández et al., die unter dem Namen "Stability Comparison of Dimensionality Reduction Techniques Attending to Data and Parameter Variations" [GFVLD13] veröffentlicht wurde, stellt eine Studie über die Stabilität, Robustheit und der Leistungsfähigkeit einiger dieser Dimensionsreduktionsverfahren vor. Vor allem fokussiert sich diese Studie auf die Algorithmen hinter den Verfahren und auf die verschiedenen Parameter. In der Regel haben diese Eigenschaften einen großen Einfluss auf das Endergebnis der Verfahren. Für die Analyse wurde eine große Gruppe der Verfahren mit künstlich erstellten und realen Daten durchprobiert. Dabei wurden vor allem die sich ergebenden Variationen der Visualisierungen untersucht, die sich bei Änderung verschiedener Parametereingaben ergaben. Eines der vorgestellten und untersuchten Verfahren in der Arbeit von García-Fernández et al. ist das im Rahmen dieser Bachelorarbeit verwendete t-SNE Verfahren. Der Grund hinter der Entscheidung für dieses Verfahren in dieser Bachelorarbeit hängt vor allem mit der Erhaltung der Struktur zusammen. Die Erhaltung der Strukturen der Datenobjek-

te im hochdimensionalen Raum auf der zweidimensionalen Projektion gehört zu den wichtigsten Eigenschaften eines Dimensionsreduktionsverfahren.

4. Konzept

Im folgenden Kapitel wird das Konzept des interaktiven Ansatzes beschrieben, welches im Rahmen dieser Arbeit entwickelt wurde. Dem Nutzer soll durch diesen Ansatz eine visuelle Analyse von Präpositionen ermöglicht werden. Dazu gehören die Entwicklung einer geeigneten zweidimensionalen interaktiven Visualisierung der Präpositionskontexte und das Konzipieren von Interaktionsmethoden, die eine genauere Exploration der Präpositionskontexte erlauben.

4.1. Datentransformation

Um das Weiterarbeiten mit hochdimensionalen Datenobjekten zu vereinfachen, werden diese wie die Textdokumente, der in Kapitel 3 beschriebenen Arbeiten, in hochdimensionale Vektoren umgewandelt. Diese Vektordarstellung ermöglicht beispielsweise einen direkten Vergleich von Datenobjekten. Für die vorliegende Arbeit wäre es daher hilfreich die Präpositionen als hochdimensionale Vektoren zu repräsentieren.

Die Idee hinter dieser Vektordarstellung basiert auf dem *Vector Space Model* [MRS08], welches für die Informationsbeschaffung bei Dokumenten konzipiert wurde. Der allgemeine Aufbau eines Dokumentvektors hätte das folgende Schema:

$$Dokument_j = (Wort_{1j}, Wort_{2j}, Wort_{3j}, ..., Wort_{nj})$$

Jede Dimension im Vektor entspricht einem $Wort_{ij}$ im Dokument und um die Werte für diese zu bekommen, werden beispielsweise die Häufigkeiten benutzt, d.h. Anzahl an Vorkommen eines Wortes in einem Dokument. Ein aus einem Satz bestehende Dokument *Die Häuser verdecken die Bäume.* könnte man als folgenden Vektor darstellen:

$$\vec{v} = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{matrix} \Leftarrow \text{die} \\ \Leftarrow \text{Häuser} \\ \Leftarrow \text{Bäume} \\ \Leftarrow \text{verdecken} \end{matrix}$$

Das Wort *die* tritt zweimal auf und die anderen Wörtern kommen im Beispieldokument nur einmal vor. Überträgt man dieses Modell auf die gegebenen Präpositionskontexte dieser Arbeit so ergibt sich für jedes Kontextwort eine Dimension im hochdimensionalen Vektorraum:

$$Präposition_j = (Kontextwort_{1j}, Kontextwort_{2j}, ..., Kontextwort_{nj})$$

	gehen	**vermuten**	**greifen**	**nennen**
ab	2360	17	213	279
in	209787	3123	13406	20191
durch	22368	127	391	475
mittels	69	2	11	9

Tabelle 4.1.: Auszug der Häufigkeitswerte in den Präpositionsvektoren

Die folgende Tabelle 4.1 zeigt ein Auszug der gegebenen Daten, die als Werte für die einzelnen Dimensionen bzw. Kontextwörter eingesetzt werden können. Die Werte dieser Tabelle sind ebenfalls Häufigkeitswerte, d.h. Information darüber welche Wörter mit welchen Präpositionen im gleichen Kontext auftauchen. Die verwendeten Daten bieten neben den Häufigkeitswerten auch *LMI*-Werte, welche in Kapitel 2 näher erläutert wurden. Die Tabelle zeigt beispielsweise, dass die Präpositionen *ab* und *durch* eher in einem Kontext mit dem Verb *gehen* vorkommen als in einem mit dem Verb *vermuten*.

In der ersten Spalte befinden sich die Namen der jeweiligen Präpositionen und die oberste Zeile zeigt ein Auszug der in den Daten gegebenen Kontextwörter. Unterhalb der Kontextwörter stehen immer die zugehörigen Häufigkeitswerte bezüglich der Präposition, d.h. mögliche Werte für die jeweiligen Dimensionen eines Präpositionsvektors.

4.2. Ideen zur 2D-Projektion

4.2.1. Dimensionsreduktion

Für die Darstellung der Präpositionskontexte bzw. der hochdimensionalen Präpositionsvektoren im zweidimensionalen Raum wird ein Verfahren zur Dimensionreduktion benötigt. Die im Kapitel 3 erwähnte Arbeit von Wise et al. [WTP+95] verwendet für ihre zu analysierenden hochdimensionalen Textdokumentvektoren beispielsweise *Multidimensional Scaling* (MDS).

Solche Verfahren formen einen hochdimensionalen Datensatz in einen zwei- oder dreidimensionalen Datensatz, so dass die Datenpunkte folglich als Punktwolke abgebildet werden können. Das Ziel jeder Dimensionsreduktion ist die Struktur der hochdimensionalen Daten in der niederdimensionalen Visualisierung so gut wie möglich zu erhalten. Es gibt eine Vielzahl an Methoden zur Dimensionsreduktion, allerdings unterscheiden sich diese vor allem darin, welche Art von Struktur sie beibehalten.

Zu den bewährtesten linearen Verfahren gehören beispielsweise das *Principal Components Analysis* (PCA) und Verfahren aus dem schon genannten *Multidimensional Scaling* (MDS) [GFVLD13]. Diese Verfahren sind vor allem darauf bedacht in der niederdimensionalen Visualisierung die Datenpunkte, die sich im hochdimensionalem Raum unähnlich sind, weit voneinander entfernt darzustellen. Dadurch gehen allerdings Strukturen verloren. Aus diesem Grund wurden nicht-lineare Verfahren zur Dimensionsreduktion entwickelt, deren Ziel vor allem die Erhaltung der lokalen Struktur der Daten im niederdimensionalen Raum sein sollte. Diese Verfahren konnten jedoch nur mit Testdaten gute Ergebnisse erzielen, denn die Versuche mit hochdimensionalen Daten aus der realen Welt waren

weniger zufriedenstellend [MH08]. Die Beibehaltung von lokalen und globalen Strukturen mit diesen Verfahren erwies sich somit als schwierig.

Zu diesen nicht-linearen Methoden gehört beispielsweise das *Stochastic Neighbor Embedding* (SNE). Das Verfahren t-SNE ist eine optimierte Variation vom SNE, die als Ziel hat sowohl lokale als auch globale Strukturen der hochdimensionalen Daten so gut wie möglich in der niederdimensionalen Visualisierung zu erhalten. Wegen dieser verbesserten Erhaltung der Struktur im niederdimensionalen Raum im Vergleich zu den anderen Verfahren, fiel für diese Arbeit die Wahl auf das t-SNE [MH08].

Ähnlichkeitsmatrix

Für die Umsetzung der zweidimensionalen Projektion der Päpositionskontexte muss somit zuerst einmal mit Hilfe der erzeugten Präpositionsvektoren die Ähnlichkeiten zwischen den Präpositionen ermittelt werden.

Hierfür wird allerdings eine kleine Änderung am t-SNE vorgenommen, denn in der Regel nutzt das Verfahren als Maß für die Ähnlichkeit zwischen hochdimensionalen Datenobjekten, Wahrscheinlichkeiten, d.h. ein Wert zwischen 0 und 1. Als Ähnlichkeitsmaß wird im Rahmen dieser Arbeit jedoch das Kosinus-Ähnlichkeitsmaß (Werte im Intervall von -1 bis 1) verwendet, auf welches bereits im Kapitel der Grundlagen genauer eingegangen wurde. Die Ähnlichkeit zwischen zwei Präpositionen erhält man kurz gefasst durch die Berechnung des Kosinus vom geschlossenem Winkel der beiden Vektoren.

Diese resultierenden Vergleichswerte werden in einer Ähnlichkeit- bzw. Distanzmatrix gespeichert. Die folgende abgebildete Matrix soll den grundlegenden Aufbau einer solchen Ähnlichkeitsmatrix veranschaulichen:

$$\begin{array}{c} \\ P_1 \\ P_2 \\ \vdots \\ P_n \end{array} \begin{array}{c} \begin{array}{cccc} P_1 & P_2 & \dots & P_n \end{array} \\ \begin{pmatrix} 0 & c_{1,2} & \dots & 1,n \\ c_{2,1} & 0 & \dots & c_{2,n} \\ \vdots & \vdots & \ddots & \dots \\ c_{n,1} & \dots & c_{n,j-1} & 0 \end{pmatrix} \end{array}$$

Alle P_i stellen in diesem Fall die entsprechenden Präpositionen dar. Der Wert $c_{2,1}$ repräsentiert somit den Ähnlichkeitswert der beiden Präpositionen P_2 und P_1. Der Algorithmus von t-SNE ermittelt hier die zweidimensionalen Datenpunkte nicht wie sonst durch die selbst berechneten Wahrscheinlichkeiten, sondern mit den gerade beschriebenen Ähnlichkeitswerten.

4.2.2. Prototypen der Visualisierung

In diesem Abschnitt werden die mit Hilfe des t-SNE Verfahrens erzeugten verschiedenen Visualisierungen verglichen. Das Verfahren lieferte abhängig von den eingesetzten Werten für die Berechnung der Ähnlichkeitsmatrix verschieden zufriedenstellende Ergebnisse.

Die Ähnlichkeit zwischen Präpositionsvektoren wurde in der ersten prototypischen Visualisierung mit den Häufigkeitswerten der Kontextwörter berechnet. Die sich daraus ergebende Projektion

der zweidimensionalen Daten aus der Berechnung des t-SNE Verfahrens, wird im oberen Teil der Abbildung 4.1 gezeigt. Die einzelnen Punkte, die jeweils eine Präposition darstellen, sind hier sehr regelmäßig auf der Bildfläche verteilt. Aus dieser prototypischen Visualisierung lassen sich nur sehr schwierig Schlüsse über Gemeinsamkeiten zwischen den Präpositionen ziehen.

Da in der ersten Visualisierung die Präpositionen zu regelmäßig verteilt waren, wurde für die folgende prototypische Visualisuerung, welche im unteren Teil in Abbildung 4.1 zu sehen ist, eine Gewichtung der Kontextwörter bezüglich der Präpositionen, in welche diese enthalten sind, vorgenommen. Die Relevanz der einzelnen Kontextwörter in einem Präpositionsvektor aus einer Menge von Präpositionsvektoren, sollten durch diese Gewichtung w_{ij} berücksichtigt werden. Hierfür wurde das *tf-idf*-Maß [MRS08] verwendet, welches eigentlich zur Beurteilung der Relevanz von Wörtern in Dokumenten aus einem Dokumentkorpus konzipiert wurde:

$$w_{ij} = tf_{ij} * idf_i$$

Die tf_{ij} *(term frequency)* gibt die Vorkommenshäufigkeit an, d.h. wie häufig ein Wort i im Dokument j erscheint. Die inverse Dokumenthäufigkeit idf_i *(inverse document frequency)* misst die allgemeine Bedeutung eines Wortes für die Gesamtmenge der betrachteten Dokumente. Die inverse Dokumenthäufigkeit *idf (inverse document frequency)* hängt allerdings nicht vom einzelnen Dokument, sondern von der Anzahl N der Dokumente im Korpus ab.

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

Das n_i steht für die Anzahl der Dokumente, die das Wort i beinhalten. Diese für Dokumente entwickelte Gewichtung wurde für eine möglich bessere Visualisierung auf die Präpositionsvektoren übertragen. Für das tf_{ij} werden hierfür, die gegebenen Häufigkeitswerte der Kontextwörter bezüglich der Präpositionen eingesetzt. Der andere Faktor idf_i für die Gewichtung wird unter Verwendung der Anzahl aller Präpositionsvektoren N und der Anzahl an Präpositionsvektoren n_i die in den Daten ein Wert für das Kontextwort i enthalten, ermittelt. Diese Visualisierung scheint schon eher eine Struktur aufzuweisen als die ohne Gewichtung, z.B. erkennt man, dass die Punktdarstellungen der Präpositionen *seit* und *bis* näher aneinander abgebildet sind als zu gewissen anderen Punkten. Dennoch sind die Punktdarstellungen der Präpositionen nach wie vor zu regelmäßig verteilt und ermöglichen noch keinen guten Analyseansatz für Präpositionen.

Alternativ zu den Häufigkeitswerten können auch die gegebenen LMI-Werte der Kontextwörter für die Umsetzung der zweidimensionalen Projektion verwendet werden. Die in das t-SNE eingesetzte Ähnlichkeitsmatrix hat hierfür die Ähnlichkeiten zwischen den Präpositionen mit den LMI-Werten zu den Kontextwörtern berechnet statt mit den Häufigkeitswerten. Die Abbildung 4.2, auf der diese Visualisierung mit den LMI-Werten dargestellt wird, ist deutlich anschaulicher als die beiden prototypischen Visualisierungen davor. Es ist beispielsweise gut zu erkennen, dass die Präpositionen *ab*, *bis* und *seit* nahe beieinander abgebildet sind. Die genannten Präpositionen, die vor allem für Angaben im temporalen Gebrauch genutzt werden, wurden in Nähe des anderen abgebildet, da sie scheinbar häufig in einem ähnlichem Kontext verwendet werden.

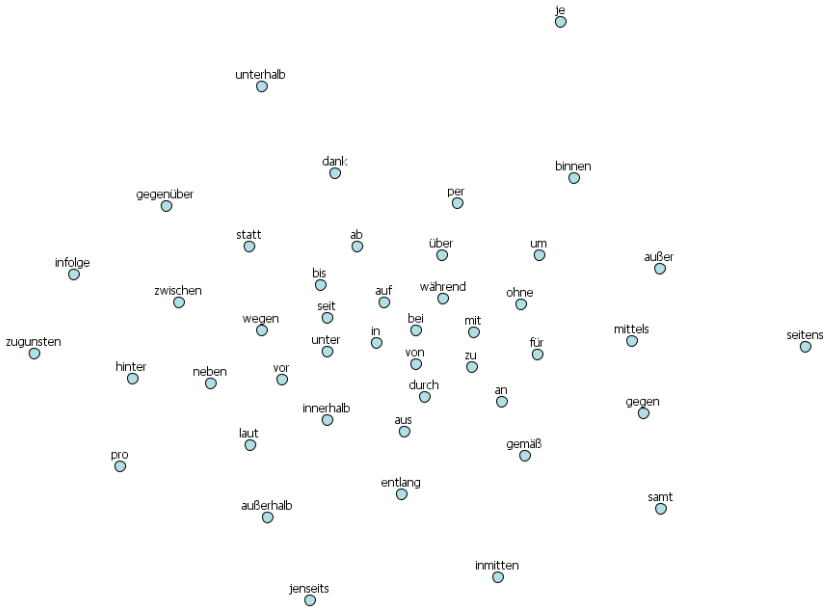

Abbildung 4.1.: Visualisierung mittels einer Ähnlichkeitsmatrix, die durch Häufigkeitswerte der Kontextwörter berechnet wurde (oberes Bild); tf-idf-Gewichtung der Häufigkeitswerte für Berechnung der Ähnlichkeitsmatrix verwendet (unteres Bild)

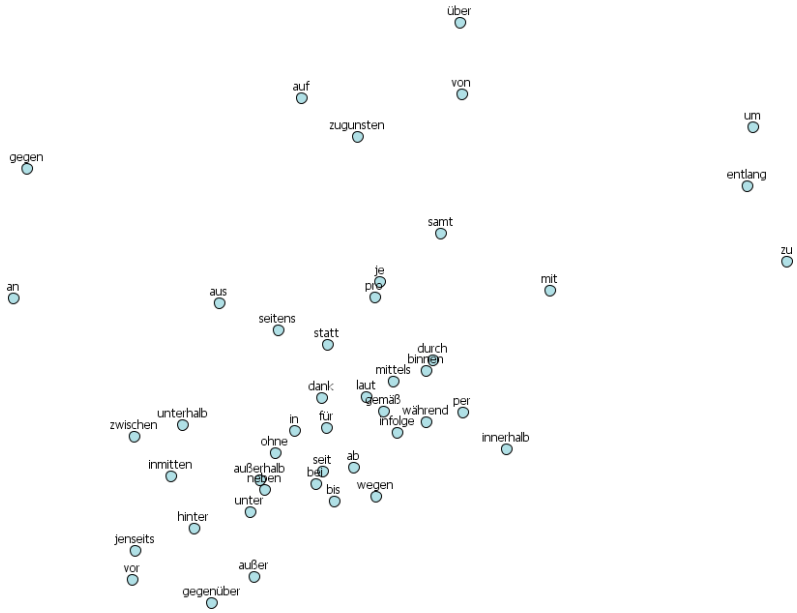

Abbildung 4.2.: Visualisierung mittels einer Ähnlichkeitsmatrix, welche durch LMI-Werte der Kontextwörter berechnet wurde

4.3. Interaktionsmethoden

Im folgendem Unterkapitel werden die wichtigsten, für eine genauere Exploration der Präpositionen, erstellten Interaktionskonzepte und die Überlegungen dahinter erläutert.

4.3.1. Anzeige der "echten" Abstände

In den vorherigen Abschnitten wurde beschrieben wie man durch das t-SNE Verfahren eine zweidimensionale Darstellung der Präpositionskontexte erreichen kann. Allerdings liefert das t-SNE nur ein grobes Abbild der wirklichen Distanzen bzw. Ähnlichkeiten zwischen den Präpositionen. Dem Nutzer soll daher ermöglicht werden, sich für eine bestimmte Präposition die tatsächlichen Abstände zu den anderen anzeigen zulassen. Die grundlegende Idee für diese Interaktionsmethode stammt aus der Arbeit "Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis" von Chuang et al. [CRMH12], auf die im vorherigen Kapitel 3 näher eingegangen wurde.

Die "echten" Abstände für eine bestimme Präposition werden aus der Ähnlichkeitsmatrix gefolgert, d.h. die "echten" Ähnlichkeitswerte werden aus der entsprechenden Zeile in der Matrix entnommen. Diese aus der Ähnlichkeitsmatrix entnommenen Werte werden zuerst invertiert, um sie als Distanzen

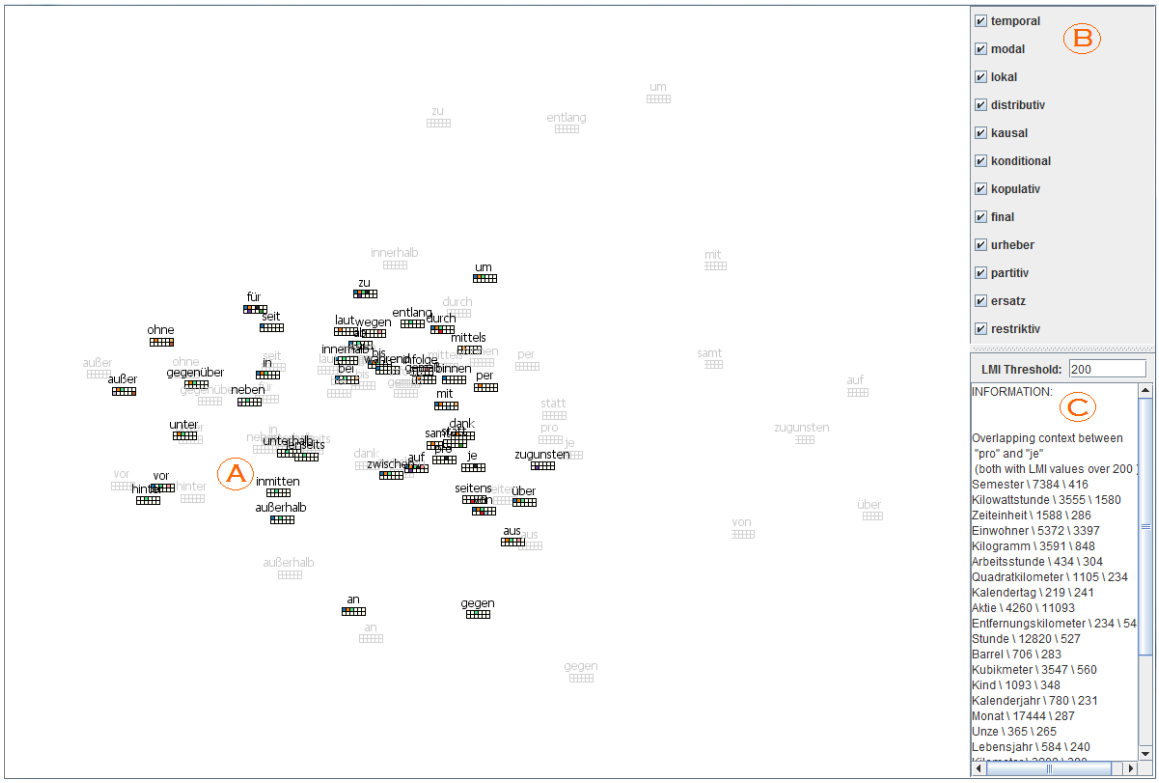

Abbildung 4.3.: Ausschnitt des entwickelten interaktiven Analyseansatzes in dem die "echten" Abstände für die Präposition *inmitten* zu den anderen angezeigt wird

des hochdimensionalen Raums zu nutzen. Somit ergibt sich beispielsweise aus einer Ähnlichkeit von 0.9 zwischen zwei Präpositionen eine Distanz d_{hd} von 0.1 oder aus einer Ähnlichkeit von 0.2 eine Distanz d_{hd} von 0.8. Da die Werte auch nach der Inversion nur im Intervall von -1 bis 1 liegen, werden diese nicht direkt als die Abstände von der fokussierten Präposition aus zu jeder anderen Präposition verwendet. Für die Normalisierung werden die *k* nächsten Nachbarn der fokussierten Präposition im zweidimensionalen Raum verwendet, ähnlich wie beim bekannten *k-Nearest-Neighbor*-Algorithmus (KNN) vorgegangen wird. Bei den anfänglichen Versuchen die Interaktion ohne einer *k-Nearest-Neighbor*-Normalisierung zu realisieren, waren die "echten" Abstände der Präpositionen oft so groß, dass einige Präpositionen sogar außerhalb des Displays des Prototypen positioniert wurden und somit für den Nutzer nicht mehr sichtbar gewesen wären. Ein Versuch darunter war es für die Normalisierung den Mittelwert aller Abstände zwischen den Präpositionen im zweidimensionalem Raum zu verwenden. Die folgende Methode wurde letzten Endes für die "echten" Abstände d_{2d} zwischen den Präpositionen ermittelt:

$$d_{2d} = (\frac{d_{hd}}{mean_{hd}}) * mean_{2d-kNN}$$

Für die Berechnung der "echten" Abstände d_{2d} werden neben den berechneten Distanzen d_{hd}, noch der Mittelwert $mean_{hd}$ aller hochdimensionalen Distanzen und der Mittelwert der euklidischen

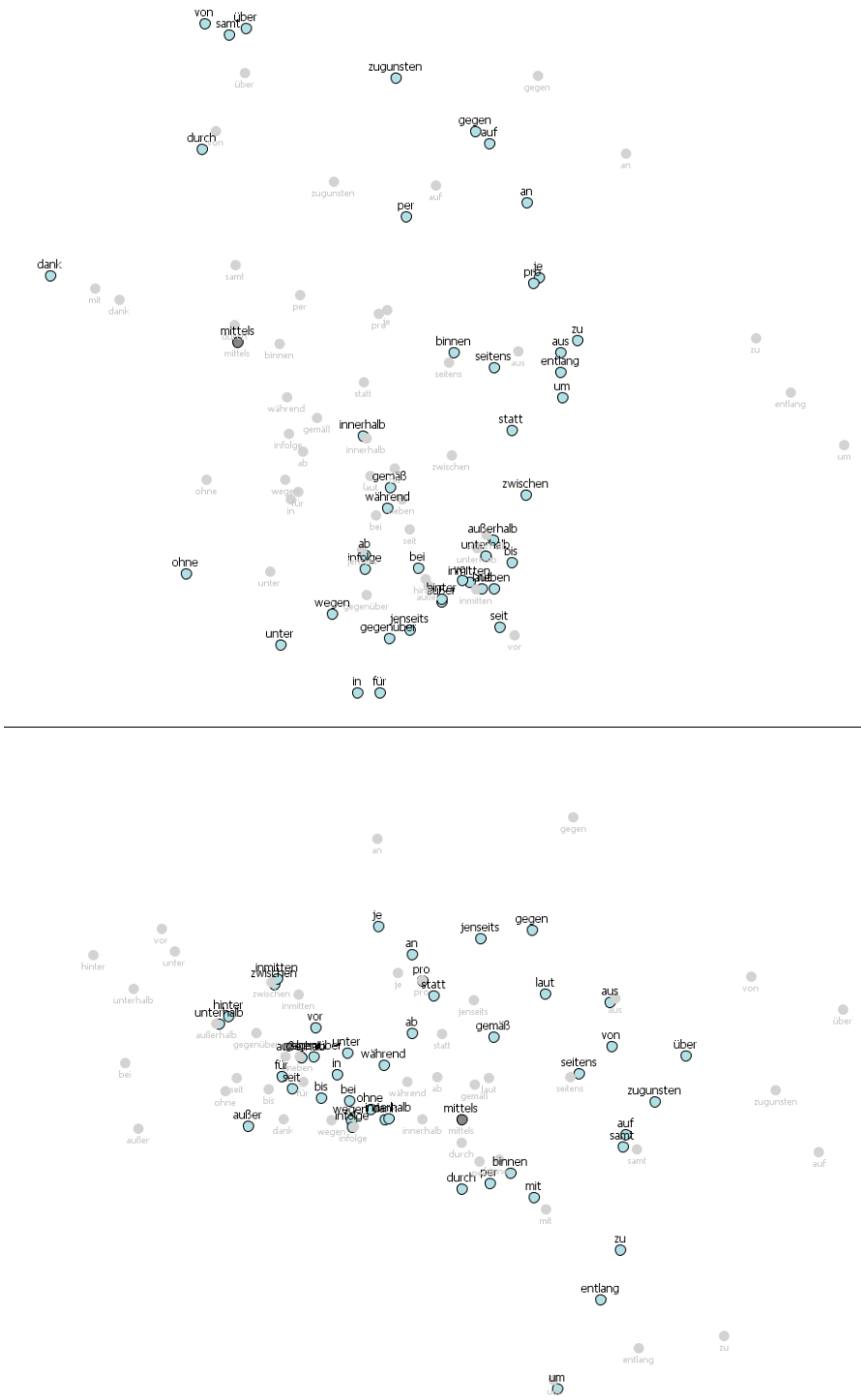

Abbildung 4.4.: Anzeige der tatsächlichen Abstände für die Präposition $mittels$ zu den anderen: Distanzen d_{hd} aus den Werten einer, durch LMI-Werte ermittelte, Ähnlichkeitsmatrix (oberes Bild); Distanzen d_{hd} aus den Werten einer, durch Häufigkeitswerte ermittelte, Ähnlichkeitsmatrix (unteres Bild)

Distanzen von den *k-Nearest-Neighbor*-Präpositionen zur fokussierten Präposition miteinbezogen. Als die *k-Nearest-Neighbors* wurden hier die *5* am nächsten stehende Präpositionen zur fokussierten Präposition aus der zweidimensionalen Projektion gewählt.

In der Abbildung 4.4 sind zwei Prototypen zu sehen in der jeweils die "echten" Abstände zur Präposition *mittels* angezeigt werden. Die grau hinterlegten Punkte in der Abbildung zeigen die Positionen der Präpositionen, die sich nach Anwendung des t-SNE Verfahrens ergeben. Für die obere Visualisierung der Abbildung 4.4 wurden die Distanzen d_{hd} aus den Werten einer Ähnlichkeitsmatrix ermittelt, die mit den LMI-Werten berechnet wurde. Allerdings ist die daraus resultierende Anzeige der "echten" Abstände unter Verwendung der oben beschriebenen Methode weniger zufriedenstellend. Die Punkte entfernen sich zu stark von ihren ursprünglichen Positionen. Die untere Visualisierung der Abbildung 4.4 ermittelt die Distanzen d_{hd} aus den Werten einer Ähnlichkeitsmatrix, die mit den Häufigkeitswerten der Kontextwörter berechnet wurde. Hier wurde ebenfalls die oben beschriebene Methode verwendet und diese Visualisierung hat sich als geeigneter erwiesen, da die Präpositionen sich nicht mehr so stark von ihren ursprünglichen Positionen entfernen. Der mit einem *A* gekennzeichneten Teil der Abbildung 4.3 stellt die schlussendliche Umsetzung der beschriebenen Interaktionsmethode dar. Diese Interaktion zeigt in der Abbildung 4.3 die "echten" Abstände für die Präposition *inmitten* zu den anderen an.

4.3.2. Analyse der Ambiguität

Um eine Analyse der Ambiguität von Präpositionen zu ermöglichen, entstand die Idee in der Visualisierung auch die jeweiligen zugehörigen semantischen Klassen farblich anzeigen zu lassen. Jeder semantischen Klasse wird dafür eine Farbe zugeordnet. Da eine Präposition aus den gegebenen Daten maximal zu sechs verschiedenen semantischen Klassen gehören kann, war die erste Überlegung eine Darstellung mit sechs "Klötzen" zu wählen. Die Wahl auf eine Darstellung mit zwölf "Klötzen" in der jede semantischen Klasse auch eine bestimmte Position zugeordnet wird, entstand durch die Überlegung auch einen Vergleich der Ambiguitäten zwischen den Präpositionen zu ermöglichen.

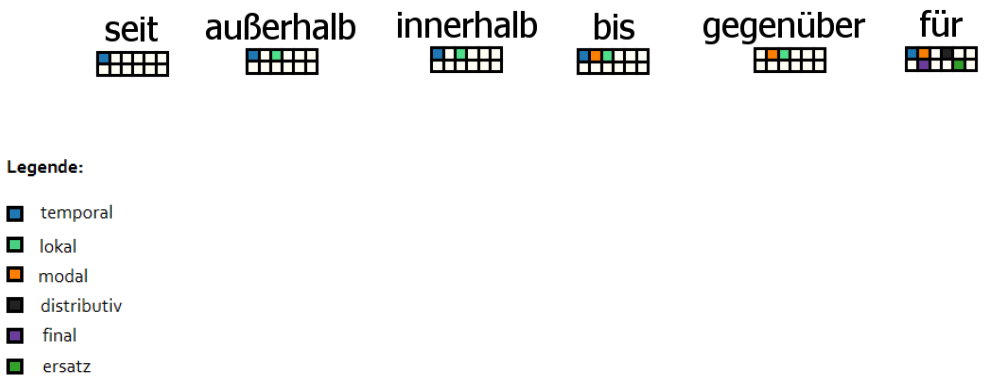

Abbildung 4.5.: Idee für den Vergleich der Ambiguität von Präpositionen

Die visuelle Analyse soll somit durch das Erkennen von Mustern unterstützt werden. Zusätzlich sollte ein interaktives Ein- und Ausschalten der einzelnen semantischen Klassen für den Nutzer möglich sein, falls für diesen nur bestimmte Klassen bei der Analyse im Fokus stehen. In der Abbildung 4.5 erkennt man beispielsweise sofort das gleiche Farbmuster für die Präpositionen *außerhalb* und *innerhalb*. Beide Präpositionen werden für temporale und lokale Zwecke verwendet. Auch dass die Präpositionen *bis* und *gegenüber* zwei semantische Klassen gemeinsam haben und *für* zu mehr Klassen gehört als alle andere abgebildeten, ist ebenfalls sofort erkennbar. Diese Farbmuster ermöglichen es weitere Unterschiede und Gemeinsamkeiten zwischen Präpositionen schnell festzulegen. In der Abbildung 4.3 ist die Umsetzung des zuvor beschrieben Interaktionskonzepts (siehe *B*) zu finden.

4.3.3. Vergleich gemeinsamer Kontextwörter

Eine weitere Interaktionsmethode zur Unterstützung der visuellen Analyse von Präpositionen soll die Anzeige von Kontextwörtern mit den zugehörigen LMI-Werten sein. Dem Nutzer soll ermöglicht werden zwei Präpositionen in der Visualisierung zu wählen, um die gemeinsamen Kontextwörter angezeigt zu bekommen. Zusätzlich dazu soll es dem Nutzer auch möglich sein, einen Grenzwert einzugeben, um die Ausgaben der Kontextwörter zu filtern, d.h. gibt der Nutzer die Zahl 50 ein, so erhält dieser nur Kontextwörter zu den Präpositionen die ein LMI-Wert größer oder gleich 50 haben.

In der Abbildung 4.3 stellt der Teil, der mit einem *C* gekennzeichnet ist, die Umsetzung der beschriebenen Interaktion dar. Der realisierte Prototyp hat hier die gemeinsamen Kontextwörter von den, in diesem Fall, gewählten Präpositionen *je* und *pro* ausgegeben. Der Grenzwert wurde in diesem Fall auf 200 gesetzt, d.h. dass nur Kontextwörter mit LMI-Werten über 200 angezeigt werden. Der Vergleich durch Kontextwörter ermöglicht es Schlüsse zu ziehen in welchen Kontext beide auftreten.

5. Implementierung

Dieses Kapitel beschreibt die Umsetzung des im vorherigen Kapitel vorgestellten interaktiven Visualisierungskonzepts. Wichtige Komponenten des entstandenen Ansatzes, werden hierfür genauer beschrieben.

5.1. Realisierung der interaktiven Visualisierung

Allgemein

Der prototypische Ansatz für diese Arbeit wurde komplett in der Programmiersprache Java implementiert. Für die zweidimensionale Visualisierung der Präpositionskontexte und den interaktiven Funktionalitäten, wurde unter anderem das Software-Werkzeug *Prefuse* verwendet. Der Nutzer kann sich mit dem Prototypen eine zweidimensionale Visualisierung der hochdimensionalen Präpositionskontexte anzeigen lassen. Hier hat er die Möglichkeit sich diese Visualisierung entweder mit den Verben oder den Nomen als Kontextwörter erstellen zu lassen. Da das zur Dimensionsreduktion verwendete t-SNE Verfahren nur ein verzerrtes Abbild der tatsächlichen Abstände zwischen den Präpositionen liefert, ist es dem Nutzer auch möglich sich die "echten" Abstände einer Präposition zu den anderen anzeigen zu lassen. Diese Interaktion wird durch eine Animation und durch die Anzeige der Anfangspositionen in grauer Farbe unterstützt. Durch das Anklicken einer Präposition wird die Animation angehalten und die Präpositionen bleiben auf der Position, auf welcher sie sich zu diesem Zeitpunkt befunden haben. Wählt der Nutzer eine weitere Präposition in der Visualisierung aus, so werden die gemeinsamen Kontextwörter beider angeklickten Präpositionen und deren LMI-Werte vom Prototyp ausgegeben. Es werden nur Kontextwörter ausgegeben, die größer gleich dem Grenzwert sind, die der Nutzer beliebig setzen kann. Die Darstellung der Position der Präpositionen wurde wie im vorherigen Kapitel vorgenommen, umgesetzt (siehe Abbildung 5.1). Der Nutzer kann entscheiden, welche semantische Klassen er bei der Visualisierung farblich angezeigt haben möchte, indem er die entsprechenden Checkboxen auswählt. Die in den nächsten Abschnitten vorgestellten Klassen sollen zu einem Grundverständnis der Realisierung des Prototyps verhelfen.

Umsetzung der Präpositionsvektoren

Die *Google Guava Library* wurde eingebunden, da diese eine spezielle Implementierung einer Tabelle, namens *HashBasedTable*, enthält. Diese *HashBasedTable* basiert im wesentlichen auf der folgenden verschachtelten Hashtabelle: *HashMap<R, HashMap<C, V»*. Das Listing 5.1 zeigt in vereinfachter Form wie diese *HashBasedTable* verwendet wurde, um die für diese Arbeit notwendigen Präpositionsvektoren zu erstellen. Durch die Methode *create()* wurde eine neue *HashBasedTable* erzeugt, die

Listing 5.1 Gekürzte Darstellung der Klasse PrepContextVectors

```
// create a new HashBasedTable
Table<String, String, Integer> prepContextTable = HashBasedTable.create();

// update frequency values
if (prepContextTable.contains(prep, context)) {
    prepContextTable.put(prep, context, frq);
} else {
    prepContextTable.put(prep, context, 0);
}
```

als Zeilennamen die Präpositionen, als Spaltennamen die Kontextwörter und als Tabelleneinträge die Häufigkeitswerte bzw. die LMI-Werte verwendet. Anschließend wird die Tabelle mit den Werten, aus den gegebenen Daten für diese Arbeit, gefüllt. Die Methode *contains(prep, context)* überprüft das Vorkommen eines Wortpaares aus einer Präposition und einem Kontextwort und trägt für den Fall *true* den gegeben Wert dazu ein. Ein großer Vorteil dieser *HashBasedTable* ist der vereinfachte Zugriff auf die Tabellenwerte über die entsprechenden Zeilen- und Spaltennamen. Das kann beispielsweise von Nutzen sein, wenn die gemeinsamen Kontextwörter und deren Werte von zwei Präpositionen ermitteln werden sollen.

Umsetzung der Visualisierung

Für die Visualisierung von Daten stellt *Prefuse* unter anderem die Tabelle *prefuse.data.Table* zur Verfügung, wie im Kapitel 2 schon näher erläutert. Im Listing 5.2 wird in gekürzter Form gezeigt, wie das Schema dieser Tabelle für den Prototypen dieser Arbeit aussieht. In jeder Zeile der Tabelle ist somit mindestens eine Präposition und deren Koordinaten für die Darstellung einer zweidimensionalen Projektion enthalten.

Listing 5.2 Gekürzte Darstellung der Klasse Prep2DTable

```
public class Prep2DTable extends Table {

    public Prep2DTable() {

       // setup table schema for display
       this.addColumn("Preposition", String.class);
       this.addColumn("x-Coordinate", double.class);
       this.addColumn("y-Coordinate", double.class);
    }
}
```

Im Listing 5.2 ist eine Anpassung der Klasse *Display* für den Prototypen zu sehen, wenn auch nur in einer verkürzten und vereinfachten Form. Die Tabelle wird der Klasse *PrepDisplay* übergeben. Diese erzeugt aus jeder Zeile ein *VisualItem* und ermöglicht die Visualisierung dieser *VisualItems* auf dem Bildschirm. Durch die Nutzung von *Actions* werden Eigenschaften der Visualisierung festgelegt. Die *Action AxisLayout* weist den *VisualItems* Positionen entlang einer einzigen Dimension (x oder y) zu. Die x- und y-Koordinaten werden hier aus der übergebenen Tabelle entnommen. Eine weitere

Listing 5.3 Gekürzte Darstellung der Klasse PrepDisplay

```
public class PrepDisplay extends Display {

   public static Visualization vis = new Visualization();
   public String prepPoints = "prepPoints";

   public PrepDisplay(Table data) {
      super(vis);

      // setting up the visualized data
      vis.add(prepPoints, data);

      AxisLayout x_axis = new AxisLayout(prepPoints, "x-Coordinate", Constants.X_AXIS,
          VisiblePredicate.TRUE);

      AxisLayout y_axis = new AxisLayout(prepPoints, "y-Coordinate", Constants.Y_AXIS,
          VisiblePredicate.TRUE);

      // actions to process the visual data
      ColorAction strokeColor = new ColorAction(prepPoints, VisualItem.STROKECOLOR,
          ColorLib.rgb(0, 0, 0));

      // these actions are combined to an ActionList
      // and linked to the Visualization
      ActionList draw = new ActionList();
      draw.add(x_axis);
      draw.add(y_axis);
      draw.add(strokeColor);
      vis.putAction("draw", draw);

      // launching the visualization
      vis.run("draw");
   }
}
```

Action ist die *ColorAction*, mit welcher man die Färbung der *VisualItems* bestimmen kann. Im Listing 5.3 sieht man die Anwendung der *ColorAction* für die Färbung der Umrandungen (*STROKECOLOR*) der *VisualItems*. Eine sogenannte *ActionList* repräsentiert mehrere *Actions*. Durch Methoden wie *run(String action)*, *runAfter(String before, String after)*, *alwaysRunAfter(String before, String after)* oder *runAt(String action, long startTime)*, kann festgelegt werden zu welchem Zeitpunkt bzw. in welcher Reihefolge die entsprechenden *Actions* oder *ActionLists* ausgeführt werden sollen.

Renderer für die Anzeige der semantischen Klassen

Wie in Kapitel 2 schon beschrieben, liegt für diese Arbeit eine Datei vor mit Information, die eine ambige Zuordnung der Präpositionen zu semantischen Klassen enthält. Im vorherigen Kapitel wurde eine Idee beschrieben, wie die Anzeige der semantischen Klassen pro Präposition aussehen könnte. Da *Prefuse* keinen *Renderer* in dieser Form für die *VisualItems* beinhaltet, musste ein *Renderer* implementiert werden, welches diese bestimmte Darstellung mit den zwölf "Klötzen" pro Präposition darstellen

Listing 5.4 Gekürzte Darstellung der Klasse SquaredShapeRenderer

```
public class SquaredShapeRenderer extends ShapeRenderer {

   @Override
   public void render(Graphics2D g, VisualItem vi) {

            ItemShapeContent[] shapes = getRawShapeElements(vi);

            if (shapes!=null){
                  GraphicsTools.drawGraphics(g, vi, shapes, getStroke(vi),  getRenderType(vi));
   }
}
```

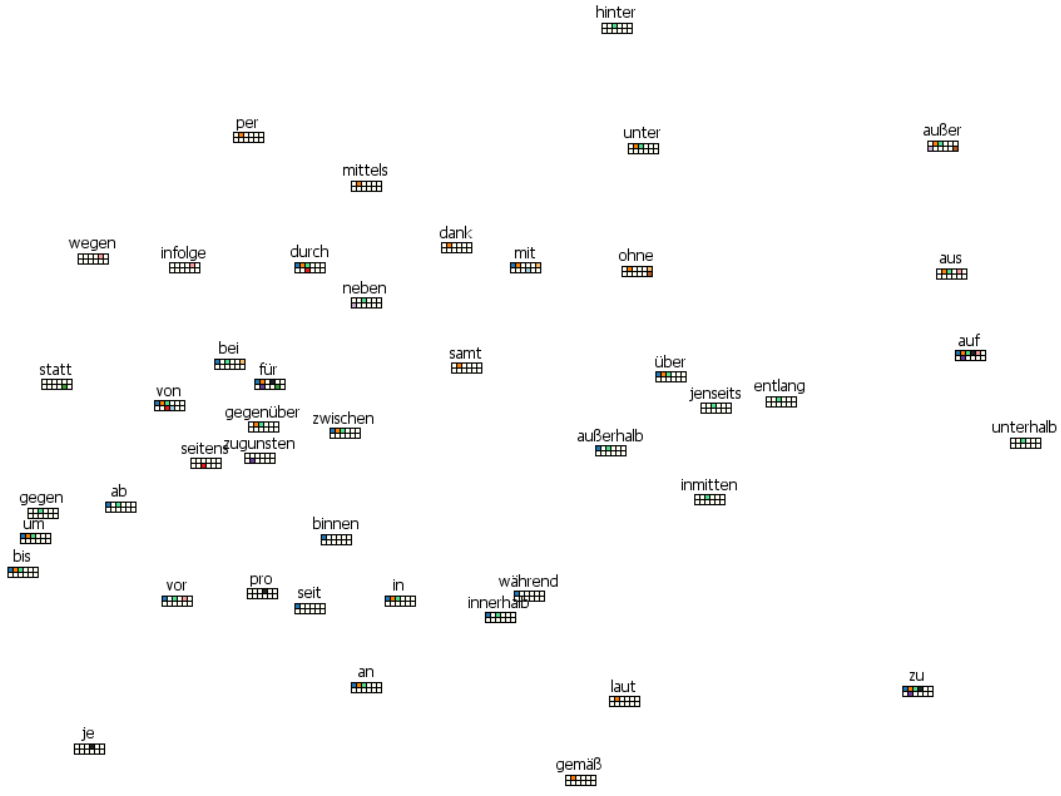

Abbildung 5.1.: Zweidimensionale Visualisierung der Präpositionen mit ihren zugehörigen semantischen Klassen

kann. In der Standardeinstellung von *Prefuse* kann man einem *VisualItem* immer nur eine Farbe und eine Form (*Shape*) zuordnen. Die, für die Präpositionen, gewünschte Darstellung erfordert jedoch mehr als eine Farbe und eine Form, denn sie besteht im Prinzip aus zwölf Quadraten, die jeweils eine andere Farbe annehmen können. Um diese zwölf Quadrate pro *VisualItem* zur realisieren, musste somit ein eigener *Renderer* entwickelt werden. Für die Realisierung des Prototyps entstand somit die Klasse *SquaredShapeRenderer*, die von der *Prefuse*-Klasse *ShapeRenderer* erbt. In dieser Klasse wird die Methode *render(Graphics2D g, VisualItem vi)* von der *Prefuse*-Klasse *Renderer* überschrieben (siehe Listing 5.4), da diese nicht mehrere Formen zeichnen kann. In der ursprünglichen *render*-Methode wurde die Methode *paint(Graphics2D g, VisualItem item, Shape shape, BasicStroke stroke, int typ))* von der *Prefuse*-Klasse *GraphicsLib* aufgerufen. Diese musste ebenfalls neu implementiert werden um, das "Rendern" von mehreren Formen zu ermöglichen. Der neuen Methode *GraphicsTools.drawGraphics(g, vi, shapes, getStroke(vi), getRenderType(vi)* können mehrere Formen zum zeichnen übergeben werden. Die Anzahl, Anordnung und Form der zwölf "Klötze" werden in der Methode *getRawShapeElements* des *SquaredShapeRenderer* festgelegt. Die daraus resultierende Visualisierung für die Präpositionen ist in Abbildung 5.1 zu finden.

6. Experten-Feedback

In diesem Kapitel wird die Evaluation des interaktiven Analyseansatzes beschrieben. Dafür wurde die Meinung von Experten im Bereich der Sprachanalyse eingeholt.

6.1. Beschreibung der Durchführung

Zu Beginn der Evaluation wurde den teilnehmenden Experten die verschiedenen Funktionalitäten des entstandenen Prototyps vorgeführt. Anschließend wurde den Teilnehmern ein Fragebogen vorgelegt um eine Einschätzung der Nutzbarkeit des Prototyps zu erhalten. Der Fragebogen ist im Anhang dieser Arbeit zu finden. Im Wesentlichen sollte dieser Fragebogen vor allem als Leitfaden für Meinungen und Verbesserungsvorschläge von Seiten der Teilnehmer dienen. Im ersten Teil des Fragebogens sollten Aussagen, wie beispielsweise *Die 2D-Visualisierung der Präpositionskontexte ist für eine erste Analyse nützlich.* mit einem Wert aus einer 5-stufigen Skala von *trifft gar nicht zu* zu *trifft voll zu* bewertet werden. Im zweiten Teil des Fragebogens sollten allgemeine Fragen, wie beispielsweise *Welche weiteren Interaktionen wären nützlich für die Analyse von Präpositionen?*, beantwortet werden.

6.2. Erkenntnisse

Durch das Experten-Feedback konnten mehrere Erkenntnisse bezüglich der Nutzbarkeit gewonnen werden. Nach der Meinung der Sprachanalyseexperten ist die zweidimensionale Visualisierung der Präpositionskontexte für eine erste Analyse hilfreich, jedoch nicht für eine tiefer gehende Analyse. Auch die Anzeige der semantischen Klassen pro Präposition ist laut den Experten bei der Analyse von Nutzen, jedoch würde man sich wegen der hohen Anzahl an semantischen Klassen selten alle auf einmal anzeigen lassen. Eine Visualisierung mit kreisförmigen Glyphen für die Präpositionen statt der Anzeige der semantischen Klassen, wurde nicht als besser empfunden, außer für den Fall, dass die semantischen Klassen für die Analyse mal nicht relevant sind. Als eines der nützlichsten Funktionalitäten wurde die interaktive Anzeige der "echten" Abstände genannt, da diese eine tiefere Analyse der Präpositionen ermögliche. Die Experten erklärten, dass im Normalfall bei der Analyse von Präpositionen nur ein Teil dieser ausgewählt werden, um anschließend deren Abstandswerte zueinander in einer Tabelle o.Ä. zu vergleichen. Aufgrund dessen wurde von den Experten die Interaktion als ein hilfreiches Werkzeug für die Analyse der Abstände zwischen Präpositionen gesehen. Die, bei der Anzeige der "echten" Abstände, als grau hinterlegten ursprünglichen Positionen der Präpositionen stellten eine leichte Hilfe bei der Analyse dar. Die Ausgabe gemeinsamer Kontextwörter von zwei Präpositionen, mit der interaktiven Grenzwerteingabe bezüglich der LMI-Werte, wurde ebenfalls als eines der nützlichsten Funktionalitäten bewertet. Ausschlaggebend für die Analyse ist

für die Experte vor allem die Ausgabe der Kontextwörter selbst, weniger die vereinzelten LMI-Werte. Da sich die LMI-Werte als weniger hilfreich für den Vergleich zweier Präpositionen erwiesen, kam von Seiten der Experten die Idee auf, möglicherweise *Pointwise Mutual Information*-Werte zu den Kontextwörtern anzeigen zu lassen.

Für die Experten stellte der interaktive Ansatz im Ganzen ein gutes Werkzeug für die Analyse von Präpositionen dar. Auch Verbesserungsvorschläge wurden für den Fall einer Erweiterung des Ansatzes geäußert. Ein Vorschlag war das Analysewerkzeug generischer zu entwickeln, so dass die Möglichkeit, noch weitere Daten für die Analyse einsetzen zu können, entsteht. Das Abspeichern der zweidimensionalen Visualisierungen und ein interaktiver Wechsel der verschiedenen Darstellungen der Präpositionen waren weitere Verbesserungsvorschläge.

7. Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde ein interaktiver Ansatz für die Analyse von Präpositionen entwickelt. Dazu gehörte die Entwicklung einer geeigneten zweidimensionalen Visualisierung der Präpositionskontexte und von Interaktionsmethoden, die eine genauere Exploration der Präpositionskontexte erlauben sollen.

Um das Weiterarbeiten mit den hochdimensionalen Präpositionskontexten zu vereinfachen werden diese in hochdimensionale Vektoren umgewandelt. Für die Umsetzung der zweidimensionalen Projektion der Präpositionskontexte werden zuallererst mit Hilfe der erzeugten Präpositionsvektoren die Ähnlichkeiten zwischen den Präpositionen ermittelt. Diese Werte werden anschließend in einer Ähnlichkeitsmatrix gespeichert. Dem für diese Arbeit ausgewählten Verfahren t-SNE wird die Ähnlichkeitsmatrix übergeben, damit anschließend die Berechnung der zweidimensionalen Projektion der hochdimensionalen Daten erfolgen kann. Da das t-SNE nur ein grobes Abbild der wirklichen Ähnlichkeiten zwischen den Präpositionen liefert, ist es dem Nutzer unter anderem durch Interaktion möglich, sich für eine bestimmte Präposition die "echten" Abstände zu den anderen anzeigen zulassen. Die Umsetzung des interaktiven Ansatzes wurde mit Hilfe des Software-Werkzeugs *Prefuse* realisiert.

Für die Evaluation des interaktiven Ansatzes wurde die Meinung von Experten im Bereich der Sprachanalyse eingeholt. Die Experten empfanden die interaktive Anzeige der tatsächlichen Abstände und die Ausgabe der gemeinsamen Kontextwörter von zwei Präpositionen, als die nützlichsten Interaktionen für die Analyse.

Ausblick

Von Seiten der in der Evaluation teilnehmenden Experten kamen schon die ersten Ansatzpunkte für zukünftige Erweiterungen des erstellten Prototyps, wie beispielsweise, das Abspeichern der zweidimensionalen Visualisierungen zu ermöglichen.

Neben den in der Evaluation entstandenen Verbesserungsvorschlägen, gäbe es noch einen weiteren Ansatzpunkt zur Optimierung der aktuellen Implementierung in Bezug auf die Visualisierung der Präpositionen. In den zweidimensionalen Projektionen kommt es teilweise zur einer Überlappung der Labels von den visualisierten Präpositionskontexten. Vor allem die Anzeige der "echten" Abstände einer Präposition zu den anderen führt manchmal zu mehreren Überlappungen der Labels. Dadurch wird die Übersichtlichkeit der Visualisierung eingeschränkt. Da in vielen Informationsvisualisierungen die Labels einen wesentlichen Bestandteil darstellen, um visualisierte Daten zu verstehen, haben sich Luboschik et al. in ihrer Arbeit "Particle-Based Labeling: Fast Point-Feature Labeling without Obscuring Other Visual Features" mit verschiedenen Ansätzen bezüglich der Platzierung von Labels

beschäftigt. Einer ihrer Ansätze vermeidet die Überlappung der Labels mit Hilfe einer Spirale. In diesem Ansatz wird eine Spirale festgelegt, die aus einer bestimmten Anzahl an Punkten besteht. Bei einer Überlappung zweier Labels beispielsweise wird eines der Labels mit Hilfe Spirale neu positioniert. Dabei wird jeder Punkt der Spirale von innen nach außen durchgegangen bis eine neue Position gefunden wird, bei der keine Überlappung der Labels mehr stattfindet. Allerdings gehen dadurch die exakten Position der Präpositionslabels verloren, deshalb sollte man um die Aussagekraft der Visualisierung zu erhalten, die neu platzierten Labels durch Linien mit den Glyphen der Präpositionen verbinden.

A. Anhang

Fragebogen - Tool für die Analyse von Präpositionen

EVALUATION					
	trifft gar nicht zu	trifft eher nicht zu	teils/teils	trifft eher zu	trifft voll zu
Die 2D-Visualisierung der Präpositionskontexte ist für eine erste Analyse nützlich. **(zu Bild 1 bzw. 2)**	☐	☐	☐	☐	☐
Die Darstellung der Präpositionen mit ihren zugehörigen semantischen Klassen ist in dieser Form hilfreich für die Analyse. **(zu Bild 1)**	☐	☐	☐	☐	☐
Die kreisförmigen Glyphen für die Darstellung der Position von Präpositionen sind im Vergleich zur oben genannten Darstellung geeigneter für die Analyse. **(zu Bild 2)**	☐	☐	☐	☐	☐
Die interaktive Anzeige der tatsächlichen Abstände von Präpositionen trägt stark zu Analyse bei. **(zu Bild 3)**	☐	☐	☐	☐	☐
Die Anzeige der ursprünglichen Positionen ist hilfreich solange die genauen Abstände für eine ausgewählte Position angezeigt werden. **(zu Bild 3)**	☐	☐	☐	☐	☐
Die angezeigten Kontextwörter zu den beiden ausgewählten Präpositionen helfen bei der Interpretation der Analyse. **(zu Bild 4)**	☐	☐	☐	☐	☐
Die einzelnen LMI-Werte der Präpositionen zu den Kontextwörtern sind von Bedeutung für die Analyse. **(zu Bild 4)**	☐	☐	☐	☐	☐

1/4

Wie führen Sie die Analyse von Präpositionen normalerweise durch?

Welche weiteren Interaktionen wären nützlich für die Analyse von Präpositionen?

Wo sehen Sie Probleme oder Raum für Verbesserungen bezüglich der Präpositionsvisualisierung?

Was gefällt Ihnen besonders am Tool?

Zusätzliche Kommentare:

Bild 1: Darstellung der Präpositionen mit ihren zugehörigen semantischen Klassen (durch verschiedene Farben gekennzeichnet)

Bild 2: Kreisförmige Glyphen für die Darstellung der Position von Präpositionen

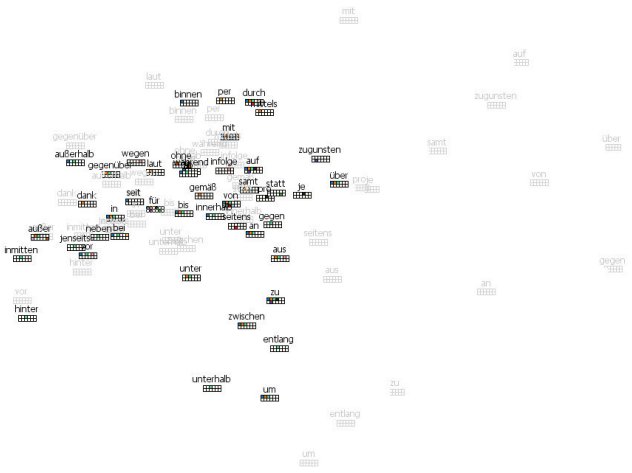

Bild 3: Anzeige der tatsächlichen Abstände von den Präpositionen zu „für“ und ihrer ursprünglichen Positionen (grau hinterlegt)

Bild 4: Anzeige der zugehörigen Kontextwörter für beide Präpositionen mit einem eingegebenen LMI-Grenzwert von 500

4/4

Literaturverzeichnis

[Bä07] K. Bäßler. Entwicklung eines Tools zur Gewinnung und Visualisierung von vernetzten Bibliographie-Daten. Technischer Bericht, Institut für Systemarchitektur, Technische Universität Dresden, 2007. (Zitiert auf Seite 16)

[BND+11] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner. Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLoS ONE Vol. 6.3, 2011. (Zitiert auf Seite 21)

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman. *Readings in Information Visualization: Using Vision to Think.* Morgan Kaufmann Publishers Inc., 1999. (Zitiert auf den Seiten 7 und 11)

[COL] Association Measures. URL http://www.collocations.de/AM/index.html. (Zitiert auf Seite 13)

[CRMH12] J. Chuang, D. Ramage, C. D. Manning, J. Heer. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012. (Zitiert auf den Seiten 7, 20, 22 und 30)

[GFVLD13] F. J. García-Fernández, M. Verleysen, J. A. Lee, I. Díaz. Stability Comparison of Dimensionality Reduction Techniques Attending to Data and Parameter Variations. EuroVis Workshop on Visual Analytics using Multidimensional Projections, 2013. (Zitiert auf den Seiten 22 und 26)

[HAF13] N. Heulot, M. Aupetit, J.-D. Fekete. ProxiLens: Interactive Exploration of High-Dimensional Data using Projections. EuroVis Workshop on Visual Analytics using Multidimensional Projections, 2013. (Zitiert auf den Seiten 7, 21 und 22)

[Hee04] J. M. Heer. Prefuse - A Software Framework for Interactive Information Visualization. Technischer Bericht, Computer Science Division, University of California, Berkeley, 2004. (Zitiert auf den Seiten 7 und 16)

[HWV09] F. van Ham, M. Wattenberg, F. B. Viégas. Mapping Text with Phrase Nets. IEEE Transactions on Visualization and Computer Graphics. 2009. (Zitiert auf den Seiten 7, 12 und 13)

[KKEM10] D. Keim, J. Kohlhammer, G. Ellis, F. Mansmann. *Mastering the Information Age - Solving Problems with Visual Analytics.* Eurographics Association, 2010. (Zitiert auf Seite 12)

[LNG] Computerlinguistik: Was ist das eigentlich? URL http://www.ims.uni-stuttgart.de/studium/interessierte/leitfaden. (Zitiert auf Seite 9)

[MH08] L. van der Maaten, G. Hinton. Visualizing Data using t-SNE. Journal of Machine Learning Research 9. 2008. (Zitiert auf den Seiten 14 und 27)

[MRS08] C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008. (Zitiert auf den Seiten 14, 25 und 28)

[SJ09] A. Sears, J. Jacko. *Human-Computer Interaction: Design Issues, Solutions, and Applications.* Taylor & Francis, 2009. (Zitiert auf Seite 11)

[SPI] IN-SPIRE Visual Document Analysis. URL http://in-spire.pnnl.gov/videos. (Zitiert auf den Seiten 7 und 19)

[ULP] Universität Leipzig. Informationsvisualisierung. URL http://www.informatik.uni-leipzig.de/bsv/homepage/sites/default/files/Infovis_1-intro.pdf. (Zitiert auf Seite 11)

[War00] C. Ware. *Information Visualization - Perception for Design.* Morgan Kaufmann Publishers, 2000. (Zitiert auf Seite 11)

[WTP+95] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow. Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. IEEE Information Visualization '95. 1995. (Zitiert auf den Seiten 19 und 26)

Alle URLs wurden zuletzt am 30. 09. 2014 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift