Verfahren zur Klassifizierung mehrsprachiger eMail-Anfragen zum Zwecke der automatischen Antwortgenerierung

Johannes Graën
johannes.graen@studenten.ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstraße 12
70174 Stuttgart

Studienarbeit Nr. 123

Beginn:	1. Juni 2011
Ende:	30. August 2011
Betreuer & Prüfer:	PD Dr. Ulrich Heid

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig verfaßt und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

Stuttgart, den 30. August 2011	
Ort, Datum	Unterschrift

Hinweis

Aus Datenschutzgründen wurden in allen Originalbeispielen die personenbezogenen Daten soweit durch * ersetzt, daß keine Rückschlüsse auf die zugehörige Person mehr möglich sind, die Art der ersetzten Daten (Name/Vorname, Zimmer-/Kontonummer, IP-/MAC-Adresse, …) aber weiterhin erkennbar bleibt. Automatische, maschinelle Anonymisierung zeichnet sich durch vollständige Ersetzungen der betroffenen Zeichenketten aus.

Zusammenfassung

Die vorliegende Arbeit beschreibt anhand eines spezifischen Falles wie bei einer Organisation eingehende eMail-Anfragen automatisch aufbereitet, klassifiziert und zur Antwort-Generierung genutzt werden können.

Die Struktur der eingehenden Texte wird dabei u.a. mithilfe eines Parsers bestimmt und aus den relevanten Teilen werden Wörter extrahiert. Statistische Verfahren nutzen diese, um die Anfrage bzgl. ihrer Sprache zu klassifizieren. Die dazu benötigten Vergleichsdaten stammen aus dem Korpus bereits beantworteter Anfragen.

Aus den extrahierten Wörtern der Texte wird pro Sprache eine hierarchische Cluster-Struktur generiert. Die jeweils größten kohärenten Cluster beschreiben ein Anliegen in Form von mit Wörtern verknüpften Häufigkeiten.

Unter Zuhilfenahme benutzerspezifischer externer Daten sowie gegebenenfalls auch menschlicher Interaktion können im Anschluß zu dem jeweiligen Anliegen eine Antwort und - sofern erforderlich - auch ein oder mehrere Aktionen generiert werden.

Die auf diese Art als korrekt bestätigten sowie manuell verworfenen Antworten können genutzt werden, um die Qualität der Klassifizierung durch stetig verfeinerte Kalibrierung der Parameter zu erhöhen. Vermittels manueller Klassifizierung könnten sich bei Bedarf so auch neue Klassen hinzufügen lassen.

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Der Verein Selfnet e.V.[1] betreibt ein Hochgeschwindigkeitsnetz, das sieben Stuttgarter Studentenwohnheime mit insgesamt über 2000 Netzwerkanschlüssen (i.d.R. einer pro Zimmer) untereinander sowie mit dem Internet verbindet. Die Netznutzer (Vereinsmitglieder) treten primär persönlich während der Sprechstunden mit den Netz- und Dienstbetreibern - d.h. der Aktivenschaft - in Kontakt. Zur Kommunikation von Informationen wird die vereinseigene Webseite sowie vor allem das Medium eMail verwendet. Außerhalb der Sprechstunden kommt seit 2003 ein *Issue-Tracking-System* (Siehe Vincent u. a. (2005), Kapitel 1) zum Einsatz, das per eMail ankommende Anfragen entgegennimmt und den in Frage kommenden Bearbeitern erlaubt, die Beantwortung der Anfragen zu koordinieren, den Kommunikationsverlauf (für gewöhnlich als *Ticket* bezeichnet) einzusehen und auf vorherige Kommunikationen zurückzugreifen.

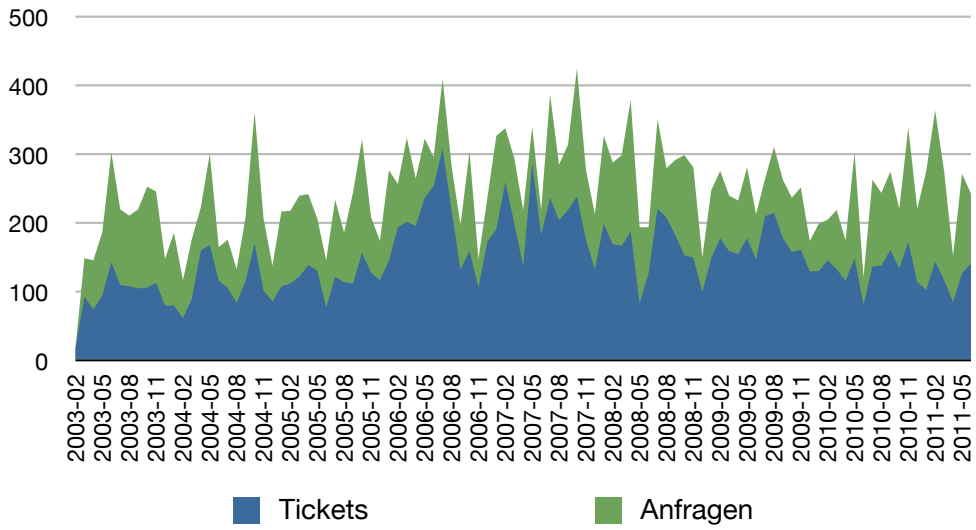

Abbildung 1: Anfrageaufkommen über den Einsatzzeitraum des Issue-Tracking-System (Tickets und eingehende Anfragen)

Im Schnitt erreichen das System monatlich 241 Anfragen, die 32 Stunden nach ihrem Eintreffen beantwortet werden, wobei 139 davon erstmalige An-

[1] Webseite des Vereins: `http://www.selfnet.de/`

fragen und der Rest Antworten auf von Bearbeitern geschriebene Antworten sind (siehe Abb. 1)[2].

Auf diese Weise sind bis Mai 2011 über 15.000 Anfragen beantwortet worden; die Anzahl der insgesamt geschriebenen Antworten liegt bei über 25.000.

Die Bearbeitungsdauer einer Anfrage - also die Zeit, die nötig ist, um nach dem Lesen einer Anfrage eine entsprechende Antwort verfassen zu können - beträgt im günstigsten Fall mehrere Minuten (wenn beispielsweise nur eine allgemeingültige Information erfragt wird), kann sich aber auch durch Rücksprache mit anderen Personen oder die Konsultierung schriftlicher Unterlagen sowie durch die Erstellung von Formularen zu Dokumentationszwecken über mehrere Stunden erstrecken. Eine Erfassung der in die Beantwortung der Anfragen investierten Arbeitszeit findet nicht statt, allerdings sind mehrere Bearbeiter mehrmals pro Woche damit beschäftigt, die aktuellen eMail-Anfragen zeitnah zu beantworten, so daß diese Tätigkeit einen nicht unwesentlichen Teil der Vereinsarbeit ausmacht.

Ein Teil der Anfragen sind spezieller Natur und erfordern eine individuelle Antwort, ein Großteil aber läßt sich einem jeweiligen Schema folgend abarbeiten. Dieser wiederum besteht aus:

1. Anfragen, die einen allgemeinen Informationsbedarf zum Ausdruck bringen (beispielsweise nach Öffnungszeiten, Formularen oder Vorgehensweisen zu möglichen Aktionen bzgl. des Vereins);

2. Anfragen, die einen spezifischen Informationsbedarf zum Ausdruck bringen (beispielsweise bzgl. eingegangener Zahlungen der Mitgliedsgebühr oder zur vereinsseitigen Konfiguration des jeweiligen Anschlusses);

3. Anfragen, die eine Aktion ausdrücken und eine Reaktion des Vereins bezwecken (beispielsweise Kündigung der Mitgliedschaft, Änderung individueller Daten oder Widerspruch gegen die Einzugsermächtigung).

Im Gegensatz zu Anfragen der ersten Art, erfordern die beiden letzteren ein Konsultieren der Mitglieder-Datenbank, vor allem auch, um die jeweils anfragende Person zu identifizieren, da Informationen aus jener Datenbank aus Datenschutzgründen nicht an Dritte herausgegeben werden dürfen. Gegebenenfalls ist in diesen beiden Fällen auch das Nachschlagen archivierter Informationen oder das Archivieren einer Aktion nötig.

Die Beantwortung aller wiederholt vorkommender - also nicht spezieller - Anfragen, die keine Interaktion mit nicht online verfügbaren Daten - i.d.R.

[2]Jeweils Median-Werte über die Daten von Februar 2003 bis Mai 2011.

dem Archiv in Papierform - erfordern, ist prinzipiell automatisierbar, sofern es gelingt, diese zweifelsfrei zu klassifizieren. Als Datenbasis kann hierbei zum einen die Anfrage selbst, aber auch die Mitglieder-Datenbank dienen, die alle persönlichen Informationen zur Verwaltung der Mitgliedschaft und alle technischen Daten zur Konfiguration des Netzwerkanschlusses enthält. Die Anfragen mit manuellem Handlungsbedarf lassen sich halbautomatisch beantworten, indem der Bearbeiter bestimmte Aktionen ausführt und bestätigt, so daß anschließend eine entsprechende Antwort erstellt werden kann.

Die Art der wiederholt auftretenden Kommunikationsziele bleibt i.d.R. über einen längeren Zeitraum konstant, auf die gesamten Betriebsjahre des Issue-Tracking-Systems gesehen werden allerdings einige Themen im Laufe der Zeit obsolet und andere, neuartige, kommen auf. Diese Dynamik kann Auswirkungen auf die Klassifizierung der Anfragen haben.

1.2 Ziele und Anforderungen

Alle eingehenden Anfragen sollen ihrem Charakter nach klassifiziert werden, wobei bei Unterschreiten eines Qualitätsgrenzwertes die Anfrage als nicht klassifizierbar gekennzeichnet werden soll. Dieser Wert kann aufgrund menschlicher Interaktion im laufenden Betrieb in Form von Akzeptieren oder Ablehnen vorgeschlagener Antworten in beide Richtungen korrigiert werden.

Die Klasse der allgemeinen Informationsanfragen kann mit vordefinierten Antworten (*Canned Text*[3]) zufriedenstellend gelöst werden. Im Falle der Aktionen kann je nach der vom Bearbeiter gewählten Lösung wiederum ein vorgefertigter Text - gegebenenfalls unter Einbeziehung spezifischer Daten - gewählt werden.

Eine Antwort auf das Erfragen individueller Informationen sollte, sofern diese nicht mangels Identifikation zurückgewiesen werden muß, die gewünschte Information beinhalten. Eine komplette Generierung eines Antworttextes - wie beispielsweise von der Meaning Text Theory (Mel'čuk (1988)) skizziert - übersteigt bei weitem die Erfordernisse und rechtfertigt hier den Aufwand nicht. Auch sollten bei der Texterzeugung die Bedürfnisse des durchschnittlichen Fragestellers beachtet und deshalb auf Einfachheit - sowohl des Vokabulars als auch der Textstruktur und Syntax - geachtet werden.[4] Es bietet sich an, Textbausteine absatzweise zusammenzufügen und - falls erforderlich - den Text auf oberflächenmorphologischer Ebene anzupassen. Um die Natürlichkeit des Textes zu gewähren, könnten mit Strukturattributen versehene Textalternativen zur Verfügung gestellt werden.

[3]Siehe Reiter und Mellish (1993).

[4]Siehe Abb. 31 im Anhang als Beispiel.

1.3 Das verwendete Issue-Tracking-System

Das im Verein eingesetzte Issue-Tracking-System nennt sich *Request Tracker*[5] (im folgenden *RT* genannt) und stammt von der Firma *Best Practical*[6]. Es besteht aus einer Datenbank[7] und - darauf aufbauend - einer Web-Oberfläche und einer eMail-Schnittstelle zur Interaktion mit den Benutzern (detailliertere Informationen finden sich in Vincent u. a. (2005), Kapitel 8).

Eingehende eMails erhalten den Status „neu“ bzw. „offen“, wenn sie sich auf über das *RT* geschriebene Antworten beziehen. Die Zuordnung erfolgt über eine eindeutige Nummer pro Kommunikation, die im Betreff der eMails angegeben ist. Die Gesamtheit aller eMails, die unter der gleichen Nummer zusammengefaßt sind, ist für gewöhnlich ein *Ticket*.[8] Auf der Web-Oberfläche kann man sehen, welche eMails gerade auf Bearbeitung warten. Die Bearbeiter haben die Möglichkeit, ein Ticket einer Person (also auch sich selbst) zuzuordnen, um dieser das Vorrecht auf eine Antwort zu erteilen.

Das *RT*-System bietet weiterhin eine „Wissensdatenbank“, in der aus bereits erteilten Antworten sogenannte *stock answers* (entspricht *Canned Text*) extrahiert werden können, auf die daraufhin bei späteren Antworten zurückgegriffen werden kann.

[5]Webseite des Systems: `http://bestpractical.com/rt/`

[6]Webseite des Unternehmens: `http://bestpractical.com/`

[7]Siehe Abb. 33 im Anhang.

[8]Da sich mehrere Tickets zu einem zusammenfassen lassen, kann ein Ticket sich auch über mehrere solcher Nummern erstrecken.

2 Preprocessing

2.1 eMail-Standards

Eine eMail besteht laut RFC 5322[9] aus einem Kopf- und einem Rumpfbereich (*Header* resp. *Body*). Beide basieren auf einzelnen Zeilen mit maximaler Länge von 998 Zeichen, wobei jede Zeile die Länge von 78 Zeichen nicht überschreiten sollte.[10] Im *Header* entspricht jede Zeile einem Eintrag, mit Ausnahme überlanger Einträge, die sich auch über mehrere Zeilen erstrecken können (siehe Abb. 18), wogegen ein Umbruch im *Body* i.d.R. auch einer neuen Zeile im Text entspricht. In Bezug auf den textuellen Inhalt der eMail kann beim Auftreten eines Umbruchs nicht eindeutig zwischen einem gewollten Umbruch des Textes und einem automatischen Umbruch durch Überschreiten der 78-Zeichen-Grenze unterschieden werden.

Der MIME-Standard (*Multipurpose Internet Mail Extensions*) für eMails nach RFC 2045, RFC 2046, RFC 2047, RFC 4288, RFC 4289 sowie RFC 2049[11] beschreibt, wie in dem *Body* vormals ausschließlich aus 7-bit ASCII-Zeichen[12] bestehender eMails verschiedenartige Dokumente und auch Texte anderer Kodierung untergebracht werden sollen. Auf diese Weise ermöglicht MIME dem Empfänger einer eMail, diese exakt nach den Vorgaben des Senders zu interpretieren.

Eine eMail mit MIME-*Body* (Abb. 19 stellt ein Beispiel dar) besteht gewöhnlicherweise aus mehreren Teilen, die beliebige Dokumente (Typ[13] und Formatierung der Daten sowie die Daten selbst) enthalten können, aber auch selbst wiederum eine MIME-Nachricht sein können, so daß sich eine Baumstruktur ergibt, bei der die Knoten MIME-Informationen und die Blätter die eigentlichen Dokumente enthalten.

2.2 Typische Struktur einer eMail

In der RT-Datenbank wird eben diese Baumstruktur abgebildet. Am häufigsten anzutreffen ist eine MIME-Nachricht vom Typ „multipart/alternative“ (die Unterknoten sind inhaltlich äquivalent; der Empfänger bzw. sein Programm entscheidet, welcher zur Darstellung ausgewählt wird). Diese enthält ein Dokument des Typs „text/plain“ (unformatierter Text) und eines

[9] `http://tools.ietf.org/html/rfc5322` (30. August 2011)

[10] Nicht jedes eMail-Programm hält sich an diese Vorgabe.

[11] `http://tools.ietf.org/rfc/index` (30. August 2011)

[12] ASCII-Zeichen (*American Standard Code for Information Interchange*) zwischen 0x20 (Leerzeichen) und 0x7e (Tilde)

[13] `http://www.iana.org/assignments/media-types/index.html` (30. August 2011)

vom Typ „text/html“ (HTML[14]-formatierter Text), der neben dem reinen Text beispielweise Farben, Schrifttypen und -schnitte sowie Hyperlinks unterstützt (siehe Abb. 2).

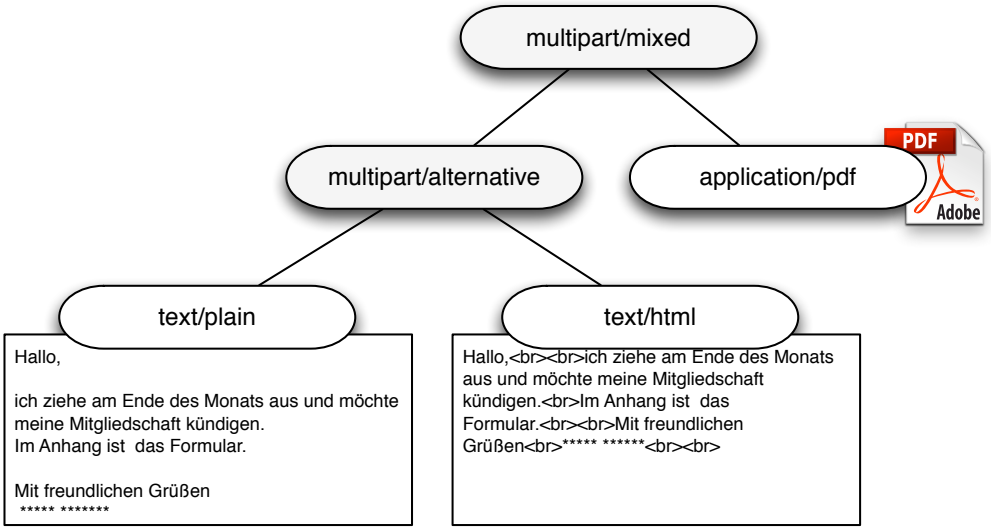

Abbildung 2: Struktur einer MIME-eMail (Beispiel aus der Datenbank)

Da beide Teile als äquivalent gekennzeichnet sind, reicht es aus, einen der beiden zu betrachten, um an den textuellen Inhalt der eMail zu gelangen, und da der Anteil an eMails, die entweder reinen Text enthalten oder eine MIME-Nachricht vom Typ „multipart/alternative“ mit wiederum eingebettetem reinen Text bei mehr als 98% liegt, beschäftige ich mich im folgenden ausschließlich mit Inhalt vom Typ „text/plain“.

Wie im Beispiel in Abb. 20 im Anhang zu sehen ist, unterteilt sich der *Body* in verschiedene Abschnitte, und zwar Zitate (Zeile 11 bis 62), Signaturen (64 bis 66) und reguläre Textblöcke (1 bis 9). Leere Zeilen lassen sich nicht immer eindeutig einem Block zuordnen, werden von jedem Autor individuell gehandhabt (von keiner einzigen bis zu einer Mehrzahl als Abstandshalter) und besitzen somit für die weitere Verarbeitung keine Aussagekraft, weshalb ich sie im Preprocessing-Schritt vor der weiteren Verarbeitung eliminiere.

Ebenso werden zu Beginn der Vorverarbeitung absenderspezifische und außersprachliche Daten wie Namen, eMail-Adressen und Hyperlinks durch entsprechende Symbole ersetzt. Außer dem Datenschutz-Aspekt bringt dieses

[14]Spezifikation der aktuellen Version: `http://www.w3.org/TR/1999/REC-html401-19991224/` (30. August 2011)

Vorgehen auch zwei computationelle Vorteile mit sich: Zum einen erleichtert die Markierung durch Symbole die Strukturanalyse der eMail (s.u.), zum anderen steuern diese Daten keinerlei relevante Information in Bezug auf Sprach- bzw. Anliegen-Klassifizierung bei, können also unbetrachtet bleiben.

2.3 Flacher Parser zur Strukturbestimmung

Jeder Zeile wird anhand von Mustervergleichen mit einfachen *Regulären Ausdrücken*[15] ein Typ zugeordnet, wobei eine Zeile, auf die mehrere Muster passen würden, dem Typ des komplexeren Musters zugewiesen wird. So bekommen alle Zeilen, zu denen keines der aufwendigeren Muster paßt, letztendlich einen allgemeinen Typ zugewiesen, dessen Muster zwangsläufig zu jeder Zeile kompatibel ist.

Muster, die sich aus dem vorhandenen eMail-Korpus ableiten lassen, sind beispielsweise[16]:

- Die Zeile beginnt mit einem Zitatzeichen „>“ (q_t);
- Die Zeile endet mit einem Doppelpunkt (t_2);
- Die Zeile enthält einen Ausdruck in runden Klammern (t_5);
- Die Zeile enthält ein Symbol wie beispielsweise den Vornamen des Autors (t_4);
- Die Zeile enthält außer Zwischenraumzeichen (*whitespaces*) nur einen Block mit mehreren Vorkommen typischer Separatoren-Zeichen (s_0).

Auf diese aus den Korpusdaten abgeleite Art und Weise klassifiziere ich jede Zeile mit einem von insgesamt 13 Mustern und erstelle eine Liste der jeweiligen Typen als symbolische Repräsentation der eMail.

Ebenfalls aus dem Korpus abgeleitet habe ich Regeln, in welchen Kontexten, welche Typen vorzufinden sind und daraus einen Parser konstruiert, der mit der oben beschriebenen Liste als Eingabe bei Erfolg eine Beschreibung der Struktur der analysierten eMail zurückliefert. Bei einem Testset aus 46759 Listen konnten 99,91% erfolgreich geparst werden[17].

[15] Siehe Kleene (1956).

[16] Die Symbole dafür, die der Parser aus Abb. 16 und der Regelsatz in Abb. 17 im Anhang verwenden, sind in Klammern angegeben.

[17] Das Testset wurde gewonnen, indem aus 46809 aus dem Korpus mittels Mustervergleichen gewonnenen Listen manuell alle falsch negativ erkannten Listen entfernt wurden; wie beispielsweise solche, die von Spam-eMails herrührten.

Da die Struktur keinerlei Rekursion aufweist, läßt sich der Parser (siehe Abb. 16) ebenfalls ohne Rekursion mithilfe hintereinandergeschalteter *Regulärer Ausdrücke* als Ersetzungsregeln (siehe Abb. 17) implementieren. Durch das fortlaufende Ersetzen der zwischenzeitig erzeugten Symbole entsteht zuletzt eine flache Struktur, die dicht aus Blöcken der Kategorien „Text“, „Zitat“ sowie „Signatur“ besteht, wobei mindestens einer davon vom Typ „Text“ sein muß. Falls dies nicht zutreffen sollte, wird der bisherige Parse zu Analysezwecken als Fehler zurückgeliefert.

Während Zitate für gewöhnlich Äußerungen Anderer sind, werden Signaturen häufig von Anbietern kostenloser eMail-Dienste (sogenannten Freemail-Anbietern) zu Werbezwecken genutzt. Zitate der vom Verein versandten eMails sind zwar in den meisten Fällen ein verläßlicher Indikator für das Anliegen des Absenders, allerdings liegen diese sowieso intern vor und bei Antworten auf eben jene läßt sich auch aus dem Betreff auf die eigentliche eMail schließen. Zu guter Letzt gibt es auch Fälle, in denen das Mitglied (wohl aus Gründen der Bequemlichkeit) mit einem neuen Anliegen auf irgendeine alte eMail-Kommunikation antwortet, so daß Zitate als zuverlässiges Kriterium ausscheiden. Was bleibt sind die *Text*-Blöcke als einzige authentische Äußerung des Absenders bzgl. seines Anliegens.

3 Sprach-Klassifizierung

3.1 Mögliche Verfahren

Die Mehrzahl der eingehenden eMails ist - wenn man sie auf die *Text*-Blöcke, also den vom Absender selbst verfassten Teil reduziert - einsprachig in deutsch oder englisch gehalten[18]; zur besseren Vergleichbarkeit beziehe ich auch die in sehr geringer Frequenz vorkommenden eMails in spanisch und französisch mit ein.[19] Zum Bestimmen der Sprache, in der der jeweilige Text gehalten ist, kommen mehrere Möglichkeiten in Betracht:

1. Suche in *Stoppwort*-Listen[20]
2. Vergleich der durchschnittliche Wortlängen
3. Vergleich des Anteils an Groß- und Kleinbuchstaben
4. Suche in Lexika
5. Betrachtung der Zeichenverteilung (N-Gramme)

Stoppwort-Listen leisten an sich sehr gute Dienste, wortweise Überschneidungen zwischen den Sprachen[21] können aber zu Problemen führen, wenn die Anzahl an Wörtern (im Sinne von Zeichenketten zwischen Leerzeichen bzw. Satzzeichen) gering[22] ist. Die Möglichkeiten (2) und (3) lassen sich für eine Unterscheidung zwischen deutsch und englisch bei Texten ab einer gewissen Länge gut verwenden, im vorliegenden Fall führen aber beide zu sehr schlechten Ergebnissen, zumal Groß- und Kleinschreibung im eMail-Verkehr keine große Rolle spielt und die deutschen Texte vieler ausländischer Mitglieder sowohl mit Korpus-Texten als auch mit denen deutscher Muttersprachler verglichen im Schnitt einen geringeren Wert an Großbuchstaben aufweisen.

Die Verwendung externer Lexika gestaltet sich problematisch, weil einerseits signifikante Terme der Domänen Vereinsmitgliedschaften und Netzwerktechnik dort nicht enthalten sind[23] und andererseits Ausdrücke der einen

[18] Nicht selten erreichen den Verein auch eMails, deren unterschiedliche Teile beispielsweise in deutsch, englisch und spanisch verfasst sind.

[19] Im Folgenden kommen die Abkürzungen nach ISO 639-1 für die Sprachen zum Tragen.

[20] Für eine Definition siehe Manning, Raghavan und Schütze (2008), Kapitel 2.2.2.

[21] Beispiele: $du_{de} \leftrightarrow du_{fr}$; $es_{de} \leftrightarrow es_{es}$; $was_{de} \leftrightarrow was_{en}$; $me_{en} \leftrightarrow me_{fr} \leftrightarrow me_{es}$.

[22] Mehr als 10% der Anfragen bestehen aus weniger als sechs Wörtern, der Median liegt allerdings bei 36,5.

[23] Im Gegensatz zum englischsprachigen WordNet bietet für deutsche Wörter CanooNet, wenn die direkte Suche zu keinem Treffer führt, in den meisten Fällen eine korrekte Analyse an, mithilfe des *Unknown Word Analyzer, Lemmatizer & Recognizer* (`http://www.canoo.com/languageexperts/products/toolkit/\#unknown`).

oder anderen Sprache als universelle Terme in allen Sprachen Verwendung finden[24]. Um hier eine Wörterbuchsuche anwenden zu können, bedarf es also speziell angepaßter Wörterbücher. Da aber für das vorliegende Korpus keine Annotationsdaten vorhanden sind und ich die automatische Klassifizierung aufgrund des mit manueller Vorgehensweise verbundenen Zeitaufwands als Prämisse gewählt habe, stelle ich dieses Verfahren zugunsten der automatischen Generierung passender Wörterbücher zurück.

Möglichkeit (5), die Analyse des N-Gramm-Vorkommens weist keinen der Nachteile der anderen Verfahren auf, verspricht im Gegenteil sogar eine gewisse Robustheit gegen die in unserem Korpus zuhauf vorkommenden Rechtschreibfehler, da hierbei nicht mehr die Wörter als Ganzes, sondern die Einheiten, aus denen sie bestehen, Gegenstand der Betrachtung sind.

3.2 Klassifizierung durch Vergleich der Unigramm-Verteilung

Als ersten Schritt habe ich aus dem Europarl-Korpus[25] für jede der vier zu unterscheidenden Sprachen den Text von sämtlichen Meta-Informationen (u.a. auch den jeweiligen Sprechernamen) befreit, und die Anzahl der Unigramme aller auftretenden Buchstaben in ihre relative Verteilung umgerechnet.[26] Wie man in Abb. 3 sehen kann, treten deutliche Unterschiede in der Verteilung hervor. Spanisch und französisch weisen die größte Ähnlichkeit auf, was aufgrund der typologischen Verwandtschaft der beiden Sprachen naheliegend ist, jedoch treten auch bei diesen beiden auffällige Abweichungen wie bei der Verteilung von „a“, „t“ oder „o“ auf.

Mithilfe dieser Unigramm-Verteilung läßt sich die Sprach-Klassifizierung anhand folgender an die Euklische Distanz[27] angelehnte Distanzwerte vornehmen:

$$\tau(\mathfrak{p}, \mathfrak{q}) = \left(\sqrt{\frac{\sum\limits_{i \in (\mathfrak{p}_\gamma \cap \mathfrak{q})} (\mathfrak{p}_i - \mathfrak{q}_i)^2}{\max(|\mathfrak{p}|, |\mathfrak{q}|)}} \right) \tag{1}$$

$\mathfrak{p}$ sind hierbei die aus dem Korpus extrahierten Verteilungswerte einer bestimmten Sprache und $\mathfrak{q}$ steht für die zu untersuchende Verteilung ei-

[24] Beispielsweise $Bankleitzahl_{en}$ (17% der Vorkommen des Wortes in englischen eMails), $Mitgliedsnummer_{en}$ (13%), $ethernet_{de}$ (49% der Vorkommen des Wortes in deutschen eMails), $policy_{de}$ (42%) oder $membership_{de}$ (7%).

[25] Siehe Koehn (2005); Webseite des Korpus: `http://www.statmt.org/europarl/`

[26] Das Ergebnis ist für jede Sprache γ jeweils ein Vektor $\mathfrak{p}_\gamma \in \mathbb{R}^N$ mit pro Sprache variierendem N.

[27] Siehe Manning, Raghavan und Schütze (2008), Kapitel 6.3.

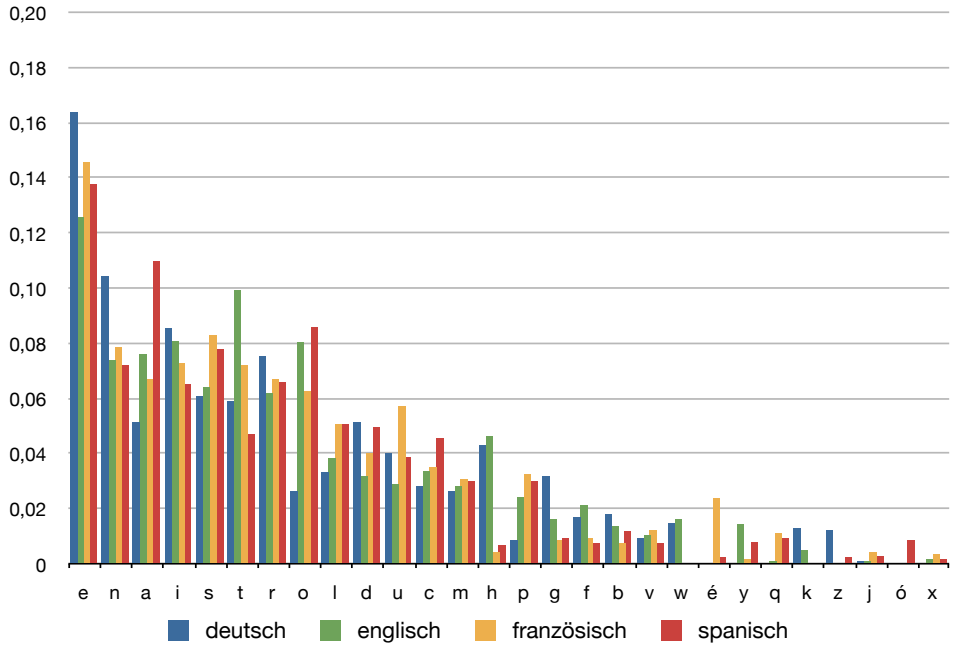

Abbildung 3: Verteilung der in den jeweiligen Sprachen am häufigsten auftretenden Zeichen (Die Abdeckung der Abbildung liegt zwischen 97,86% für deutsch und 99,99% für englisch).

nes bestimmten Textes. Die gemäß der Ähnlichkeit der Zeichenverteilungen wahrscheinlichste Sprache l_1 für $\mathfrak{q}$ ist dann diejenige (aus der Gesamtheit der Sprachen Γ), für die τ den größten Wert liefert:

$$l_1(\mathfrak{q}) = \arg\max_{\gamma \in \Gamma} \tau(\mathfrak{p}_\gamma, \mathfrak{q}) \tag{2}$$

Da $\mathfrak{q}$ aufgrund der Kürze der Texte[28] verglichen mit $\mathfrak{p}_\gamma$ regelmäßig weniger Dimensionen aufweisen wird (man spricht von $\mathfrak{q}$ deshalb auch als einem *sparse vector*), hängt die Anzahl der Summanden hauptsächlich von $\mathfrak{q}$ ab. Um Sprachen, deren Zeichenvorrat den anderer Sprachen beinhaltet (z.B. gilt $|\mathfrak{p}_{fr}| > |\mathfrak{p}_{en}|$), deshalb keinen Vorteil zuteil werden zu lassen, ist es nötig, den als Distanz gewonnen Wert vor dem Vergleichen mittels Division durch die Dimension des höherdimensionalen Vektors zu normalisieren.

[28]Bei einer durchschnittlichen Wortlänge von 8,27 Buchstaben und 36,5 Wörtern pro Text stehen rund 300 Buchstaben der Stichprobe je etwa 15 Mio. der Zielpopulation gegenüber.

3.3 Klassifizierung mit neugewonnenem Lexikon

Obiges Verfahren ergibt angewandt auf alle verfügbaren Texte eine Verteilung, wie sie in Abb. 4 dargestellt ist. Mein subjektiver Eindruck und ein stichprobenartiger Vergleich zeigen, daß Falschklassifizierung i.d.R. zugunsten der weniger stark vertretenen Klassen stattfand. Das bedeutet, daß dieses Verfahren dazu tendiert, jeder Klasse scheinbar unabhängig vom ermittelten Wert bestimmte Texte zuzuordnen. Wie eine Kontrollprobe der im Verdacht stehenden eMails zeigt, geschieht dies hauptsächlich bei sehr kurzen Texten, also Vektoren deutlich geringerer Dimension als die der Vergleichsvektoren.

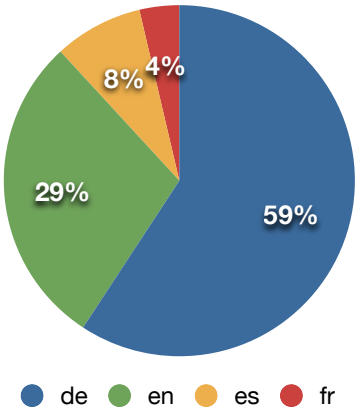

Abbildung 4: Ergebnisse der Sprachklassifizierung mittels Distanzvergleich zwischen Unigrammverteilungen.

Statt nun die bisherigen Ergebnisse zu verwerfen und mit Bi- und Trigrammen erneut anzusetzen, erschien es mir erfolgversprechender, die klassifizierten Daten zu nutzen, um daraus ein mit Verteilungswerten annotiertes Lexikon zu konstruieren, so daß die Fehler bei Verwendung der Unigrammverteilung sich auf statistischem Wege nivellieren würden. Dazu habe ich die ohnehin schon vorhandenen Daten über Wörter und Anfragen, in denen sie vorkommen, samt Häufigkeiten mit der in (2) ermittelten Sprache verknüpft und global, d.h. für alle im Lexikon vorkommenden Wörter, die jeweiligen Zuweisungswahrscheinlichkeiten zu einer bestimmten Sprache γ ermittelt. Die in (3) beschriebene Funktion $\xi_\gamma(\lambda)$ liefert einen Indikator für die relative

Häufigkeit, mit der ein bestimmtes Wort λ in den den Texten der Sprache γ ($\bigcup\gamma$) auftritt.[29]

$$\xi_\gamma(\lambda) = \frac{|\lambda \in \bigcup\gamma|}{|\bigcup\gamma|} \tag{3}$$

Die so gewonnenen Vergleichswerte werden in einem Zwischenschritt (4) normalisiert und zur Verstärkung der Differenzen quadriert. Für die Sprache γ mit den vergleichsweise häufigsten Vorkommen eines Wortes λ in bezug zu ihrer Gesamtwortzahl, gilt $\varsigma_\gamma(\lambda) = 1$.[30]

$$\varsigma_\gamma(\lambda) = \left(\frac{\xi_\gamma(\lambda)}{\max\limits_{\gamma' \in \Gamma} \xi_{\gamma'}(\lambda)} \right)^2 \tag{4}$$

Die von $\varsigma_\gamma(\lambda)$ zurückgelieferten Werte ergeben über alle Wörter eines bestimmten Textes $\vec{q}$ aufsummiert einen Wert, der die Gesamtwahrscheinlichkeit eines Textes zu einer Sprache γ zu gehören beschreibt. Die Sprache γ mit dem aufsummiert größten Wert entspricht am wahrscheinlichsten der des Textes:

$$l_2(\vec{q}) = \arg\max_{\gamma \in \Gamma} \sum_{\lambda \in \vec{q}} \varsigma_\gamma(\lambda) \tag{5}$$

In Tabelle 1 sind Wahrscheinlichkeitsverteilungen für unterschiedliche Arten von Texten dargestellt. Während der Text zu Beispiel (a) bis auf wenige Ausnahmen einsprachig in französisch gehalten ist und problemlos richtig eingeordnet werden kann, besteht der Text zu (b) zur Hälfte aus einem fremdsprachigen Zitat, was die Häufigkeitsverteilung der vorkommenden Zeichen dermaßen verfälscht, daß obige Formel den Text weder der einen noch der anderen verwendeten Sprache zuordnet, sondern stattdessen spanisch präferiert. Im zweiten Schritt, unter Einbeziehung des Lexikons, wird die vorherige Einordnung allerdings zufriedenstellend korrigiert, was hauptsächlich auf das Vorkommen signifikanter, häufig verwendeter, Wörter wie *ich*, *nicht* oder *habe* zurückzuführen ist.

[29] Gerundete Beispielwerte für $\xi_\gamma(\lambda)$:

	λ_{send}	λ_{bitte}	λ_{est}	$\lambda_{formular}$	$\lambda_{support}$
γ_{de}	0,00070	0,15302	0,00011	0,02088	0,04402
γ_{en}	0,03025	0,01076	0,00027	0,00197	0,03296
γ_{es}	0,00412	0,00606	0,00039	0,00249	0,00955
γ_{fr}	0,00180	0,02192	0,00567	0,00825	0,04551

[30] So gilt $\varsigma_{\gamma_{en}}(\lambda_{send}) = \varsigma_{\gamma_{de}}(\lambda_{bitte}) = \varsigma_{\gamma_{fr}}(\lambda_{est}) = \varsigma_{\gamma_{de}}(\lambda_{formular}) = \varsigma_{\gamma_{fr}}(\lambda_{support}) = 1$, wogegen beispielsweise $\varsigma_{\gamma_{en}}(\lambda_{formular}) = 0{,}09$ und $\varsigma_{\gamma_{de}}(\lambda_{support}) = 0{,}97$ ist.

		γ_{de}	γ_{en}	γ_{es}	γ_{fr}
a)	$P(l_1(\mathfrak{q}) = \gamma)$	18,78%	0,00%	20,66%	**60,56%**
	$P(l_2(\vec{q}) = \gamma)$	16,00%	14,58%	4,53%	**64,89%**
	ΔP	-2,78%	+14,58%	-16,13%	+4,33%
b)	$P(l_1(\mathfrak{q}) = \gamma)$	0,00%	19,55%	**40,26%**	40,19%
	$P(l_2(\vec{q}) = \gamma)$	**41,39%**	10,68%	21,21%	26,72%
	ΔP	+41,39%	-8,87%	-19,05%	-13,47%
c)	$P(l_1(\mathfrak{q}) = \gamma)$	0,00%	**54,70%**	7,54%	37,76%
	$P(l_2(\vec{q}) = \gamma)$	**80,84%**	10,26%	8,66%	0,24%
	ΔP	+80,84%	-44,44%	+1,12%	-37,52%

Tabelle 1: Beispielwerte für Wahrscheinlichkeitsverteilung mit l_1 und l_2 (Die Texte, aus denen $\mathfrak{q}$ und $\vec{q}$ für (a), (b) und (c) gewonnen wurden, befinden sich im Anhang in den Abb. 21, 22, resp. 23).

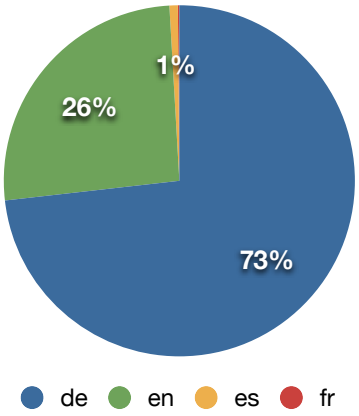

Abbildung 5: Ergebnisse der Sprachklassifizierung mithilfe des vorher generierten Lexikons.

Das Beispiel (c) schließlich stellt einen Extremfall dar, insofern der Text aus nur zwei Wörtern bzw. zehn unterschiedlichen Zeichen besteht. Die Klassifizierung $l_1 = \lambda_{en}$ mag darauf zurückzuführen sein, daß der Buchstabe t sowohl in Abb. 23, als auch in der Verteilung in Abb. 3 für das Englische überdurchschnittlich häufig vertreten ist; allein ist das Verfahren erst dann erfolgversprechend, wenn ausreichend viele Daten vorhanden sind, um eine

repräsentative Verteilung zu erhalten. Ein möglicher Indikator, wann eine solche erreicht ist, könnte der Einfluß neuer Daten auf die Gesamtverteilung sein. Für meine Zwecke spielen allerdings die Falschklassifizierungen im ersten Schritt keine Rolle, da sie im zweiten Schritt in den allermeisten Fällen wieder korrigiert werden. Die Verteilung der erneut klassifizierten Texte ist in Abb. 5 zu sehen. Eine manuelle Stichprobe von 100 Texten (je 50, die als deutsch bzw. englisch klassifiziert wurden) ergab 50 korrekte Zuweisungen bei den deutschen Texten und 49 bei den englischen.[31]

[31]Die einzige Ausnahme stellt ein Liebesgedicht auf Hindi dar, das uns fälschlicherweise zugesandt wurde, in dem aber auch vereinzelte englische Wörter vorkommen.

4 Klassifizierung der Anliegen

4.1 Problembeschreibung

Nachdem die Sprache der Anfragen trotz vorhandener Schwierigkeiten wie Orthographie oder der Kürze der Texte in den allermeisten Fällen korrekt bestimmt werden konnte, spielen bei der Klassifizierung der Anliegen neben diesen Problemen auch andere Aspekte eine Rolle:

1. Die Formulierungen für ein und dasselbe Anliegen können stark variieren; nicht selten sind sich die Mitglieder ihres Anliegens selbst gar nicht bewußt, und der Bearbeiter muß versuchen, vom Text der eMail auf das Anliegen zu schließen (Siehe Abb. 24 bzw. 25 im Anhang).

2. Neben der Rechtschreibung beinflußt auch die Worttrennung, welche Teile des Textes als Wort erkannt werden; dies wiederum beeinflußt eine wortbasierte Klassifizierung.

3. In wenigen Fällen wird in einer eMail mehr als ein Anliegen ausgedrückt. Diesen Fall habe ich zwar von vornherein von der Liste der Fälle ausgenommen, die ich bearbeiten möchte, allerdings müßte dazu sichergestellt sein, daß entsprechende Texte nicht fälschlicherweise in eine der enthaltenenen Klassen fallen.

Mein erster Ansatz sah vor, alle nicht im Lexikon gefundenen Wörter mittels eines um phonologische Regeln erweiterten Levenshtein-Vergleichs (Vgl. Levenshtein (1965)) auf ein im Lexikon vorhandenes Wort abzubilden.[32] Erste Versuche ergaben aber ein geringes Korrekturpotenzial bei gleichzeitig großem Rechenaufwand.[33] Auch der Aufbau einer Ersetzungstabelle mit allen möglichen Falschschreibungen erschien mir nicht zielführend, aufgrund

[32]Phonologische Regeln wären dabei in Form sprachspezifischer Gewichte zum Tragen gekommen, die bei der Berechnung der gewichteten Levenshtein-Distanz solchen Fällen weniger Gewicht zugeordnet hätten, deren lautsprachlicher Unterschied gering gewesen wäre, wie beispielsweise „Miete“ vs. *„Mite“, „Austrittserklärung“ vs. *„Austrittserklerung“, „schon“ vs. *„shon“ und „Vertrag“ vs. *„Vetrag“. Des Weiteren könnte eine Berücksichtigung der unterschiedlichen Tastaturbelegungen bei der Definition der spezifischen Gewichte dazu beitragen, Wörter richtig zuzordnen, deren Autor über keine Umlaute und kein „ß“ verfügt und die Ersetzungsregeln dafür nicht kennt; gleiches gilt entsprechend auch für die Berücksichtigung von mit „ae“, „oe“, „ue“ sowie „ss“ umschriebenen Sonderzeichen in deutschsprachigen Texten. In Extremfällen wird in Ermangelung der deutschen Tastaturbelegung „ß“ auch durch die Homoglyphen „β“ bzw. „B“ ersetzt.

[33]Eine phonologische Beurteilung der jeweils betrachteten Ersetzung bedarf einer vorhergehenden morphologischen Analyse des Wortes, da die Regeln je nach Kontext (silben- oder wortinitial bzw. -final) jeweils andere wären.

der Variation bei bereits in den vorhandenen Texten vorkommenden falsch geschriebenen Wörtern.

4.2 Clusteranalyse mittels tf-idf-basierter Ähnlichkeitswerte

Ein anderer Ansatz kommt ohne vorherige Korrektur eventueller Fehler aus: Eine Clusteranalyse der vorhandenen Texte aufgrund der Ähnlichkeit der enthaltenen Wörter hilft, den Fehler eines einzelnen Wortes gegenüber einem sonst übereinstimmenden Texte zu nivellieren, so daß neue Texte einem bereits gefundenen Cluster zugeordnet werden können. Mögliche Clusteranalyseverfahren sind entweder flach[34] oder hierarchisch[35]. Bei einer flachen Analyse bedarf es einer vorher definierten festen Anzahl an Clustern, auf die die vorhandenen Elemente verteilt werden, bei hierarchischen Verfahren werden Cluster bei sinkender Ähnlichkeit in größere Cluster zusammengefaßt bzw. bei steigenden Ähnlichkeitswerten in kleinere Cluster aufgeteilt (je nach Sicht der Verfahrensweise). Dabei konstituiert zuallererst jedes Element sein eigenes Cluster mit maximaler Ähnlichkeit[36]. Spätestens bei der minimalen Ähnlichkeit von 0 befinden sich alle Element in ein und demselben Cluster.

4.2.1 Ähnlichkeitsbewertung

Die einfachste Möglichkeit, die Ähnlichkeit zweier Texte zu beschreiben, ist der Jaccard-Koeffizient (Siehe Jaccard (1901) oder auch Manning und Schütze (1999), Kapitel 8.5). Er ist definiert als die Anzahl der Elemente, die in beiden Texten p und q übereinstimmen, in Relation zur Anzahl der insgesamt vorkommenden Elemente:[37]

$$J(p,q) = \frac{|P \cap Q|}{|P \cup Q|} \tag{6}$$

Für die Beispiele aus Abb. 6 ergibt sich $J(p_1,p_2) = J(p_2,p_3) = \frac{2}{4} = 0{,}5$, da insgesamt jeweils vier Wörter vorkommen, wovon jeweils zwei in beiden

[34] Siehe Manning, Raghavan und Schütze (2008), Kapitel 16.

[35] Ebd., Kapitel 17.

[36] Die Ähnlichkeit eines Clusters zu sich selbst ist per Definition 1, was gleichzeitig den Maximalwert für Ähnlichkeiten darstellt.

[37] Diese und folgende durch Kleinbuchstaben repräsentierte Texte sind jeweils als *bag of words* (siehe Manning, Raghavan und Schütze (2008), Kapitel 6) zu verstehen, die einzelne Wörter so oft enthalten, wie sie im Originaltext vorzufinden sind, wogegen die Darstellung mittels Großbuchstaben auf eine Menge verweist, in der jedes Element distinkt ist.

`Anbei meine Austrittserklärung.`

$p_1 \quad = \quad \{\ \lambda_{anbei},\ \lambda_{meine},\ \lambda_{\text{austrittserklärung}}\ \}$

`Anbei die Austrittserklärung.`

$p_2 \quad = \quad \{\ \lambda_{anbei},\ \lambda_{die},\ \lambda_{\text{austrittserklärung}}\ \}$

`Anbei die Einzugsermächtigung.`

$p_3 \quad = \quad \{\ \lambda_{anbei},\ \lambda_{die},\ \lambda_{\text{einzugsermächtigung}}\ \}$

Abbildung 6: Minimalbeispiel dreier fiktiver Texte. λ_x repräsentiert hier und im weiteren dasjenige Wort, das durch die Zeichenkette „x“ beschrieben wird.

Texten vertreten sind (Siehe Abb. 7), und $J(p_3, p_1) = \frac{1}{5} = 0{,}2$, bei nur einem übereinstimmenden Wort bei einer Anzahl von fünf.[38]

Generell ist der Jaccard-Koeffizient ein gutes Maß für die relative Anzahl der übereinstimmenden Wörter, allerdings behandelt er alle Wörter gleich, so daß dem Textpaar (p_1,p_2), das in Bezug auf das Anliegen eine sehr hohe Ähnlichkeit aufweist (es unterscheidet sich in puncto Pronomen/Determinante) der gleiche Wert zugeordnet wird wie dem Paar (p_2,p_3), welches als erheblich verschiedener anzusehen ist (den Unterschied machen hier Nomina aus).

$$
\begin{array}{lccccccc}
 & & \textit{anbei} & \textit{meine} & \textit{die} & \textit{austrittserklärung} & \textit{einzugsermächtigung} & \\
\vec{p}_1 = (& 1 & 1 & 0 & 1 & 0 &)^{\mathrm{T}} \\
\vec{p}_2 = (& 1 & 0 & 1 & 1 & 0 &)^{\mathrm{T}} \\
\vec{p}_3 = (& 1 & 0 & 1 & 0 & 1 &)^{\mathrm{T}}
\end{array}
$$

Abbildung 7: Die Texte aus dem Beispiel in Abb. 6 als Vektoren dargestellt.

Um diesen inhaltlichen Unterschied auch zur Geltung zu bringen, ist eine Gewichtung der in die Bewertung eingehenden Elemente nötig. Von einem

[38] Die Funktion J ist kommutativ, da Schnitt- und Vereinigungsmenge es auch sind; von daher gilt $J(p_2, p_1) = J(p_1, p_2)$ usf.

Part-Of-Speech-Tagging der Texte muß aufgrund der gegebenen Besonderheiten (keine einheitliche Syntax, viele unbekannte Wörter, …) abgesehen werden. Stattdessen bietet sich hier die *inverse Dokumentfrequenz* $idf_\gamma(\lambda)$[39] an. Sie liefert zu jedem Wort λ einen Wert, der ein Maß dafür ist, wie häufig oder selten dieses Wort in der Gesamtheit aller Texte einer Kollektion γ (hier: einer Sprache) vorkommt. So weisen *Stoppwörter* einen kleinen idf-Wert auf, wogegen er für die Inhaltswörter umso größer ist, je seltener sie sind. Die Funktion ist wie folgt definiert:

$$idf_\gamma(\lambda) = \log \frac{|\bigcup \gamma|}{|\lambda \in \bigcup \gamma|} \tag{7}$$

Für die Wörter aus dem Beispiel in Abb. 7 ergeben sich für die Sammlung der deutschsprachigen Anfragen (γ_{de}) folgende Werte:

λ	$idf_{\gamma_{de}}(\lambda)$
λ_{anbei}	1,53
λ_{meine}	0,63
λ_{die}	0,44
$\lambda_{\text{austrittserklärung}}$	1,62
$\lambda_{\text{einzugsermächtigung}}$	2,10

Abbildung 8: idf-Werte für die Wörter aus Abb. 6 mit $|\gamma_{de}| = 28150$.

Eine Erweiterung des Jaccard-Koeffizienten um die spezifischen idf-Werte der einzelnen Wörter läßt sich wie folgt definieren:

$$J'_\gamma(p,q) = \frac{\sum\limits_{\lambda \in P \cap Q} idf_\gamma(\lambda)}{\sum\limits_{\lambda \in P \cup Q} idf_\gamma(\lambda)} \tag{8}$$

Diese Funktion angewandt auf die obigen Beispieltexte ergibt folgende Werte:

$$J'_\gamma(p_1, p_2) = 0{,}7465$$
$$J'_\gamma(p_2, p_3) = 0{,}3456$$
$$J'_\gamma(p_3, p_1) = 0{,}2417$$

[39] Siehe Manning, Raghavan und Schütze (2008), Kapitel 6.2.

Die Texte p_1 und p_2 sind sich also gemäß der Definition von J' deutlich ähnlicher, als die beiden anderen Kombinationen. Wenn man jetzt noch die *Termfrequenz* $tf(\lambda, p)$[40], also die Anzahl der Vorkommen eines Wortes in einem Text, berücksichtigt, ergibt sich folgende Ähnlichkeitsfunktion:

$$J''_\gamma(p,q) = \frac{\sum\limits_{\lambda \in p \cap q} \min(tf(\lambda,p), tf(\lambda,q)) \cdot idf_\gamma(\lambda)}{\left[\begin{array}{r} \sum\limits_{\lambda \in p} [tf(\lambda,p) \cdot idf_\gamma(\lambda)] \\ + \sum\limits_{\lambda \in q} [tf(\lambda,q) \cdot idf_\gamma(\lambda)] \\ - \sum\limits_{\lambda \in p \cap q} [\min(tf(\lambda,p), tf(\lambda,q)) \cdot idf_\gamma(\lambda)] \end{array}\right]} \tag{9}$$

Diese auf den ersten Blick etwas umständlich erscheinende Erweiterung von J' weist eine große Übereinstimmung mit der Ähnlichkeitsfunktion von Manning, Raghavan und Schütze (2008) (10) auf:

$$sim(p,q) = \frac{\vec{V}(p) \cdot \vec{V}(q)}{|\vec{V}(p)||\vec{V}(q)|} \tag{10}$$

Die Funktion $\vec{V}(x)$ bildet hierbei den Text x als *bag of words* auf einen Vektor im Vektorraum $\mathbb{N}^N$ mit $N = |\bigcup \gamma|$ ab.[41] Für die Funktion in (10) ergibt sich also:

$$\begin{aligned} sim(p,q) &\widehat{=} \frac{\sum\limits_{\lambda \in \bigcup \gamma} [tf(\lambda,p) \cdot idf_\gamma(\lambda)] \, [tf(\lambda,q) \cdot idf_\gamma(\lambda)]}{\sqrt{\sum\limits_{\lambda \in p} [tf(\lambda,p) \cdot idf_\gamma(\lambda)]^2} \sqrt{\sum\limits_{\lambda \in q} [tf(\lambda,q) \cdot idf_\gamma(\lambda)]^2}} \\ &= \frac{\sum\limits_{\lambda \in \bigcup \gamma} tf(\lambda,p) \cdot tf(\lambda,q) \cdot idf_\gamma(\lambda)^2}{\sqrt{\sum\limits_{\lambda \in p} [tf(\lambda,p) \cdot idf_\gamma(\lambda)]^2} \sqrt{\sum\limits_{\lambda \in q} [tf(\lambda,q) \cdot idf_\gamma(\lambda)]^2}} \end{aligned} \tag{11}$$

[40] Ebd. Ein *bag of words* kann als Multimenge verstanden werden, was die formale Definition vereinfacht, Funktionen wie tf obsolet macht und die Verwendung von Mengenoperatoren rechtfertigt. Zugunsten der intuiven Lesbarkeit der folgenden Formeln, bleibe ich bei der tradierten Darstellung.

[41] Im Folgenden sei $\vec{x}$ implizit als Vektordarstellung von x verstanden.

Bei den in Abb. 6 gegebenen Minimalbeispielen ergibt sich im Vergleich zu den Werten der Funktion J'' hinsichtlich der Größenverhältnisse kaum ein Unterschied:[42]

$$\begin{aligned} sim(p_1, p_2) &= 0{,}9440 \approx 1{,}26 \cdot J''_\gamma(p_1, p_2) \\ sim(p_2, p_3) &= 0{,}4222 \approx 1{,}22 \cdot J''_\gamma(p_2, p_3) \\ sim(p_3, p_1) &= 0{,}3825 \approx 1{,}58 \cdot J''_\gamma(p_3, p_1) \end{aligned}$$

Die Verwendung des Minimums der Termfrequenzen der zu vergleichenden Texte hat allerdings zur Folge, daß - im Gegensatz zur Winkelberechnung zweier längennormalisierter Vektoren bei $sim(p, q)$ - ein Dokument p nicht den Maximalwert 1 ($\hat{=}\ 0°$) beim Vergleich mit einer um den Faktor n ($n \in \mathbb{N}; n \geqq 2$) gestreckten Kopie seiner selbst erhält. Sei beispielsweise $\vec{p}_4 = 2 \cdot \vec{p}_1 = (2\ 2\ 0\ 2\ 0)^{\mathrm{T}}$, dann gilt $sim(p_1, p_4) = 1$, aber für J'' ergibt sich wegen der Verwendung des Minimalwertes beider tf-Werte entsprechend $J''(p_1, p_2) = 0{,}5$. Der Unterschied ist in Abb. 9 deutlich zu erkennen.

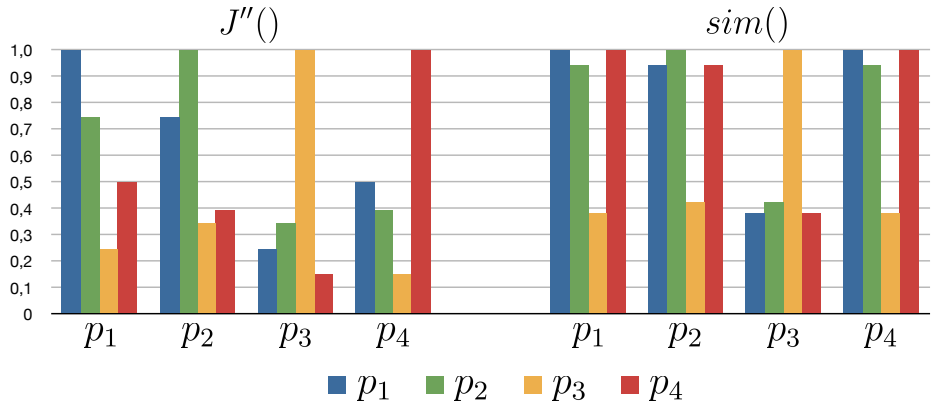

Abbildung 9: Vergleich der Ähnlichkeitswerte der Texte aus Abb. 6 bzw. 7 sowie p_4 für beide Funktionen J'' und sim. Für den Vergleich eines Textes p mit sich selbst gilt aufgrund der Definition der Funktionen immer $J''(p, p) = sim(p, p) = 1$.

Der Fall, daß ein Text genau n-mal den Inhalt eines anderen Textes enthält, ist aber rein hypothetischer Natur. In den meisten Fällen sind es vor

[42] Wie in Abb. 9 exemplarisch zu sehen ist, tendiert sim dazu, bei gleichen Vergleichstexten höhere absolute Werte zu liefern als J''. Für die Rangliste der bewerteten Textpaare ist dies allerdings nicht von Belang.

allem Stoppwörter wie Präpositionen, Pronomen und Determinantien, deren Termfrequenz 1 übersteigt. Erhöht sich die Termfrequenz eines Wortes in einem Text, verlängert sich der Basisvektor der entsprechenden Dimension während alle anderen sich der Normalisierung wegen entsprechend verkürzen. Der entstehende Einheitsvektor weicht nun um einen geringfügigen Winkel vom Vektor des unveränderten Textes ab. Diese Abweichung ist in Abb. 10 für je ein Wort von hoher und niedriger Aussagekraft (hoher bzw. niedriger idf-Wert) dargestellt. Deutlich erkennbar ist, daß die Kurve für das Wort „austrittserklärung“ mit einem idf-Wert von 1,62 einen grundsätzlich anderen Verlauf nimmt als die von „meine“ mit 0,63.[43]

Demgegenüber verhält sich J'' sehr regelmäßig; die Werte der Funktion sind durchgängig degressiv fallend und hohe idf-Werte wirken sich entsprechend negativer aus als niedrige (d.h. ein in einem der Texte häufiger vorkommendes Pronomen hat beispielsweise weniger Einfluß als ein spezifischeres Wort, dessen Anzahl nicht übereinstimmt). Während im Bereich des *Information Retrieval* Texte unterschiedlichster Länge betrachtet werden, die wie in Abb. 9 zu sehen auch Ähnlichkeitswerte von 1 aufweisen können, ist im vorliegenden Fall die Länge der Texte durchaus relevant für die Zuordnung zu anderen Texten. So sind eMails mit konkreten Anliegen i.d.R. sehr kurz gehalten (siehe Abb. 20 bis 25 im Anhang), während längere Texte - sofern es sich nicht um *Spam* handelt - generell eher auf spezielle Anfragen hindeuten, welche der Idee der Clusteranalyse folgend sowieso erst zu einem späten Zeitpunkt, d.h. bei einem niedrigen Ähnlichkeitswert, in ein größeres Cluster integriert werden sollten.

Die Funktion J' verwendet lediglich die idf-Werte der in den Texten vorkommenden Wörter und läßt deren Frequenzen dabei unberücksichtigt. Bei kürzeren Texten sind Termfrequenzen größer 1 ohnehin meist nur bei Stoppwörtern anzutreffen, von daher würde in jenen Fällen J' als Ähnlichkeitsfunktion ausreichen. Wenn die Termfrequenz für jedes Wort der Kollektion entweder 0 oder 1 ist, verhält sich J'' identisch zu J' (siehe J'' in Abb. 11 für identische Termfrequenzen); bei längeren Texten mit größeren Termfrequenzen erfolgt eine stufenweise Reduzierung der Ähnlichkeit in Abhängigkeit

[43]Eine genaue Betrachtung der Funktion zeigt, daß für $\Delta n \to \infty$ jede Kurve sich einem Wert annähert, der zu dem idf-Wert des betrachteten Wortes proportional ist und daß die Kurvenkrümmung dabei ebenso von diesem abhängt (für Wörter, deren idf-Wert über 1 liegt (wie der von „austrittserklärung“ oder „anbei“), ergibt sich ein degressiv fallender Verlauf, kleinere idf-Werte haben einen zuerst progressiv fallenden Verlauf zur Folge, was übertragen auf die Funktion *sim* bedeutet, daß ein geringer Unterschied in der Termfrequenz weniger Einfluß auf den Funktionswert hat, wenn es sich um Wörter geringerer Aussagekraft handelt, und bei großen Unterschieden eben jene den Wert deutlicher absenken als es aussagekräftigere Wörter tun würden. Beispielsweise gilt: $sim(p_1, p_1') \approx sim(p_1, p_1'')$ mit $p_1' = p_1 + 5 * \lambda_{meine}$ und $p_1'' = p_1 + 31 * \lambda_{\text{austrittserklärung}}$.

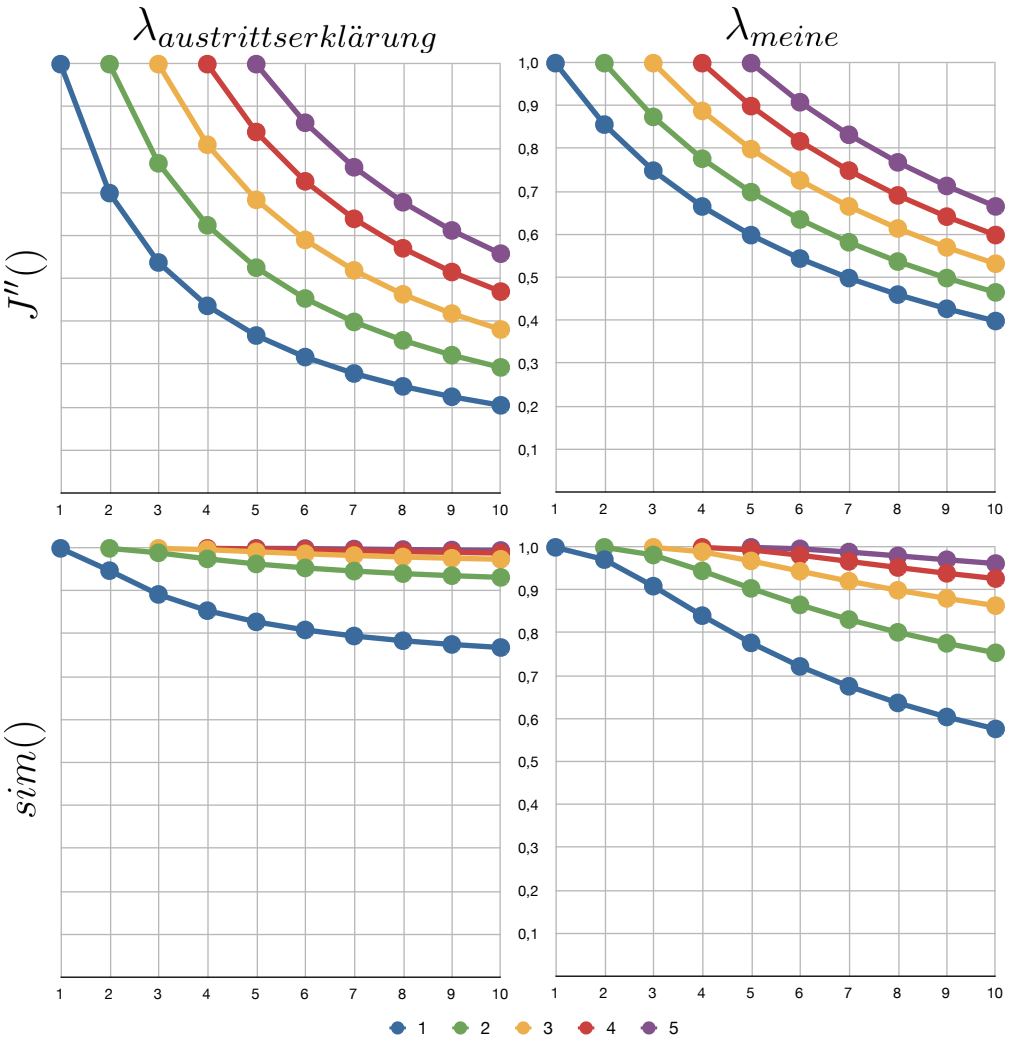

Abbildung 10: Darstellung der Veränderungen des Funktionswertes, der sich für J'' und sim ergibt, wenn sich beim Eingabevektor $\vec{p}_1$ aus Abb. 8 die Termfrequenz von $\lambda_{\text{austrittserklärung}}$ bzw. λ_{meine} ändert. Die Termfrequenz des einen Vektors ist auf der x-Achse aufgetragen, die des anderen durch den jeweiligen Graphen gegeben. Manning, Raghavan und Schütze (2008) schlagen vor, als eine Verbesserung der Formel die Termfrequenz logarithmisch herunterzuskalieren; in diesem Fall würde die Funktion sim in jedem Fall höhere Ähnlichkeitswerte liefern.

von der Differenz der Frequenz eines Wortes sowie der durchschnittlichen Frequenz über beide Texte.[44]

[44]Im allgemeinen gilt, daß die gleiche Differenz eine desto schwächere Auswirkung auf das Ergebnis zur Folge hat, je größer die Durchschnittsfrequenz in beiden Texten ist (siehe auch Abb. 10).

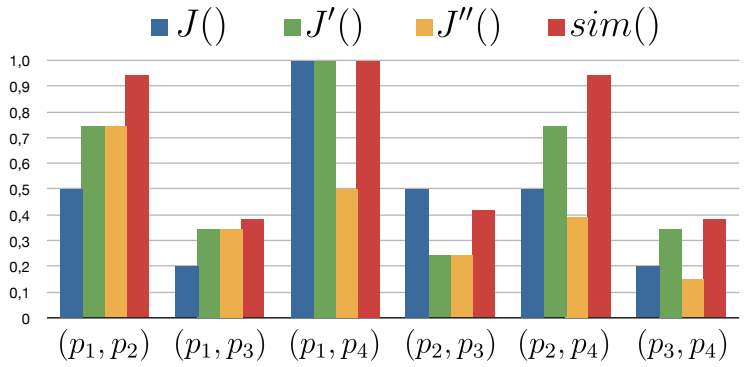

Abbildung 11: Vergleich der Ergebnisse verschiedener Ähnlichkeitsfunktionen für jede Kombination der Beispieltexte p_1 bis p_4.

4.2.2 Hierarchische Clusteranalyse

Aufgrund fehlender Informationen über die Anzahl der Cluster[45] fiel meine Wahl auf ein hierarchisches Verfahren. Solche Verfahren werden wenn sie den Hierarchie-Baum von den Blättern (den einzelnen Texten) her zur Wurzel hin aufbauen *agglomerativ*, wenn sie mit der Wurzel beginnend jedes Cluster in kleinere Untercluster unterteilen *divisiv* genannt (siehe Manning, Raghavan und Schütze (2008), Kapitel 17). Da jedes der verschiedenen Verfahren zur Clusteranalyse auf die Werte der Ähnlichkeitsfunktion zurückgreift, waren diese zuerst zu berechnen. Dabei ergeben sich für die Gesamtheit aller vorhandenen Anfragen ca. 40 Mio. Vergleichswerte auf deutsch und 8,5 Mio. auf englisch; die verwendete Untermenge aus etwa 30% zufällig ausgewählten Texten ergibt nur noch 3,9 Mio. Werte für deutsch und 0,8 Mio für englisch. Wie in Abb. 12 zu erkennen ist, weisen die meisten Textpaare eine sehr geringe Ähnlichkeit auf.[46]

Aus diesem Grund und weil an einem kompletten Cluster-Baum kein Bedarf besteht[47], gehe ich von den Paaren mit der höchsten Ähnlichkeit aus, verwende also ein agglomeratives Verfahren. Das am wenigsten rechenintensive Verfahren, das ohne Neuberechnung der Ähnlichkeitswerte auf Cluster-

[45] Die Mengen korrespondierender Anliegen der beiden betrachteten Sprachen können bzgl. ihrer relativen Größe stark voneinander abweichen; in vielen Fällen gibt es zu einer Menge in der jeweils anderen Sprache keine Entsprechung, so daß keine übereinstimmende Anzahl an Clustern zu erwarten ist.

[46] Der Median liegt bei beiden Sprachen zwischen 0,02 und 0,04.

[47] Im Weiteren werde ich Ähnlichkeitswerte $\leq 0{,}1$ nicht weiter betrachten; Stichproben ergaben einen Bereich von 0,23 bis 0,26, der im allgemeinen in Bezug auf das formulierte Anliegen korrekt zusammengefaßte Cluster von inkorrekten trennt.

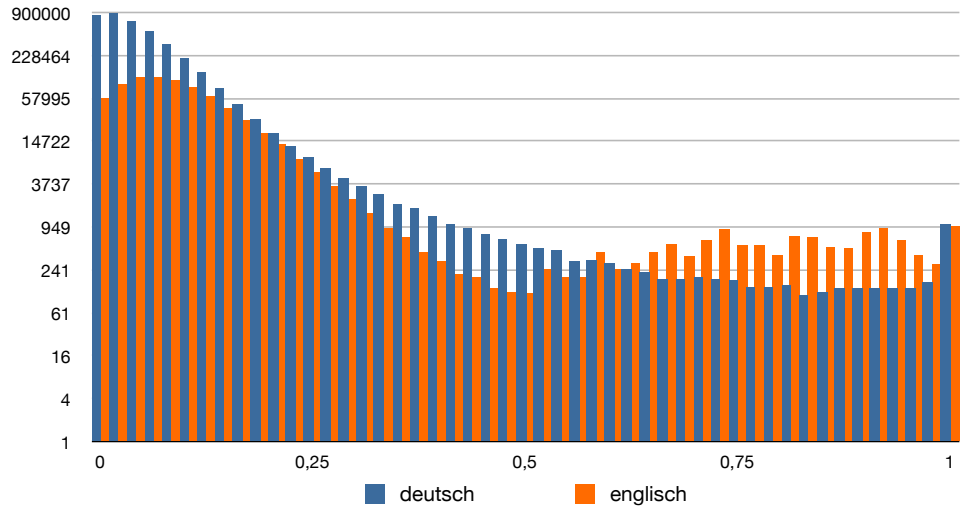

Abbildung 12: Verteilung der berechneten Ähnlichkeitswerte (die Anzahl ist logarithmisch aufgetragen).

Basis auskommt ist *Single Link*[48]. Der für gemeinhin als Nachteil betrachtete Effekt des *Chaining*[49] kann durchaus einen positiven Nebeneffekt haben, und zwar dann, wenn über „Ketten“ von Texten andere Texte - aufgrund einer hohen Ähnlichkeit zu diesen - hinzugefügt werden, deren Wörter aber nicht mit denen am anderen Ende der Kette übereinstimmen (Ein Beispiel dafür befindet sich in Abb. 26 im Anhang).

Der eigentliche Clustering-Algorithmus greift linear auf die nach Ähnlichkeitswert absteigend sortierten Textpaare zu und integriert diese in die bestehende Clusterstruktur:

1. Sind beide Texte noch nicht Teil eines anderen Clusters wird ein neues Cluster der Größe zwei mit eben diesen Texten angelegt.

2. Ist einer der beiden Texte bereits Bestandteil eines oder mehrerer Cluster wird dasjenige dieser Cluster zusammen mit dem anderen Text zu einem neuen Cluster zusammengefaßt, das nicht Bestandteil eines anderen Clusters ist.

3. Wenn beide Texte in einem Cluster enthalten sind, werden diese analog zu (2) in einem neuen Cluster zusammengeführt, sofern sich die Cluster-

[48]Siehe Manning, Raghavan und Schütze (2008), Kapitel 17.2.
[49]Ebd.

Strukturen nicht überschneiden, d.h. beide Texte sich bereits im selben Cluster befinden.

Nach Durchlaufen des Clustering-Algorithmus über die beschriebene Untermenge der Anfragen, ergeben sich 1157 Cluster bei den englischsprachigen Texten und 2716 bei den deutschsprachigen. Diese gehen zu rund 50% auf (2), und zu je 25% auf (1) und (3) zurück, d.h. die Cluster, die sich aus einem bereits bestehenden Cluster sowie einem noch nicht verarbeiteten Text zusammensetzen, überwiegen. Von der Cluster-Struktur her bleiben bei Erreichen der festgelegten Grenze von 0,1 in beiden Sprachen 35 bzw. 31 unverbundene Cluster, von denen je eines sehr groß ist (ca. 2500 bzw. 1000 Elemente) und die restlichen über jeweils weniger als 40 Elemente verfügen.[50]

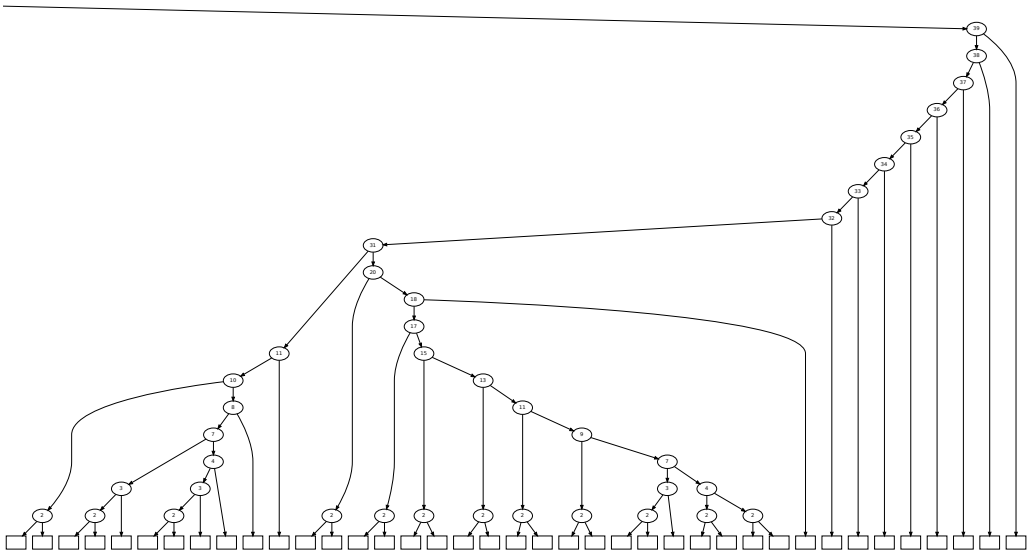

Abbildung 13: Ausschnitt aus einer Clusteranalyse als Dendrogramm.

Die größte Schwierigkeit ist im Anschluß herauszufinden, an welcher Stelle einzelne Äste des Cluster-Baumes abgeschnitten werden sollten, so daß der entsprechende Ast lediglich die zu einem Anliegen zugehörigen Texte umfaßt[51], dies für den nächsthöheren Knoten aber nicht mehr gilt. Eine mögliche Heuristik ist, jeden Ast abzuschneiden, für den die Differenz zwischen den Ähnlichkeitswerten des aktuellen und des nächsthöheren Knotens einen bestimmten Wert überschreitet. Eine andere betrachtet das Verhältnis der

[50] In diesen fanden sich bei einer manuellen Überprüfung keine relevanten Anfragen.

[51] Es ist denkbar, daß sich zwei oder mehr Cluster zum gleichen Anliegen an unterschiedlichen Stellen des Cluster-Baumes befinden und erst zusammengeführt werden, wenn bereits zu einem anderen Anliegen gehörende Cluster in den Baum integriert worden sind.

Anzahl der Elemente des aktuellen Knotens zur Differenz der Anzahl beider Knoten, d.h. eine Vereinigung zweier großer Cluster würde dispräferiert und stattdessen lieber das kleinere abgetrennt.

Ein Vergleich der verschiedenen Heuristiken mit manuell markierten „Sollbruchstellen“ ergab kaum Übereinstimmungen. Es scheint als wären die Anzahl der Elemente eines Clusters und sein Ähnlichkeitswert einem Text eines anderen Clusters gegenüber nicht ausreichend, um eine zuverlässige Entscheidung bzgl. des Abtrennens dieses Clusters vom Rest des Baumes treffen zu können. Bei der vorhandenen Anzahl an Clustern läßt sich diese Aufgabe mithilfe einfacher Heuristiken im Vorfeld stark vereinfachen[52] und anschließend an einen menschlichen Bewerter deligieren.[53]

4.3 Verbesserung des Verfahrens mithilfe individueller Stoppwörter

Beim manuellen Bewerten der heraussortierten Cluster fiel auf, daß die Texte einzelner kleinerer Cluster kaum inhaltliche Ähnlichkeit aufwiesen, aber alle vom selben Autor stammten. Trotz des Bereinigens der Texte im Preprocessing-Schritt um den aus dem Header stammenden Namen des Autors und des Nichtbeachtens von Signaturen ist es möglich, daß diese Daten - oder Fragmente davon - sich nach der Bereinigung noch im Text wiederfinden.[54] Diese einmal indizierten Wörter werden daraufhin von allen weiteren Verarbeitungsschritten zur Identifikation von Anliegen in die Berechnungen mit einbezogen, und da sie meist eine niedrige Termfrequenz haben,[55] ist ihr idf-Wert umso höher und wirkt somit mit einem i.d.R. höheren Gewicht für ein Clustering nach Autor als nach Anliegen.

Meine Grundannahme diesbezüglich war, daß sich diese beim Clustering Fehler verursachenden Wörter zur Auswahl aller Texte eines Autors analog verhalten wie *Stoppwörter* zur gesammten Kollektion. Deshalb habe ich eine Gewichtungsfunktion ψ definiert, die für ein Wort, einer Kollektion (d.h. alle Texte einer Sprache γ) und eine Funktion Σ, die diese Kollektion auf

[52] Wenn man annimmt, daß sich die relevante Anzahl an Texten im Bereich von $(5, 50)$ bewegt und Cluster, die nur Texte der Ähnlichkeit 1 - also identische Texte - enthalten, nicht relevant sind, reduziert sich die Anzahl der zu betrachtenden Cluster auf lediglich 230 bzw. 145.

[53] Abb. 26, 27 und 28 im Anhang demonstrieren den Inhalt manuell selektierter Cluster. Abb. 34 zeigt ein Dendrogramm mit manuell als korrekt bewerteten Clustern.

[54] Manche Autoren verwenden auch Grußformeln, die es leichtmachen, die von ihnen verfaßten Texte mit hoher Genauigkeit zu identifizieren.

[55] Mit Ausnahme relativ weit verbreiteter Namen wie beispielsweise Daniel, Christian oder Stefanie.

eine Teilmenge reduziert (hier von der Gesamtheit aller Text in γ auf die Gesamtheit aller Texte eines Autors α in γ), einen Wert zurückliefert, der beschreibt, wie prominent jenes Wort in der Kollektion des Autors verglichen mit der Gesamtkollektion ist:

$$\psi_\gamma(\lambda, \Sigma_\alpha) = 1 - \frac{|\lambda \in \bigcup \gamma| - |\, \lambda \in \Sigma_\alpha(\bigcup \gamma)|}{|\lambda \in \bigcup \gamma|} \tag{12}$$

Für jedes Wort λ, das nicht Bestandteil der Texte des Autors $\Sigma_\alpha(\bigcup \gamma)$ ist, ergibt sich $\psi_\gamma(\lambda, \Sigma_\alpha)$ zu 0. Sind alle Texte in denen das Wort λ vorkommt, Werke des einen zu betrachtenden Autors, liefert $\psi_\gamma(\lambda, \Sigma_\alpha)$ 1 zurück. Versuche ergaben 0,1 als geeigneten Wert zur Abgrenzung dieser „persönlichen Stoppwörter“, d.h. wenn 10% oder mehr der Vorkommen eines Wortes in den Texten des betrachteten Autors zu finden sind, gilt dieses als eines seiner *Stoppwörter*[56] und wird vor der weiteren Verarbeitung aus seinen Texten entfernt.[57] Die folgende Funktion filtert anhand dieses Kriteriums alle individuell definierten*Stoppwörter* des jeweiligen Autors aus einem gegebenen Textvektor heraus:

$$\Upsilon_\gamma(p, \alpha) = \forall\ \lambda \in p\colon \psi_\gamma(\lambda, \Sigma_\alpha) \not\geq 0{,}1 \tag{13}$$

Damit läßt sich nun die ursprüngliche Definition von J'' erweitern:

$$J''_{\gamma,\alpha}(p, q) = \frac{\sum\limits_{\lambda \in \Upsilon_\gamma(p,\alpha) \cap \Upsilon_\gamma(q,\alpha)} \min(tf(\lambda, p), tf(\lambda, q)) \cdot idf(\gamma, \lambda)}{\left[\begin{array}{r} \sum\limits_{\lambda \in \Upsilon_\gamma(p,\alpha)} [tf(\lambda, p) \cdot idf(\gamma, \lambda)] \\ + \sum\limits_{\lambda \in \Upsilon_\gamma(q,\alpha)} [tf(\lambda, q) \cdot idf(\gamma, \lambda)] \\ - \sum\limits_{\lambda \in \Upsilon_\gamma(p,\alpha) \cap \Upsilon_\gamma(q,\alpha)} [\min(tf(\lambda, p), tf(\lambda, q)) \cdot idf(\gamma, \lambda)] \end{array}\right]} \tag{14}$$

Eine globale Verbesserung anhand der dadurch neu erstellten Clusteranalyse läßt sich nur schwer belegen, da schon kleine Änderungen in der Ähnlichkeitsbewertung große Folgen auf die Clusterbildung haben können. Allerdings sind die Veränderungen im Kleinen ziemlich überzeugend, was Anlaß zur Vermutung gibt, daß sich diese auch entsprechend auf die Clusterstrukturen auswirken wird. Die drei Beispiele in Abb. 14 stammen aus der Kollektion der deutschsprachigen Anfragen.

[56] Siehe auch Abb. 33 im Anhang.

[57] Das Verfahren funktioniert nicht für Mitarbeiter, die ein Vielfaches an Texten, verglichen mit einem durchschnittlichen Fragesteller, produzieren. Diese Texte sind allerdings vom *Issue-Tracking-System* i.d.R. bereits als Antwort auf eine Anfrage markiert worden und gehören somit nicht zur Kollektion der erstmaligen Anfragen.

```
Im Anhang die Austrittserklärung.
```
$p_5 \quad = \quad \{\ \lambda_{im},\ \lambda_{anhang},\ \lambda_{die},\ \lambda_{\text{austrittserklärung}}\ \}$

```
Hallo,
im Anhang meine Austrittserklärung.
MfG
B***** S******
```
$p_6 \quad = \quad \{\ \lambda_{hallo},\ \lambda_{im},\ \lambda_{anhang},\ \lambda_{meine},\ \lambda_{\text{austrittserklärung}},\ \lambda_{mfg},\ \lambda_1,\ \lambda_2\ \}$

```
Hallo,
anbei meine Austrittserklärung!
MfG
F****** S********
```
$p_7 \quad = \quad \{\ \lambda_{hallo},\ \lambda_{anbei},\ \lambda_{meine},\ \lambda_{\text{austrittserklärung}},\ \lambda_{mfg},\ \lambda_3,\ \lambda_4\ \}$

Abbildung 14: Drei aus der Kollektion entnommene Beispiele. λ_1 bis λ_4 sind die Wörter, die den Eigennamen aus p_6 und p_7 entsprechen.

Der Vergleich der Ähnlichkeitswerte in Abb. 15 zeigt, daß das Herausfiltern der Eigennamen mittels individueller Stoppwörter eine deutliche Verbesserung bewirkt[58] - in diesen Fällen fast eine Verdoppelung der Werte -, so daß davon auszugehen ist, damit diese Fälle früher in einem Cluster zusammenfassen zu können als bislang. Insgesamt reduziert sich die Anzahl der Cluster auf der Höhe von 0,1 von 35 auf 3 in der Kollektion der deutschen und von 31 auf 7 in der Kollektion der englischen Texte, dabei handelt es sich wiederum um ein Cluster mit nahezu allen Texten und wenige kleinere, die in diesem Fall auf englischer Seite ausschließlich *Spam* und auf deutscher ausschließlich automatisierte Rechnungen enthalten - also keine Texte, die für die weitere Verarbeitung von belang wären.

	J''_{γ}	$J''_{\gamma,\alpha}$
(p_5, p_6)	0,3094	0,6250
(p_5, p_7)	0,1302	0,2138
(p_6, p_7)	0,1899	0,4834

Abbildung 15: Vergleich der Ähnlichkeitswerte ohne und mit Einbeziehung der individuellen Stoppwörter der jeweiligen Autoren.

[58]Der große Unterschied der Werte rührt vor allem von den hohen idf-Gewichten der Eigennamen her.

4.4 Indirekte Clusteranalyse über Ähnlichkeitswerte der Antworten

Wesentlich homogener als die Anfragen sind die Antworten formuliert. Trotz geringer Abweichungen hinsichtlich Ausführlichkeit und Stil wird per Konvention[59] auf die gleichen Begrifflichkeiten zurückgegriffen, wenn es um die Benennung von Gegenständen, Vorgängen oder Institutionen geht. Teilweise ist die „richtige“ Wortwahl auch schon durch die Satzung des Vereins oder anderweitige rechtliche Bestimmungen vorgegeben. In solchen Fällen werden den eigentlichen Begriffen noch umschreibende Erläuterungen zur Seite gestellt, falls die Anfrage den Eindruck erweckt, ihr Verfasser könnte u.U. Probleme mit dem Verständnis der jeweiligen Fachtermini haben. Aufgrund der teilweise vorgegebenen Wortwahl und der vergleichsweise geringen Anzahl an Autoren der Antworten, war meine Annahme, daß eine Clusteranalyse auf eben jenen Antworten deutlich bessere Ergebnisse liefern müßte als die der Anfragen.

In der Tat sind die Cluster auch bei eher geringen Ähnlichkeitswerten (0,26 und 0,24 bei Abb. 29 bzw. 31) thematisch passend, wenn auch nicht so homogen, wie es ein Standardtext wäre. Wenige Cluster sind komplett unbrauchbar wie in Abb. 30 zu sehen.[60] In einigen Fällen - wie in Abb. 31 und 32 - ist die Ähnlichkeit der zu dem gebildeten Cluster gehörenden Anfragen offensichtlich, in anderen basiert die formulierte Antwort nicht nur auf dem Text der Anfrage, sondern auch auf eventuellen vorhergehenden Kommunikationen (mitunter ist der Anfragende sogar persönlich bekannt), an die eMail angehängten Dateien (meist ausgefüllten Formularen), sowie den Informationen aus der eingangs erwähnten Mitglieder-Datenbank. Des weiteren besteht zwischen Anfragen und Antworten (selbst bei Ausschluß derjenigen Anfragen, die mehrere Anliegen zum Ausdruck bringen) keine **1:1**-, sondern eine **n:m**-Beziehung, d.h. es ist nicht der Fall, daß eine Art von Anfrage immer eine bestimmte Antwort zur Folge hat; die Frage, warum keine Inter-

[59] Feste Regeln, wie genau auf die Anfragen zu antworten ist, gibt es im Gegensatz zu den Sprachregelungen vieler Unternehmen im Verein keine. Als Referenz für Formulierungen stehen bereits beantwortete Anfragen zur Verfügung.

[60] Die beiden angeführten Texte unterscheiden sich darin, daß der eine das Gegenteil des anderen besagt. Inhaltlich gesehen weisen sie deshalb eine sehr geringe Ähnlichkeit auf. Aus Sicht der Clusteranalyse anhand eines *bag of words* allerdings zählt die Negation der Gesamtaussage nur in Form des Vorhandenseins eines Negationspartikels in einem Text, der in dem anderen nicht vorkommt. In Fällen, in denen ein Inhaltswort die Negation trägt (z.B. „entsperren“), geht (9) auf Seite 20 folgend der inhaltliche Unterschied mit den idf-Gewichten beider Wörter in das Ähnlichkeitsmaß ein, was in Relation zu eventuell vorhandene Inhaltswörtern auch nicht ausreicht, um den Ähnlichkeitswert beider Texte deutlich zu reduzieren.

netverbindung bestehe, kann je nach Fall beispielsweise auf einen fehlenden Mietvertrag, eine ungedeckte Bankverbindung, einen Verstoß gegen die Nutzungsbedingungen oder einfach einen technischen Fehler zurückzuführen sein. Entsprechend kann ein und dieselbe Anfrage in Abhängigkeit davon, wer sie wann stellt, ganz unterschiedliche Antworten hervorrufen. Andersherum gilt auch nicht, daß eine Anwort nur auf eine Art von Anfragen gegeben wird; die Aussage „ist eingetragen“ kann sich als Antwort auf einen Mietvertrag, eine Einzugsermächtigung, einen Antrag auf „ruhende Mitgliedschaft“ usw. beziehen.

Obige Aspekte lassen den Schluß zu, daß eine Clusterbildung ausschließlich anhand der gegebenen Antworten nicht zum Erfolg führen kann. Eine Überprüfung der in Kapitel 4.2 und 4.3 gewonnen Cluster anhand der gegebenen Antworten wäre jedoch denkbar. Allerdings ist die in Kapitel 4.2.2 skizzierte manuelle Überprüfung der relevanten Cluster mit weniger Aufwand verbunden und per Definition[61] genauer, so daß ich diesen Ansatz nicht weiter verfolge.

[61] Von einer manuellen Kontrolle kann bei einer nur geringen Anzahl an zu kontrollierenden Elementen eine nahezu 100%ige Exaktheit erwartet werden.

5 Bewertung und Verbesserungsmöglichkeiten

5.1 Grenzen der Verfahren

- Es ließ sich kein Weg finden, automatisch die korrekten von den inkorrekten Clustern zuverlässig zu trennen. Hier war eine manuelle Bewertung von Nöten.

- Trotz vorhergehender Filterung war nicht zu verhindern, daß einzelne Wörter hohen idf-Gewichts aber ohne Bezug zum Anliegen (wie beispielsweise Eigennamen)[62] die Clusteranalyse beeinflussten.

5.2 Verbesserungsmöglichkeiten

Filterung

- **Anrede** und **Grußformel** können anhand ihrer Position (Anfang sowie Ende des Textes bzw. auch eines *Text*-Blockes) identifiziert und dann von der weiteren Verarbeitung ausgeschlossen werden. Dazu müßten entweder bereits beim Parsen (siehe Kapitel 2.3) entsprechend definierte Muster genutzt oder aber pro Wort eine Liste der Vorkommen in Form von Zeilennummern übergeben werden, mithilfe derer anhand einer separaten Clusteranalyse der entsprechenden Zeilen extrahiert werden könnte, was als Anrede bzw. Grußformel zählt. Grundsätzlich befinden sich die eigentlichen Informationen im beschreibenden Fließtext, also dem Part zwischen Anrede und Gruß, weshalb dieses Verfahren die Relevanz des zu analysierenden Textes erhöhen würde.

- Bestimmte Begriffe wie **Wochentage**, **Monatsnamen** oder die Bezeichnungen der **Wohnheime** beeinflussen die Clusteranalyse dahingehend, daß bei Fehlen aussagekräftigerer Wörter (mit hohem idf-Gewicht) Texte beispielsweise nach dem Wohnheim des Absenders in Clustern zusammengefaßt werden könnten, wobei dieses von ihm nur als Referenz zur Identifizierung seiner Person genannt worden sein mag. Eine (eventuell auch manuell erstellte) **Stoppwort-Liste** solcher Begriffe dürfte die Qualität der Cluster noch ein wenig verbessern. Des Weiteren würde die Identifizierung der Referenzen von Zeit und Ort einer späteren automatischen Beantwortung einen Teil der dabei benötigten Daten liefern.

[62] Wenn es sich beispielsweise um den Namen des Angesprochenen handelt, greift die Filterfunktion aus Kapitel 4.3 nicht.

- In vielen Fällen sind die eingehenden Anfragen Reaktionen auf eine von unserem System automatisch versandte eMail. **Zitate** werden als Ergebnis des Parsens gleich zu Beginn herausgefiltert, sofern beim Zitieren korrekte Zitierformen verwendet wurden.[63] Die Information darüber, welche eMail[64] Anlaß zum Antworten gab, geht dabei bislang verloren. Sofern der Betreff nicht verändert wurde, läßt sich aus diesem leicht auf die zitierte eMail rückschließen.[65] Das Wissen um die zitierte eMail kann aber von Nutzen sein, wenn es darum geht, fehlerhaft formatierte eMails zu „säubern"; in ungünstigen Fällen gelangen nämlich Teile der zitierten eMail in einen *Text*-Block und werden so als Teil der Anfrage weiterverarbeitet.[66] Unter Verwendung einer zur zitierten eMail gehörenden Wortliste ließe sich dieser Fehler korrigieren.

Korrektur/Anreicherung der Eingabedaten

- Ein **phonologischer Vergleich** der Wörter auf Basis der Levenshtein-Distanz mit den bereits im Lexikon vorhandenen Wörtern[67] wie in Kapitel 4.1 beschrieben dürfte einige Fehler der zu analysierenden Texte zu korrigieren helfen - insbesondere, wenn es sich bei den Autoren nicht um deutsche oder englische Muttersprachler handelt. Denkbar wäre auch, diese Vergleiche auf das konstruierte Lexikon anzuwenden, um so regelmäßige **Falschschreibungen** herauszufinden.

- Deutschsprachige Texte wiesen bei der Clusteranalyse höhere Ähnlichkeitswerte auf als englische, wenn relevante Wörter enthalten waren.

[63] Kleinere Abweichungen wie Umbrüche, die zu vereinzelten Wörtern in nicht als Zitat gekennzeichneten Zeilen führen, können anhand der resultierenden Struktur erkannt und korrigiert werden.

[64] Die Anzahl der automatisch versendbaren eMails ist begrenzt, da diese als *Templates*, also Texte mit Platzhaltern, vorliegen.

[65] In Kapitel 2.3 erwähne ich als weiteres Gegenargument das nicht inhaltlich motivierte Anworten auf eMails. So erreichen uns als Antwort auf die Begrüßungs-eMail, in der wir den Verein vorstellen, allerlei Anliegen, die keinen Bezug dazu aufweisen.

[66] Mitunter ändert sich hierbei auch die Sprach-Klassifizierung von englisch zu deutsch, da die automatisch versandten *Templates* immer zweisprachig gehalten sind und die deutsche Sprache Komposition bei der Wortbildung bevorzugt, so daß - relativ gesehen - wenigen Wörtern mit höherem idf-Gewicht eine größere Anzahl mit niedrigerem idf-Gewicht gegenübersteht. Wie in der Definition von $\xi_\gamma(\lambda)$ in (3) auf Seite 13 zu sehen, geht die Größe des Lexikons in die Berechung der Wahrscheinlichkeit ein; die englische Sprache weist mit 34317 Wörtern ein größeres Lexikon auf als die deutsche mit 28150.

[67] Hierbei sollten natürlich nur diejenigen als Vergleichswerte herangezogen werden, die mit einer nicht zu geringen Frequenz im Lexikon vorkommen, und bei im Sinne der Levenshtein-Distanz gleichwertigen Alternativen wäre ein geringerer idf-Wert ausschlaggebendes Kriterium.

Ein gewichtiger Grund hierfür ist die **Wortstruktur** von Fachtermini. Während für die deutsche Sprache die Komposition das Mittel der Wahl ist, wenn es um Fachtermini geht, tendiert das Englische eher zur Bildung von **Mehrwortausdrücken** (*multiword expressions*, *MWE*).[68] Diese neigen wiederum zur Variation u.a. in Abhängigkeit von der Herkunft des Autors.[69] So kann die Wahl des jeweiligen Ausdrucks für einen Referenten (Mietvertrag, Hausmeister, Überweisung, …) davon abhängen, ob es sich beispielsweise um einen Amerikaner, Engländer, Inder oder Chinesen handelt.

- In einigen Fällen führt fehlende Kenntnis der Orthographie (meist der deutschen) dazu, daß relevante Wörter nicht in die Berechnung der Ähnlichkeit eingehen.[70] Ein ***Stemming*** der Eingabewörter könnte dem entgegenwirken. Allerdings können Fälle mit Stammänderung bei der Flexion (*„überweist“ beispielsweise) nicht ohne Aufwand korrekt behandelt werden. Externe Ressourcen wie ein Lexikon oder mit morphologischen Daten trainierte Korrekturregeln in Form *Regulärer Ausdrücke* könnten von Nutzen sein.

- An der Stelle, an der die Zeilen in verschiedenartige Blöcke unterteilt worden sind, selektiert ein *Regulärer Ausdruck* aus den verbleibenden Zeichen die Wörter, indem er Buchstaben[71] von Nicht-Buchstaben trennt. Hier gibt es ein großes Verbesserungspotential, da durch Nicht-Buchstaben (wie Apostroph oder Bindestrich) verbundene Teile eines Wortes bislang separat behandelt wurden.[72]

Ähnlichkeitsfunktion

- Da die **Länge eines Wortes** (mit Ausnahme von Abkürzungen) i.d.R. proportional zu seinem Informationsgehalt ist[73], würde diese als Bestandteil der Ähnlichkeitsbewertung neben der *inversen Dokumentfre-*

[68] Vergleiche „Einzugsermächtigung“ und „direct debit authorization“.

[69] Für „Mietvertrag“ finden sich in der Kollektion der englischsprachigen Texte die *MWE*s „tenancy agreement“, „rental agreement“, „tenancy contract“ und „rental contract“ in etwa gleicher Häufigkeit. Andere *MWE*s wie „lease agreement“ oder „room contract“ treten seltener auf.

[70] Es handelt sich hierbei hauptsächlich um Plural- und Partizipbildung: *„Formulars“ statt „Formulare“, *„überweist“ statt „überwiesen“.

[71] Alle Grapheme, die in *UTF-8*-Kodierung als Buchstabe markiert sind.

[72] So wird „don't“ zu „don“ und „t“ und „E-Mail“ zu „e“ und „mail“ verarbeitet.

[73] Funktionswörter wie Artikel, Präposition und Pronomina (die nicht identisch mit den *Stoppwörtern* sein müssen, aber im allgemeinen eine große Überschneidung mit diesen aufweisen) sind in vielen Sprachen einsilbig.

quenz ebenfalls zur Präferenz der relevanteren Wörter beitragen. Die Unterteilung der Wörter in **Silben** ist ein nicht triviales Unterfangen. Das dafür benötigte morphophonologische Wissen in Form von Regeln ist in weiten Teilen das gleiche, das auch Bestandteil der oben beschriebenen Erweiterung des Levenshtein-Vergleichs wäre. Denkbar ist aber auch die bloße Anzahl an Zeichen oder die Anzahl an Vokalen und Konsonanten als Maß zu verwenden.[74]

- Einzelne Cluster zeichnen sich dadurch aus, daß sie bei einer hohen Ähnlichkeit komplett gegensätzliche Anliegen aufweisen, Abb. 30 im Anhang ist so ein Fall. Für die Ähnlichkeitsfunktion unterscheiden sich Negationswörter wie „nicht“ oder „kein“ nicht von anderen, die ähnliche tf- und idf-Werte aufweisen.[75] Da die Möglichkeiten, eine Aussage zu negieren, äußerst facettenreich sind[76], gibt es keinen Ansatz, alle zu behandeln. Eine mögliche Vorgehensweise zur Korrektur dieses Typs von Falschklassifizierungen wäre, die Wörter, deren Negationscharakter bekannt ist, zu markieren und diesen als zusätzliche, höher gewichtete Dimension bei der Ähnlichkeitsberechnung miteinzubeziehen.[77]

Clusteranalyse

- Bislang kommt *Single Link* als Verfahren zur Clusteranalyze zum Einsatz, da es von allen bekannten Verfahren bei vorberechneten Ähnlichkeitswerten am performantesten ist und sich der Negativeffekt des *Chaining* in diesem Fall als nützlich herausgestellt hat. Es besteht durchaus die Möglichkeit, daß andere Verfahren bessere Ergebnisse liefern. So würde die Verwendung von *GAAC* (*group-average agglomerative clustering*)[78] kompaktere Cluster produzieren, dabei dem *Chaining* aber entgegenwirken.

[74]Um letzteres zu ermitteln, könnte das Wort als Zeichenkette mithilfe *Regulärer Ausdrücke* heuristisch auf Vokal- und Konsonantensymbole reduziert werden.

[75]Der idf-Wert von „nicht“ liegt bei 0,5, der von „kein“ bei 1,27. Damit kommt „nicht“ etwa gleich häufig vor wie die bestimmten Artikel und könnte auch als *Stoppwort* betrachtet werden.

[76]Negation kann zum einen durch Negationspartikel, Quantoren, usw. ausgedrückt werden, zum anderen durch morphologische Prozesse wie Derivation der Wortbedeutung inhärent sein („Unwillen“, „irrelevant“, etc.).

[77]Ein großes Problem, das dabei zu lösen sein wird, sind die unterschiedlichen Arten, wie verschiedene Sprachen Negation interpretieren. Teilweise treten selbst innerhalb verschiedener Ausprägungen (Hochsprache, Dialekt, Umgangssprache, …) einer einzigen Sprache Unterschiede auf: „I didn't sign anything.“ entspricht inhaltlich „I didn't sign nothing.“, es handelt sich jedoch bei ersterem um einfache und bei letzterem um doppelte Negation.

[78]Siehe Manning, Raghavan und Schütze (2008), Kapitel 17.3.

6 Der Weg zur Beantwortung der Anfragen

Eine generierte Antwort soll - wie eingangs erwähnt - vor allem einfach formuliert und damit leicht verständlich sein, auch für diejenigen, die weder über gute Deutsch- noch Englischkenntnisse verfügen. Die Wortwahl ist - wie in Kapitel 4.4 beschrieben - teilweise durch Bestimmungen wie das Vereinsrecht etc. vorgegeben, was i.d.R. im Widerspruch zur Anforderung der leichten Verständlichkeit steht. Im Normalfall reicht es aus, beide Begrifflichkeiten synonym einzuführen; manchmal folgt dem kurzen offiziellen Text auch eine längere umschreibende Erläuterung.

Bislang werden automatische eMails als *Templates*, also Texte mit Platzhaltern realisiert. Bei Reiter und Mellish (1993) heißt diese Repräsentation von Generierungsdaten *canned text with embedded KB (knowledge base) references* (*EKR*). Die nächstkompliziertere Form nennen sie *case frames with textual case fillers* (*TCF*). Diese stellt im Grunde einen *semantischen Frame* à la FrameNet[79] samt Füllern dar und übersteigt bei weitem die gegebenen Anforderungen. Die Verwendung von *EKR* bzw. *Templates* erfüllt neben obigen Anforderungen auch die Prämisse, daß Personen ohne explizites linguistisches Wissen in der Lage sind, vorhandene *Templates* zu modifizieren und neue zu erstellen.

Der komplette Ablauf vom Eingang der eMail bis zu derer Beantwortung stellt sich dann bei Verwendung von *Templates* wie folgt dar:

1. Eine eMail erreicht das *Issue-Tracking-System* und wird von ihm in ihre Bestandteile (wie in Kapitel 2.2 beschrieben) zerlegt.

2. Die Preprocessing-Komponente bekommt den unformatierten Text übergeben und unterteilt ihn mithilfe des Parsers aus Kapitel 2.3 in Blöcke unterschiedlichen Typs.

3. Aus den *Text*-Blöcken werden Wörter extrahiert, die zusammen mit der Anzahl ihres Vorkommens einen Vektor bilden.

4. Mittels eines Ähnlichkeitsmaßes wie beispielsweise der Euklidischen Distanz werden die in Kapitel 4.2.2 als korrekt markierten Cluster mit

[79] Siehe Baker, Fillmore und Lowe (1998) sowie Ruppenhofer u. a. (2005).

dem Vektor des neuen Textes verglichen und entsprechend des Ergebnisses sortiert.[80]

   (a) Falls ein Cluster sich bzgl. seiner Ähnlichkeit zu dem neuen Text deutlich von den anderen abhebt, wird der Text diesem Cluster zugeordnet.

   (b) Gibt es mehrere Cluster mit ähnlich hohen Ähnlichkeitswerten, werden diese einem Benutzer zur Auswahl überlassen.

   (c) Überschreitet kein Cluster mit dem ihm gegenüber ermittelten Ähnlichkeitswert ein festgelegtes unteres Limit, wird der Text als nicht zuordenbar gekennzeichnet.

5. Ist eine Zuordnung geschehen (automatisch oder manuell), hängt das weitere Vorgehen von der Art der Anfrage ab (siehe auch Kapitel 1.1). Im allgemeinen Fall findet eine Interaktion mit der Mitglieder-Datenbank und/oder einem menschlichen Benutzer statt, der die nicht automatisierbaren Schritte übernehmen und das Ergebnis zurückmelden muß.

6. Sind alle für das vorliegende Anliegen definierten Bedingungen erfüllt, wird das zugehörige *Template* mit den ermittelten Daten befüllt und versandt. Gegebenenfalls lassen sich einzelne generische Abschnitte der Antworten in Abhängigkeit des in der Mitglieder-Datenbank abgebildeten Zustand an- bzw. abschalten.

Der die Cluster erzeugende Analyseprozess ist sehr ressourcenintensiv[81], von daher ist abzuwägen, ob und wenn ja wann ein erneuter Durchlauf nötig ist. Dieser könnte u.U. eine komplett andere Clusterstruktur zur Folge haben und damit neben der manuellen Kontrolle der Cluster auch ein Neuverknüpfen aller Cluster mit den durch Bedingungen und *Templates* charakterisierten Anliegen erfordern. Eine einfachere Alternative dazu ist das manuelle Gruppieren von Texten zu einem Cluster und seine Verknüpfung zu einem Anliegen, sobald ein solches neu hinzukommen ist.

[80] Das einfachste Vorgehen ist sicherlich der Vergleich mit den Zentroiden der einzelnen Cluster. Bei Verwendung von *Single Link* als Clusteranalyseverfahren kann es jedoch vorkommen, daß der zu vergleichende Text identisch mit einem ist, der zu einem Cluster mit ausgeprägtem *Chaining* gehört, jedoch eine höhere Ähnlichkeit zum Zentroiden eines anderen aufweist, da dieser „näher“ liegt. In diesem Fall böte sich für den Vergleich entsprechend wieder *Single Link* als Verfahren an, das als Ähnlichkeitsmaß des Textes zu einem Cluster die Ähnlichkeit des maximal ähnlichsten Elementes dieses Clusters verwendet.

[81] Sowohl hinsichtlich des Festplattenspeicherplatzes als auch der Rechenzeit.

7 Schlußbetrachtung

Eingehende eMails werden mithilfe *Regulärer Ausdrücke* und eines Parsers aufbereitet und verschiedene Klassifizierungsverfahren sorgen für die korrekte Einordnung bezüglich Sprach und Anliegen. Bis auf eine manuelle Bewertung der einmalig erzeugten Clusterstruktur verläuft die Bearbeitung automatisch.[82]

Die Strukturbestimmung der eMails liefert die gewünschten Ergebnisse und schlägt mit 0,09% (wie in Kapitel 2.3. zu sehen) erstaunlich selten fehl. Obwohl der Mechanismus zur Wortermittlung extrem einfach konzipiert ist, liefert die darauf aufbauende Clusteranalyse Ergebnisse, die mehr als zufriedenstellend sind. Dazu kommt, daß sich der für allgemeinhin als negativ betrachtete *Chaining*-Effekt des *Single-Link*-Verfahrens hinsichtlich der intendierten Ähnlichkeit der Anfragen offenbar positiv auf die resultierenden Cluster auswirkt.

Nicht geplant war hingegen der manuelle Eingriff bei der Qualitätsbewertung der erzeugten Cluster. Möglicherweise kann einer der Vorschläge aus Kapitel 5.2 - insbesondere die Verwendung eines anderen Clusteranalyseverfahrens - hier Abhilfe schaffen. Ebenso war angedacht, der Clusteranalyse weniger Raum (und Zeit) zu geben und dafür am Ende eine komplette Prozessierungskette am Laufen zu haben. So ist allerdings die Qualität der erzeugten Cluster recht hoch; und die Verwendung von *Canned Text* oder *Templates* zur Generierung der Antworten ist gegenüber der Clusteranalyse ein vergleichsweise geringer Aufwand. Ein systematischer Vergleich des aktuellen Zustand mit der erzielten Leistungssteigerung jeder der aufgelisteten Verbesserungsmöglichkeiten ist im Gegensatz dazu kaum realisierbar.

Alles in allem überwiegt aus meiner Sicht der Erkenntnisgewinn in der Tiefe in einem Modul (hier Kapitel 4) den Effekt, den eine vollständige Verkettung mehrerer eher oberflächlicherer Module (wie Kapitel 3.2 beispielsweise) gehabt hätte. Die zur Integration in das bestehende *RT*-System noch nötigen Schritte sind eher handwerklicher Natur. Vom erstmaligen Einsatz bis zur Bewertung der Qualität des Gesamtsystems muß ein Vielfaches der Anzahl möglicher Anliegen als Antworten vorliegen, d.h. um sichere Aussagen über eventuelle Fehler treffen zu können, ist eine intensive Nutzung des Systems über einen Zeitraum von einigen Monaten erforderlich.

[82] Bei der eigentlichen Antworterstellung, die durch diese Arbeit nicht abgedeckt wird, ist menschliche Interaktion unumgänglich wie in Kapitel 1.1 geschildert.

A Anhang

A.1 Skripte und Grammatiken

```
#!/bin/sed -rnf

s/([[:alnum:] ]*)/>>\1<</
s#(qt )+#<QB>&</QB>#g
s#</QB>t0 <#qt </QB><#g
s#(t2 |q1 (t2 )?|qa )<QB>#<QB>qt_ \2#g
s#</QB>fx0 #qt_ </QB>#
s#((t[02345] )+)#<TB>&</TB>#g
s#(s0 )<TB>((t[02345] )+)</TB>(s0 )#<QR>\1\2\4</QR>#g
s#(s[01] )<QR>#<QR>\1#
s#s1 s1 <TB>([^<]*)</TB>s1 s1 #<QH>s_ s_ \1s_ s_ </QH>#g
s#</TB>s1 s1 <TB>#s_ s_ #g
s#(s1 ){1,2}<TB>((t[02345] )+)</TB>(s1 ){1,2}#<QS>\1\2\4</QS>#g
s#t0 </TB>s0 <TB>(t[04] [^<]+)</TB>fx0 #</TB><QJ>fx_ t_ \1fx_ </QJ>#g
s#((t5 )?fx1 (s0 )?)<TB>([^<]+)</TB>(<QS>([^<]+)</QS><TB>([^<]+)</TB>)?(fx0 )#<QA
    >\1\4\6\7\8</QA>#g
s#((t5 )?fx1 (s0 )?)<TB>([^<]+)</TB>s1 <TB>([^<]+)</TB>?(fx0 )#<QD>\1\4 s_\5\6</
    QD>#g
s#</QA><QA>##g
s#s0 <QA>#<QA> s0#g
s#(q1 |fx0 |qa )<TB>([^<]*)</TB>#<QE>qt_ \2</QE>#g
s#(fx[01] |s[01] )<QB>#<QB>qt_ #g
s#</QB><QB>##g
s#(fx0 (fx0 )?|fx1 (fx1 )?)<TB>((t[02345] )+)</TB>#<QX>\1\4</QX>#g
s#<QB>([^<]+)</QB>(<TB>(t0 )</TB>)?<QA>([^<]+)</QA>#<QC>\1\3\4</QC>#g
s#s[01] (s[01] )?(<(TB|QE)>([^<]+)</\3>)?<<#<SG>sg_ \1\4</SG><<#g
s#s[01] (s[01] )?(<(TB|QE)>([^<]+)+</\3>)#<SG>sg_ \1\4</SG>#g
s#</SG>s0 #sg_ </SG>#g
s#</SG><SG>##g
s#s0 <QE>#<QE>s_ #g
s#<QE>([^<]+)</QE><QB>([^<]+)</QB>#<QT>\1\2</QT>#g
s#(s[01] )<TB>([^<]+)</TB>(s[01] )#<QY>\1\2\3</QY>#g
s#(<QB>[^<]*?)<..>([^<]+)</..>#\1\2#g
s#<(..)></\1>##g
s#>>((<TB>[^<]+?</TB>|<SG>[^<]+?</SG>|<Q.>[^<]+?</Q.>)*)(<TB>[^<]+?</TB>)((<TB
    >[^<]+?</TB>|<SG>[^<]+?</SG>|<Q.>[^<]+?</Q.>)*)<<#<ROOT>\1\3\4</ROOT>#gp
s#>>(.*)<<#<STOP>\1</STOP>#gp
```

Abbildung 16: Flacher Parser für die Struktur einer eMail in *sed*. Die regulären Ausdrücke entsprechen - bis auf die Verwendung von Rückwärtsreferenzen (*backreferences*) - dem Typ 3 der Chomsky-Hierarchie (siehe Chomsky (1956)).

$$
\begin{aligned}
Q_B &\rightarrow q_t+ \\
Q_B &\rightarrow Q_B\ t_0 \\
Q_B &\rightarrow (t_2|q_1\ t_2?|q_a)\ Q_B \\
Q_B &\rightarrow Q_B\ fx_0 \\
TB_x &\rightarrow (t_0|t_2|t_3|t_4)*\ t_0 \\
TB &\rightarrow (t_0|t_2|t_3|t_4)+ \\
Q_R &\rightarrow s_0\ TB\ s_0 \\
Q_R &\rightarrow (s_0|s_1)\ Q_R \\
Q_H &\rightarrow s_1\ s_1\ TB\ s_1\ s_1 \\
TB &\rightarrow TB\ s_1\ s_1\ TB \\
Q_S &\rightarrow s_1\ s_1?\ TB\ s_1\ s_1? \\
TB\ Q_J &\rightarrow TB_x\ s_0\ TB\ fx_0 \\
TB &\rightarrow TB_x \\
Q_A &\rightarrow t_5?\ fx_1\ s_0?\ TB\ (Q_S\ TB)?\ fx_0 \\
Q_D &\rightarrow t_5?\ fx_1\ s_0?\ TB\ s_1\ TB\ fx_0 \\
Q_A &\rightarrow Q_A\ Q_A \\
Q_A &\rightarrow s_0\ Q_A \\
Q_E &\rightarrow (q_1|fx_0|q_a)\ TB \\
Q_B &\rightarrow (fx_0|fx_1|s_0|s_1)\ Q_B \\
Q_B &\rightarrow Q_B\ Q_B \\
Q_X &\rightarrow (fx_0\ fx_0?|fx_1\ fx_1?)\ TB \\
Q_C &\rightarrow Q_B\ TB?\ Q_A \\
SG &\rightarrow (s_0|s_1)\ (s_0|s_1)?\ (TB|Q_E)?\ \$ \\
SG &\rightarrow (s_0|s_1)\ (s_0|s_1)?\ (TB|Q_E) \\
SG &\rightarrow SG\ s_0 \\
SG &\rightarrow SG\ SG \\
Q_E &\rightarrow s0\ Q_E \\
Q_T &\rightarrow Q_E\ Q_B \\
Q_Y &\rightarrow (s_0|s_1)\ TB\ (s_0|s_1) \\
S &\rightarrow (TB|SG|Q_*)*TB(TB|SG|Q_*)*
\end{aligned}
$$

Abbildung 17: Die Regeln, die in Abb. 16 beschrieben sind. Da sie einmalig in der obigen Reihenfolge sequentiell angewandt werden und somit keine Rekursion zustande kommt, handelt es sich hierbei nicht um eine formale Grammatik im eigentlichen Sinne.

A.2 eMails

```
Return-Path: <root@gw.selfnet.de>
Delivered-To: web138p1@server01.vorias.net
Received: from mail.selfnet.de (mail.selfnet.de [141.70.124.2])
        by server01.vorias.net (Postfix) with ESMTP id B7D8F648008
        for <johannes@apocalypsys.net>; Wed, 30 Jun 2004 02:17:38 +0200 (CEST)
Received: from lisa.selfnet.de (gw.selfnet.de [141.70.124.46])
        (using TLSv1 with cipher EDH-RSA-DES-CBC3-SHA (168/168 bits))
        (No client certificate requested)
        by mail.selfnet.de (Postfix) with ESMTP id 470E47DA6
        for <johannes@apocalypsys.net>; Wed, 30 Jun 2004 02:17:36 +0200 (CEST)
Received: (from root@localhost)
        by lisa.selfnet.de (8.12.5/8.12.5/Submit) id i5U0HZXw001318
        for johannes@apocalypsys.net; Wed, 30 Jun 2004 02:17:35 +0200
Message-Id: <200406300017.i5U0HZXw001318@lisa.selfnet.de>
Subject: Selfnet e.V. - Kontoauszug - VNR 1489
From: kassenwart@selfnet.uni-stuttgart.de (Kassenwart Selfnet e.V.)
Date: Wed, 30 Jun 2004 02:17:35 +0200 (CEST)
Reply-To: support@selfnet.uni-stuttgart.de
To: johannes@apocalypsys.net
X-Mailer: fastmail [version 2.5 PL6]
X-Spam-Checker-Version: SpamAssassin 2.63 (2004-01-11) on server01.vorias.net
X-Spam-Status: No, hits=-99.0 required=5.0 tests=RCVD_IN_ORBS,
        USER_IN_ALL_SPAM_TO autolearn=no version=2.63
X-Spam-Level:
```

Abbildung 18: Beispiel eines eMail-Headers

```
This is an OpenPGP/MIME signed message (RFC 2440 and 3156)
--------------enigC69946BF0978F12AC97A6694
Content-Type: multipart/mixed;
 boundary="------------030309010000000902080109"

This is a multi-part message in MIME format.
--------------030309010000000902080109
Content-Type: text/plain; charset=us-ascii; format=flowed
Content-Transfer-Encoding: 7bit

protokoll nr. 5

--------------030309010000000902080109
Content-Type: application/pdf;
 name="2004-07-21-Selfnet-Vorstandssitzung-Protokoll_aktive.pdf"
Content-Transfer-Encoding: base64
Content-Disposition: inline;
 filename="2004-07-21-Selfnet-Vorstandssitzung-Protokoll_aktive.pdf"

JVBERi0xLjQNCiXk9tzfDQoxIDAgb2JqDQo8PCAvTGVuZ3RoIDIgMCBSDQogICAvRmlsdGVy
IC9GbGF0ZURlY29kZQ0KPj4NCnN0cmVhbQ0KeJzdXMlu3EYQvQvQP/AcIGOyN5KAEUAjaYIc
nQjIB2QDgjiBfcnvh5yZrqbqVReLM6PACXyRW+yt9npVrU9NO/2LY9z5Zgiu+fxL8+NXzZ/3
d5+Ov2ibz7/d3+1f7u/6MH2R/Lgbmpefm3eHrun8bmxefm3ex/ab5uX3+7tx+qKdf/u+d6eR

⋮

CjAwMDAwODI0NjIgMDAwMDAgbg0KMDAwMDA4MjU4MyAwMDAwMCBuDQowMDAwMDgyNjQwIDAw
MDAwIG4NCjAwMDAwODI2ODIgMDAwMDAgbg0KMDAwMDA4Mjk2MSAwMDAwMCBuDQowMDAwMDgz
MDIxIDAwMDAwIG4NCnRyYWlsZXINCjw8IC9TaXplIDIwDQogICAvUm9vdCAxOCAwIFINCiAg
IC9JbmZvIDE5IDAgUg0KPj4NCnN0YXJ0eHJlZg0KODMyNDcNCiUlRU9GDQo=
--------------030309010000000902080109--

--------------enigC69946BF0978F12AC97A6694
Content-Type: application/pgp-signature; name="signature.asc"
Content-Description: OpenPGP digital signature
Content-Disposition: attachment; filename="signature.asc"

-----BEGIN PGP SIGNATURE-----
Version: GnuPG v1.2.4 (MingW32)

iD8DBQFBT1ganDei4azjmoERAkt5AJ48s3JHJfY9B89vYlFImGFaZl8T8QCcDuQG
X1UYcoEOkUqAJdsWaFGI7co=
=4Kh1
-----END PGP SIGNATURE-----

--------------enigC69946BF0978F12AC97A6694--
```

Abbildung 19: Beispiel einer MIME-Nachricht

```
Hallo,

ich hatte auf dem Antrag eigentlich angegeben, dass bis einschließlich Juni die
    Mitgliedschaft ruhen soll.

Könnt ihr das bitte noch ändern?

DANKE

******* ********

-------- Original-Nachricht --------
> Datum: Mon, 29 Nov 2010 20:57:46 +0100 (CET)
> Von: "Support Selfnet e.V." <support@selfnet.de>
> An: "******* ********" <*******@gmx.net>
> Betreff: VNR 1234 - Ruhende Mitgliedschaft

> (English version below)
>
> ============= Deutsche Version ========================================
>
> Hallo *******,
>
> wir haben gerade deine beantragte ruhende Mitgliedschaft in die
> Datenbank eingetragen. Die Mitgliedschaft ruht ab dem 01.02.2011.
> Ab dem 01.06.2011 bist du wieder passives Mitglied bei
> Selfnet.
>
> Mehr Informationen ueber die ruhende Mitgliedschaft findest du in
> unserer FAQ: http://www.selfnet.de/faq#ruhende_mitgliedschaft
>
> Bei Fragen wende dich an: support@selfnet.de
>
> Mit freundlichen Gruessen
>
> Selfnet Support Team
> support@selfnet.de
> http://www.selfnet.de
>
> -- Diese E-Mail wurde automatisch erstellt und verschickt. --
>
>
> ============= English version ========================================

:
:

> -- This email was generated and sent automatically. --
>

--
Neu: GMX De-Mail - Einfach wie E-Mail, sicher wie ein Brief!
Jetzt De-Mail-Adresse reservieren: http://portal.gmx.net/de/go/demail
```

Abbildung 20: Beispiel einer eMail vom Typ *text/plain*

```
Bonjour,
Nous vous informons que nous avons reçu, en date du 02/08/2010,
instruction de MLE M****** S****** de vous régler par virement la somme de
15,66 EUR par crédit de votre compte DE XXXX XXXX 0122 6273 00 .
Motif de l'opération :
Betreff:  VNR: 101**
Zahlung Mitgliedsbeitrag_S******
Référence end-to-end : VNR: 101**
Virement transmis dans le cadre du service CyberMUT
CAISSE FEDERALE DE CREDIT MUTUEL
Ceci est un message automatique, merci de ne pas répondre.
```

Abbildung 21: eMail-Text, der von beiden Verfahren korrekt erkannt wird.

```
Warum darf ich der betrag nicht überweisen? ich habe
momentan keine zugang zu Internet
« Si tu n as aucune foi en toi-même tu es doublement vaincu dans la course de la
    vie. Avec la foi tu as gagné avant même d avoir commencé ».
Marcus Garvey
D******** K*****
Allmandring 10/**
70569 Stuttgart
0175*******
```

Abbildung 22: Gemischtsprachlicher Text, der mittels Vergleich der Unigramm-Verteilung zu einer falschen Klassifizierung führt.

```
Selfnet Austritt
```

Abbildung 23: Minimaler eMail-Text.

```
hiermit reiche ich meine Kündigung aus dem Verein Selfnet e.V. ein.
```

```
aufgrund meines Auszuges aus dem Bauhäusle möchte ich meinen Internetanschluss
    kündigen.
```

```
als Anlage schicke ich Ihnen die Austrittserklärung für Selfnet e. V. Stuttgart.
```

```
Hiermit kündige ich meine Mitgliedschaft im Selfnet e.V. Stuttgart zum Ende
des Monats August.
```

```
Ich bewege mich aus meinem Zimmer und morgen muss meine Internet abzubrechen.
```

```
anbei die gescannte Kündigung der Mitgliedschaft zum 31.05.2011 weil ich aus dem
    Zimmer ausziehe.
```

```
hiermit kündige ich meine mitgliedschaft. Anbei ist die schrifliche Kündigung.
```

```
Das Kündigungsformular befindet sich als pdf im Anhang.
```

Abbildung 24: Auszüge aus eMails, die das Anliegen beinhalten, aus dem Verein austreten zu wollen.

```
Ich wohnet früher im Allmandring1 10D Zi:35. Ich habe schon 6 wochen umgezogen.
    Aber dieser Monate habe ich noch 21euro bezahlen? Ich habe schon zur ichren
    Fargen,dass Selfnetvertrag gleich wie Wohnenvertrag endet.
```

```
ich wohne nicht mehr im Wohnheim und brauche den Internetverbindung nicht
```

```
Ich habe in Pfaffenwaldring 48F 2** gewohnt.Ich habe am 29.10.2010 ausgezogen und
     das Zimmer abgegeben.
```

```
bitte stornieren Sie mein Selfnet Dauerauftrag und Abzüge von meinen Konto - ich
    wohne seit Jahren nicht mehr in Allmandring (oder in Deutschland überhaupt)
    und mein Auftrag ist schon langst abgelaufen aber ich sehe immernoch Selfnet
    Regelmassig auf mein Bankkonto.
```

Abbildung 25: Auszüge aus eMails, die indirekt das Anliegen beinhalten, aus dem Verein austreten zu wollen.

```
Hallo Selfnet-Team,
ich werde am 28.03.2011 wieder in mein altes Zimmer einziehen. Es wäre nett wenn
    ihr mir meinen Anschluss wieder öffnen könntet. Danke im Voraus!
Viele Grüße,
**** ****
```

```
Hallo liebes Selfnet Team,
ich werde wie geplant am 01.09.2010 wieder in mein Zimmer einziehen und möchte ab
     diesem Datum wieder meinen Internetzugang nutzen.
Vielen Dank und eine gute Zeit.
Mit freundlichen Grüßen
**** ****
```

```
Hallo!
Ich werde zum 01.01.09 wieder in mein altes Zimmer 206**, Heilmannstr. 4a
    einziehen und bitte hiermit darum, meinen Internetanschluss ab da dann wieder
     zu öffnen.
Mit freundlichen Grüßen
A*** S******
```

```
Hallo,
ich ziehe am 25.09. wieder in mein altes Zimmer (Heilmannstraße 4b, 508**) ein.
**** ****
```

```
Hallo,
da meine ruhende Mitgliedschaft bald endet, nun eine Nachricht von mir. Ich ziehe
     am 30.4 wieder in mein altes Zimmer 18B 04 im Allmandring I ein.
Vielen Dank und Grüße,
****
```

```
Hi,
ich ziehe am 12.10.2008 wieder in mein altes Zimmer ein.
Bitte schaltet meinen Anschluss ab diesem Tag wieder frei.
Mfg
****
```

```
Hallo Selfnet Team,
ich bitte euch meinen Anschluss ab 01.05.09 wieder freizuschalten.
Viele Grüße
**** ****
```

Abbildung 26: Auszug aus einem Cluster der Größe 37.

```
Hallo,
hiermit kündige ich meine Mitgliedschaft bei Selfnet e.V. zum 30. April. Vielen
    Dank
Mit freundlichen Grüßen
**** ****
```

```
Hallo Selfnet,
hiermit kündige ich meine Mitgliedschaft zum 31.01.2011
Mit freundlichen Grüßen,
**** ****
```

```
Hallo Selfnet-Team,
wegen Umzug Ende des Monats kündige ich hiermit meine Mitgliedschaft bei "Selfnet
    ".
Mit freundlichen Grüßen
**** ****
```

```
im Anhang das Formular mit dem ich meine Mitgliedschaft zum Ende des Monats
April kündige
```

```
Hallo,
Hiermit kündige ich meine Mitgliedschaft im Selfnet e.V. Stuttgart zum Ende des
Monats April 2011. Ich habe das Formular beigefügt.
Viele grüße
M. J. ****
```

```
Hallo Selfnet,
hiermit kündige ich meine Mitgliedschaft bei Selfnet zum 28.02.2010.
Anbei füge ich die unterschriebene Austrittserklärung.
Gruß,
****
```

```
Hallo,
Anbei erhalten sie meine Kündigung.
Mit freundlichen Grüßen
**** ****
```

```
Sehr geehrte Damen und Herren,
anbei sende ich Ihnen die Austrittserklärung.
Original liegt in Ihrem Postfach.
```

```
Sehr geehrte Damen und Herren,
ich möchte zum 31.12. aus dem Verein austreten.
Beste Grüße
**** ****
```

```
Hallo Selfnet,
hiermit Kündige ich die Mitgliedschaft bei Selfnet e.V. weil ich ausgezogen
bin.
Name: **** A*** ****
Mit freundlichen Grüßen
```

Abbildung 27: Auszug aus einem Cluster der Größe 32.

```
Hallo
ich bin **** ****
vor zwei wochen habe ich geld zu ihnen überwiesen, und bis jetzt habe ich kein
    internet zu hause
vielen dank
```

```
Hallo,
Ich habe das Geld schon überwiesen, bitte überprüft es und sagt mir, ob ihr das
    Geld schon habt.
Vielen Dank !
**** ****
```

```
Hallo,
ich habe schon die Überweisung seit letzte Freitag gemacht. Und Jetzt habe ich
    kein Internet im Zimmer.Bitte, können sie das Internet wiedergeben.
```

```
Hallo,
ich habe die Überweisung schon Dienstag gemacht, aber habe ich jetzt kein
    Internet zu Hause.
Ich brauche es wichtig fürs Wochenende.
Danke.
**** ****
```

```
Hallo,
ich habe vorgestern (am Dienstag 09.02.2010) mit einer Überweisung fürs
Internet bezahlt aber trotzdem habe ich jetzt keinen Internetzugang. Können
Sie mir, bitte, wieder Internet einschalten?
Mein Betreff Nr : VNR97**
```

```
Hallo
Ich habe die Rechnung heute bezahlt. Können Sie mir den Internetzugang wieder
    erlauben?
Freundliche Grüße
```

Abbildung 28: Alle Texte eines Clusters der Größe 6.

```
Hallo N***,
Wenn es nur um die Eintragung deiner neuen Mietzeiten geht, reicht es,
wenn du eine Kopie des neuen Mietvertrags zukommen lässt. Das kannst du
per Post oder E-Mail erledigen, oder persönlich in unseren Briefkasten
in Vaihingen werfen oder in der Sprechstunde abgeben (heute ist in
Vaihingen support).
Viele Grüße
****
```

```
Hallo A***,
Da du bereits bei uns Mitglied bist und die Mitgliedschaft unbefristet
ist, reicht es für die Verlängerung des Internetanschlusses aus, wenn
du uns den neuen Mietvertrag als Kopie zukommen lässt (per Post oder in
unseren Briefkasten werfen, du kannst ihn auch einscannen und uns per
Mail senden).
Viele Grüße und fröhliche Weihnachten
**** ****, Selfnet Support-Team
```

```
Hallo J****,
Für eine Verlängerung deines Internetanschlusses müssen wir einen
gültigen Mietvertrag vorliegen haben. Wenn du nicht vorbeikommen kannst
reicht es auch wenn du eine Kopie deines Mietvertrages in unseren
Briefkasten wirfst, oder einscannst und uns per E-Mail zusendest.
Mit freundlichen Grüßen
Selfnet Support Team
```

Abbildung 29: Cluster mit Antworttexten zum Thema „neuer Mietvertrag“.

```
Das Geld ist noch nicht eingegangen.
Viele Grüße, ****
```

```
Hallo C****,
dein Geld ist heute bei uns eingegangen.
Viele Grüße
****
```

Abbildung 30: Negativbeispiel für die Clusteranalyse der Antworttexte.

```
Hallo C**** M***** B****,
Du hast deine Mitgliedschaft im Verein Selfnet nicht gekündigt und bist
daher noch immer Mitglied unseres Vereins. Die Mitgliedschaft ist
unabhängig von deinem Mietvertrag und muss daher separat gekündigt
werden.
Da du offensichtlich kein Interesse mehr am Verein hast werde ich deine
Mitgliedschaft nun beenden. Der Vereinsaustritt ist immer zum
Monatswechsel möglich, ich werde ihn daher für Ende Mai 2011 eintragen.
Eine rückwirkende Kündigung ist nicht möglich.
Der bereits eingezogene Mitgliedsbeitrag für Juni wird dir auf dein
Konto überwiesen.
Viele Grüße,****
```

```
Hallo Y****,
Hier liegt offensichtlich ein Missverständnis vor, die Mitgliedschaft im
Verein Selfnet endet NICHT automatisch wenn dein Mietvertrag endet. Die
Mitgliedschaft endet nur, wenn du sie kündigst.
Da wir keine Kündigung von dir bekommen haben bist du immer noch Mitglied
unseres Vereins und musst somit den Mitgliedsbeitrag bezahlen.
Ich werde nun deinen Mitgliedschaft zum Monatsende beenden. Du wirst dann
den Mitgliedsbeitrag für Dezember (7 Euro) zurückerstattet bekommen.
Viele Grüße,
****
```

```
Hallo Y****** Z****,
Die Mitgliedschaft im Verein Selfnet e.V. endet nicht automatisch, wenn
du aus dem Wohnheim ausziehst, da wir überhaupt nicht wissen, wann du
ausziehst (Der Mietvertrag, den du uns gezeigt hast läuft bis März
2011). Du musst deine Mitgliedschaft in unserem Verein kündigen,
ansonsten bleibt du Mitglied.
Da du uns nun mitgeteilt hast, dass du ausgezogen bist werde ich deinen
Vereinsaustritt zum Ende des laufenden Monats veranlassen, den
Mitgliedsbeitrag für Dezember erhälst du dann von uns zurück.
Viele Grüße,
****
```

```
Hallo V*****,
Wir buchen noch immer den Mitgliedsbeitrag von deinem Konto ab, da du
deine Mitgliedschaft in unserem Verein bisher nicht gekündigt hast und
somit immer noch Mitglied im Verein Selfnet e.V. bist. Die
Mitgliedschaft ist unabhängig vom Mietvertrag und endet nur, wenn sie
gekündigt wird.
Da offensichtlich kein Interesse deinerseits am Verein Selfnet mehr
besteht werde ich deine Mitgliedschaft nun beenden, rückwirkend ist
dies jedoch nicht möglich.
Wenn du eine Rückerstattung der in der Zwischenzeit angefallenen
Mitgliedsbeiträge wünschst, so kannst du bei unserem Vereinsvorstand
( **** ) darum bitten. Über eine außerordentliche
Rückerstattung kann nur der Vorstand entscheiden. Da wir keine
Kündigung von dir erhalten haben war die Einziehung der
Mitgliedsbeiträge rechtens und es wäre reine Kulanz, wenn wir eine
Rückerstattung vornehmen. Ich kann daher nicht garantieren, dass der
Vereinsvorstand einer Rückerstattung zustimmen wird.
Viele Grüße,
**** (Selfnet Support-Team)
```

Abbildung 31: Cluster mit Antworttexten zum Thema „Mitgliedschaft unabhängig vom Mietvertrag”(4 von 6 der enthaltenen Texte).

```
Guten Tag,
Am 2te Mai, habe ich von Selfnet €21 von meiner Konto rausgenommen. Mein
Vertrag dauerte bis ende April, deswegen verstehe ich nicht warum dieses
Geld Sie genommen haben.
Ich bitte Ihnen eine Erklärung.
Meine alte Addresse war:
Allmandring 20D, **.
Vielen Dank,
C**** **** B****
```

```
Guten Tag,
Ich wohnet früher im Allmandring1 10D Zi:**. Ich habe schon 6 wochen
umgezogen.Aber dieser Monate habe ich noch 21euro bezahlen?
Ich habe schon zur ichren Fargen,dass Selfnetvertrag gleich wie
Wohnenvertrag endet.
Warum bezahle ich noch?
Danke!
MfG
y****
```

```
Sehr geehrte Damen und Herrn.
mein Name ist Y****** Z****.Ich habe in Pfaffenwaldring 48F 2** gewohnt.Ich habe
    am 29.10.2010 ausgezogen und das Zimmer abgegeben.
Ich moechte fragen,ob ich die Internetgebuhr fuer November und Dezember bezahlt
    habe.Wenn ich sie bezahlt habe,koennen Sie die Gebuehr zurueck ueberweisen?
Ich habe mein Konto des BW Bank gekuendigt.
Mein neues Konto ist:
Y****** Z****
Konto Nr. 1185***
BLZ:60070024
Deutsche Bank
Mit freundlichen Gruessen
Y****** Z****
```

```
Sehr geehrte Damen/Herren,
bitte stornieren Sie mein Selfnet Dauerauftrag und Abzüge von meinen Konto -
ich wohne seit Jahren nicht mehr in Allmandring (oder in Deutschland
überhaupt) und mein Auftrag ist schon langst abgelaufen aber ich sehe
immernoch Selfnet Regelmassig auf mein Bankkonto.
BZW
KTO 122627***
SELFNET E.V.,20110423-57**
LG
V***** S*****
```

Abbildung 32: Die Anfragen zu den Antworten aus Abb. 31.

A.3 Diagramme

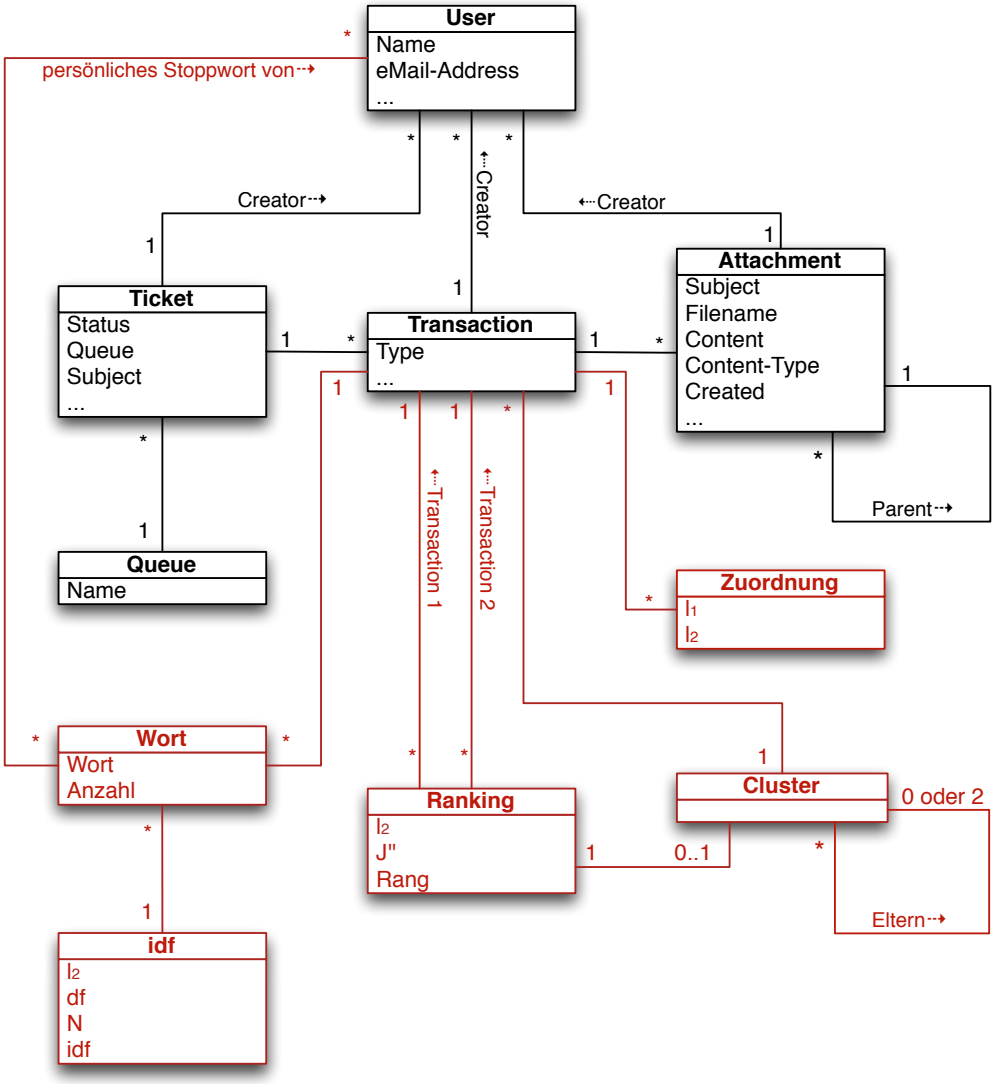

Abbildung 33: Der für die weitere Verarbeitung der eMail-Daten relevante Teil der Datentruktur des *Request Trackers* (schwarz), sowie die darauf aufbauenden von mir hinzugefügten Strukturen (rot).

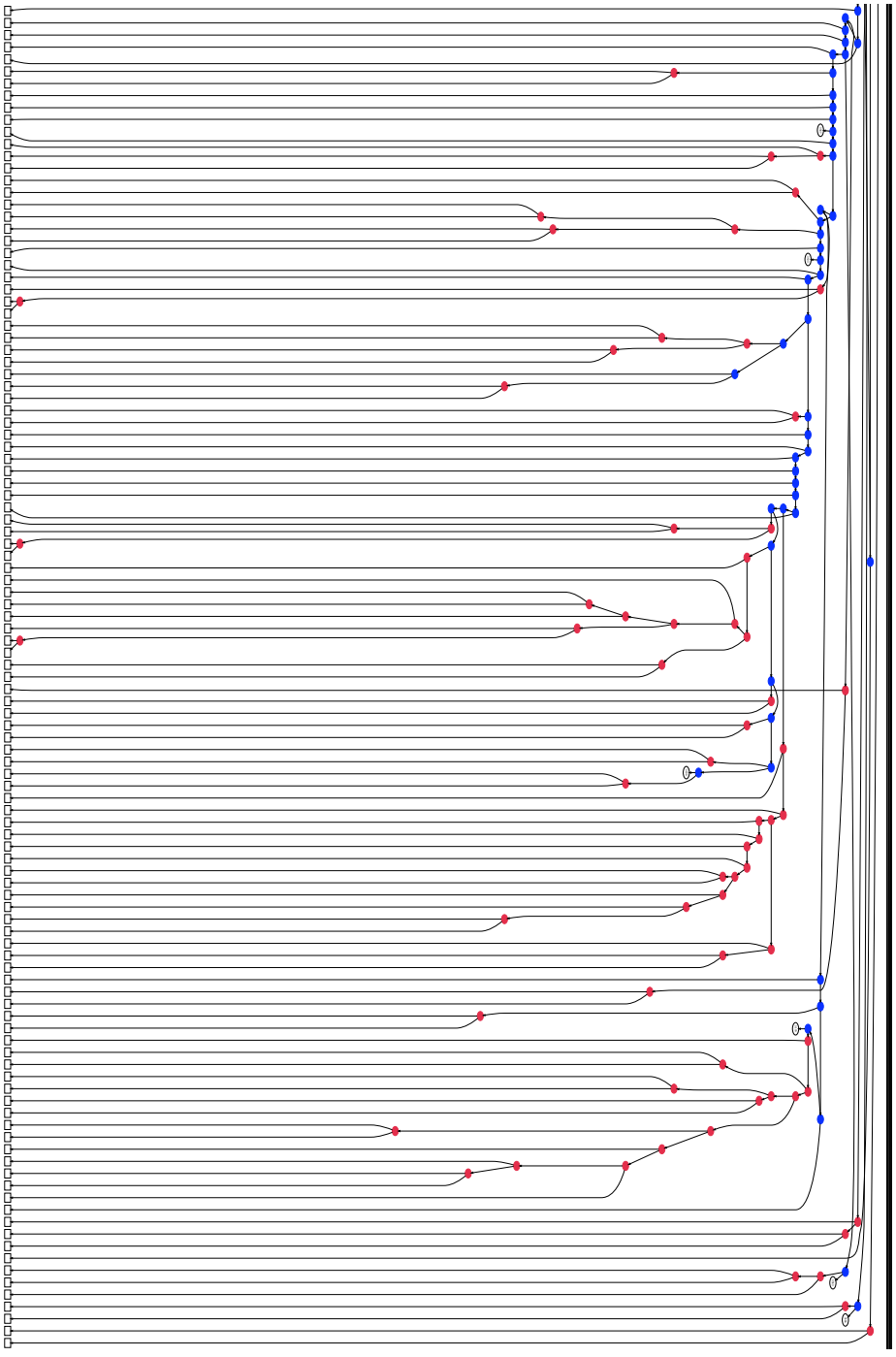

Abbildung 34: Ausschnitt aus dem Cluster-Baum der deutschsprachigen Texte als Dendrogramm. Die bei der manuellen Kontrolle als korrekt bewerteten Cluster sind rot markiert; nicht verwertbare Zweige werden in einem einzigen weißen Knoten zusammengefasst. Die Ähnlichkeitswerte nehmen von links (1,0) nach rechts ($\approx$ 0,3) ab.

Literatur

Baker, C.F., C.J. Fillmore und J.B. Lowe (1998). „The Berkeley FrameNet Project“. In: *Proceedings of the 17th international conference on Computational linguistics*. ACL '98. Montreal, Quebec: Association for Computational Linguistics, S. 86–90. URL: `http://framenet.icsi.berkeley.edu/papers/acl98.pdf` (besucht am 30. 08. 2011).

Chomsky, N. (1956). „Three models for the description of language“. In: *IRE Transactions on Information Theory* 2, S. 113–124. URL: `http://www.chomsky.info/articles/195609--.pdf` (besucht am 30. 08. 2011).

Jaccard, P. (1901). „Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura“. In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, S. 547–579.

Kleene, S.C. (1956). „Representation of Events in Nerve Nets and Finite Automata“. In: *Automata Studies*. Hrsg. von C.E. Shannon und J. McCarthy. Princeton: Princeton University Press, S. 3–40.

Koehn, P. (2005). „Europarl: A Parallel Corpus for Statistical Machine Translation“. In: *Conference Proceedings: the tenth Machine Translation Summit*. AAMT. Phuket, Thailand: AAMT, S. 79–86. URL: `http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf` (besucht am 30. 08. 2011).

Levenshtein, V.I. (1965). „Binary Codes with correction of deletions, insertions and substitution of symbols“. In: *Doklady Akademii Nauk SSSR* 163.4, S. 845–848.

Manning, C.D., P. Raghavan und H. Schütze (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Manning, C.D. und H. Schütze (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Mel'čuk, I.A. (1988). *Dependency syntax: theory and practice*. SUNY series in linguistics. New York: State University Press of New York.

Reiter, E. und C.S. Mellish (1993). *Optimizing the costs and benefits of natural language generation*. DAI research paper. Edinburgh: Edinburgh University. URL: `http://www.csd.abdn.ac.uk/~ereiter/papers/ijcai93b.pdf` (besucht am 30. 08. 2011).

Ruppenhofer, J. u. a. (2005). *FrameNet II: Extended Theory and Practice*. Techn. Ber. ICSI. URL: `http://framenet.icsi.berkeley.edu/book/book.pdf` (besucht am 30. 08. 2011).

Vincent, J. u. a. (2005). *RT essentials*. Essentials Series. Sebastopol: O'Reilly.