Regelbasierte Suche in Textdatenbanken mit Nichtstandardisierter Rechtschreibung (Rule Based Search in Text Databases with Non-Standard Orthography)

Thomas Pilz *(pilz@informatik.uni-duisburg.de)*
University of Duisburg-Essen
Prof. Dr. Wolfram Luther
(luther@informatik.uni-duisburg.de)
University of Duisburg-Essen
Prof. Dr. phil. Ulrich Ammon
(ammon@uni-duisburg.de)
University of Duisburg-Essen
Prof. Dr.-Ing. Norbert Fuhr
(fuhr@uni-duisburg.de)
University of Duisburg-Essen

In this paper we describe our interdisciplinary project in support of the conservation of cultural heritage, especially for the German reception of Nietzsche. We present a rule based fuzzy search-engine which allows retrieval of text data independently of its orthographical realization. The rules used are derived from statistical analyses, historical works, linguistic principles and professional administration. Our web based tool aims at experts as well as interested amateurs. In addition to its present features, further functions are currently worked out that include automatic rule derivation and a finer result classification via a generalized Levenshtein similarity measure.

Dans cette note, nous décrivons notre projet interdisciplinaire concernant l'édition électronique de la réception allemande des idées et de l'œuvre de Nietzsche. Nous avons centré un point d'intérêt sur la création d'un moteur de recherche accessible dans le Web. Celui-ci permet la recherche floue, phonétique et par troncation nécessaire au traitement de la plupart des textes numérisés écrits avant la réforme de l'orthographe en Allemagne en 1901/02. Le logiciel est basée sur un algorithme qui déduit pour chaque nom, verbe et adjectif toute orthographe possible selon un système de principes ou règles linguistiques cités dans la littérature historique ou dérivés en collaboration avec des spécialistes. Plusieurs autres options sont prévues y compris une dérivation automatique des règles et une classification des résultats basée sur une mesure de similarité généralisée de Levenshtein.

Im Kontext eines Digitalisierungsprojekts zur Nietzsche-Rezeption aus den Jahren 1865-1945, das seit mehreren Jahren in Duisburg in Zusammenarbeit mit dem Nietzsche-Kolleg in Weimar verfolgt wird [BM02, BM03], beschäftigt sich das von der Deutschen Forschungsgemeinschaft geförderte RSNSR-Projekt mit der Erforschung und Entwicklung eines linguistischen Regelsystems, einer Transformations-methodik und zeitabhängiger Filter zur Unterstützung der Suche in Textdokumenten in nichtstandardisierter Rechtschreibung.

Es wurde bereits eine Java-basierte Suchmaschine erstellt, welche es durch einen neu entwickelten phonetischen Regelsatz ermöglicht, auf Texten, die mehrere hundert Jahre vor der Rechtschreibvereinheitlichung des Jahres 1901 verfasst wurden, eine Suche mittels orthographisch genormter Schlagwörter durchzuführen (vgl. Abbildung 1) [P03]. Durch Einführung eines Abstandsbegriffs [ZD96] sind verschiedene Stufen der Ähnlichkeit realisiert. Außerdem erlaubt der Algorithmus durch einen zusätzlichen speziellen Regelsatz auch die Suche nach Wörtern, welche durch OCR-Software fehlerhaft erkannt wurden. Die Suchmaschine ist in das online-verfügbare HTML-basierte Nietzsche-Archiv integriert.

Mit der regelbasierten Suche verfolgen wir einen anderen Ansatz als viele große Wörterbuchprojekte. Indem nicht mit statischen Wortlisten gearbeitet wird, erhoffen wir uns eine höhere Trefferquote, besonders bei Texten mit stark variierender Schreibung. Zusätzlich wird der Arbeitsaufwand durch manuelle Eintragung von Wort-Relationen vermieden. Andererseits hoffen wir durch Grundlagenforschung, besonders in den Bereichen der Phonem-Graphem-Struktur des Deutschen, der unscharfen Suche und der Ähnlichkeitsmetriken, einen Wortabstandsbegriff zu definieren, der sowohl eine größtmögliche Differenzierung unterschiedlicher als auch Zusammenfassung äquivalenter Wörter ermöglicht [A98].

Neben der Anwendung als Suchmaschine sind auch Einsatzpunkte im Vergleich oder der temporal-lokalen Einordnung von Texten denkbar. Zentraler Betrachtungszeitraum sind für uns die Jahre 1700-1900. Eine spätere Ausweitung des Regelsatzes auch auf frühere Zeitabschnitte ist durchaus möglich.

Im Einzelnen verfolgt das Projekt die folgenden Ziele:

- Entwicklung von Zeit- und Ortsfiltern für phonetische Regeln, Revision der Regeln aus der Textbasis und aus statistischen Analysen, Nutzung eines Kontrollwörterbuchs gegen Homonymhäufung.

- Entwicklung eines neuen adäquaten Abstandsbegriffs auf der Basis eines modifizierten graphematischen und phonetischen Levenshtein-Ähnlichkeitsmaßes, Berücksichtigung typischer Erfassungsfehler, Entwicklung von Unschärfeskalen.
- Integration der Suchmaschinen in das Nietzsche-Projekt und in andere Systeme wie das Deutsche Rechtswörterbuch oder das Projekt Deutsch Diachron Digital, Entwicklung von Regelsätzen und Erweiterung der Suchmaschine auf (früh-)neuhochdeutsche Archive.

Hauptsächliche Arbeitspunkte sind zur Zeit

- eine Verbesserung des Tools in Hinsicht auf Effizienz
- Grundlagenforschung zum regelbasierten Ansatz
- Untersuchungen zur Levenshtein-Distanz
- ein Vergleich regelbasierter mit Wörterbuch-basierter Suche
- eine Einbringung der Suchmaschine in andere Projekte.

Mittelfristig wird eine verbesserte Realisierung mit einem Java-Frontend, einem Web-Server und einer modernen XML-basierten Archivlösung angestrebt [FGG02], die auch in vergleichbaren Digitalisierungsprojekten Anwendung finden kann.

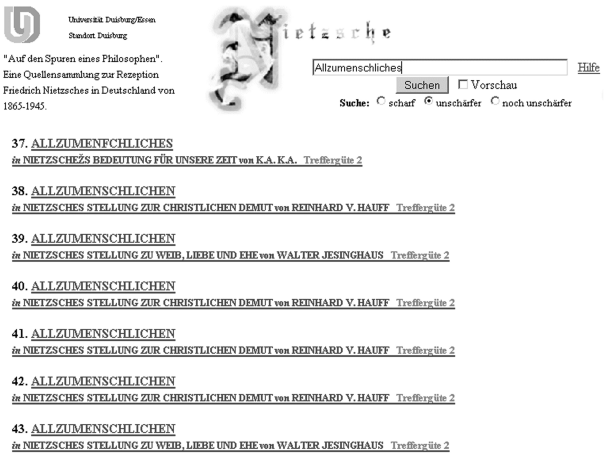

Abbildung 1: Webinterface zur unscharfen Suche

Arbeitsmethodik

Die zu Grunde liegenden Texte der bisher bearbeiteten Nietzsche-Rezeption reichen teilweise bis in das Jahr 1865 zurück. Dadurch, dass in diesen Dokumenten Zitate noch deutlich älteren Ursprungs verwendet werden, treten zusätzlich Formen auf, die bereits zu dieser Zeit obsolet waren.

Die Verwendung eines historischen Wörterbuchs wurde zunächst nicht ins Auge gefasst, da ein solches zwar einen Teil der auftretenden Wörter effizient zu indizieren vermag, eine zumindest annähernd vollständige Erkennung allerdings nicht ermöglichen kann. Dies beruht auf der enormen Divergenz prinzipiell möglicher Schreibungen, die mit dem Alter der Texte immer weiter ansteigt. Es ist ja gerade Merkmal der Texte vor der Rechtschreibvereinheitlichung von 1901/02, dass — zumindest prinzipiell — jede Schreibung möglich war, auch wenn diese nicht immer realisiert wurde.

Allerdings sind diese Allographe durch die Zugehörigkeit von Graphemen zu bestimmten Lautklassen, die sogenannte Graphem-Phonem-Korrespondenz, beschränkt: Jedes Phonem der deutschen Sprache lässt sich nur durch eine endliche Menge von Graphemen realisieren. Unter der Annahme, dass ein heutiges Wort sich in seiner Lautstruktur prinzipiell nur unerheblich von seinem historischen Gegenstück unterscheidet, kann dieses mittels Variation seiner Phonem-Realisierungen rekonstruiert werden. Diese Annahme lässt sich immerhin für das Neuhochdeutsche — und damit für einen Zeitraum von rund 600 Jahren — bestätigen, wenn auch mit einer in der historischen Tiefe abnehmenden Sicherheit.

Durch eine Untersuchung, mittels welcher Grapheme damals eine der wahrscheinlich heute noch gültigen Aussprache ähnliche Lautung zu erreichen war, können Repräsentationsfehler vermieden werden.

Grundlage der schon entwickelten Suchmaschine sind somit phonetische Regeln, welche die in dem betrachteten Zeitraum möglicherweise auftretenden Schreibweisen nachbilden. Durch Kombination weniger Regeln ergibt sich bereits eine erstaunlich realistische Transformationsvariation. Betrachtet man nun noch die weiteren Variationsmöglichkeiten abhängig von der Initial-, Medial- oder Finalstellung der Grapheme, so empfehlen sich keine Arbeitsmethoden, die wie bei vergleichbaren Realisierungen auf zentraler Verwendung eines Wörterbuchs basieren.

Die zu jener Zeit vorkommenden Allographe konnten dafür aus Arbeiten von Rechtschreibreformern wie Adelung und Schottelius indirekt abgeleitet werden. Indem diese forderten, dass Schreibung A vermieden und durch Schreibung B zu ersetzen sei, belegen sie die Existenz der Phonemrealisierung A. Die vor allem im 16. und 17. Jh. aufkommenden Normierungsbestrebungen und die endgültige Übernahme des Hochdeutschen für den gesamten deutschsprachigen Raum haben glücklicherweise eine Vielzahl gut dokumentierter linguistischer Arbeiten hervorgebracht.

Eine eingehende Untersuchung der Aussprache des heutigen Standarddeutschen konnte weitere produktive Regeln hervorbringen. Das grundlegende Schreibprinzip alphabetischer Schriftsysteme '"Schreibe, wie du sprichst, und sprich, wie du schreibst!'" (phonologisches Prinzip der Rechtschreibung) [K86] hält hierbei damals wie heute die tatsächlich anwendbare Gesamtzahl der Allographe in einem überschaubaren Rahmen.

Durch eine modulare Erweiterung des Regelsatzes um spezielle OCR-Probleme betreffende Produktionen, etwa die Fehlerkennung des <s> in Fraktur durch <f>, wurde der Suchmaschine weitere Funktionalität verliehen.

Auf Basis beliebiger HTML-Dokumente werden in der bereits existierenden Version mittels des verwendeten Regelsatzes zwei Varianten jedes Schlagwortes gebildet 'unschärfer' und 'noch unschärfer'. Die erste verwendet nur einen Teil der Regeln, welche die erwartungsgemäß häufigsten Unterschiede betreffen. Die zweite Variante berücksichtigt alle Produktionen inklusive möglicher OCR-Fehler. Diese Unterteilung wurde aufgrund der resultierenden Homonymhäufung getroffen: Durch die Schlagworttransformationen fallen umso mehr Wörter zusammen, je umfangreichere Regeln verwendet werden. Diesem Phänomen wird mit einem Kontrollwörterbuch zu begegnen sein. Aus den drei Repräsentationen jedes indizierten Wortes sowie aus dessen Position innerhalb des entsprechenden Dokumentes werden mittels eines JAVA-Programms die Tabellen in einer *MySQL*-Datenbank mit Daten versehen.

Hauptarbeitspunkte

Beim Anlegen der Wort-Tabellen für ein Dokument erscheint es naheliegend, neben dem Wort auch eine phonetische Realisierung desselben abzulegen und dann bei der Suche nur wenige ‚nahe' Wörter zu berücksichtigen. Allerdings ist es äußerst schwierig, eine korrekte phonetische Realisierung zu einem vorgegebenen Wort zu finden. Es müssen daher Bewertungsmethodiken entwickelt werden, die bestimmen, welche Wörter zu dem Suchwort passen. Dabei können Gesetzmäßigkeiten der Phonetik und Graphematik, aber auch der Wahrnehmungspsychologie wertvolle Hinweise liefern.

Wenn wir bei der Suche Transformationen regelbasiert berechnen sowie relevante Regeln *on the fly* für einen konkreten Text auswählen oder generieren und fakultativ validieren, gelangen wir zu einem schlanken, anpassungsfähigen und letztlich auch tragfähigeren Werkzeug, das die Verwendung von Wörterbüchern (mit mehr oder weniger 'modernen' Einträgen) auf ein Minimum begrenzt und damit die Abhängigkeit von der Vollständigkeit des Wörterbuchs aufbricht. Zusätzlich sollen Regeln zu OCR-Fehlern für eine Suche dazugeschaltet oder aber ganz abgeschaltet werden können. Wichtig für unseren Ansatz ist dabei eine effiziente Verwendung einer weiterentwickelten Levenshtein-Distanzfunktion. Um eine klare Trennung zwischen OCR-Fehlern und Allographen zu ermöglichen, sollen diese anders gewichtet werden als Abweichungen phonetisch naher Schreibweisen. Berücksichtigung bei der Gewichtung finden sollte die Anzahl der angewendeten Regeln, um die Schreibung zu erreichen, wie auch ihre Relevanz. Dabei geht allerdings die Symmetrie einer Distanzfunktion verloren, da die Ableitungen i.a. nicht umkehrbar sind.

Die Suchmaschine behandelt Anfragen in folgender Art und Weise: In einem Vorverarbeitungsschritt werden Sprache, Zeit und Ort der zu suchenden Dokumente bestimmt, woraus sich die anzuwendenden Regelsätze ergeben. Die Suchterme einschließlich etwaiger Wildcards werden dann durch Anwendung der Regelsätze und unter Berücksichtigung einer parallel zu entwickelnden verallgemeinerten Levenshtein-Ähnlichkeitsmetrik [C&D, 2000] in die internen Suchbedingungen übersetzt; dabei können durch Vorgabe eines Schwellwertes und / oder Ausnutzung eines Kontrollwörterbuchs unwahrscheinliche Varianten ausgeschlossen werden, bevor die eigentliche Suche durchgeführt wird. Die Suchergebnisse werden dann nach absteigender Ähnlichkeit geordnet ausgegeben.

Im Rahmen der nächsten Arbeitsschritte soll die existierende Suchmaschine bezüglich Retrievalqualität und Funktionalität verbessert werden. Um eine hohe Anzahl relevanter Dokumente zu finden, also den Recall zu erhöhen, sollen möglichst alle Flexionsformen und Schreibvarianten eines Suchwortes bei der Suche berücksichtigt werden. Hierzu müssen durch Anwendung der entsprechenden Regelsätze alle Varianten eines Anfragewortes erzeugt werden, mit denen dann im Dokumentenbestand gesucht wird. Da der Regelbestand sehr dynamisch ist, können nicht, wie sonst insbesondere in experimentellen Information-Retrieval-Systemen üblich, die Dokumente schon beim Einfügen in die Datenbasis entsprechend indexiert werden, sondern die Expansion der Suchwörter muss zum Retrievalzeitpunkt erfolgen. Um die Antwortzeiten trotzdem gering zu halten, müssen noch entsprechend effiziente Verfahren implementiert werden.

Durch diese Vorgehensweise können sehr viele Dokumente gefunden werden, wovon aber auch viele nicht relevant sind. Nur durch eine entsprechende Rangordnung der Retrievalantworten kann eine hohe Präzision des Suchergebnisses gewährleistet werden. Liegen zu einem Suchbegriff Wörterbucheinträge vor, so sollen diese Angaben bei der Suche mit berücksichtigt werden. Schreibweisen, die auch im Wörterbuch auftauchen, erhalten dann ein höheres Gewicht als andere Varianten.

Zur Erweiterung der Suchfunktionalität soll die Suchmaschine um gängige Suchoperatoren erweitert werden. Bei der Eingabe von Einzelwörtern soll Trunkierung erlaubt werden, und mehrgliedrige Begriffe sollen mit Hilfe von Kontextoperatoren (Wortabstandssuche) spezifiziert werden können. Mehrere Suchbedingungen sollen wahlweise durch Boolesche Konnektoren verknüpft oder in Form einer linearen Anfrage als Menge von möglicherweise gewichteten Bedingungen spezifiziert werden können. Hier muss die Retrievalfunktion dann die Gewichtungen eines Dokumentes bezüglich der

einzelnen Suchbedingungen passend verrechnen. Hierzu können wir uns an die Definition der Semantik der von uns entwickelten Anfragesprache XIRQL anlehnen [FG01].

Bibliografie

Ammon, U. *Variationslinguistik/ Linguistics of Variation/ La linguistique variationelle.* Tübingen: Niemeyer (Sociolinguistica 12), 1998.

Biella, D., E. Dyllong, H. Kaiser, W. Luther, and Th. Mittmann. "Wege zur digitalen Erfassung der Nachwirkung Nietzsches in Deutschland von 1865-1945. Ein Arbeitsbericht zum Duisburger Retrodigitalisierungsprojekt." Kolloquium "Vom Umgang Nietzsches mit Büchern zum Umgang mit Nietzsches Büchern", Weimar 23.09.-25.9.2002, erscheint in einem Sammelband.

Biella, D., E. Dyllong, H. Kaiser, W. Luther, and Th. Mittmann. "Edition électronique de la réception de Nietzsche des années 1865 à 1945." *Proceedings of ICHIM03, Paris,* . 8.-12. Sept. 2003.

Camps, R., and J. Daudé. "Improving the efficacy of approximate personal name matching." *Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems (NLDB'03).* 2003. Accessed 2005-04-13. <http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>

Fuhr, N., and K. Großjohann. "XIRQL: An XML Query Language Based on Information Retrieval Concepts." *ACM Transactions on Information Systems* 22 (2004): 313-356. Accessed 2005-04-13. <http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>

Fuhr, N., N. Gövert, and K. Großjohann. "HyREX: Hypermedia Retrieval Engine for XML." *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval.* New York: ACM, 2002. 449. Accessed 2005-04-13. <http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>

Keller, R. *Die Deutsche Sprache und ihre historische Entwicklung.* Hamburg: Helmut Buske Verlag, 1986.

Pilz, Th. *Unscharfe Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung am Beispiel von Frakturtexten zur Nietzsche-Rezeption.* Staatsexamensarbeit, Universität Duisburg-Essen, 2003.

Zobel, J., and P. Dart. "Phonetic String Matching: Lessons from Information Retrieval." *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96).* Ed. H.-P. Frei, D. Harman, P. Schäuble and R. Wilkinson. New York: ACM, 1996. 166-172. Accessed 2005-04-13. <http://www.lsi.upc.es/dept/techreps/ps/R03-9.ps.gz>