Institut für Informatik

Bachelorarbeit

Identifizierung unterschiedlicher Dialekte in deutschen Romanen des 16 bis 20 Jahrhunderts

Autor: Irina Leikam
engelrade@icloud.com

Betreuer: Prof. Dr. Frank Puppe,
Markus Krug, M.Sc.

Abgabedatum: 09.02.2016

I Abstract

With the actual spreading of global Internet access, text is available not only in a great number of languages, but in many dialects forms. Automatic treatment of these texts is necessary for further language processing. This Bachelor thesis focuses on automatically identifying the dialect of a text, given as a sample of paragraphs, and demonstrates what technologies can be employed to improve Automatic Dialect Recognition.
This paper describes three approaches to the task of automatically identifying the dialect a text is written in. It compares the success of each approach from a set of sentences in 12 german dialects. The three techniques to investigate were chosen: dictionary word recognition, bigram/trigram based recognition and Naïve Bayes Classifier. Each method was implemented (using Java language with UIMA-Framework), by training the model on roughly 17 to 19 kilobytes of text. The length of the sentence samples has been varied, to see how performance was affected.
According to the N-Gram-Based Text Categorization research of William B. Cavnar and John M. Trenkle, bigram/trigram approaches have been successfully employed to language identification, achieving a 99.8% correct classification rate. At first, the effectiveness of this model was analyzed, applying it here to the dialect identification. The results showed the bigram/trigram based identification to be very poor, about 0%, since the test samples seemed to contain the same bigrams or trigrams.
The best results were achieved with the Naïve Bayes Classifier and the word recognition model, applying them to training corpus with the sentence length of more then thirty characters. The Naïve Bayes Classifier was 92,7% successful, being surpassed by the word recognition approach. For the word classification approach, optimal success (80,8%) was reached on text samples with 5 to 90 characters, whereas 98% success could be gained on samples of more than 30 characters.
The detail description of all obtained results and possible improvements will be discussed in the evaluation.

II Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Wissenschaft und Technik haben in den letzten Jahrhunderten eine rasante Entwicklung gezeigt. Weltweit steigt die Zahl der Internetbenutzer immer weiter an und so werden technische Neuerungen, wie Computertechnik und Internet immer öfter benutzt. Mit der Entwicklung der Internettechnologien in der ganzen Welt wird das Internet also immer mehr zu einem multikulturellen Medium. Man findet im Internet fast alle möglichen Informationen, jedoch nicht in jeder Sprache. Infolge steigt Interesse an der automatischen Extraktion, Klassifikation und Filterung von unstrukturierten Texten. Es gibt viele Online-Tools, z.B. Google Translate, Bing (Microsoft) Translation, Yahoo! Babelfisch. Diese Dienste, die den Text erkennen und übersetzen können, sind nicht immer überzeugend, aber in vielen Fällen zuverlässig. In den meisten Fällen ermöglichen diese Tools hohe Erkennungsrate der Sprachen. Es gibt aber eine Menge von Nuancen, die nicht immer berücksichtigt werden. Es bleibt weiterhin eine substantielle Anzahl von fremden Wortformen, z.B. alte Schreibungen, OCR-Fehler, Slang oder Dialektwörter, die sogar für einen Muttersprachler nicht immer eindeutig sind. Beim Übersetzen erkennt Google Translate den bayrischen Dialekt als Deutsch, aber es findet keine Übersetzung statt.
Die automatische Identifizierung der Sprachen für elektronische Dokumente, die reguläuren Text enthalten, kann als gelöstes Problem betrachtet werden. Aufgrund der Existenz der verschiedenen Dialekte, reicht es allerdings nicht aus, sich nur auf Sprachenidentifizierung zu beschränken. Es gibt mehrere Gründe dafür:

1. Popularität der Dialekte in deutscher Sprache.
2. Auftreten der Dialekte in der Literatur wegen Fehlens einer Standardsprache im 16 - 20 Jahrhundert.

In dieser Arbeit wird der Schwerpunkt auf die Dialekterkennung gelegt. Die Identifizierung der Dialekte gilt als eine der zusätzlichen Forschungsgebiete, die eine Erweiterung für die Sprachenidentifizierung schaffen. Man erkennt gleich, auf der Identifizierung der Dialekte eines elektronischen Dokumentes bauen weitere Verarbeitungsschritte auf, von der automatischen Übersetzung bis zur Autokorrektur. Dieses Gebiet ist recht jung und aus theoretischer Sicht sehr interessant, da noch kein Ansatz gefunden wurde, der das Problem befriedigend löst.
Die Arbeit ist durch 3 konkrete Punkte motiviert:

1. Nichterkennung der veralteten Wortformen führt zur falschen Identifizierung der Sprache.

2. Verschlechterung der Sprachenerkennung aufgrund der Existenz der Dialektwörter.
3. Ein 100% reines Dialektwörterbuch ist in vielen Fällen nicht ohne weiteres verfügbar.
4. Dynamische Wahl der Tooling-Software beim Parsing, statt einem Standardparser kann ein bestimmter Parser verwendet werden, der nicht für alle, sondern nur für bestimmte Dialekte hohe Erkennungsrate aufweist.

Identifikation der Dialekte elektronischer Textdokumente ist eine der wichtigsten Stufen in vielen Prozessen maschineller Textverarbeitung. Diese Arbeit beschäftigt sich mit der Erkennung der Dialekte bei der Verarbeitung von Dokumenten, die in einem Dialekt verfasst wurden.

1.2 Aufbau der Arbeit

Die Arbeit wird in zwei Hauptteile gegliedert. Der erste Teil besteht aus Kapiteln 1-5, in denen theoretische Grundlagen zum Thema Dialekterkennung dargelegt werden.
Das erste Kapitel beschreibt durch welche Punkte diese Arbeit motiviert ist.
Im zweiten Kapitel werden grundlegende Begriffe präsentiert. Dieses Kapitel stellt Arbeiten vor, die sich mit der Identifikation in Texten befasst haben.
Im dritten Kapitel wird die Identifizierung der Sprachen mit Bezug auf Dialekte erläutert und einen kurzen Überblick in die verwendeten Daten gegeben.
Das vierte Kapitel erläutert kurz das UIMA-Framework.
Das fünfte Kapitel stellt drei Methoden zur Dialekterkennung vor, einen wortbasierten, einen N-Gramm basierten Ansatz und Naive Bayes-Klassifikator. Vor- und Nachteile der drei Techniken werden miteinander verglichen und diskutiert.
Der zweite Teil der Arbeit stellt die Implementierung der Methoden zur Identifizierung der Dialekte dar.
Im sechsten Kapitel wird gezeigt, wie die im letzten Kapitel vorgestellten Ansätze in Java implementiert werden. Zu diesem Zweck wird das UIMA-Framework verwendet. Dieses Kapitel vergleicht drei Ansätze und befasst sich mit den im Prozess der Entwicklung aufgetretenen Problemen. Der Vergleich des Goldstandardes mit dem Lösungsvorschlag der vorgestellten Ansätzen wird in der Evaluierung (6) präsentiert. Fehleranalyse wird durchgeführt und die wichtigsten Ergebnisse werden diskutiert.

2 Grundlegendes zur Dialektidentifikation

Dieses Kapitel führt den Begriff der Dialektidentifikation ein. Es werden theoretische Grundlagen beschrieben und schafft einen Überblick über die wissenschaftlichen Beiträge auf diesem Gebiet.

2.1 Dialekterkennung als Teilschritt des IE

Informationsextraktion (IE) beschreibt das Verfahren, unstrukturierte Informationen aus elektronischen Texten maschinell zu verarbeiten. Mit anderen Worten beschäftigt sich Informationsextraktion mit der Suche nach relevanten Informationen in großen Textmengen. Die Ausgangsdaten für die IE sind hauptsächlich in Form von unstrukturierten elektronischen Dokumenten gegeben.
Unter der automatischen Identifikation der Dialekte wird in der vorliegenden Arbeit ein Ablauf verstanden, in dem der Dialekt eines elektronischen Textdokuments mithilfe eines entwickelten Programms erkannt wird. Automatische Dialekterkennung wird in der Fachliteratur als Automatic Dialect Recognition bezeichnet.
Hauptsächlich verläuft der Prozess der automatischen Dialektidentifikation in zwei Stufen. In der ersten Phase lernt das Programm die typischen Wörter, bzw. Merkmale eines Dialektes. Für diese wichtige Phase wird eine Sammlung von in elektronischer Form vorliegenden Dokumenten benötigt, die bezüglich deren Dialekte bereits vorklassifiziert sind. Die Dokumente müssen dabei alle Dialekte abdecken, die später durch das System identifiziert werden sollen. In der zweiten Phase werden Dialekte aller Textdokumente bestimmt, die vom System klassifiziert werden müssen. Mittels eines Klassifikationsverfahrens wird das Testdokument bzw. Testroman dem entsprechenden Dialekt zugeordnet. Eine detaillierte Beschreibung der darin angewandten Methoden werden in den nachfolgenden Kapiteln dargestellt.

2.2 Stand der Forschung

Mit der automatischen Identifikation der Sprachen beschäftigten sich einige Arbeiten, indem Ergebnisse der Evaluierungen unterschiedlicher Ansätze präsentiert wurden. In diesem Zusammenhang sind vor allem die Arbeiten von B. M. Schulze, C. Souter, G. Grefenstette, W. B. Cavnar & J. M. Trenkle zu erwähnen. In diesen Arbeiten werden meistens N-Gramm-Techniken präsentiert, die gute Erkennungsquoten erreichen.
G. Grefenstette[Gregory Grefenstette, 1995] beschränke sich dabei auf Sprachen mit lateinischem Alphabet. Die N-Gramm-Analyse wird verwendet, um die Frage zu beantworten, wie wahrscheinlich ein Satz in einem bestimmtem Dialekt verfasst wurde.
In der Arbeit von W. B. Cavnar & J. M. Trenkle [William B.Canvar and John M.Trenkle, 1994]

wird ein N-Gramm basierter Ansatz zur Textkategorisierung beschrieben. Dieses Verfahren erzielt 99,8% für korrekte Einstufung der Usenet-Nachrichtengruppen-Artikel, die in verschiedenen Sprachen geschrieben wurden. Typischerweise waren diese Trainingssätze 20 KB bis 120 KB lang (20 480 bis 122 880 ASCII-Zeichen). Insgesamt lieferte das System seine beste Leistung mit 400 N-Grammen. Falsch klassifiziert wurden nur 7 Artikel von 3478, was eine Gesamtklassifikationsrate von 99,8% ergeben hat.
M. J. Martino & R. C. Paulsen[Michael John Martino, Robert Charles Paulsen et al., 2001] haben normierte Häufigkeit des Auftretens der Wörter (NFO) verwendet um indoeuropäische Sprachen zu unterscheiden. Jede Tabelle präsentiert eine Liste mit den Wörtern und deren zugehörige Häufigkeit des Auftretens für jede Sprache. Z.B. in Französisch beträgt NFO für “que“ 24,8. In Spanisch, ist NFO für “que“ 21,4. Aus den Berechnungen für das Wort “que“, ist zu erkennen, dass bei der Prüfung elektronischer Dokumente, die jeweils in Französisch und Spanisch geschrieben wurden, die Wahrscheinlichkeit, dass diesem Dokument Französisch zugeordnet wird, ist 24.8 / (2l.4 + 24,8) = 0,537, und für Spanisch 21,4 / (21,4 + 24,8) = 0,463. Diese Technik der Identifikation hängt stark davon ab, wieviele Wörter mit den Wörtern aus den Tabellen ubereinstimmen. Dem Dokument wird die Sprache mit der höchsten akkumulierten Summe zugeordnet.
P. Sibun & J. C. Reynar[Penelope Sibun, Jeffrey C.Reynar, 1996] haben Kullback-Leibler-Divergenz verwendet um achtzehn Sprachen zu klassifizieren. Die KL-Divergenz wird auch relative Entropie genannt. In Englisch, folgt der Buchstabe “u“ typischerweise nach dem Buchstaben “q“. Der Buchstabe “ q“ kann nach dem Buchstaben “v“ nicht gefunden werden. Einige Buchstaben-Kombinationen sind nicht möglich, andere erscheinen häufiger. Kullback-Leibler-Divergenz bezeichnet ein Maß für die Unterschiedlichkeit zweier Warscheinlichkeitsverteilungen. Die kleinste Distanz zwischen den Verteilungen der Sprache und dem Testdokument entspricht der Sprache, in der das Dokument verfasst ist. Bei der mehr als 2000 Zeilen zeigte diese Technik zwischen 81,5 und 100% Erfolgsquote. Erkennungsrate variiert aufgrund der Anzahl (1 bis 20) und des Types (Unigramm oder Bigramme) der N-Gramme. Bei einer Anzahl von vierzig Bigrammen wurde eine korrekte Erkennung von 100 % erreicht.
C. Souter et al. (1994)[Clive Souter, 2015] hat mehrere Verfahren zur Identifikation von neun Sprachen (Friesian, Englisch, Französisch, Gälisch, Deutsch, Italienisch, Portugiesisch, Serbokroatisch und Spanisch) vorgestellt, unter anderem ist Identifikation auf der Basis von typischen Sonderzeichen (z. B. ů, ä, ß, ï usw.) oder Identifikation mithilfe der Wahrscheinlichkeit der unikalen Buchstabenkombinationen. Bei diesen Verfahren wird dem Algorithmus rund 100 Kilobytes von Text mit den Sätzen verschiedener Länge übergeben, um zu sehen, wie die Leistung durch Variieren der Länge der Textproben, beeinträchtigt wird. Diese Methoden haben nicht besonders hohe Erkennungsrate erwie-

sen und werden nicht oft in der Praxis eingesetzt. Wie erwartet, erwies sich die Methode als nicht erfolgreich, insgesamt erreichte eine Erfolgsquote von 24%, weil in vielen Fällen das Testmaterial keine der einzigartigen Buchstabenfolgen enthielt. In ihrer Forschung wurden auch Bigramme und Trigramme verwendet, um indoeuropaischen Sprachen außer Französisch und Englisch zu klassifizieren. Bei dem Bigrammmodell wurde ein optimales Ergebnis (nahezu 100%) auf Textproben von 200 Zeichen oder mehr erreicht, wohingegen bei den Trigrammmethoden 100% Erkennungsquote an Proben von mehr als 175 Zeichen erzielt wurde. Für sehr lange Testdateien war das Trigrammmodell fast zu 100% erfolgreich bei der Identifizierung von 45 Sprachen. Die Erkennung des Portugiesischen war jedoch sehr schlecht. Es gibt keine Bigramme, die eindeutig dem Portugiesischen zugeordnet werden konnten. Mit der Erhöhung der Länge des Testmaterials steigt auch die Erkennungsrate. Oben genannte Techniken und Methoden zur automatischen Identifizierung sind universell verwendbar.

Nach Zaidan & Callison-Burch [Omar F. Zaidan, Chris Callison-Burch, 2012], ähnelt sich Dialekterkennung weitgehend der Erkennung der Sprachen, d.h. diese können sowohl für Erkennung der Sprache als auch für Erkennung der Dialekte angewendet werden. Diese zwei Techniken zur automatischen Identifikation, nämlich, der wortbasierte und der N-Gramm basierte Ansatz werden in den nächsten Kapiteln dieser Arbeit ausführlicher präsentiert.

Es gibt fundamentale Arbeiten, die sich mit der Erkennung der Dialekte beschäftigten: Fadi Biadsy (2009, 2010, 2014), Emmanuel Ferragne und Francois Pellegrini (2007), Fatiha Sadat & Farnazeh Kazemi und Atefeh Farzindar (2014).

Fadi Biadsy [Fadi Biadsy, Julia Hirschberg et al., 2009] verwendete die Trigramm-Technik zur Identifikation der arabischen Dialekte. Die untersuchten Dialekte sind Levantinisch-Arabisch, Irakisch-Arabisch, Golf-Arabisch und Ägyptisch-Arabisch. Insgesamt wurden 9 x 4 = 36 Trigramme gesammelt. Aus den Ergebnissen war zu sehen, die beste Erkennungsquote unter den vier Dialekten erreichte Ägyptisch (F-Maß von 94%), gefolgt von Levantinisch (F-Maß von 84%). Eine nicht besonders hohe Identifizierungsrate wurde bei der Erkennung der Dialekte Golf und Irakisch (F-Maß von 68,7%, 67,3%) festgestellt. Diese Verwechselung ist konsistent, da Irakisch als ein Unterdialekt des Golfs betrachtet wird.

In der wissentschaftlichen Arbeit von Fatiha Sadat, Farnazeh Kazemi und Atefeh Farzindar [Fatiha Sadat, Farnazeh Kazemi et al., 2014] wurden ein N-Gramm-Ansatz und ein Bayes-Klassifikator für die Identifizierung der arabischen Dialekte angewandt. Experimentelle Ergebnisse zeigten, dass der Naive Bayes-Algorithmus basierend auf Bigrammmodell achtzehn arabische Dialekte mit einer beträchtlichen Gesamtgenauigkeit von 98% ermittelte.

In der Arbeit [Emmanuel Ferragne, Francois Pellegrino, 2004] nahmen Wissenschaftler Emmanuel Ferragne and Francois Pellegrini in ihren Experimenten dialektische Unterschiede in der Struktur des Vokalsystems der englischen Dialekte in Betracht, weil die Anzahl der Vokale sowie deren Häufigkeit des Auftretens nicht gleich sei. Auswertung basierte auf 13 Dialekten der englischen Sprache im Gebiet der britischen Inseln. Korrekte Identifizierung wurde bis zu 90% erreicht.
Aufgrund der verschiedenen Struktur der Sprachen ist es gelungen mit Trigrammmodellen Erkennungsrate auf 100% zu steigern. Obwohl Dialekte einer Sprache eine ähnliche Struktur haben, konnten arabische Dialekte zu 98% korrekt identifiziert werden. Dialekte der englischen Sprachen erreichten nur 90% des Erfolges. Es lässt darauf schließen,dass die Erkennungsrate nicht nur von der Technik abhängt, sondern auch von den zu klassifizierenden Dialekten. Automatische Identifizierung der deutschen Dialekte wurde nach meinem besten Wissen noch nicht durchgeführt. Die aus dieser Arbeit fließende Ergebnisse werden in der Evaluierung (6) präsentiert.

3 Sprachen und Dialekte

3.1 Identifiziererung der Sprachen im Bezug auf Dialekte

Die automatische Identifizierung der Sprachen für elektronische Dokumente ist in den letzten Jahren durch die Zunahme von multilingualen elektronischen Textsammlungen zu einem unentbehrlichen Werkzeug geworden. Die Identifizierung der gängigen Sprachen verläuft meistens problemlos, da die Häufigkeitsverteilung der Buchstaben stark von der jeweiligen Sprache abhängt. Fast jede Sprache hat irgendein Alleinstellungsmerkmal, mit dem man den Text einer Sprache zuordnen kann. Die gängigen Sprachen besitzen Sonderzeichen, Ligaturen, besondere Schriftzeichen und viele spezielle Diakritika. Dadurch lässt sich die Sprache des Textes leicht identifizieren. Für die Identifizierung der Sprachen haben sich vor allem N-Gramm Ansätze etabliert. Viele Autoren haben diese Technik für die Identifizierung der Sprache untersucht und halten diese für besonders erfolgreich [William B.Canvar and John M.Trenkle, 1994].
Der Dialekt einer Sprache ist ein sehr wichtiges expressives Mittel, um Identität einer gesellschaftlichen Gruppe Ausdruck zu verleihen. Demnach treten Dialekte sehr oft in der Literatur auf. Nach Zaiden und Callison-Burch [Omar F. Zaidan, Chris Callison-Burch, 2012], ist die Aufgabe der Dialektidentifikation, die Entwicklung einer Software, die in der Lage ist, festzustellen, ob ein Satz oder ein Text dialektalen Content enthält. Weiterhin betonen die beiden Wissenschaftler, dass das Problem der Dialektidentifikation in vielerlei Hinsicht der Sprachenerkennung äquivalent ist. Erkennung der Dialekte gehört zu einer eng verwandten Gruppe einer Sprache, ist damit ein schwieriger Fall der Sprachenerkennung. In ihrer Forschung zur Identifikation der arabischen Dialekte wurde ein N-Gramm basierter Ansatz angewandt. Da Dialekt funktional eingeschränkt und nicht standardisiert ist, hat er keinen offiziellen Status. Die Identifizierung der Texte, die im Dialekt verfasst wurden, stellt ein Problem dar. Dialekte einer Sprache teilen die gleiche Schrift, haben ähnliche Schreibweisen und somit sind nicht einfach voneinander trennbar. Im Gegensatz zur automatischen Sprachenerkennung, stellen N-Gramm Modelle im Bereich der Dialekterkennung noch einen relativ neuen Ansatz dar. Die Zuverlässigkeit der Resultate in dieser Arbeit wird in der Evaluierung (6) besprochen.

3.2 Sprachwissenschaftliche Kriterien zur Abgrenzung der Dialekte und Sprachen

Eines der wichtigsten und schwierigsten Probleme in der linguistischen Forschung ist die Frage, wie sich eine Sprache von einem Dialekt oder wie man zwei Dialekte einer Sprache abgrenzen kann.
Trotzt der Aussage von Robert Hinderling : *„Von den grammatischen Besonderheiten*

her ist das Eigengepräge des Bairischen gegenüber dem Schriftdeutschen so stark, dass allein diese Tatsache genügt, dem Bairischen den Status einer eigenen Sprache zu verleihen.“ [1] wird Bairisch als Dialekt der deutschen Sprache eingestuft. Im Rahmen dieser Arbeit werde ich mich mit der Problematik „Automatische Abgrenzung der Dialekte“ mehr beschäftigen. Zuordnung der Wörter zu einem Dialekt wird oft nach teilweise widersprechenden Überlegungen durchgeführt.
Schwäbisch zählt zu den alemannischen Dialekten, Bairisch - zu den bairischen Dialekten, jedoch gehören beide Dialekte zur deutschen Sprache und besitzen einige Ähnlichkeiten. Da die meisten Dialekte nicht standardisiert sind, sind Variationen noch viel geläufiger. Es gibt Wörter, die zu beiden Dialekten gehören, z.B. “*Jawoll*“, “*Jeld*“ oder “*Jeschichte*“ (volle Liste findet man im Anhang 1). In diesen Fällen braucht man einige Merkmale der entsprechenden Dialekte, damit sichere Erkennung stattfinden kann, z. B. die Kombination „*sch*“ vor den Buchstaben „*t*“, „*d*“, „*b*“, „*p*“, „*g*“, „*k*“ , sowie die Buchstabenkombination „*äi*“ sind typisch, besonders für den schwäbischen Dialekt.

[1] https://de.wikipedia.org/wiki/Bairische_Dialekte

3.3 In dieser Arbeit verwendete Daten

Über 1500 in deutscher Sprache geschriebene Romane wurden zur Verfügung gestellt. Zunächst wird nach den Romanen gesucht, die den größten Prozentanteil an fremde Wörter enthalten. Mit Hilfe des UIMA-Frameworks [2] (4) ist es möglich auf der Basis von regulären Ausdrücken Entitäten oder Wörter innerhalb der direkten Rede auszufiltern, deren Formen in deutscher Sprache nicht erkannt werden. Zum Beispiel, Wörter „*hòb, hòsd, hòd, ham, habts*“ sind verschiedene Wortformen von einem Lexem mit Lemma „*haben*“. Solche nicht erkannten Wortformen werden in den entsprechenden Dialektwörterbüchern gespeichert. Drei verschiedene Varianten der Klammern werden berücksigtigt:

```
"≫.*?≪", "\".*?"\", "<.*?>"
```

Der Punkt im regulären Ausdruck steht für ein beliebiges Zeichen, und der folgende Stern ist ein Quantifizierer, der viele beliebige Zeichen erlaubt. Es wurden nicht erkannte Wörter (Dialektwörter) innerhalb der direkten Rede gefunden, und deren summierte Anzahl durch die gesamte Wörteranzahl dividiert. Zwei Tabellen im Anhang 3 stellen Romane dar, in denen relative Häufigkeit des Auftretens der nicht erkannten Wörter (Dialektwörter) am größten ist. Es wurden insgesamt 28 Romane ausgesucht mit einem Dialektanteil zwischen 3.2% und 15.2%. Bairisch, Berlinerisch, Mecklenburgisch, Schwäbisch und Schlessisch werden laut Wikipedia als Dialekte der deutschen Sprache eingestuft. [3] Jedoch ist es nicht immer möglich den Dialekt eindeutig zu bestimmen. Im Roman von Speckmann Diedrich geht es um keinen Dialekt, sondern um Plattdeutsch - eine verbreitete westgermanische Sprache in Norddeutschland und im Osten der Niederlande, die eine Vielzahl unterschiedlicher Dialektformen besitzt und sich aus dem Altsächsischen entwickelt hat [4].
Aus den ausgesuchten Romanen wurde ein Goldstandard erstellt. Zunächst wurden zwei Trainingskorpusse gebildet. Erster enthält 204 Sätze in verschiedener Länge, die in 12 verschiedenen Dialekten geschrieben wurden. Der andere Text enthält nur die Sätze, die mehr als 30 Zeichen enthalten, um zu schauen wie die Erkennungsrate sich verändert. Dritter Trainingskorpus enthält nur in Hochdeutsch geschriebenen Text, um sicherzustellen, dass Deutsch nicht als Dialekt erkannt wird.

[2] https://uima.apache.org
[3] https://de.wikipedia.org/wiki/Deutsche_Dialekte
[4] https://de.wikipedia.org/wiki/Niederdeutsche_Sprache

4 Einführung in UIMA-Framework

Die Abkürzung UIMA steht für „Unstructured Information Management Architecture“ und dient zur Analyse, Filtern und Wissensextraktion. UIMA kann in Form eines von Apache frei verfügbares Frameworks genutzt werden. UIMA kann verwendet werden für Informationen, die in beliebigen Formaten vorliegen. In dieser Arbeit handelt es sich um „unstrukturierten Daten“ in Form von Text, wobei UIMA nicht auf Text beschränkt, sondern offen für Daten aller Art ist.

5 Methoden und Techniken zur Erkennung der Dialekte

In diesem Abschnitt werden aus der wissenschaftlichen Literatur bekannte und meist verbreitete Ansätze zur Identifikation der Dialekte beschrieben und miteinander verglichen. Es gibt rund zehn Methoden, die wichtigsten sind Merkmalklassifizierung, Wortbasierter Ansatz, N-Gramm-Techniken und ein Naive Bayes-Klassifikator.

5.1 Merkmalserkennung

Fast jeder Dialekt hat bestimmte Merkmale. Typisch für den tschechischen Dialekt ist beispielsweise der Buchstabe „*ů*“. Im schweizerischen Dialekt taucht der Buchstabe „*ß*“ nicht auf, sondern nur „*ss*“. Die Merkmalserkennung nutzt in der Regel mehrere Erkennungsverfahren nacheinander. Die Ergebnisse aller genutzten Verfahren muss die Software dann vergleichen und gewichten. Im schwäbischen Dialekt wird meistens „*oi*“ statt „*ei*“ *(hoim)* verwendet. Diese Technik ist nicht immer treffsicher bei der Identifizierung der Dialekte, da meistens Dialekte einer Sprache die gleichen Sonderzeichen besitzen.

5.2 N-Gramm-Ansatz

Jeder Dialekt hat eigene Prinzipien, wie die neuen Wörter gebildet werden. Häufigkeitsanalyse der Buchstabenkombinationen ist eine der Möglichkeiten um Dialekte voneinander zu unterscheiden. Das System nutzt die Wahrscheinlichkeit gängiger Buchstabenfolgen für die Dialekte, die erkannt werden sollen. Die meisten Dialektidentifikationssysteme dieser Art arbeiten mit Bigramm- und Trigrammfolgen. Möchte man z.B. automatisiert den Dialekt ermitteln, so gibt es hier einen Algorithmus. In jeder Sprache werden einzelne Buchstabenkombinationen häufiger benutzt. Mit einer Tabelle, die für jeden Dialekt erstellt wurde, kann man schließlich die Gewichtungen und Wahrscheinlichkeiten ausrechen, dass der Text in einem bestimmten Dialekt verfasst wurde.
In der Abbildung 1 sind die meist vorkommenden Bigramme der deutschen Sprache aufgezeichnet. Aus dem Diagramm ist ersichtlich, dass mehrere Dialekte (Mechlenburgisch, Altbairisch, Plattdeutsch, Tschechisch und Schwedisch) ähnliche Bigrammereihenfolge (“*en*“,

“*er*“, “*ch*“) haben. Aufgrund dessen, fällt es schwer, den Text einem bestimmten Dialekt zuzuordnen. Seltene Bigramme findet man in der Abbildung 2. Es gibt unikale Bigramme,

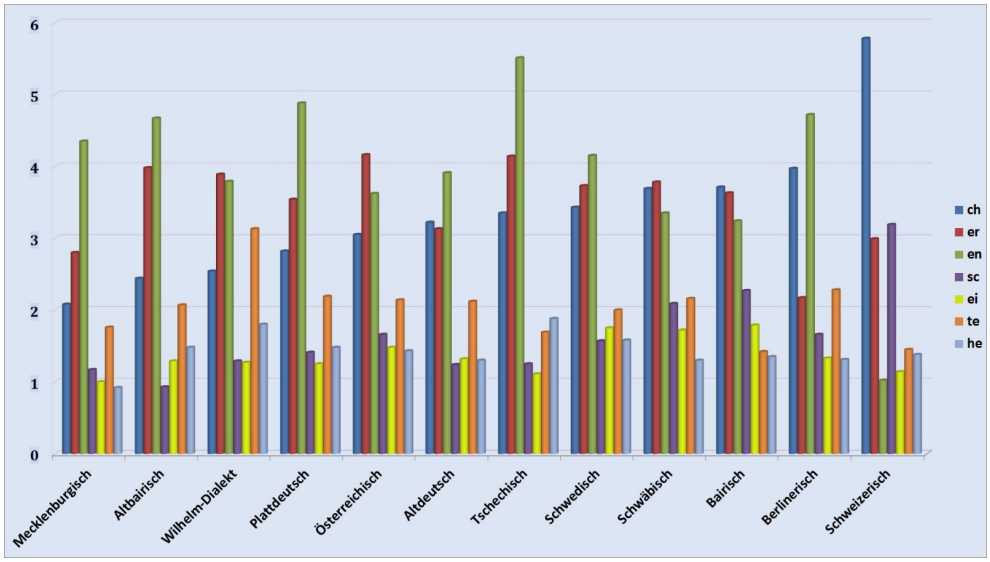

Abbildung 1: Sieben meist vorkommende Bigramme in zwölf deutschen Dialekten

die nur für einen bestimmten Dialekt gekennzeichnet. Z.B. “*je*“ und “*ee*“ kommen nur im Berlinerischen vor. Man trifft das Bigramm “*ee*“ sehr selten. Es kommt aber nicht immer vor, dass eine der unikalen Bigramme im zu untersuchenden Text auftaucht, was zufolge hat, dass diese “Merkamalserkennung“ (das Vorkommen des “*ee*“ nur im Berlinerischen Dialekt) nicht zuverläßig ist.

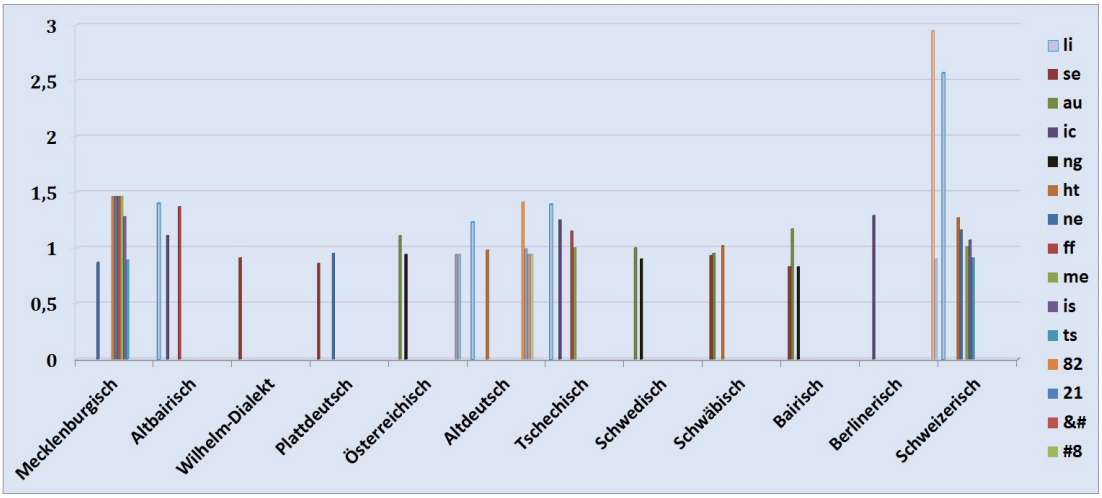

Abbildung 2: Seltene Bigramme der deutschen Dialekte

Je mehr Tabellen mit N-Grammfolgen vorhanden sind, um so mehr Dialekte werden erkannt, allerdings um so länger dauert auch die Generierung der Wahrscheinlichkeiten für jeden Dialekt.

Folgende Tabelle präsentiert eine Liste von Bigrammen für mecklenburgischen und altbairischen Dialekte, Plattdeutsch und deren relative Häufigkeit in Prozent. Aus mehr als 300 Bigramme wurden die zwanzig häufigsten ausgewählt. Aus der Tabelle ist zu entneh-

Mecklenburgisch	Altbairisch	Plattdeutsch
en 4.35	en 4.67	en 4.88
er 2.80	er 3.98	er 3.54
ch 2.08	ge 2.93	ch 2.82
te 1.76	ch 2.44	te 2.19
st 1.61	te 2.07	ge 1.83
82 1.46	et 2.02	**st 1.58**
21 1.46	he 1.48	he 1.48
&# 1.46	li 1.40	**sc 1.41**
#8 1.46	re 1.38	in 1.25
ge 1.44	be 1.38	ei 1.25
17 1.28	ff 1.37	ie 1.11
sc 1.17	le 1.29	el 1.11
re 1.12	ei 1.29	de 1.11
in 1.07	in 1.25	nd 0.99
ei 1.00	**st 1.15**	le 0.98
el 0.95	ic 1.11	re 0.96
he 0.92	an 1.06	**ne 0.95**
we 0.89	nd 1.02	be 0.95
an 0.88	de 1.02	es 0.92
ne 0.87	**sc 0.93**	se 0.86

men, dass z.B. Mecklenburgischer Dialekt und Plattdeutsch über ähnliche Reihenfolge der N-Gramme verfügen. Folgender Satz “a Schlawiner, a Tanzmoasta!“ enthält die gleiche Reihenfolge von Bigrammen: “*ta*“ , “*zm*“ , “*r,*“ , **“st“** , “*nz*“, **“sc“**, **“ne“**, “*oa*“, “*mo*“, “*wi*“. Der Satz konnte nicht korrekt identifiziert werden, da die gleiche Reihenfolge von Bigrammen mit zwei Dialekten (Mecklenburgisch und Plattdeutsch) überreinstimmt, zwar: “*st*“, “*sc*“, “*ne*“.

5.3 Wortbasierter Ansatz

Bei dieser Methode verwenden die Programme große Datenbanken mit unterschiedlichen Wörtern. Stimmt das gefundene Wort mit einem Wort aus der Datenbank überein, gilt es als erkannt. Das klingt zwar einfach, ist in der Praxis aber zeitintensiv. Im besten Fall kennt das trainierte System alle Wortformen der Dialektes, die erkannt werden sollen und deren Frequenz im Trainingskorpus. Im einfachsten Fall enthält das Lexikon nur hochfrequente Wortformen. Es kann allerdings beliebig mit weniger häufigen Wortformen ergänzt werden, um die Erkennungsrate für kurze Dokumente zu erhöhen.

Oben erwähnte Techniken werden oft für Identifizierung der Sprachen eingesetzt. Erkennungsquote dieser Techniken bei der Dialekterkennung wird in Evaluierung (6) diskutiert.

5.4 Vorteile und Nachteile der N-Gramm und wortbasierten Ansätze

5.4.1 N-Gramm Ansatz

Beim Erstellen der Tabellen mit N-Buchstabenfolgen hat diese Methode eine geringe Dateigröße, da meistens 300 N-Grammfolgen für jeden Dialekt ausreichend sind. Dadurch wird eine schnelle Laufzeit erreicht. Der N-Gramm Ansatz hat viele Vorteile, aber auch einige Nachteile. Bei Flexionen kommt es oft zu Problemen: z.B. bei Umlauten: beim Wort „Bücher“ gibt es keine eindeutige Zerlegung „ü“(ue) oder „üc“(uec). Der größte Nachteil ist, dass diese Technik nur bei langen Texten zuverlässig ist. Die Ergebnisse der N-Gramm-Techniken sind bei der Unterscheidung sehr nah verwandter Dialekte nicht besonders treffsicher (d.h. Dialekte mit ähnlichen Bigrammen).
In der Abbildung 3 sind Bigramme aufgelistet, die nicht in jedem Dialekt vorkommen. Berlinerisch und Österreichisch lassen sich nicht einfach unterscheiden, obwohl diese Dialekte nicht nur abweichende Bigrammreihenfolge, sondern auch völlig verschiedene Bigramme aufweisen (im Bairischen gibt es unikale Bigramme: “*je*“ und “*ee*“, im Österreichischen: “*ra*“ und “*g’*“). Wenn in einem Satz nur zwei Bigramme vorkommen: “*en*“ und“*ch*”, kommt es wieder zu Schwierigkeiten bei der Dialektzuordnung, da beide Bigramme in der gleichen Reihenfolge in diesen zwei Dialekten zu finden sind. Klassifizierung einzelner Dokumente lassen sich in Systemen, denen dieser Ansatz zugrunde liegt, nicht ohne weiteres erkennen und eliminieren.

5.4.2 Wortbasierter Ansatz

Das Training des wortbasierten Ansatzes ist aufwändiger als bei dem oben genannten Ansätzen, da für eine sichere Identifizierung der Dialekte die Wörterbücher manuell zusammengefasst und gefiltert werden sollen, um Fremdwörter weitgehend aus den Wortlis-

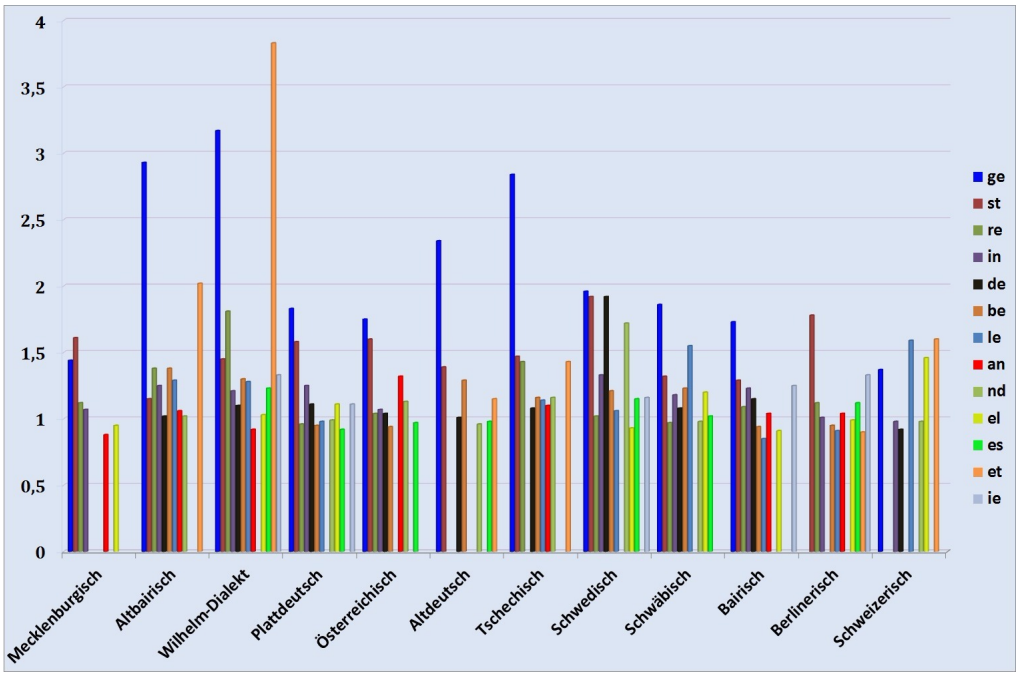

Abbildung 3: Bigramme, die nicht in jedem Dialekt vorkommen

ten zu eliminieren. Während solche Wortformen bei langen Dokumenten keine größeren Probleme bereiten, können Sie bei sehr kurzen Dokumenten leicht zu einer Fehlklassifikation führen. Ein anderer Nachteil ist, dass man ziemlich umfangreiche Wörterbücher für jeden Dialekt braucht. Die Erkennungsrate wächst mit der Größe der Wörterbücher für jeden Dialekt.

5.5 Probleme bei Ambiguität der Dialektwörter

Ambiguität kann sich auf vielen Bedeutungsebenen beziehen. Es existiert zwei Hauptarten von lexikalischer Ambiguität: Homonymie und Polysemie. Bei der Erkennung der Dialekte soll das Problem der Polysemie (Mehrdeutigkeit) besonders untersucht werden. Von dem Begriff der Ambiguität wird gesprochen, wenn einem Wort mehrere Bedeutungen zugeordnet sind. In der Sprachverarbeitung liegt die Schwierigkeit meistens in der Disambiguierung von mehrdeutigen und unbekannten Wörtern. Die Erscheinung der verschiedenen Bedeutungen eines Wortes ist eng verwandt mit der Entstehung der Dialekte. Das gleiche Wort kann sich sowohl auf deutsche Sprache als auch einen Dialekt beziehen:
Schwelle (Berliner Dialekt): die eigene Schwester [5]
Schwelle (Hochdeutsch): in den Türrahmen eingepasster, etwas erhöht liegender Balken als unterer Abschluss einer Türöffnung; Türschwelle [6]

[5] http://www.spreetaufe.de/berlinerisch-berliner-jargon/woerterbuch-berlinisch-p-z/#S1
[6] http://www.duden.de/rechtschreibung/Schwelle

In diesem Fall ist die Zuordnung eines Wortes zur deutschen Sprache oder zu einem Dialekt problematisch. Analyse der syntaktischen Struktur ist in diesen Fällen notwendig,

5.6 Naive Bayes-Klassifikator

Naive Bayes-Klassifikator ist in der Textanalyse weit verbreitet. Der Bayes-Klassifikator ist eine statistische Methode, welche jedes Objekt der Klasse zuordnet, bei der die wenigsten Kosten entstehen. Ihre guten Klassifizierungseigenschaften werden insbesondere zur Klassifikation von E-Mails oder beim Einsatz in Spam-Filtern eingesetzt. Die Formel des Bayes' Klassifikators ist bezüglich der Dialekterkennung folgendermaßen definiert:

$$c = \arg\max_{c \epsilon C} \Big[\log \frac{D_c}{D} + \sum_{i \epsilon Q} \log \frac{W_i c + 1}{|V| + L_c} \Big] \tag{1}$$

c	Dialekt mit der höchsten Warscheinlichkeit
D_c	= 1, da jedes Dialektworterbuch in einem Dokument erfasst ist
D	Anzahl von Dialektwörterbüchern
$\|V\|$	eindeutige Wörter im Wörterbuch
L_c	Anzahl aller Wörter im Dialektwörterbuch
$W_i c$	Häufigkeit des Wortauftretens im Wörterbuch (0(nicht vorhanden) oder 1(vorhanden))
Q	alle inclusive wiederholende Wörter

Bei der Klassifikation wird auch auf Disambiguitäten eingegangen, wenn ein Wort in mehreren Wörterbüchern vorhanden ist, wird die Wahrscheinlichkeit für jeden Dialekt berechnet und der Dialekt mit der größten Wahrscheinlichkeit als Ergebnis ausgegeben. Hier ist noch ein kleines Beispiel:
Es gäbe 3 Dialektwörterbücher: Dialekt1, Dialekt2 und Dialekt3, gefüllt mit den entsprechen Wörtern.

- [Dialekt1] Ambopoäng gsehgn Hausl;
- [Dialekt2] Hellbrunen hett herümmerführt;
- [Dialekt3] Letst hett erlabet;

Nun rechnen wir die Wahrscheinlichkeit des Satzes „Hellbrunen hett Hausl“ für jeden Dialekt:
Dialekt1: $\log \frac{1}{3} + \log \frac{1}{6} + \log \frac{1}{6} + \log \frac{2}{6}$ = -2,51
Dialekt2: $\log \frac{1}{3} + \log \frac{2}{6} + \log \frac{2}{6} + \log \frac{2}{6}$ = -2,2
Dialekt3: l $\log \frac{1}{3} + \log \frac{1}{6} + \log \frac{2}{6} + \log \frac{1}{6}$= -2,51
Daraus ist auszulesen, dass der oben genannte Satz mit der größeren Wahrscheinlichkeit im Dialekt2 verfasst wurde. Für eine erfolgreiche Implementierung des Klassifikators mussten alle Dialektwörterbücher mit einander vergleichen und doppelte Wörter eliminiert werden.

6 Evaluierung

6.1 Programmiersprache und die Entwicklungsumgebung

Unter der Benutzung des UIMA-Frameworks und der Programmiersprache JAVA werden Methoden und Algorithmen entwickelt, um Dokumente nach Sätzen und Wörtern zu durchsuchen und diese einem Dialekt zugewiesen. Apache OpenNLP Bibliothek unterstützt die gängigsten NLP Aufgaben wie Tokenisierung und Segmentierung der Sätze. Diese Schritte werden in der Regel erforderlich, um erweiterte Textverarbeitungsdienste aufzubauen. Das Ziel der praktischen Arbeit ist drei Ansätze zu implementieren, zwar: Wortbasierter und N-Gramm basierter Ansatz, sowie Naive Bayes-Klassifikator, und zu untersuchen wie die Trefferquote aufgrund der Satzlänge variiert. Folgende Jar-Bibloitheken werden benutzt:
opennlp-tools-1.6.0.jar
uima-tools.jar
uimafit-core-2.1.0.jar
uimaj-core-2.8.1.jar

6.2 Erstellung des Wörterbuches

Mit Hilfe von UIMA-Frameworkes wurden 28 Romane nach unbekannten Wörtern gesucht, solche ausgefiltert und in dem zugehörigen Dialektwörterbuch zusammengefasst. Aus allen Dialektwörterbüchern wurden Stoppwörter entfernt, um die Identifizierungsrate der Dialekte zu steigern. Folgende Liste präsentiert Anzahl der enthaltenen Wörter in jedem der zwölf Dialektwörterbücher.

Dialektname:	Anzahl der Wörter:
Altbairisch	3264
Altdeutsch	8870
Bairisch	19018
Berlinerisch	4242
Mecklenburgisch	8484
Österreichisch	4831
Plattdeutsch	5062
Schwäbisch	5892
Schwedisch	1884
Schweizerisch	1409
Tschechisch	8724
Wilhelm_Dialekt	2992

6.3 Goldstandard

Zunächst wurden zwei Trainingskorpusse gebildet. Erster enthält 204 in verschiedener Länge Sätze, die in 12 verschiedenen Dialekten geschrieben wurden, insgesamt 2943 Wörter. Der anderer Text enthält nur 60 Sätze, die mehr als 30 Zeichen enthalten (insgesamt 3143 Wörter), um zu schauen wie die Erkennungsrate sich verändert.Den Sätzen wurde Dialekt per Hand mithilfe vom Programm „Athen“ zugeordnet.
Zwei weitere Trainingskorpusse für Hochdeutscherkennung wurden erstellt. Erster Trainingskorpus enthielt 41 in Hochdeutsch geschriebene Sätze. Zweiter beinhaltet gemischen Text, bestehend aus 12 Hochdeutsch- und 12 Dialektsätzen, um zu schauen, ob Hochdeutsch nicht als Dialekt erkannt wird.

6.4 Experimentaufbau und -durchführung

Aus rund 1500 Romanen wurden welche ausgesucht, die den grössten Anteil von Fremdwörtern enthielten. Es wurden insgesamt 28 Romane mit einem Dialektanteil zwischen 3.2% und 15.2% (ausgewählte Romane sind im Anhang 3 zu finden) gefunden. Jedem Roman wurde ein Dialekt per Hand zugewiesen, in dem er verfasst wurde. Mit Hilfe von Uima-Framework wurden nicht erkannte Fremdwortformen (“unknown“ Lemmata) pro Roman ausgefiltert und in jeweiligem Dialektwörterbuch gespeichert. Aus erstellen Wörterbüchern wurden 1804 Stoppwörter eliminiert um die Erkennungsrate der Dialekte zu steigern, alle Dialektwörterbücher wurden danach miteinander verglichen um doppelte Einträge zu entfernen, dies ist notwendig für Naive Bayes-Klassifikator.
Aus den ausgewählten Romanen wurden zwei Trainingskorpusse mit den verschiedenen Satzlängen gebildet, um zu schauen wie die Erkennungsrate sich verändert. Jeder Trainingskorpus enthielt Sätze, die in einem der zwölf Dialekten geschrieben wurden. Jeder Satz wurde mit einem Dialekt annotiert, um festzustellen ob Dialekt der implementierten Ansätzen mit dem Goldstandard übereinstimmt.
Zunächst wurde der wortbasierte Ansatz implementiert, jedes nicht erkannte Wort aus einem eingegebenen Satz wurde mit den Wörtern jeder der zwölf unterstützten Dialekte verglichen; falls das Wort in einer der Listen gefunden wurde, wurde der Zähler für die jeweilige Sprache um eins erhöht. Es wurde angenommen, dass das Dokument im Dialekt verfasst wurde, die den höchsten Zählerstand hatte.
Aus den zwölf erstellten Dialektwörterbüchern konnten Bigramme und Trigramme gebildet werden, davon wurden zwanzig der meist vorkommenden ausgewählt und mit der entsprechenden Auftretenshäufigkeit absteigend gespeichert (alle Bigramme der zwölf Dialekte sind im Anhang 2 aufgelistet). Die Reihenfolge der Bigramm-Häufigkeiten wurde mit den Bigramm-Häufigkeiten jedes Dialektes verglichen. Je mehr Bigramme in der richtigen Reihenfolge sind, desto wahrscheinlicher, dass der Satz im bestimmten Dialekt verfasst

wurde.
Naive Bayes-klassifikator wurde nach der Formel aus Abschnitt 5.6 implementiert.

6.5 Auswertung der Ansätze zur Dialektidentifikation

6.5.1 Wortbasierter Ansatz

Auswertung №1 (Sätze verschiedener Länge). Der wortbasierte Ansatz bekam als Eingabe nur eine Datei in XMI-Format, die untersucht werden musste. Altbairischer, Altdeutsch, schweizerischer und tschechischer Dialekte wurden zu 100% richtig identifiziert. Auf dem zweiten Platz steht der schwedische Dialekt mit der Erkennungsrate von 90%. Die oben aufgezählten Dialekte erreichten hohe Erkennungsrate, weil jeder Dialekt fast keine Übereinstimmungen mit einander hat. Wilhelmer Dialekt sowie Bairisch wurden zu 50% korrekt erkannt. Aus den 28 in bairisch geschrieben Sätzen wurden nur 14 korrekt identifiziert. Aus den Ergebnissen ist zu entnehmen, dass meistens 2 Dialekte erkannt werden, zwar bairisch und österreichisch. Man stellt sich die Frage: ist Bayrisch und Österreichisch derselbe Dialekt? In Österreich gibt es einige Dialekte, die mit dem bairischen ähnlich klingen, aber trotzdem sehr verschieden sind. Die Wurzeln bairisch-österreichischer Gemeinsamkeiten rühren daher, dass das Deutsche und das österreichische relativ eng verwandte Sprachen sind und dem gleichen Zweig der germanischen Sprachfamilie angehören. Viele österreichische Wörter, die an sich aus dem Dialekt stammen, sind längst in österreichische Standardsprache übernommen und dort akzeptiert, gelten aber in Deutschland entweder als fremd oder dialektal (z.b. bairischer Dialekt). Die Erkennungsrate des Dialektes hängt von der Länge des Textes ab. Der Satz „Is scho recht!“ kann nicht evaluiert werden, das hier nur das Wort „scho“ als Dialektwort erkannt wird, kommt aber in mehreren Dialekten vor: bairisch, schweizerisch, schwäbisch und österreichisch, was 33% beträgt. In diesen Fällen sind beide Dialekte ohne weitere Ansätze nicht einfach trennbar. Österreichisch wurde aber zu 82% identifiziert. (23 aus 28 wurden korrekt erkannt), gefolgt vom schwäbischen Dialekt - 79 %. Mecklenburgisch wurde zu 75% richtig erkannt, Berlinerisch sowie Plattdeutsch - zu 72 %. Im Roman von Meinhold Wilhelm, nämlich „Die Bernsteinhexe Maria Schweidler“ wurden meistens alte Wortformen verwendet, die mit allen Dialekten eng verwandt sind, deshalb betrug die Erfolgsquote nur 50 %. Insgesamt lieferte der Wortbasierter Ansatz eine korrekte Erkennungsrate von 80,8 % (siehe Abbildung 4).

Auswertung №2 (ab 30 Zeichen pro Satz). Höhere Erkennungsraten können erwartet werden, wenn Sätze, die weniger als 30 Zeichen, eliminiert werden.

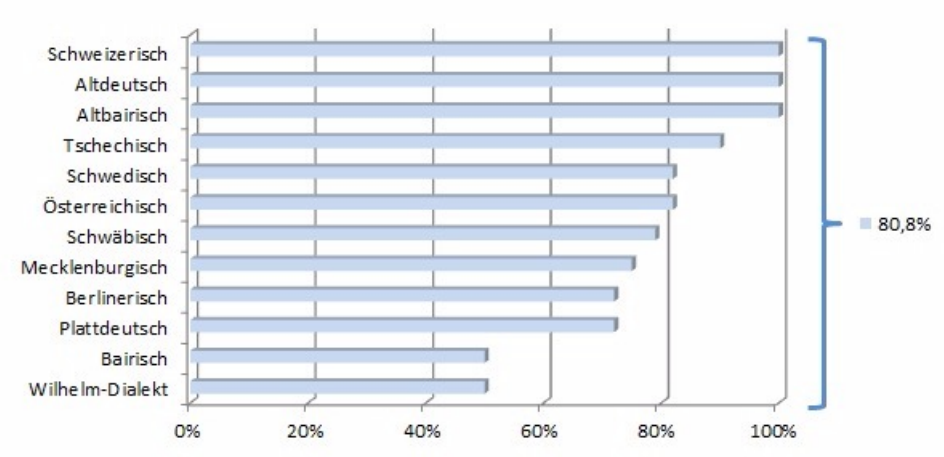

Abbildung 4: Erkennungsrate des wortbasierten Ansatzes

Es wurden etwa 60 Sätze ausgewählt, die zwischen 30 und 90 Zeichen enthalten. Eine korrekte Identifizierung des Dialektes hat zu 98% stattgefunden. Ein Satz aus 59 wurde falsch klassifiziert:
"Net lang, so geht dir"s Lichtle aus, Ond steht bei Uehrle still im Haus, Jetzt, Menschekind: waas soll dees Ganz, Oh, glaub: die Welt ischt Gaukeltanz, Ischt bunter Traum, e Schattespiel, Du Närrle, gelt?"'
Die in diesem Satz vorkommenden Wörter treten in mehreren Dialektwörterbüchern auf. Im bairischen und schwäbischen erreicht die Trefferquote 17% (6 (Anzahl der erkannten Dialektwörter) / 34 (Gesame Wörteranzhal im Satz) = 0,1764).
dir's: [schwedisch, schwäbisch, plattdeutsch, bairisch, wilhelm_dialekt, österreichisch]
waas: [schwäbisch, bairisch]
dees: [schwäbisch, bairisch]
ischt: [schwäbisch, bairisch]
e: [schwedisch, schwäbisch, bairisch, altbairisch, schweizerisch]
gelt: [schwedisch, tschechisch, schwäbisch, plattdeutsch, bairisch, mecklenburgisch, österreichisch, altbairisch, schweizerisch].

Auswertung №3 (Sätze, die nur in Hochdeutsch geschrieben wurden). In diesem Abschnitt wird Hochdeutsch auf Dialekte untersucht. In 44 aus 47 Sätzen wurde kein Dialekt gefunden. 3 Sätze konnten nicht richtig identifiziert werden, da in den Sätzen, die in Hochdeutsch geschrieben sind, kamen auch einige Dialektwörter vor. Im Satz "Auf der Straße wurde es mir ganz bitter im Mund" wurde das Wort "bitter" gefunden, das nicht nur in Hochdeutsch, sondern auch im Plattdeutsch vorkommt. Da Dialektwörterbücher aus den im Dialekt geschrieben Romanen gesammelt wurden, treten auch fremde Wort-

formen auf. Für bessere Erkennungsrate wurden Stoppwörter eliminiert. Es könnten auch Wörter entfernt werden, die keine Relevanz für die Erfassung des Dokumentinhalts besitzen: Städtenamen, gebräuchliche Vor- und Nachnamen, Zahlen oder Präpositionen. In vielen Fällen ist es nicht einfach, da eine volle Liste meistens nicht verfügbar ist.

Auswertung №4 (Sätze, die nicht nur in Hochdeutsch, sondern auch in Dialekt verfasst wurden). Zunächst wurden zwölf Hochdeutsch- und zwölf Dialektsätze zusammengemischt. Nachdem der wortbasierter Ansatz angewandt wurde, war zu sehen, dass alle Sätze richtig identifiziert wurden.

6.5.2 N-Gramm basierter Ansatz

Bigramme sowie Trigramme wurden aus den zuvor erstellten Dialektwörterbüchern mithilfe von JAVA extrahiert. Zwanzig häufigste N-Gramme pro Dialekt sind in der entsprechenden Dateien „Bigramme“, „Trigramme“ aufgelistet. Die Zahlen sind als Wahrscheinlichkeit in Prozent angegeben, wie oft BIgramm im Dialekt vorkommt. Die Erkennungsrate bei Bigramm- und Trigrammansätzen beträgt nahezu 0%. Aus 204 Sätzen wurde kein einziger Satz korrekt erkannt. 134 davon konnten nicht identifiziert werden. Kein Dialekt wurde in 21 Sätzen festgestellt. In 49 Sätzen wurden Sätze falsch klassifiziert. Obwohl N-Gramm Techniken gute Erkennungsraten für die Identifikation der Sprachen liefern, ist für die die Erkennung der Dialekte nicht zuverlässig. In der Abbildung 5 ist zu sehen, dass Trigramme “*sch*“, “*che*“ nicht nur in jedem Dialekt zu finden sind, sondern die Häufigkeit des Auftrettens übereinstimmt.
Abgesehen davon, gibt es auch unikale Trigramme, die in der Abbildung 6 aufgelistet sind. Solche Trigramme kommen aber nicht in jedem Satz vor, der auf Dialekte untersucht werden soll, deshalb kann man sich auf unikale Buchstabenkombinationen nicht verlassen. N-Gramm basierte Ansätze haben eine Erkennungsrate von nahezu 0% gezeigt bei der Erkennung der Dialekte.

6.5.3 Naive Bayes-Klassifikator

Bei 204 Sätzen verschiedener Länge zeigte der Bayes-Klassifikator eine korrekte Erkennung von 66,5%. Aus 51 Sätzen, die falsch klassifiziert wurden, wurden 56,9% nicht richtig identifiziert und etwa 43% konnten nicht erkannt werden. Eine Erkennungsrate konnte auf 92,7% gestiegen werden, indem nur Sätze mit mehr als 30 Zeichen berücksichtigt wurden. Bei der Erkennung des deutschen Hochdeutschtextes erfolgte die Zuordnung des Satzes zu

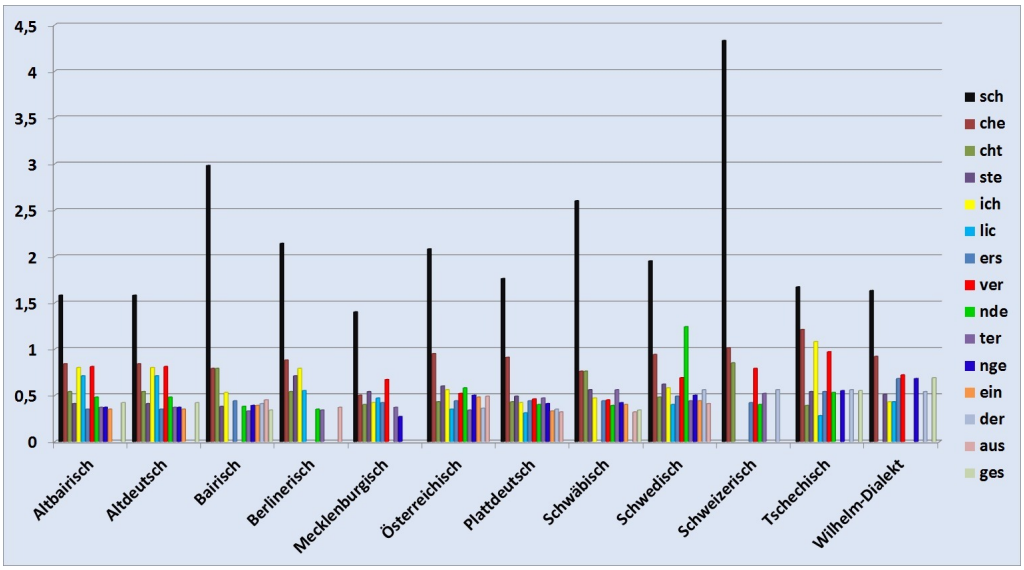

Abbildung 5: Trigramme, die in jedem Dialekt vorkommen

einem Dialekt in 35 Sätzen problematisch. In 12 Sätzen wurde falscher Dialekt zugewiesen, z.B. im Satz “Ich konnte mit meinen kleinen Beinen kaum mehr weiter.“ wurde bairischer Dialekt nicht richtig erkannt. Falsche Klassifizierung erfolgte, weil bairischer Dialekt das Wort “weiter“ enthält. Wörterbücher können nicht immer zu 100% ausgefiltert werden.

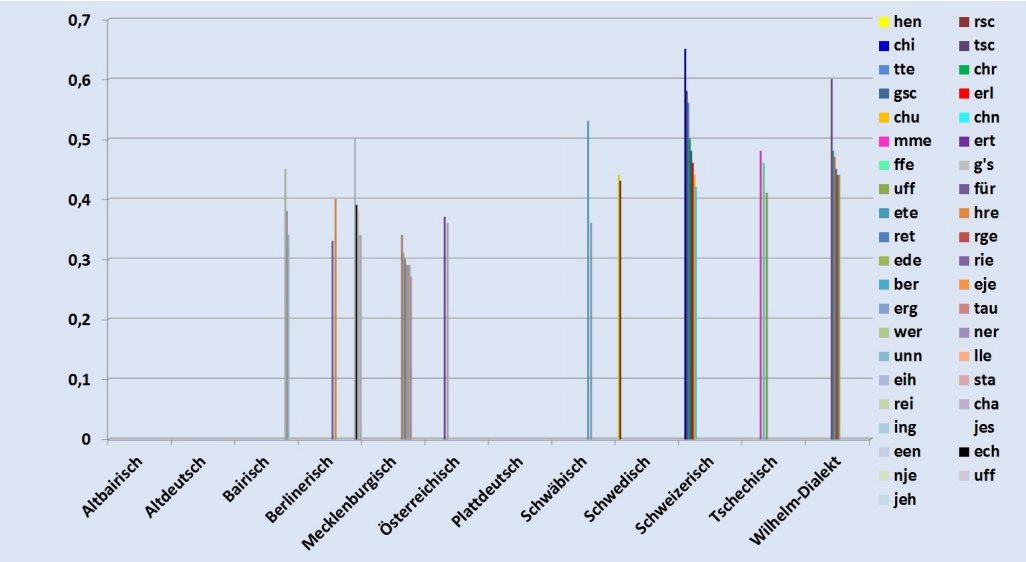

Abbildung 6: Unikale Trigramme der Dialekte

6.6 Fehleranalyse

Wörterbücher wurden mithilfe von Uima-Framework aus Romanen halbautomatisch gesammelt, was zufolge hat, dass:

1 Wörter mit Schreibfehlern wurden in den Wörterbüchern gespeichert, obwohl sie als solche nicht existieren. Das führt zum unnötigen Wachstum der Wörterbücher herbei.

2 Polysemie der Wörter, nicht erkannte Wörter gehören nicht immer zu Hochdeutsch, sondern auch zu einigen Dialekten, z.B. das Wort "Tankstelle" hat seine eigene Bedeutung im Berlinerischen: eine Gaststätte oder Kneipe.[7]

3 Wörterbücher enthalten Wörter, die entfernt werden können, da es keine Relevanz für die Erfassung des Textinhalts besteht: gebräuchliche Vor- und Nachnamen, Zahlen oder Stoppwörter.

4 Im besten Fall sollen Wörterbücher nicht nur hochfrequente Wörter, sondern auch alle deren Wortformen enthalten. Da Dialekte als Umgangssprache meistens betrachtet werden, gibt es keine standardisierten schriftlichen Sprachformen.

5 N-Gramm Ansätze sind für die Erkennung der Dialekte nicht geeignet, da Bi- und Trigramme vieler Dialekte deutscher Sprache ähnlich sind. Man könnte versuchen Dialekte nach linguistischen Kriterien näher untersuchen, um bestimmte Merkmale aufzufinden, das ist nicht im Rahmen dieser Arbeit vorgesehen.

6.7 Fazit

Die beste Erkennungsrate von 98% wurde mit dem wortbasierten Ansatz erreicht. Auf dem zweiten Platz steht Naive Bayes-Klassifikator mit 92,7%. Beide Algorithmen benötigen beim Input Sätze, deren Länge mehr als 30 Zeichen ist. Bei der kleineren Anzahl an Zeichen sinkt auch die Erkennungsrate entsprechend auf 81,8% und 66,5%. Da alle Ansätze auf der Basis der Wörterbücher aufgebaut sind, es ist theoretisch möglich, die Erkennungsrate bei der Hochdeutschklassifizierung zu erhöhen, indem alle Wörterbücher gereinigt werden und nur Dialektwortformen, keine Hochdeutschwörter, enthalten sollen.

[7] http://www.spreetaufe.de/berlinerisch-berliner-jargon/woerterbuch-berlinisch-p-z

7 Literaturverzeichniss

[Omar F. Zaidan, Chris Callison-Burch, 2012] "Arabic Dialect Identification", Association for Computational Linguistics

[Fatiha Sadat, Farnazeh Kazemi et al., 2014] "Automatic Identification of Arabic Dialects in Social Media", University of Quebec in Montreal, NLP Technologies Inc.

[Gregory Grefenstette, 1995] "Comparing Two Language Identification Schemes", Xerox Research Centre Europe

[Murat Akbacak, Dimitra Vergyri et al., 2011] "Effective Arabic Dialect Classification Using Diverse Phonotactic Models", Speech Technology and Research Laboratory, SRI International

[Penelope Sibun, Jeffrey C.Reynar, 1996] "Language Identification: Examining the Issues", The Institute for the Learning Sciences

[Michael John Martino, Robert Charles Paulsen et al., 2001] "Natural Language Determination using Natural Words", International Business Machines Corporation, NY (US)

[Clive Souter, Gavin Churcher et al.,1994] "Natural Language Identification using Corpus- Based Models", School of Computer Studies

[Clive Souter, 2015] "Natural Language Identification using Corpus-Based Models", University of Leeds

[Emmanuel Ferragne, Francois Pellegrino, 2004] "Rhythm in Read British English: Interdialect Variability", Laboratoire Dynamique Du Langage, Lyon, France

[William B.Canvar and John M.Trenkle, 1994] "N-Gram-Based Text Categorization", Environmental Research Institute of Michigan

[Daniel Jurafsky and James H. Martin, 1999] "Speech and Language Processing", ISBN

[Fadi Biadsy, Julia Hirschberg et al., 2009] "Spoken Arabic Dialect Identification Using Phonotactic Modeling", Association for Computational Linguistics

8 Anhang

Gleiche Wörter im Schwäbischen und im Bairischen

Alsdann Aranjuez Backsteinkäs Bauernbursch Bescht Bessers Bildchen Bischt Bitt Brav Bsonders Bua Bursch Bussel Darf Dees Dei Derf Dolden Dr Drin Droben Drüben Durscht End Ernscht Ernstes Erscht Erstens Flegelhaftigkeit Fleischmarken Flickschneider Frag Fratz Freid Freilein Freili Freind Freud' Fürschten G'schäft Gebirg Gehrock Geischt Gelt Geschicht Geschreibsel Gewinsel Gewölk Glei Glöcklein Glückes Gnad Gschicht Gschäft Gsell Hab' Hat's Hauptsach Hei Hemdärmel Herein Herreden Heuen Hie Hilf Hingerissen Hiobspost Hm Hohenzollern Hoscht Hu Händ Häusel Höch Höh Höll Hör' Hüh Jawoll Jeld Jemüt Jeschichte Jesses Jetz Johr Justav Kamme Kaschpar Katz Kavaliere Kirch Klass' Klauben Kneifer Kommandostimme Kommis Krankenbette Kreuzganges Krischt Köpf Köpf' Küch Leg Lenau Leut' Luschtbarkeit Ma Mach' Malör Maschin Met Mieder Minischter Mutters Möcht Mögli Mühl Naa Nach'n Nanu Nee Net Nich Numero Näh Nämlich Oha Oho Olja Pfeif Pfenning Pfui Pscht Rasch Red Reisemantel Roheit Rosel. Rosels Ruh Röck S' Sach Schand Scheltworte Schlafkammer Schlossergeselle Schläg Scho Schul Schul' Schurzfell Schöns Se Sedan Seel Seht Sell Sie's Sieh Sofort Solchene Sonnenscheine Sorg Spielkameradin Sprach Stell Sterngucker Stimm Stimmlein Straf Straß Straßenkot Strohbündeln Studiosus Stücker Säck Sünd Teifel Täg Uff Ui Un Und's Vadruß Vatter Veilchenstrauß Vergiß Verstehe Viecher Vortritt Waas Wart Wart' Weibsleute Wenn's Wenns Werd Werg Widersetzlichkeit Wie's Willscht Willst Wird's Wolkenballen Wär Zibeben Zuerscht Zung Zähn a! all's alledem alls alsdann ander andre andres arms as auf'm auf'n aufhorchend aufm aus'm ausbitten ausgschaut auße bald's begehrlich begütigte behandschuhten beschte bi bischt bissel bitt bleib' brauch' bravs bringscht by bäuerische d' dahier dahinein dee dees dei dene denk' derf derfen dern det deutli dir's do druff du's dumms där e eenen eener ei einemmal er's ergebenster erlaub' erscht erschte fahrt's fescht find freili freit fufz'g fui für'n g'macht gangen ge geb gegäben geh' gehn gehscht gel gelt geredt gereuen gesehn getruckt gfallt gfehlt gfragt ghabt gholt ghör ghört glang glei gleißenden glernt gläsernen gmacht gnade gnau gnug goldiger goldnen graad grad grad' gradaus graden gsagt gschaut gscheit gscheiter gschieht gschwind gspielt gspürt gstellt gsund guet gwohnt hab' hagern han han's has hatt hausbackene haußе hei heimgingen heit heite heiterm hexenhaft hi hinaufkommen hm ho holdselig holet hoscht hot hät hätt hättscht hör' hüstelnd i i's ich's ick ihr's imstand is isch ischt janz jar jetz jut keen keene keener kennscht keune klaube kohlschwarzen koin komm' kriagt kunnt käm lad leit liaber liabs liebreich liebs luschtige läben lär ma mach' macha man's mang mei mei'm meim mein' mer mi mir's miserable mueß mähr möcht mögli natürli nauf nee nei ner net net's netts netzte nich nischt nit no num nunter nüber o ob's obern olle ollen

ooch pfaucht pfui pommerschen ratet red richt s' sag' sagscht schad schau' scheen scheene schen schenierlich scheunt schließli schneidt scho scho. schwär schö se seh' seh'n sei'. seim selbigen sell sich's sie's siehste so'n soll' sollscht spitzige steh' stehn sähr talab tan tret tät uf uff un unsern unsre ver verbauert verlegenes vor'm vor's wat wenn's werd weuß wie's wien willscht wir's wir wirkli wo's wohi wohl. woll wolln wurd' wußt wär wär' würd wüßt z' zruck zugehn zulieb zumut zwee

Bigramme

Altbairisch	Altdeutsch	Bairisch	Mecklenburgisch	Plattdeutsch	Österreichisch
en 4.67	en 3.91	ch 3.71	en 4.35	en 4.88	er 4.16
er 3.98	ch 3.22	er 3.63	er 2.80	er 3.54	en 3.62
ge 2.93	er 3.13	en 3.24	ch 2.08	ch 2.82	ch 3.05
ch 2.44	ge 2.34	sc 2.27	te 1.76	te 2.19	te 2.14
te 2.07	te 2.12	ei 1.79	st 1.61	ge 1.83	ge 1.75
et 2.02	hr 1.41	ge 1.73	82 1.46	st 1.58	sc 1.66
he 1.48	st 1.39	te 1.42	21 1.46	he 1.48	st 1.60
li 1.40	ei 1.32	he 1.35	&# 1.46	sc 1.41	ei 1.48
re 1.38	he 1.30	st 1.29	#8 1.46	in 1.25	he 1.43
be 1.38	be 1.29	ie 1.25	ge 1.44	ei 1.25	an 1.32
ff 1.37	sc 1.24	in 1.23	17 1.28	ie 1.11	nd 1.13
le 1.29	li 1.23	au 1.17	sc 1.17	el 1.11	au 1.11
ei 1.29	et 1.15	de 1.15	re 1.12	de 1.11	in 1.07
in 1.25	de 1.01	re 1.09	in 1.07	nd 0.99	re 1.04
st 1.15	ah 0.99	an 1.04	ei 1.00	le 0.98	de 1.04
ic 1.11	ht 0.98	be 0.94	el 0.95	re 0.96	es 0.97
an 1.06	es 0.98	el 0.91	he 0.92	ne 0.95	ra 0.94
nd 1.02	nd 0.96	le 0.85	we 0.89	be 0.95	g' 0.94
de 1.02	un 0.94	se 0.83	an 0.88	es 0.92	be 0.94
sc 0.93	ig 0.94	ng 0.83	ne 0.87	se 0.86	ng 0.93

Berlinerisch	Schweizerisch	Schwäbisch	Schwedisch	Tschechisch	Wilhelm-Dialekt
en 4.72	ch 5.78	er 3.78	en 4.15	en 5.51	er 3.89
ch 3.97	sc 3.19	ch 3.69	er 3.73	er 4.14	et 3.83
je 2.95	er 2.99	en 3.35	ch 3.43	ch 3.35	en 3.79
te 2.28	li 2.57	te 2.16	te 2.00	ge 2.84	ge 3.17
er 2.17	et 1.60	sc 2.09	ge 1.96	he 1.88	te 3.13
st 1.78	le 1.59	ge 1.86	st 1.92	te 1.69	ch 2.54
sc 1.66	el 1.46	ei 1.72	de 1.92	st 1.47	re 1.81
ie 1.33	te 1.45	le 1.55	ei 1.75	re 1.43	he 1.80
ei 1.33	he 1.38	st 1.32	nd 1.72	et 1.43	st 1.45
he 1.31	ge 1.37	he 1.30	he 1.58	li 1.39	ie 1.33
ic 1.29	ht 1.27	be 1.23	sc 1.57	sc 1.25	be 1.30
re 1.12	ne 1.16	el 1.20	in 1.33	ic 1.25	sc 1.29
es 1.12	ei 1.14	in 1.18	be 1.21	nd 1.16	le 1.28
an 1.04	is 1.07	de 1.08	ie 1.16	be 1.16	ei 1.27
in 1.01	en 1.02	ht 1.02	es 1.15	ff 1.15	es 1.23
el 0.99	me 1.01	es 1.02	le 1.06	le 1.14	in 1.21
be 0.95	nd 0.98	nd 0.98	re 1.02	ei 1.11	de 1.10
le 0.91	in 0.98	re 0.97	au 1.00	an 1.10	el 1.03
et 0.90	de 0.92	au 0.95	el 0.93	de 1.08	an 0.92
ee 0.90	ts 0.91	se 0.93	ng 0.90	me 1.00	se 0.91

Verwendete Daten

Name/Autor des Romanes	Dialektanteil	Dialektname
Thoma,-Ludwig_Jozef Filsers Briefwedel	15,2%	Bairisch
Reuter,-Fritz_Ut de Franzosentid	6%	Mecklenburgisch
Reuter,-Fritz_Dörchläuchting	5,8%	Mecklenburgisch
Reuter,-Fritz_De Reis“ nah Bellingen	5%	Mecklenburgisch
Christ,-Lena_Die Rumplhanni	4,3%	Bairisch
Thoma,-Ludwig_Der Wittiber	4%	Bairisch
Graeser,-Erdmann_Lemkes sel. Wwe.	3,9%	Berlinerisch
Christ,-Lena_Madam Bäuerin	3,8%	Bairisch
Christ,-Lena_Madam Bäurin	3,8%	Bairisch
Wille,-Bruno_Glasberg	3,6%	Schwäbisch
Thoma,-Ludwig_Der Ruepp	3,4%	Bairisch
Ganghofer,-Ludwig_Der Jäger von Fall	3,3%	Bairisch
Meinhold,-Wilhelm_Die Bernsteinhexe Maria Schweidler	3,3%	Schlesisch
Thoma,-Ludwig_Altaich	3,2%	Bairisch
Thoma,-Ludwig_Satiren	3,2%	Bairisch

Name/Autor des Romanes	Dialektanteil	Dialektname
Haemmerli,-Marti-Sophie_Mis Chindli	15,7%	Schweizerisch
Zesen,-Philipp-von_Adriatische Rosemund	9,9%	Altdeutsch
Andreae,-Johann-Valentin_Die chymische Hochzeit	8,5%	Altbairisch
Wickram,-Georg_Der jungen Knaben Spiegel	6,6%	Tschechisch
Wickram,-Georg_Von guten und bösen Nachbarn	4,9%	Tschechisch
Wickram,-Georg_Der Goldtfaden	4,7%	Tschechisch
Stehr,-Hermann_Leonore Griebel	3,4%	Schwedisch
Wickram,-Georg_Gabriotto und Reinhart	3,4%	Tschechisch
Adolph,-Karl_Haus Nummer 37	3,2%	Österreichisch
Speckmann,-Diedrich_Heidehof Lohe	3,2%	Plattdeutsch
Seidel,-Heinrich_Reinhard Flemmings Abenteuer zu Wasser	3,9%	Plattdeutsch

Erklärung

Hiermit versichere ich, dass ich meine Abschlussarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Datum: .....

(Unterschrift)