Universität Karlsruhe
Fakultät für Informatik
Institut für Logik, Komplexität und Deduktionssysteme
Prof. A. Waibel

SS 2003

Studienarbeit

Rekombination von Komposita in der Spracherkennung

Some German words are so long that they have a perspective. [...] These things are not words, they are alphabetical processions. [...] So he [a student of the German language] resorts to the dictionary for help, but there is no help there. The dictionary must draw the line somewhere - so it leaves this sort of words out. And it is right, because these long things are hardly legitimate words, but are rather combinations of words, and the inventor of them ought to have been killed.

Mark Twain in „The awful German language" 1880

Almut Silja Hildebrand

August 2003

Betreuer: Dr. I. Rogina

Inhaltsverzeichnis

1 Einleitung

Die deutsche Sprache weist im Gegensatz zu vielen anderen Sprachen der Welt zahlreiche Besonderheiten in der Rechtschreibung und im Umgang mit einzelnen Wörtern auf, die sich für die Spracherkennung als Probleme erweisen können. Neben der Groß- und Kleinschreibung und der komplexen Morphologie, die uns hier nicht weiter beschäftigen, findet man mit den sogenannten Komposita eine weitere besondere deutsche Spracheigenschaft.

Ein Kompositum ist ein, aus mehreren Einzelwörtern bestehendes, zusammengesetztes Wort. Am häufigsten bestehen Komposita aus zwei zusammen gesetzten Substantiven „Bundes*republik*", aber man findet auch freie Kombinationen von von Substantiven, Adjektiven und Verben (rätselraten, Blauschimmel, Fahrradreparatur, Reparaturladen usw.), die sich „endlos" miteinander kombinieren lassen.

Gerade dieses „endlos" erscheinende Zusammensetzen von Wörtern stellt die Spracherkennung vor die entscheidende Schwierigkeit in der Erkennung von Komposita. Der Wortschatz ist dadurch deutlich größer als der einer Sprache ohne Komposita, und die einzelnen Komposita kommen in Texten sehr selten vor, sodass sie nicht robust trainiert werden können.

1.1 Komposita in der Spracherkennung

In der Spracherkennung wird traditionell alles zwischen zwei Leerzeichen als Wort betrachtet. Ein Kompositum ist zwar aus mehreren Wörtern zusammen gesetzt, aber man müsste es wie ein eigenständiges Wort behandeln und es in den Wortschatz des Spracherkenners aufnehmen.

Da es durch die freie Kombinierbarkeit theoretisch annähernd unendlich viele Komposita geben kann, ist dies in der Praxis nicht möglich. Die in Texten vorkommenden Komposita sprengen den Rahmen eines praktikablen Erkennerwortschatzes. Aus diesem Grund empfiehlt es sich, nur die gängigsten Komposita, wie z.B. „Bundesrepublik" in den Wortschatz aufzunehmen und für die restlichen Komposita einen anderen Ansatz zur Erkennung zu wählen. Auf die gleiche Weise wird auch bei den heutigen deutschen Wörterbüchern verfahren.

- 5,0% Wortfehlerrate (755 Fehler)
- die meisten Kompositumfehler sind einfach reparabel[3] (ca. 350)

1.4 Überblick

In dieser Arbeit werden zwei unterschiedliche Herangehensweisen an die Problematik vorgestellt.

Kapitel 2 beschreibt einen Lösungsansatz, bei dem versucht wird, Merkmale dafür zu finden, dass zwei Wörter im Text eigentlich ein Kompositum bilden und zusammen geschrieben sein sollten. Es wird aus einer Statistik über erste und zweite Teile von Komposita in deutschen Texten eine Wahrscheinlichkeit für das Zusammensetzen von zwei Wörtern berechnet. Diese Lösung führte jedoch noch nicht zu den erwünschten Erfolgen, weshalb versucht wurde durch eine andere Methode ein besseres Ergebnis zu erreichen.

Kapitel 3 befasst sich mit einer umgekehrten Strategie und versucht über ein Ausschlussverfahren zu einer Lösung zu kommen. Es werden hier Merkmale herausgearbeitet, die gegen das Zusammensetzen von zwei Wörtern sprechen. Durch die Anwendung mehrerer Ausschlusskriterien konnte die Wortfehlerrate im Erkenneroutput deutlich verringert werden.

Eine weitere Verbesserung des Ergebnisses konnte durch die Kombination der beiden Verfahren erzielt werden.

[3] die Komponenten ergeben direkt, ohne weitere Bearbeitung zusammen gesetzt, das korrekte Wort

2 Ansatz über Wahrscheinlichkeiten für Kompositumkomponenten

Eine Möglichkeit zu entscheiden, ob zwei Wörter Teile eines Kompositums sind, ist anzugeben, wie wahrscheinlich jedes Wort einen ersten oder zweiten Teil eines Kompositums bilden kann. Dabei spielt natürlich auch die Reihenfolge, in der die Wörter als Paar auftreten, eine Rolle. So bildet das Wortpaar „Religions Zugehörigkeit" ein Kompositum während „Zugehörigkeit Religions" wohl nie zusammen gesetzt würde.

Diese Wahrscheinlichkeit berechnet sich somit daraus, ob und wie oft das jeweilige Wort als erster oder zweiter Teil eines Kompositums vorkommt.

2.1 Erstellen von Wortlisten

In diesem Verfahren wird für jedes gegebene Wort in einem Lexikon nachgeschlagen, ob und wie wahrscheinlich es eine erste oder eine zweite Kompositumkomponente sein kann.

Es wird jeweils ein Lexikon für erste und für zweite Teile von Komposita verwendet. Um diese beiden Lexika aufzubauen, werden in einer großen Textbasis OOV-Komposita nach der selben Methode zerlegt, wie für die Trainingstranskripte (siehe Abschnitt 1.1). Dann werden die Teile in das jeweilige Lexikon aufgenommen.

Die Wahrscheinlichkeit dafür, dass ein Wort z.B. der erste Teil eines Kompositums ist, berechnet sich daraus, wie oft es als solcher vorkam, unabhängig davon, mit welchem anderen Wort es kombiniert war.

2.2 Zusammensetzen der einzelnen Wörter zu Komposita

In der Erkennerhypothese wird nun jedes Wortpaar überprüft, und entschieden, ob es zu einem Kompositum zusammengesetzt werden sollte. Wenn die kombinierte Wahrscheinlichkeit dafür, dass das erste Wort w_1 ein erster Teil und das zweite Wort w_2 ein zweiter Teil eines Kompositums ist, einen bestimmten Schwellwert t übersteigt, werden sie zusammen gesetzt.

$$P_1(w_1) \cdot P_2(w_2) \quad > \quad t \quad \Longrightarrow \quad \text{zusammen setzen}$$

Dies ermöglicht auch die Kombination von Wörtern, die vorher noch nicht zusammen vorgekommen sind. Das ist wünschenswert, da es im Deutschen durchaus üblich ist, ständig neue Komposita zu erfinden.

2.3 Ergebnisse

Die Anwendung dieses Verfahrens ohne Schwellwert ($t = 0$) hat, anstatt die Wortfehlerrate zu senken, zu einem starken Anstieg der Wortfehlerrate geführt. Es wurden extrem viele Wörter fälschlicherweise zusammen gesetzt. Die Anzahl der zusammen gesetzten Wörter war ca. zehnmal so hoch wie die geschätzte Anzahl von Kompositumfehlern im Text (2150 Fehler aber 20683 zusammen gesetzte Wortpaare).

Eine leichte Verbesserung in der Wortfehlerrate im Vergleich zur unbearbeiteten Erkennerhypothese, konnte erst bei einem sehr hohen Schwellwert erreicht werden, wobei dann nur noch sehr wenige Wortpaare zusammengesetzt wurden. Die Anzahl der **z**usammen gesetzten **W**ort**p**aare (ZWP) lag dann nur bei etwa einem Zehntel der geschätzten Anzahl von Kompositumfehlern im Text. Dies war bei beiden Testdatensätzen gleichermaßen zu beobachten.

Der Schwellwert $3,95 \cdot 10^{-13}$ in den Tabellen 2.1 und 2.2 bedeutet, dass ungefähr ein Drittel der Wörter in den Listen von ersten und zweiten Teilen ausgeschlossen wird. Bei einem Schwellwert von $2,02 \cdot 10^{-11}$ werden ca. zwei Drittel der Wörter nicht mehr zum Zusammensetzen verwendet.

Schwellwert	ZWP	Wortfehlerrate
0 (keiner)	20683	44,2%
$3,95 \cdot 10^{-13}$	20040	43,3%
$2,02 \cdot 10^{-11}$	15620	37,5%
$5,00 \cdot 10^{-8}$	2219	22,7%
$5,00 \cdot 10^{-7}$	454	22,3%
$9,00 \cdot 10^{-7}$	201	22,2%
$5,00 \cdot 10^{-6}$	4	22,3%
1 (unbearbeitet)	0	22,3%

Tabelle 2.1: Erkennerhypothese: Ergebnisse mit verschiedenen Schwellwerten

Schwellwert	ZWP	Wortfehlerrate
0 (keiner)	2380	29,4%
$3,95 \cdot 10^{-13}$	2309	28,4%
$2,02 \cdot 10^{-11}$	1850	22,3%
$5,00 \cdot 10^{-8}$	366	5,9%
$5,00 \cdot 10^{-7}$	70	4,9%
$9,00 \cdot 10^{-7}$	34	4,7%
$5,00 \cdot 10^{-6}$	2	5,0%
1 (unbearbeitet)	0	5,0%

Tabelle 2.2: Transkript: Ergebnisse mit verschiedenen Schwellwerten

Offensichtlich ist es in der deutschen Sprache möglich, fast jedes Wort auch als Teil eines Kompositums zu verwenden. Dies führt zu dem Ergebnis, dass das Vorkommen der einzelnen Wörter als Teile von Komposita kein ausreichendes Kriterium für das Zusammensetzen eines Wortpaares ist.

2.4 Vergleich: Lösungsansatz von Berton

Das Aufspalten von Komposita ist bei deutschen Spracherkennungssystemen weit verbreitet. Beim wieder Zusammensetzen der Komponenten wurden aber bisher keine großen Erfolge erzielt.

In [BFRB96] wird beschrieben, wie sich das Aufteilen von Komposita auf Wortschatz, Perplexität und Erkennungsleistung eines Spracherkenners auswirkt. Die Vokabulargröße sank nach dem Aufteilen um c.a. 24%, die OOV-Rate um 30%. Beim wieder Zusammensetzen wurde dasselbe Verfahren mit Wahrscheinlichkeiten für erste und zweite Teile von Komposita verwendet, mit denselben schlechten Ergebnissen: eine Verschlechterung der Erkennungsrate um 5%. Dabei konnte ca. ein

3 Ansatz über Ausschlusskriterien für Nichtkomposita

Da Merkmale für das Zusammensetzen eines gegebenen Wortpaares offensichtlich nicht aussagekräftig waren, beschäftigt sich dieser Lösungsansatz mit Merkmalen, die gegen das Zusammensetzen sprechen.

3.1 Analyse einer Stichprobe

Diese Lösungsstrategie stützt sich auf eine genauere Betrachtung der in einer Stichprobe vorkommenden Komposita. Die Stichprobe bestand aus 500 Zeilen Erkennerhypothesen mit einer Gesamtwortfehlerrate von 21,8% bestehend aus 3,6% getrennter erkannter Komposita und 18,7% sonstiger Fehler.

Es hat sich gezeigt, dass sich die meisten der geteilt erkannten und reparablen Komposita aus zwei Hauptwörtern zusammensetzen (über 95%). Es ist in der deutschen Sprache zwar auch möglich, Adjektive und Verben aus mehreren Wörtern zu bilden (z.B. „weiterbilden" oder „blaugrün"), aber dies ist selten. Seit der Rechtschreibreform werden auch viele der bisher zusammen gesetzten Verben getrennt geschrieben, z.B. Verbindungen von Verb + Verb: *kennen lernen*, Adverb + Verb: *zusammen setzen*, von Adjektiv + Verb: *lästig fallen*, von Substantiv + Verb: *Rad fahren*.

Dass ein zusammen gesetztes Adjektiv oder Verb geteilt erkannt wird, kommt so selten vor, dass es sich nicht lohnt, sich mit dieser Ausnahme zu befassen.

Schaut man sich nun an, wann in der Erkennerhypothese zwei Hauptwörter hintereinander stehen, so findet man Paare, die zusammen geschrieben sein sollten (Komposita: 40%, 30% zusammen gesetzte Hauptwörter und 10% Bindestrichwörter), Paare, die korrekt getrennt stehen bleiben müssen (36%), und Paare, die nur durch Erkennungsfehler entstanden sind - im korrekten Text gibt es hier keine zwei Hauptwörter (24%)(siehe Abbildung 3.1).

Abbildung 3.1: Anteile aller Hauptwortpaare in der Stichprobe

Da sich keine guten Kriterien dafür finden ließen, dass ein Wortpaar zu einem Kompositum zusammengesetzt werden sollte, lag eine Analyse der Hauptwortpaare, die korrekt getrennt geschrieben sind, nahe. Außer dem Ausschluss von Verben und Adjektiven finden sich weitere Kriterien dafür, dass ein Wortpaar nicht zusammen gesetzt werden sollte.

In zwei Dritteln der Fälle ist eines der beiden Hauptwörter ein Eigenname wie z.B. „Herr Maier“ oder „Polizeidirektion Hamburg“ - aus Eigennamen werden fast nie Komposita gebildet. Am zweithäufigsten kommen zwei Hauptwörter hintereinander in Aufzählungen vor, wie z.B. „Die Antiquitäten waren zum großen Teil aus Kirchen Schlössern und Museen gestohlen worden“. Nicht zusammen geschrieben werden außerdem Abkürzungen („Telekom AG“) und sehr lange Wörter.

Es ließ sich nur für 17% der getrennt zu belassenden Hauptwortpaare (6% von allen Hauptwortpaaren) kein einfaches Ausschlusskriterium finden (siehe Abbildung 3.1).

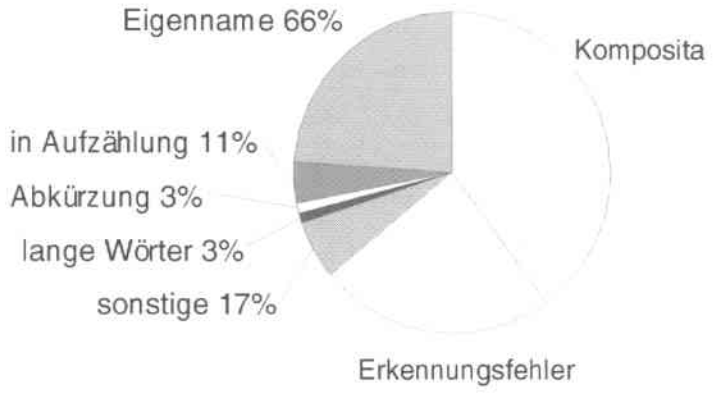

Abbildung 3.2: Hauptwortpaare: Getrenntschreibung korrekt

3.2 Hauptwörter

Um herauszufinden, welches Wort ein Hauptwort ist, liegt es erst einmal nahe, in einem Wörterbuch nachzuschlagen. Aber leider gibt es im Deutschen einige Wörter, die nicht nur Hauptwort sein können, wie z.B. „weg" und „Weg".

Daher muss man aus Statistiken und dem Kontext schließen, ob ein Wort an dieser Stelle ein Hauptwort ist. Genau das macht der Spracherkenner im Deutschen, wenn er entscheidet, ob er ein erkanntes Wort groß schreiben soll. Da dies sehr gut funktioniert, wird hier die Entscheidung des Spracherkenners übernommen, und jedes Wort, das groß geschrieben ist, wird als Hauptwort betrachtet.

3.3 Eigennamen

Eigennamen zu identifizieren ist eine komplexe Aufgabe, die den Rahmen dieser Arbeit gesprengt hätte. Einfache Ansätze, wie z.B. ein Telefonbuch zum Nachschlagen von Namen zu verwenden, ist nicht sinnvoll, da es sehr viele normale Wörter auch als Nachnamen gibt, wie z.B. „Wurst", „Schlecht" oder „Baum". Nur die häufigsten Namen aus dem Telefonbuch zu verwenden, löst dieses Problem auch nicht wirklich, weil dann in Texten oft vorkommende ausgefallene Namen wie z.B. von Politikern nicht erfasst werden. Zusätzlich werden außer Personennamen auch im Telefonbuch nicht vorkommende Namen von Firmen, Organisationen, Straßen und Orten nicht zu Komposita zusammengesetzt.

Eine Lösung für diese Aufgabe ist z.B. ein *named entity tagger*[1].

Da für diese Arbeit kein *named entity tagger* zur Verfügung stand, wurde ein von Hand erstelltes Namenslexikon verwendet (ohne Kenntnis der Testdaten). Ein Test mit einem um die Hälfte verkleinerten Namenslexikon hat gezeigt, dass die Abhängigkeit von diesem sehr guten Namenslexikon nicht sehr groß ist. Die Wortfehlerrate stieg nur im Bereich der zweiten Stelle nach dem Komma um eins.

3.4 Aufzählungen

In Aufzählungen stehen mehrere Hauptwörter (HW) hintereinander, wenn drei oder mehr gleichartige Dinge genannt werden, wie z.B. „Leinen war begehrt als das wichtigste Material zur Herstellung von Kleidung Bettwäsche Handtüchern und Tischdecken" (Der Spracherkenner setzt keinerlei Satzzeichen). Aufzählungen sind also meist der Form:

$HW \quad HW \quad \{HW\}^* \quad (\text{„und"} \parallel \text{„oder"}) \quad HW$

[1] Automatischer Markierer für Eigennamen

Es kann vorkommen, dass bei manchen der Hauptwörter Adjektive stehen („die Karte verzeichnet alle Straßen Berge Seen und größeren Flüsse"). In dieser Arbeit werden nur Aufzählungen der oben genannten einfachen Form erkannt.

3.5 Bindestriche und Abkürzungen

Es gibt im Deutschen nur sehr wenige Regeln, wann und wo ein Bindestrich gesetzt werden muss oder darf. Besonders seit der Rechtschreibreform ist man als Autor da sehr frei. In aktuellen Zeitungen sowie in den Transkripten der Nachrichtentexte, die dieser Arbeit zu Grunde liegen, werden Bindestriche fast nur dort verwendet, wo eine Abkürzung und ein Hauptwort zu einem Kompositum zusammengesetzt werden sollten, wie z.B. „SPD-Vorsitzender", „US-Truppen" oder „EU-Kommission".

Kommt eine Abkürzung als erstes Wort eines Wortpaares vor, wird also ein Bindestrich gesetzt, ist das zweite Wort eine Abkürzung (*Telekom AG*), so wird nicht zusammengesetzt.

3.6 Wortlängenbegrenzung

In deutschen Alltagstexten kommen normalerweise keine zu langen Wörter vor. Wortungetüme wie das legendäre „Donaudampfschiffskapitänsmützenbandende" sind die Ausnahme und werden zugunsten der Lesbarkeit von keinem vernünftigen Autor verwendet. Daher werden Wörter ausgeschlossen, die selbst eine bestimmte Länge (s_1) überschreiten, sowie Wortpaare, deren kombinierte Länge (s_2) zu groß ist.

Die Optimierung auf einer Kreuzvalidierungsmenge ergab die Werte $s_1 = 17$ und $s_2 = 26$.

3.7 Entscheidungsdiagramm

Der Algorithmus arbeitet die oben genannten Ausschlusskriterien als harte Entscheidungen hintereinander ab. Ein Wortpaar W_1W_2 wird nur dann zusammen gesetzt, wenn es auf keines der Ausschlusskriterien passt (siehe Abbildung 3.3).

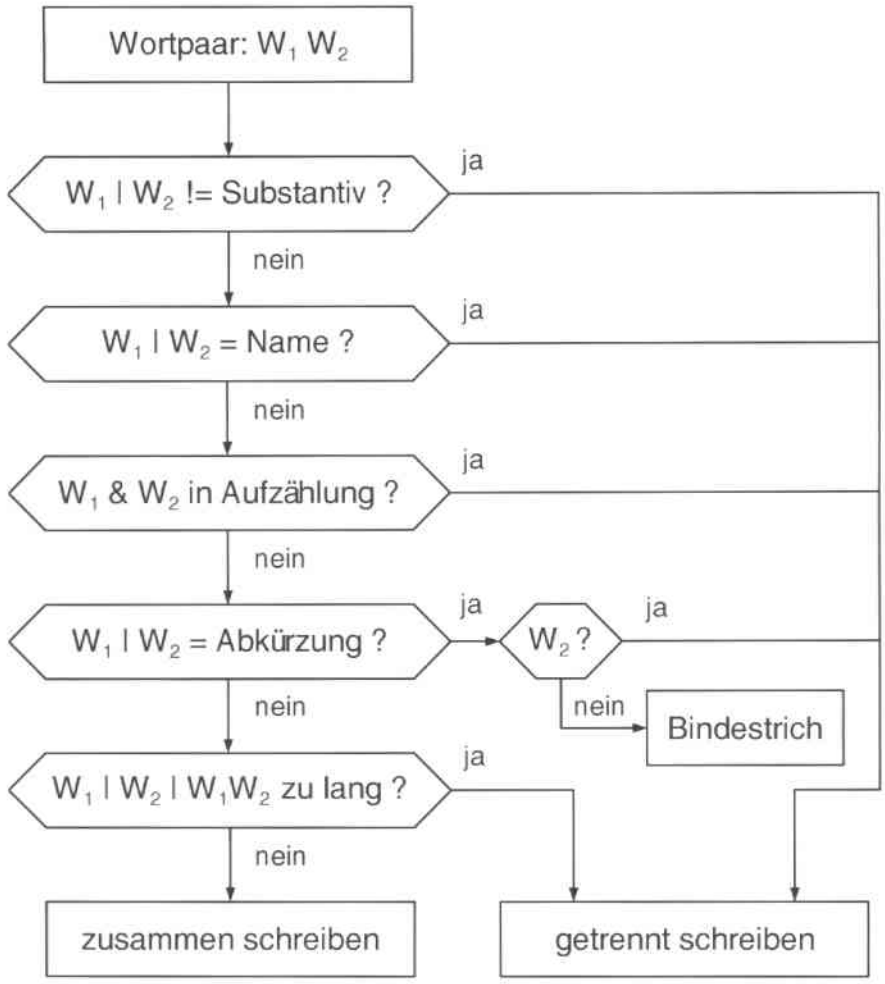

Abbildung 3.3: Entscheidungsdiagramm

3.8 Ergebnisse

Die Anwendung der verschiedenen Ausschlusskriterien nacheinander zeigt eine deutliche Verbesserung der Wortfehlerrate bei beiden Testdatensätzen.

Schon nur ein Ausschluss von Nicht-Hauptwörtern und Eigennamen und anschließendem Zusammensetzen aller restlichen Hauptwortpaare ergab eine deutliche Verbesserung der Wortfehlerrate, beim Transkript-Testdatensatz sogar fast um die Hälfte.

Die Tabellen 3.1 und 3.2 zeigen, wieviel Verbesserung jedes Ausschlusskriterium bringt, das zusätzlich angewendet wird. Nach Anwendung aller Ausschlusskriterien gelingt beim Hypothesen-Testdatensatz eine Reduktion der Kompositumfehler um 40%, wenn man den geschätzten Anteil von 3,6% Kompositumfehler zu Grunde legt, bei den Transkripten werden die Fehler sogar um zwei Drittel reduziert.

Die Anzahl der **z**usammen gesetzten **W**ort**p**aare (ZWP) liegt noch leicht über der geschätzten Anzahl von Kompositumfehlern im Text.

Kriterien (inkrementell)	ZWP	Wortfehlerrate
ursprünglich	0	22,30%
Namen	3125	21,21%
Abkürzungen und Bindestriche	3098	21,01%
Aufzählungen	2989	20,97%
Längenbegrenzung	2851	20,88%

Tabelle 3.1: Erkennerhypothese: Ergebnisse mit verschiedenen Ausschlusskriterien

Kriterien (inkrementell)	ZWP	Wortfehlerrate
ursprünglich	0	5,00%
Namen	428	2,65%
Abkürzungen und Bindestriche	410	2,08%
Aufzählungen	388	1,92%
Längenbegrenzung	365	1,70%

Tabelle 3.2: Transkript: Ergebnisse mit verschiedenen Ausschlusskriterien

3.9 Kombination beider Verfahren und Endergebnisse

Nachdem hier Wortklassen ausgeschlossen wurden, die nie zum Bilden von Komposita verwendet werden, ist es nun sinnvoll die Idee aus Kapitel 2 noch einmal aufzugreifen. Es werden jetzt, nach dem Ausschlussverfahren, zusätzlich nur Wörter zusammen gesetzt, die als erster oder zweiter Teil eines Kompositums möglich sind, d.h. in den beiden Lexika der ersten und zweiten Teile von Komposita vorkommen (siehe Abbildung 3.4).

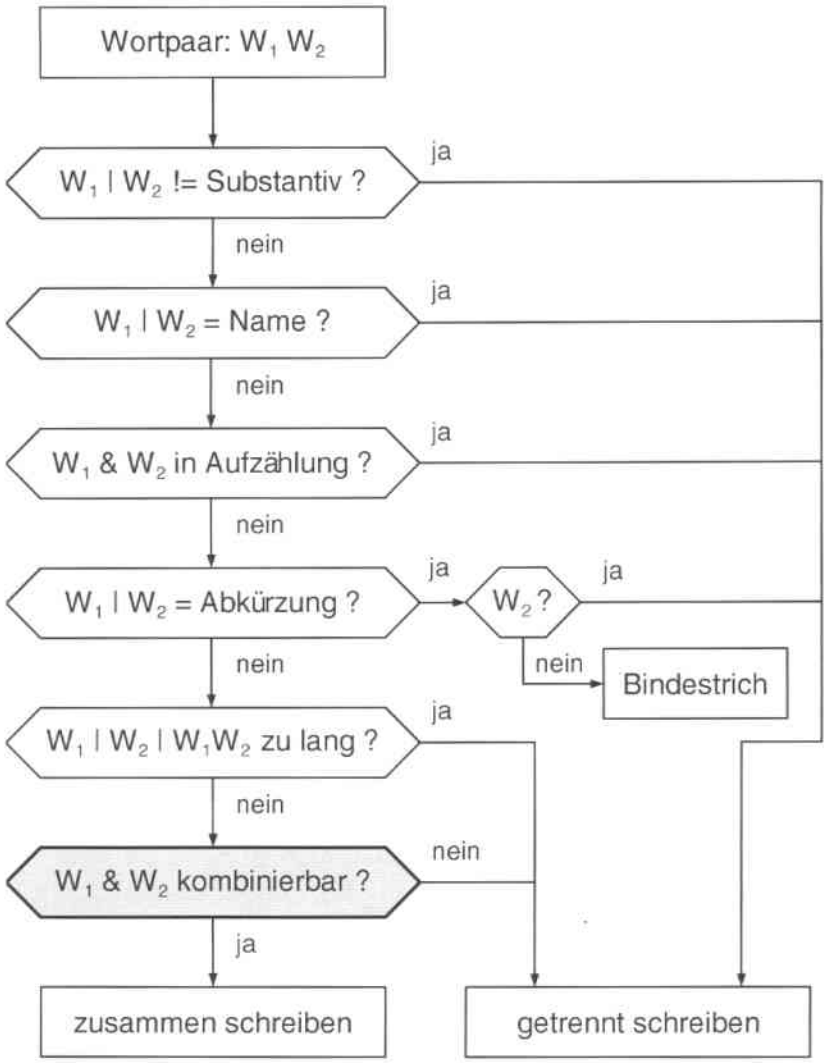

Abbildung 3.4: erweitertes Entscheidungsdiagramm

Selbst wenn alle Wörter aus den beiden Lexika zum Zusammensetzen zugelassen werden, also bei Schwellwert $t = 0$, bringt dies schon eine Verbesserung der Wortfehlerrate von 20,88% auf 20,75% bei den Erkennerypothesen und von 1,7% auf 1,54% bei den Transkripten (Abbildungen 3.5 und 3.6).

Tests auf den „Hypothesen"-Daten mit verschiedenen Schwellwerten zeigen, dass jetzt sowohl die Anzahl der **z**usammengesetzten **W**ort**p**aare (ZWP) als auch die Wortfehlerrate deutlich stabiler auf die Änderung des Schwellwertes reagieren, also im Vergleich zu Tabelle 2.1 nicht mehr so stark schwanken (siehe Tabelle 3.3). Der auf einer Kreuzvalidierungsmenge optimierte Schwellwert liegt jetzt auch wesentlich niedriger, es werden also mehr verschiedene Wörter aus den beiden Lexika zum Zusammensetzen verwendet.

Schwellwert	ZWP	Wortfehlerrate
0 (keiner)	2363	20,75%
$3,95 \cdot 10^{-13}$	2355	20,74%
$2,02 \cdot 10^{-11}$	2252	20,69%
$1,00 \cdot 10^{-10}$	2151	20,67%
$6,00 \cdot 10^{-10}$	1973	20,64%
$5,00 \cdot 10^{-9}$	1635	20,72%
1 (unbearbeitet)	0	22,30%

Tabelle 3.3: Erkennerhypothese: Ergebnisse mit verschiedenen Schwellwerten nach Ausschlussverfahren

Es werden in beiden Testdatensätzen jetzt weniger Wortpaare zusammen gesetzt, als die geschätzte Anzahl an Kompositumfehlern beträgt (jeweils ca. 90%), während die resultierende Wortfehlerrate weiter sinkt. Dies lässt den Schluss zu, dass nur noch wenige Wörter fälschlicherweise zusammen gesetzt werden.

Beim Hypothesen-Testdatensatz Verbesserte sich die Wortfehlerrate insgesamt um absolut 1,66%, was knapp die Hälfte des theoretisch Möglichen ausmacht (siehe Abbildung 3.5).

Besonders gute Ergebnisse konnten auf den Testdaten mit künstlich erzeugten Kompositumfehlern (Transkript) erzielt werden. Es wurden noch 320 Wortpaare zusammen gesetzt (bei geschätzt 350 geteilten Komposita im Text), wobei die Wortfehlerrate insgesamt um 71% gesenkt werden konnte (siehe Abbildung 3.6).

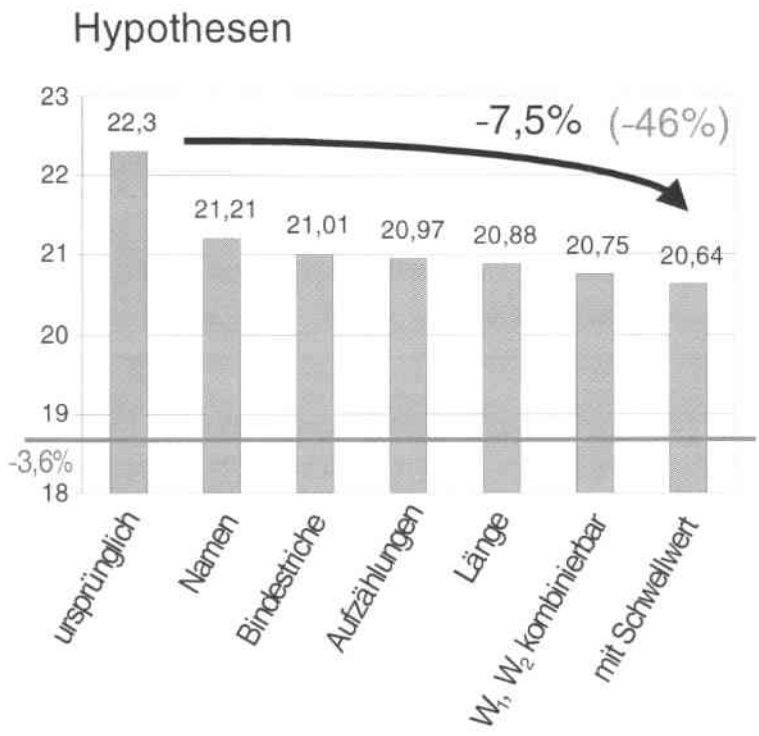

Abbildung 3.5: Hypothesen: Reduktion der Wortfehlerrate (in %)

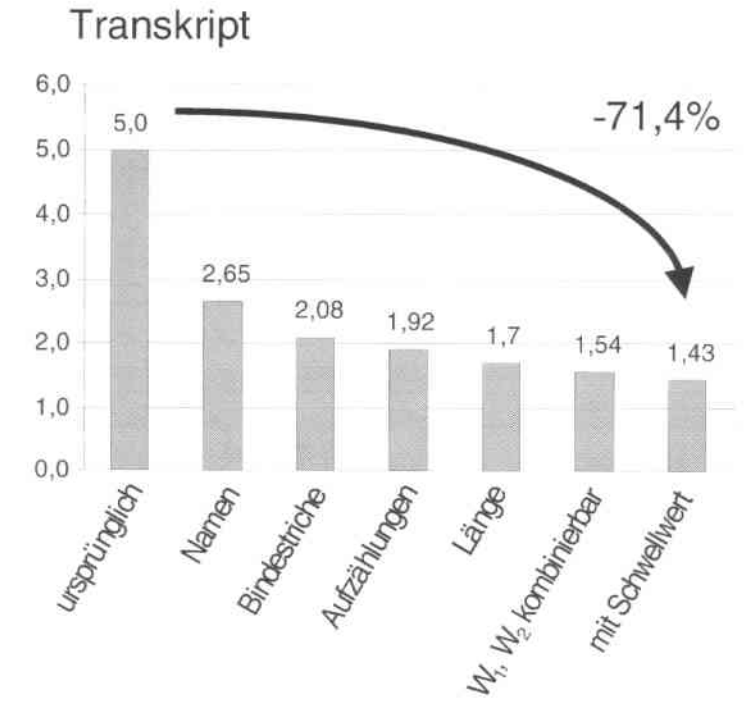

Abbildung 3.6: Transkript: Reduktion der Wortfehlerrate (in %)

4 Zusammenfassung

Die deutsche Sprache verursacht mit ihren Besonderheiten in der Spracherkennung einige Schwierigkeiten, so auch bei der Erkennung von Komposita. Komposita können gut erkannt werden, wenn man sie nicht als Einheit betrachtet, sondern ihr Bestandteile eigenständig erkennt. Dies lässt die Fragestellung offen, wie man die Teile nach der Erkennung wieder zusammen setzen kann.

In dieser Arbeit wurde deutlich, dass statistische Daten darüber, welche Wörter in deutschen Texten wie häufig zum Bilden von Komposita verwendet werden nicht genügend Anhaltspunkte dafür geben, welche Wörter im Spracherkenneroutput zusammen geschrieben werden müssen. Es können offenbar in der deutschen Sprache sehr viele Wörter als Komponenten von Komposita verwendet werden, was dazu führt, dass ein Algorithmus, der nur eine solche Statistik zur Verfügung hat, zu viele Wörter zusammen setzt.

Sehr gute Ergebnisse wurden dagegen in dieser Arbeit mit einer umgekehrten Strategie erreicht. Hierbei wurden auf Grund von Heuristiken gefundene Wortklassen von der Bildung von Komposita ausgeschlossen. Rund 50% - 70% der Kompositumfehler, je nach Testdaten, konnten so repariert werden.

Die Ausschlusskriterien werden in diesem Algorithmus alle als harte „ja oder nein"-Entscheidungen angewandt. Man könnte in Zukunft für die Klassifizierung von Namen, Aufzählungen, Abkürzungen usw. Wahrscheinlichkeiten berechnen und die Entscheidung über den Ausschluss eines Wortpaares aufgrund einer kombinierten gewichteten Wahrscheinlichkeit treffen.

Es könnten sich eventuell auch noch weitere grammatische Konstrukte wie die Aufzählung finden lassen, in denen mehrere Hauptwörter hintereinander vorkommen, um weitere Ausschlusskriterien zu definieren. Oder solche Konstrukte, in denen nie zwei Hauptwörter hintereinander stehen können, sodass klar ist dass hier zusammen gesetzt werden muss.

Literaturverzeichnis

[BFRB96] Berton, A., P. Fetter und P. Regel-Brietzmann: *Compound Words in Large-Vocabulary German Speech Recognition Systems.* 1996.