NETWORX

DIE ONLINE-SCHRIFTENREIHE DES PROJEKTS SPRACHE@WEB

André Kramer

Rechtschreibkorrektursysteme im Vergleich

DITECT versus Microsoft Word

2004

Nr. 35

@ websprache
∞ werbesprache
handysprache
Σ medienanalyse

IMPRESSUM

NETWORX ist die Online-Schriftenreihe des Projekts *sprache@web*. Die Reihe ist eine eingetragene Publikation beim Nationalen ISSN-Zentrum der Deutschen Bibliothek in Frankfurt am Main.

ISSN
1619-1021

Herausgeber
Jens Runkehl, Prof. Dr. Peter Schlobinski und Torsten Siever

Wissenschaftlicher Beirat
Prof. Dr. Jannis Androutsopoulos (Universität Hannover), für den Bereich web**sprache** & medien**analyse**.
Prof. Dr. Christa Dürscheid (Universität Zürich), für den Bereich handy**sprache**.
Prof. Dr. Nina Janich (Universität Darmstadt), für den Bereich werbe**sprache**.
Prof. Dr. Ulrich Schmitz (Universität Essen), für den Bereich web**sprache**.

Anschrift
Projekt sprache@web
Universität Hannover
Königsworther Platz 1, PF 44
30167 Hannover
Internet:
www.mediensprache.net
E-Mail:
info@mediensprache.net

Einsendung von Manuskripten
Beiträge und Mitteilungen sind an die folgende E-Mail-Adresse zu richten:
networx@mediensprache.net

Hinweis zur Manuskripteinsendung
Mit der Annahme des Manuskripts zur Veröffentlichung in der Schriftenreihe Networx räumt der Autor dem Projekt sprache@web das zeitlich, räumlich und inhaltlich unbeschränkte Nutzungsrecht ein. Dieses beinhaltet das Recht der Nutzung und Wiedergabe im In- und Ausland in gedruckter und elektronischer Form sowie die Befugnis, Dritten die Wiedergabe und Speicherung dieses Werkes zu gestatten. Unverlangt eingehende Manuskripte und Bücher werden nicht zurückgesandt.

Begutachtung
Die Begutachtung eingesandter Beiträge wird von den Herausgebern sowie den Vertretern des wissenschaftlichen Beirats vorgenommen.

Informationsstand
01. Januar 2003

ZU DIESER ARBEIT

Autor & Titel
André Kramer: Rechtschreibkorrektursysteme im Vergleich. DITECT versus Microsoft Word.
Version
1.0

Bibliografische Aufnahme
André Kramer (2004): Rechtschreibkorrektursysteme im Vergleich. DITECT versus Microsoft <http://www.mediensprache.net/networx/networx-35.pdf>. In: Net-worx. Nr. 35. ISSN: 1619-1021.

Zitiert nach Runkehl, Jens & Torsten Siever (²2001). Das Zitat im Internet. Ein Electronic Style Guide zum Publizieren, Bibliografieren und Zitieren. Hannover.

RICHTLINIEN

Umfang
1 Normseite entspricht der Größe DIN-A-4. Die Seitenzahl ist unbegrenzt.

Untergliederung
Längere Texte sollten moderat untergliedert sein; mehr als drei Untergliederungsstufen sind in der Regel nicht wünschenswert.

Versandweg
Das Manuskript soll nach Möglichkeit als Anhang einer E-Mail versendet werden (vgl. auch »Einsendung von Manuskripten« auf dieser Seite).

Adresse
Bitte mit dem Manuskript die vollständige Dienstanschrift sowie eine Telefonnummer für evtl. Rückfragen einreichen.

Korrekturverfahren
Die Redaktion behält sich Änderungswünsche am Manuskript vor.

Info zu:
→ NET.WORX-Qualität
→ NET.WORX-Homepage

INHALTSVERZEICHNIS

HINWEISE FÜR DEN BENUTZER

Dieses Internet-Dokument ist zitierbar! Diese wichtige Eigenschaft für wissenschaftliche Dokumente wird durch den vom Projekt sprache@web erarbeiteten Leitfaden ⟶🌍 »Das Zitat im Internet« erreicht. Die bibliografische Aufnahme für dieses Dokument ist ⟶🖹 hier verzeichnet; einen ⟶🌍 ShortGuide für alle wichtigen weiteren Fragen sowie nützliche Tipps zum Zitieren stehen kostenlos zum ⟶🌍 Download zur Verfügung.

Obwohl die NET.WORX als PDF-Dokumente für die Lektüre auf Papier besonders geeignet sind, unterstützen sie als Netzarbeiten natürlich auch Hyperlinks:

⟶🖹 : Link, der auf eine Textstelle innerhalb des vorliegenden Dokuments verweist. Bei einem Klick auf den Pfeil, bzw. den dahinter stehenden Begriff wird zu der entsprechenden Textstelle *innerhalb* der NET.WORX gesprungen.

⟶🌍 : Link, der auf eine Quelle im Internet verweist. Wird *bei einer bestehenden Internetverbindung* auf den Pfeil, bzw. den dahinter stehenden Begriff geklickt, wird der Nutzer mit der Quelle im Internet verbunden.

Bei direkten oder indirekten Verweisen auf fremde Internetseiten (»Links«) gilt, dass sich das Projekt sprache@web ausdrücklich von allen Inhalten aller gelinkten/verknüpften Inhalte distanziert und auch nicht für deren Inhalt verantwortlich ist. Für illegale, fehlerhafte oder unvollständige Inhalte und insbesondere für Schäden, die aus der Nutzung oder Nichtnutzung solcherart dargebotener Informationen entstehen, haftet allein der Anbieter der Seite, auf welche verwiesen wurde, nicht derjenige, der über Links auf die jeweilige Veröffentlichung lediglich verweist. Im übrigen gelten die ⟶🌍 Nutzungsbedingungen des Projekts sprache@web.

Die Herausgeber, 2004

1 EINLEITUNG

»Das aller vornemist vnd nötigst in allen Sprachen ist / das man Orthographiam helt / das ist / das man alle wörter mit jren eigenen vnd gebürlichen Buchstaben schreibe oder drücke / das man keinen Buchstaben aussen lasse / keinen zuviel neme / keinen fur den andern neme /Das einer die wörter mit buchstaben schreibe / gleich wie der ander / Item / das man die gleichlautende wörter / welche zwey ding bedeuten in jrem laut / mit sonderlichen buchstaben vnterscheide / wie die Ebreische / Griechische vnd Lateinische Sprache geordnet vnd gefasset ist.«[1]

Durch die weite Verbreitung von Martin Luthers Bibelübersetzung wurde erstmals, quasi durch einen Präzedenzfall, so etwas wie eine Leitlinie für die deutsche Rechtschreibung geschaffen. Bis zu allgemeingültigen Regeln der Orthografie, die für das Deutsche Reich erst Ende des 19. Jahrhunderts formuliert wurden, verging viel Zeit. Im 20. Jahrhundert wurde aber ein nicht mehr aufzuhaltender Prozess in Gang gesetzt, der die Schreibung im Deutschen bis in den kleinsten Bereich geregelt hat. Die Hoheit über diesen Prozess besaß bis zur Neuregelung die Dudenredaktion in Mannheim, die das Schreibverhalten der Deutschen beobachtet und über die Orthografie entschied. Dieser Prozess ist in den letzten 100 Jahren dynamisch verlaufen und daher kaum in wenige Regeln zu fassen. Ebenso wenig darf unser Zeichensystem als phonetische Wiedergabe der Lautrealisierung gesehen, die keine andere Schreibung zulässt. Ein Beispiel dafür ist die Längemarkierung der Vokale, die durch Verdopplung (*Saal*), durch Anhängen von *h* (*Bahn*) oder ohne Markierung (*Magen*) erfolgen kann. Entsprechend schwierig gestaltet sich in vielen Bereichen die Orthografie.

In der Informationsgesellschaft nimmt das geschriebene Wort einen zentralen Platz ein. Orthografie versteht sich dabei von selbst, denn wenn man nicht richtig schreibt, ist man schnell dem Verdacht ausgesetzt, auch beim Gegenstand des Textes fehlerhaft oder ungenau zu arbeiten. Das Schriftbild stellt eine Visitenkarte des Verfassers dar. Der Schreiber erwartet dabei nicht nur, dass ihn die Textverarbeitung unterstützt, viele verlassen sich mittlerweile beinahe vollkommen auf die Korrekturfunktion und vertrauen in Zweifelsfällen eher dem ›Urteil‹ der Maschine als dem eigenen gesunden Menschenverstand. Die Rechtschreibreform und ihr Ergebnis vom 1.7.1996 verstärken solche schon vorhandenen Unsicherheiten noch, da in der Bevölkerung über viele Bereiche der Reform nur ein diffuses Wissen vorhanden ist.

Korrektursysteme finden allerdings nicht nur im privaten Rahmen bei der alltäglichen Produktion von Briefen und E-Mails Anwendung, sondern ebenso bei Verlagen, Herausgebern von Zeitungen, Periodika, also überall, wo Text produziert wird. Gedruckte Texte vermitteln ein Bild von der Orthografie, das unser Schriftempfinden mehr beeinflusst als die schulische Kenntnis von orthografischen Regeln. Die großen Anbieter von Rechtschreibkorrektursoftware haben also, sofern den Vorschlägen ihrer Programme gefolgt wird, effektiv einen etwa ebenso großen Einfluss wie die Dudenredaktion oder die Teilnehmer der 3. Orthografischen Konferenz mit ihren Vorschlägen zur Reform.

Gegenstand dieser Arbeit wird die Evaluation der Möglichkeiten eines Rechtschreibkorrektursystems beim gegenwärtigen Entwicklungsstand sein. Um einen breiten Eindruck von den Fähigkeiten zu bekommen, wird ein Produkt aus dem professionellen Sektor mit einem Programm, das vor allem im privaten Rahmen eingesetzt wird, verglichen.

Beim ersten Programm, das dem Test unterzogen wird, handelt es sich um die Korrektursoftware *DITECT – Rechtschreibprüfung,* die von der Unternehmensberatung Dieckmann vertrieben wird und in vielen Verlagshäusern Verwendung

findet, unter anderem im Verlagshause Madsack für die Zeitungen *Hannoversche Allgemeine, Neue Presse* und andere. Die Referenzliste beinhaltet außerdem den Axel Springer Verlag, Bertelsmann, Sun Microsystems (StarOffice), die Frankfurter Allgemeine und die Süddeutsche Zeitung und viele andere Zeitungen, Verlage und Druckereien.[2] In jedem Falle rechtfertigt das breite Spektrum, das diese Software abdeckt, eine eingehende Untersuchung.

Zum Vergleich wird das Programm herangezogen, das wohl weltweit am häufigsten für die Textverarbeitung eingesetzt wird: *Microsoft Word*. Durch die weite Verbreitung hat die Rechtschreibkorrektur von Word einen immensen Einfluss auf das Orthografieverhalten der Deutschen. Wie oben schon angedeutet, sind die Vorschläge, die von einem Rechtschreibkorrekturprogramm gemacht werden, im Zweifelsfalle von großem Gewicht bei der Entscheidungsfindung. Dabei ist nicht klar, ob alle gemachten Vorschläge richtig sind bzw. ob alle vorhandenen Fehler tatsächlich gefunden werden.

Ziel der Arbeit ist es, soweit möglich die Funktionsweise des Korrekturprogramms aus einem umfassenden Test abzuleiten und die vorhandenen Schwachstellen aufzuzeigen. Die Hypothese, dass die tatsächliche deutsche Rechtschreibung und die von der Software angebotene Schreibung – unabhängig vom Anbieter – auseinander klaffen und in Kernbereichen die Programme systematische Fehler machen, steht dabei neben einer allgemeinen Bewertung der Korrekturleistung im Vordergrund. Gegebenenfalls können zu einigen Problemen Vorschläge zur Verbesserung der Leistung gemacht oder Probleme der Forschung benannt werden. Zumindest wird durch das Bewusstsein von den Fehlern der Umgang mit einem derartigen Programm geschärft.

Um eine professionelle Analyse der Testdaten erstellen zu können, müssen zunächst einige theoretische Grundlagen erarbeitet werden. Dies betrifft sowohl den Bereich Linguistik als auch denjenigen der Informatik. Exkurse ins Technische sind

bei der Darstellung nicht immer zu vermeiden. In diesem Sinne ist die automatische Korrektur von Rechtschreibfehlern ein klassisches Thema der Computerlinguistik. Technische und logische Probleme, die bei der zunächst recht simpel erscheinenden Rechtschreibkorrektur auftreten, werden in Kapitel 2 im Rahmen einer Erläuterung des theoretischen Aufbaus eines prototypischen Rechtschreibkorrektursystems besprochen.

Der empirische Teil der Arbeit beginnt mit der Wahl des Testkorpus. Es wurde aus aktuellen Zeitungsartikeln aus der *Neuen Presse* zusammengestellt, die im Verlagshause Madsack an hauseigenen Macintosh-Rechnern einer Rechtschreibprüfung unterzogen wurden (siehe Abbildung 1.1).

Dabei wurde vom Verfasser aus den fünf Ressorts *Seite Eins, Politik, Lokales, Wirtschaft* und *Sport* ein 50.000 Wörter umfassendes Testkorpus zusammengestellt, um ein möglichst breites Spektrum an Wortschatz und Formulierungen abzudecken. Die auftretenden Fehler und die bevorzugten Korrekturvorschläge der Programme DITECT und Microsoft Word wurden dokumentiert und anschließend analysiert. Das Ergebnis findet sich in Kapitel 3, getrennt nach den Bereichen Fehlerdetektion (3.1) und Fehlerkorrektur (3.2).

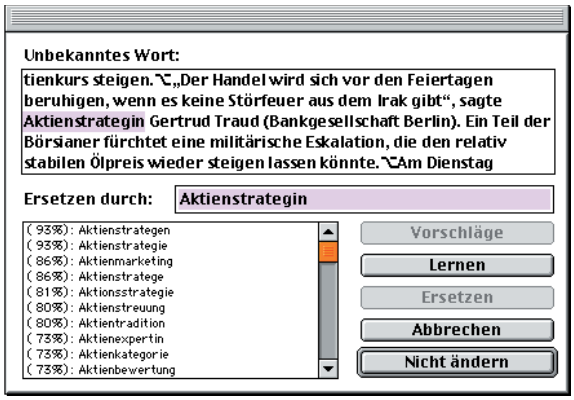

Abbildung 1.1 Rechtschreibkorrekturprogramm DITECT

In einem zweiten Testlauf werden einzelne Problembereiche herausgegriffen und isoliert anhand von Beispielkatalogen betrachtet. Es handelt sich dabei vor allem um die Bereiche Getrennt- und Zusammenschreibung und Groß- und Kleinschreibung. Beide Bereiche wurden in der Rechtschreibreform neu geregelt und sind in der Praxis nicht immer leicht anzuwenden. Die besonderen Schwierigkeiten bei der

Rechtschreibkorrektur liegen darin, dass der Fehler nicht in der genauen Zeichenfolge des jeweiligen Lexems liegt, sondern sich, beispielsweise im Falle von Substantivierungen, aus dem syntaktischen Kontext erschließt. Des Weiteren werden die klassischen Bereiche der Morphologie Flexion und Komposition behandelt. Auf diese Probleme wird in Kapitel 4 eingegangen.

Am Ende der Arbeit steht eine Bewertung der beiden untersuchten Programme. Dabei lässt sich nicht lediglich eine Aussage darüber treffen, welches Programm besser ist, sondern auch darüber, welche Design-Entscheidungen getroffen wurden und wo die einzelnen Stärken und Schwächen liegen. Davon abgehoben lässt sich sicher eine Aussage treffen, auf welchem Niveau sich die Forschung augenblicklich (oder zum Zeitpunkt des Software-Release-Datums) befindet und in welche Richtung sie sich entwickelt.

2 THEORETISCHE GRUNDLAGEN

Auf dem Gebiet der Rechtschreibkorrektur sind wie in vielen Bereichen der maschinellen Sprachverarbeitung in den letzten Jahren große Fortschritte gemacht worden, die sich auf den ersten Blick nicht ohne weiteres erschließen. Während die Benutzer von Word 7.0 oder früherer Versionen generell eher von den Korrektursystemen frustriert waren und diese aufgrund der mangelnden Fehlerdetektion nicht anwendeten, werden mittlerweile die Vorschläge der moderneren Korrekturprogramme trotz weiterhin bestehender Mängel weitgehend akzeptiert. Das Vokabular konnte beispielsweise aufgrund der erweiterten Kapazitäten der Heim-PCs in den letzten Jahren maßgeblich vergrößert werden. Word 7.0 und frühere Versionen kannten beispielsweise das Wort *Internet* noch nicht und schlugen statt dessen *Internat* vor (Klaeren 1997).

Abbildung 2.1 Rechtschreibkorrektur in Word 7.0 (Klaeren 1997)

Hier zeigt sich der Einfluss des Lexikons als primäre Quelle der Rechtschreibkorrektur. Es handelt sich zwar im Beispiel um einen Neologismus, genauer um eine Wortschöpfung aus dem Englischen. Der Begriff ist aber spätestens seit Mitte der 90er Jahre Bestandteil des allgemeinsprachlichen Wortschatzes und sollte daher im Lexikon des Programms auftreten. Solche Fehler, die den Grundwortschatz der Sprache betreffen, dürften immer seltener werden. Ein immer noch schwieriges Pro-

blem stellen aber Eigennamen dar, deren Rechtschreibung nach klar formulierten Regeln nur schwer zu erfassen ist.

Auch in strukturellen Bereichen der Rechtschreibkorrektur verbleiben noch Probleme, deren Lösung auf sich warten lässt. Dies sind beispielsweise die schon angesprochenen Gebiete Getrennt- und Zusammenschreibung und Groß- und Kleinschreibung sowie die Ergänzung mit Bindestrich, die Grammatikprüfung, Zeichensetzung und die Kompositaerkennung, wie folgendes Beispiel aus der neueren Version Microsoft Word 2000 belegt.

Die einzelnen Konstituenten *Komposita* und *Erkennung* werden einwandfrei erkannt, die Komposition zu einem zusammengesetzten Substantiv bereitet aber offenbar Probleme, obwohl nicht einmal Fugenmarkierungen wie z. B. das *s* in *Verwendungszweck* im Spiel sind. Rechtschreibkorrektur im Deutschen ist also ein schwieriger Prozess, der weit über den bloßen Abgleich mit den, in einem Lexikon gespeicherten, Wortformen hinausgeht und an die jeweilige Sprache, für die das System eingesetzt wird, angepasst werden muss.

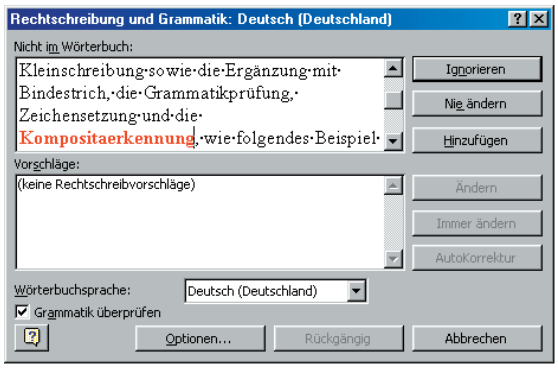

Abbildung 2.2 Rechtschreibkorrektur in Word 2000

Die Behandlung solcher Problemfälle bei der Rechtschreibkorrektur soll, soweit möglich, in diesem Kapitel geklärt werden. Zunächst wenden wir uns den Anwendungen in den vorhandenen Produkten zu. Anschließend wird dargestellt, welche Vorstellung von Fehlern der Korrektur zugrunde liegt und auf welche Weise die automatische Korrektur modelliert wird. Im zweiten Teil des Kapitels wird auf die sprachspezifischen Eigenheiten der deutschen Sprache eingegangen und der Versuch unternommen, Vorschläge zu deren Modellierung zu unterbreiten.

2.1 Die technische Realisierung

2.1.1 Anwendungsbereiche

Die automatische Rechtschreibkorrektur hat mittlerweile verschiedene Anwendungsbereiche, bei denen zwischen zwei Verfahren unterschieden werden muss. Einige Programme arbeiten interaktiv, d. h. sie machen Korrekturvorschläge und bieten dies dem Nutzer in einem Dialog an. Ein viel anspruchsvolleres Verfahren ist die automatische Korrektur, die bisher nur in einfachen Konzepten anwendungsreif ist (Kukich 1992: 379).

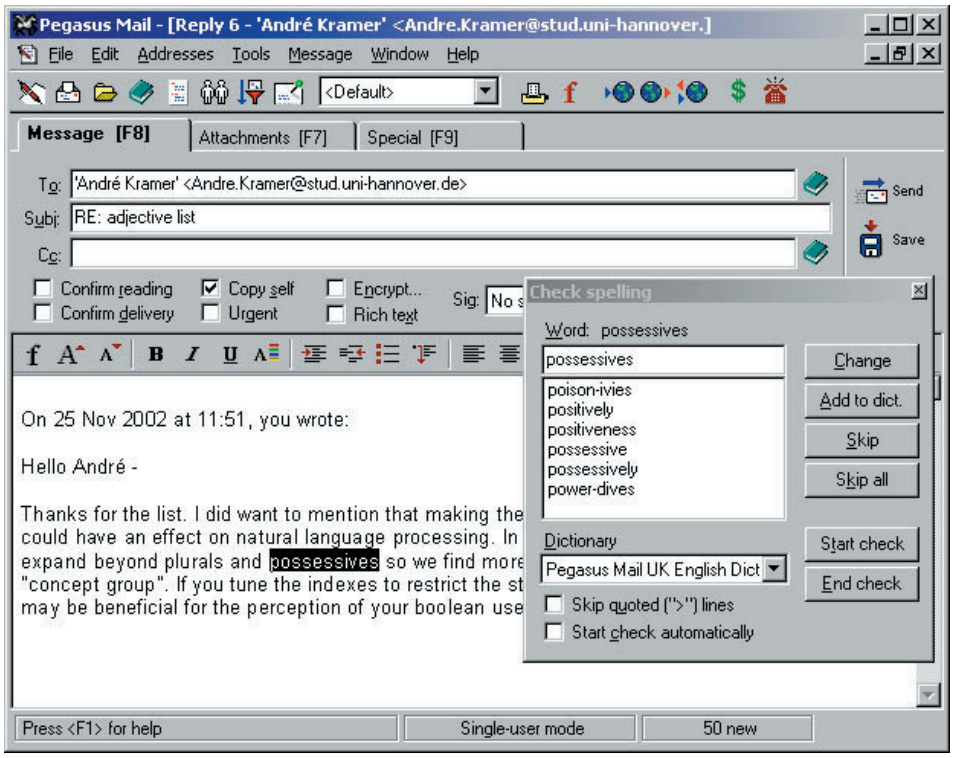

Abbildung 2.3 Interaktive Rechtschreibkorrektur bei Pegasus Mail 4.12 (Englisch)

Der wichtigste und für den Heimsektor älteste Anwendungsbereich betrifft zweifelsohne die Implementierung in Textverarbeitungsprogrammen. Damit sind in diesem Falle nicht nur die klassischen Texterzeugungsprogramme wie Microsoft

Word oder StarOffice gemeint, sondern seit neuerer Zeit auch E-Mail-Editoren und andere Programme, die natürlichsprachliche Texteingaben zu bewältigen haben.

Auch Anwendungen, die über ein Web-Interface arbeiten, benutzen Rechtschreibkorrekturprogramme, wie z. B. die Internetsuchmaschine Google. Bei wenigen Treffern auf eine Suchanfrage wartet die Suchmaschine mit Alternativvorschlägen zum eingegebenen Suchbegriff auf. Dies ist die übliche interaktive Arbeitsweise eines Korrektursystems. Wird aber überhaupt kein Dokument zurückgegeben, erfolgt die Korrektur der eingegebenen Zeichenkette automatisch und der Nutzer wird lediglich in einer kurzen Notiz oberhalb des ersten gefundenen Dokuments über die Korrektur informiert.

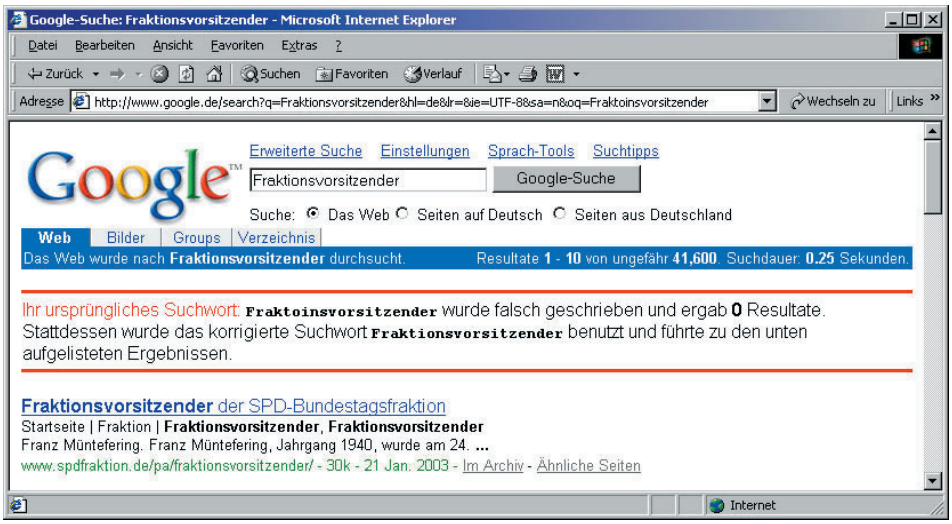

Abbildung 2.4 Automatische Rechtschreibkorrektur bei Google

Dieselben Algorithmen, die bei der Überprüfung von Texteingaben über die Tastatur implementiert werden, finden auch Verwendung in der Schrifterkennung (optical character recognition, OCR) in Programmen wie OmniPage u. a. OCR-Systeme tendieren zu höheren Fehlerraten als Texte, die über Tastatureingabe entstehen, auch wenn die Fehler unterschiedlicher Natur sind. OCR-Systeme verwechseln beispielsweise häufig *D* und *O* oder *ri* und *n*. Für diesen Bereich sind Korrekturalgorithmen unersetzlich. Selbst eine Zeichen-Erkennungsrate von 99 Prozent beispielsweise be-

deutet nur eine Wort-Erkennungsrate von 95 Prozent, wenn man annimmt, dass ein Wort fünf Zeichen umfasst. Die Zeichenfehler potenzieren sich also in der Zahl falscher Wörter.

Weitere Anwendungsgebiete sind die Handschrifterkennung in Palm-Pilot-Systemen oder im Chinesischen, wo die große Anzahl an Schriftsymbolen die Verwendung einer Tastatur unpraktisch macht (Jurafsky 2000: 143).

Streng genommen gehören in dieses Gebiet auch fachspezifische Anwendungen, die die Syntax von Programmiersprachen überprüfen. Sie zeichnen sich durch eine besonders schnelle Bearbeitung aus, da das Vokabular auf die jeweilige Programmiersprache beschränkt ist und daher klein gehalten werden kann (Kukich 1992: 379).

Bleiben wir bei den deutschsprachigen Anwendungen in der Textverarbeitung. Das gängigste und verbreitetste Textverarbeitungsprogramm ist wohl unbestritten Microsoft Word in der Office-Version 2000 oder XP, aber die Konkurrenz bringt mit StarOffice 5.4, Lotus Word Pro 9.6 und AppleWorks 6.2 leistungsfähige Konkurrenten auf den Markt, die hinsichtlich ihrer Fähigkeiten nicht hinter dem Produkt des Softwaregiganten aus Seattle zurückstehen. Alle Programme bieten integrierte Rechtschreibkorrekturfunktionen, die wahlweise nach alter oder neuer Schreibung arbeiten. Der Nutzer ist aber keineswegs auf die integrierten Funktionen angewiesen. Verschiedene Anbieter verkaufen Plugins, die sich in Word 2000 integrieren lassen. Zu nennen wären unter anderem »Deutsch Korrekt 2000« von Hexaglot für das untere Preissegment (ca. 25 Euro) und der »Primus Korrektur-Manager« von der Firma SoftEx, der erst für über 400 Euro zu haben ist.

StarOffice lässt sich mit dem untersuchten System DITECT ausrüsten. Dies ist allerdings kostenpflichtig, weshalb die Mehrzahl der Nutzer wohl auf die kostenfreie Version SpellCheck zurückgreifen wird, die im Internet unter → http://www.openoffice.org heruntergeladen werden kann.

Auch auf anderen Plattformen existieren Rechtschreibkorrektursysteme, wie für Unix das kostenlose Programm *Ispell* oder die kostenpflichtige Variante *Aspell*, die aus der früheren Version *Kspell* nach dem Entwickler Kevin Atkinson hervorging (Greve 2002). Auch der Amiga besitzt ein Rechtschreibkorrekturprogramm, das im Programm AmigaWriter implementiert ist und auf dem Unix-Programm *Ispell* basiert (Schlick 1998).

Konverterprogramme stellen eine Untergruppe der klassischen Rechtschreibkorrekturprogramme dar: Sie wandeln alte in neue Rechtschreibung um. Hier gibt es zwei Produkte, die preislich im gleichen Rahmen liegen wie die Korrekturprogramme. Der Bertelsmann Orthograf für 25 Euro ist die preisgünstige Heimalternative, während die Software »Corrigo« der Firma CLT Sprachtechnologie GmbH für ca. 500 Euro erheblich teurer und leistungsfähiger ist. Dies drückt sich beispielsweise darin aus, dass Corrigo das Problem der Zeichensetzung halbwegs beherrscht, während andere Programme davor vollständig kapitulieren müssen. Dies gilt allerdings auch nur für Fälle, in denen die neue von der alten Rechtschreibung abweicht.

Konverterprogramme sind nicht Teil dieser Untersuchung und sollen nur der Vollständigkeit halber Erwähnung finden. Im Folgenden wird die allgemeine Funktionsweise von Rechtschreibkorrektursystemen erläutert, um mithilfe dieses Wissens eine Einschätzung des Vorgangs der Fehlerkorrektur im empirischen Test möglich zu machen.

2.1.2 Fehlertypen

Für die Fehlerentdeckung oder Fehlerdetektion ist zunächst eine Kategorisierung in zwei Bereiche wichtig, zwischen denen unterschieden wird. Dies sind zum einen die Non-Word- und zum anderen die Real-Word-Errors. Non-Word-Errors bezeichnen Fehler, die nicht existierende Wörter entstehen lassen, z. B. *Graffe* für *Giraffe*.

Die Korrektur dieser Fehler kann isoliert, also ohne kontextuelle Informationen aus dem Umfeld des Wortes, erfolgen. Ich komme auf diese Problematik zurück.

Bei der anderen Fehlerkategorie handelt es sich um Real-Word-Errors. Hier werden gültige, wohlgeformte Wörter produziert, die durch typografische Fehler entstehen. Diese Wörter sind formal richtig geschrieben, stehen aber im falschen Kontext. Dies kann durch Flüchtigkeit geschehen (*Wiese – Weise*) oder auch durch Unkenntnis (*Masse – Maße*). Eine Ursache kann z. B. die phonetische Gleichheit (Homophonie) zweier Wortformen sein, wie in den Fällen *wider – wieder* oder *Seite – Saite*. Auch grammatische Fehler wie Kongruenzfehler oder falsche Wiedergabe der Rektionsbeziehungen innerhalb einer Nominalgruppe gehören zu dieser Fehlergruppe. Ein weiterer häufiger Fehler resultiert in fehlenden oder überflüssigen Leerzeichen. Solche Zeichenketten führen häufig zu gültigen Wortformen, deren Detektion besondere Probleme bereitet, da das Überschreiten der Wortgrenze bei der Fehlerdetektion in einer exponentiellen Steigerung der möglichen Formen resultiert.

Real-Word-Errors sind erheblich schwerer aufzufinden als Non-Word-Errors. Im letzteren Fall muss lediglich die ungültige, nicht im Lexikon enthaltene, Zeichenkette gefunden werden. Im Falle des Real-Word-Errors muss erkannt werden, dass die Wortform im falschen Kontext auftritt. Kontextinformation ist aber auch bei der Korrektur von Non-Word-Errors erforderlich, wenn für eine Schreibung mehrere Korrekturvarianten zur Auswahl stehen. Die Zeichenfolge **Musse* kann beispielsweise mit Änderung nur eines Buchstabens die Wörter *Russe, Muss, Messe, Masse, Muse* und *Muße* ergeben. Die richtige Auswahl ergibt sich nur aus dem Kontext.

Kukich beschreibt zusätzlich zwei Klassen von Tippfehlern, die im Kontext der Fehlerkorrektur auftreten. Die Grundlage dieser Kategorisierung ist weniger die Prozessierung durch das Korrektursystem als der Verursacher und die Motivation des Fehlers. Somit liefern diese Kategorien wichtige sekundäre Informationen über

die Natur des Fehlers, weshalb immer wieder auf sie referiert wird. Es handelt sich um

1. **Typografische Fehler,** die in erster Linie mit der Benutzung der Tastatur zusammenhängen: *Griif* statt *Griff.* Es wird angenommen, dass der Nutzer die richtige Schreibung des Wortes kennt und lediglich eine motorische Ungenauigkeit die Ursache des Fehlers ist.
2. **Kognitive Fehler,** die von Schreibern verursacht werden, die die richtige Schreibung des Wortes nicht kennen. Kognitive Fehler beinhalten phonetische Fehler: *Refarat* statt *Referat,* und Homonymfehler: *wider* statt *wieder* (Kukich 1992: 387).

Diese Kategorien sind problematisch, da die Entscheidung, ob es sich um einen typografischen oder einen kognitiven Fehler handelt, etwas willkürlich ist. Die Zeichenfolge *Refarat* muss nicht notwendigerweise ein kognitiver Fehler sein, sondern kann auch das Produkt einer motorischen Fehlleistung sein, so wie kognitive Fehler äußerlich wie typografische Fehler aussehen können.

Die wichtigste und für viele Korrekturprogramme substantielle Kategorisierung ist die Einteilung der häufigsten Tippfehler nach Fred Damerau. Diese grundlegende Einteilung geschieht nach rein formalen Kriterien und zeichnet sich durch ihren einfachen Ansatz aus. Damerau hat in seiner 1964 erschienen Studie »A Technique for Computer Detection and Correction of Spelling Errors« nachgewiesen, dass 80% aller fehlerhaft getippten Wörter auf einfachen Fehlern beruhen (*single-error misspellings*) (Damerau 1964: 175). Infolge dieser Studie beruhen große Teile der darauf folgenden Forschung auf der Korrektur einfacher Fehler:

- **Einfügung (insertion):** *Tere* für *Tee*
- **Auslassung (deletion):** *Tere* für *Teure*
- **Substitution (substitution):** *Tere* für *Türe*
- **Transposition (transposition):** *Tere* für *Teer*

In dieser Arbeit wird im Folgenden immer wieder auf diese Fehlerkategorien zurückgegriffen, da sie die elementare – gewissermaßen handwerkliche – Quelle für die meisten falsch geschriebenen Wörter darstellen.

In einer Untersuchung über die Effekte, die die Tastatureingabe auf die Fehlernatur hat, wies Grudin 1983 nach, dass typografische Fehler bei weitem die häufigste Fehlerursache bilden. Substitutionsfehler, hervorgerufen durch auf der Tastatur unmittelbar benachbarte Buchstaben, stellten 59% der Fehler bei Neulingen und immerhin noch 31% der Fehler bei erfahrenen Maschinenschreibern dar. Zählt man Fehler dazu, die auf Substitution mit Tasten in derselben Spalte oder den korrespondierenden Tasten der anderen Hand beruhen (homologe Fehler), können 83% der Substitutionsfehler bei Neulingen und 51% bei Experten als tastaturbasiert angesehen werden (Grudin, Jonathan T.: Error patterns in novice and skilled transcription typing. In: Cooper, W. E. (Hrsg.): Cognitive aspects of skilled typewriting. 1983. S. 121-139. zit. nach Jurafsky 2000: 145.).

Daneben hat es noch weitere Kategorisierungsversuche gegeben. Diese sind aber meistens umfangreicher und an die englische Sprache gebunden wie dasjenige im Vorwort *Webster's New World Misspeller's Dictionary*, das von zwölf Kategorien ausgeht und die Fehler danach einteilt, ob über eine Konsonantenverdopplung entschieden werden muss, ob der Fehler mit der Aussprache zu tun, also phonetische Gründe hat, oder ob Homonymie als Ursache in Frage kommt (Kukich 1992: 392). Zunächst ist nur die formale Einteilung Dameraus notwendig.

2.1.3 Fehlererkennung mit Hilfe eines Lexikons

Non-Word-Errors werden im Text zumeist mit Hilfe eines Lexikons erkannt. Das Wort *Refarat* z. B. würde nicht in einem Lexikon auftreten und dadurch dem Schreiber zur Korrektur vorgelegt werden. Peterson schlug 1986 vor, ein solches Lexikon klein zu halten, da umfassende Lexika seltene Wortformen enthielten, die fehlerhafte Schreibungen häufiger Lexeme verdecken könnten (im Englischen: *wont* (Gewohnheit), möglicherweise falsche Schreibung für *won't*) (Peterson 1986: 637). Damerau und Mays nahmen sich 1989 dieser Fragestellung an und untersuchten ein Korpus mit über 22 Millionen Wörtern. Mit einer Vergrößerung des Lexikons von 50.000 auf 60.000 Wörter erreichten sie eine Eliminierung von 1.348 falschen Fehlerdetektionen und nur 23 zusätzliche falsche Akzeptanzen (Kukich 1992: 384). Im Verhältnis bedeutet dies 59 Verbesserungen auf einen falsch erkannten Term aufgrund eines großen Lexikons. Die These von Peterson war somit nicht haltbar.

Nicht nur die Größe des Lexikons, sondern auch seine Beschaffenheit ist von Bedeutung. Frühere Ansätze benutzten Nachschlagewerke zur Lexikongenerierung. Walker und Amsler führten 1986 hierzu eine Untersuchung an einem acht Millionen Wörter umfassenden Korpus mit New-York-Times-Texten und dem *Merriam-Webster Seventh Collegiate Dictionary* durch. Das Ergebnis war, dass zwei Drittel (61 %) der Lexikoneinträge nicht in dem Zeitungskorpus auftraten und etwa genauso viele der Wörter aus dem Text (64 %) nicht im Lexikon verzeichnet waren (Kukich 1992: 384). Seitdem wird stärker auf anwendungsorientierte korpusbasierte Lexika zurückgegriffen.

Aufgrund der Notwendigkeit, Flexion und mit Einschränkungen auch Derivation Rechnung zu tragen, enthalten Lexika zur Fehlererkennung eine Morphologie, die eine Modellierung von Wortformen mit Hilfe von Finite-State-Transducern ermöglicht (vgl. Kap. 2.2.3). Dies sind Automaten, die die Verkettung und den Austausch von Modulen, in diesem Fall Morphemen in elektronischen Systemen modellieren.

Frühe Korrekturprogramme für die englische Sprache wie Unix Spell akzeptierten dabei alle möglichen Kombinationen von affigierten Wortformen wie **antiundog-gingly* oder **theness*. Moderne Fehlererkennungssysteme benutzen Part-Of-Speech-Tags, die zu jedem Lexikoneintrag Informationen über dessen Wortkategorien, also zumindest die Wortart enthalten. So kann vermieden werden, dass Substantive mit Verbaffixen wie dem partizipbildenden Suffix *-end* behaftet werden (Jurafsky 2000: 146). Auf die linguistische Problematik dieses Gebiets wird in Kapitel 2.2 näher eingegangen.

2.1.4 Probabilistische Modelle zur Fehlerkorrektur

Im Folgenden wird ein Modell der Fehlerkorrektur beschrieben, das weitestgehend auf die Arbeit von Kernighan et. al. aus dem Jahre 1990 zurückgeht. Vorher gab es noch die Ansätze von Yannakoudakis und Fawthrop (1983) und von Pollock und Zamora (1983), die Erwähnung verdienen. Yannakoudakis und Fawthrop suchten ein generelles Muster im Verhalten bei falscher Schreibung. Sie untersuchten 1377 Rechtschreibfehler und formulierten anschließend siebzehn heuristische Regeln, von denen zwölf auf den falschen Gebrauch von Konsonanten und Vokalen in Graphemen der englischen Sprache zurückgehen und fünf sich auf die Satzproduktion beziehen. Eine Regel besagt beispielsweise, dass der Konsonant *h* häufig in den Graphemen *ch, gh, ph* und *rh* ausgelassen wird. Die Arbeit von Pollock und Zamora zeichnet sich eher durch statistische Angaben über die Häufigkeiten von einfachen Fehlern oder bestimmte Fehlerpositionen aus (Kukich 1992: 391). Beide Ansätze sind auf die englische Sprache beschränkt und liefern keine umfassende Methode, wie mit dem Problem der Rechtschreibkorrektur umgegangen werden kann. Ihnen fehlt die Universalität.

Die Modelle, die heute in der Fehlerkorrektur angewendet werden, haben das Bayes'sche Theorem zur Grundlage, das auch in der automatischen Spracherken-

nung Anwendung bei der Modellierung von Ausspracheunterschieden findet. Erstmals setzten Bledsoe und Browning im Jahre 1959 wahrscheinlichkeitstheoretische Ansätze dieser Art für die Rechtschreibkorrektur ein. Sie modellierten Wahrscheinlichkeitswerte für die Verwechslung von Buchstaben und errechneten Übergangswahrscheinlichkeiten für alle möglichen folgenden Buchstaben. Diese sind als Markov-Wahrscheinlichkeiten bekannt. Mithilfe eines Lexikons wurde in einem zweiten Schritt eine Worthypothese erstellt (Kukich 1992: 402). Einige grundlegende Bemerkungen zu diesem Thema sind nötig, um den wahrscheinlichkeitstheoretischen Ansatz, auch Noisy-Channel-Modell genannt, in dieser konkreten Anwendung zu verdeutlichen.

Im Wesentlichen geht es bei der Fehlerkorrektur wie in der Spracherkennung um die Abbildung einer Symbolabfolge auf eine andere. Im Falle der Spracherkennung muss die phonetische Repräsentation eines Wortes auf die lexikalische abgebildet werden (vgl. zur Spracherkennung Kramer, Lehmberg, Schlobinski 2000). Dabei gibt es viele Störquellen, wie beispielsweise die Übertragung über Telefon oder Nebengeräusche (Straßen- oder Bürolärm). Bei der Fehlerkorrektur liegt der Fall ähnlich: Hier muss die unkorrekte Buchstabensequenz – als Störquelle kann hier die fehleranfällige Tastatureingabe gesehen werden – durch die korrekte ersetzt werden. Das Noisy-Channel-Modell behandelt in diesem Sinne die reduzierte Aussprache oder fehlerhafte Schreibung als Variante der lexikalischen Form. Das Ziel ist, ein Modell des Kommunikationskanals zu bauen, das es ermöglicht, die Art der Modifikation zu ermitteln und so das ursprünglich gemeinte Wort wiederherzustellen. Die Metapher des Noisy Channel stammte aus den frühen Siebzigern, als in den IBM-Laboratorien an der Spracherkennung gearbeitet wurde (Jurafsky 2000: 147).

Im Wesentlichen haben wir es hier mit zwei Ebenen zu tun, auf denen die statistischen Ansätze Verwendung finden: der Wahrscheinlichkeit, dass ein Wort *w* auftritt, und der Wahrscheinlichkeit, dass es zu einem Fehler an der Stelle *O* kommt.

Zunächst einmal wird an der Stelle, wo der Fehler auftritt, die Auftretenswahrscheinlichkeit der in Frage kommenden Wortformen *P(w)* bestimmt. Dies kann mit oder ohne Berücksichtigung des Kontextes geschehen. In jedem Fall ist dieser Wert aber, abgesehen von der Existenz des Fehlers an sich, von diesem unabhängig. Die zweite Ebene beschäftigt sich mit der Natur des Fehlers und ist daher abhängig von der Beobachtung (engl. *observation*). Sei die Beobachtung die Buchstabensequenz *Refarat*. Dann kann das falsch geschriebene Wort als Beobachtung *O,* das korrigierte Lemma *Referat* als die Wortform *w* bezeichnet werden. Später wird sich zeigen, wie die Wahrscheinlichkeit für diesen Fehler bestimmt werden kann.

Der verwendete Algorithmus ist ein Spezialfall des Bayes'schen Theorems von 1763. Bei der Bayes'schen Klassifizierung haben wir eine Beobachtung und versuchen, dieser eine Reihe von Klassen zuzuordnen, die zu der Beobachtung passt. Im Falle der Spracherkennung entspräche die Beobachtung einer Sequenz von Phonemen. Bei der Fehlerkorrektur handelt es sich um eine Buchstabensequenz, der das richtige Wort zugeordnet werden muss. Praktisch gesagt: Wie das Wort *Referat* auch ausgesprochen oder geschrieben ist, wir wollen eine Klassifikation durchführen, die es als das Lexem *Referat* wiedergibt. Dazu muss aus dem Lexikon der Eintrag oder aus dem Wortuniversum das Wort gefunden werden, das zur Beobachtung passt, also das Wort, für das *P(w|O)*, die Wahrscheinlichkeit eines möglichen Wortes in Abhängigkeit von der Beobachtung maximal ist (Jurafsky 2000: 148). Sei $\hat{w}$ unsere Annahme des korrekten Wortes *w*, *O* die Beobachtung und *V* das Vokabular, dann gilt:

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid O)$$

Die Funktion $\operatorname{argmax}_x f(x)$ besagt dabei, dass x so gewählt ist, dass f(x) maximal ist. Das wahrscheinlichste Wort an einer beobachteten Position wird also als das korrekte Wort $\hat{w}$ angenommen. Das Problem besteht darin, dass nicht klar ist, wie diese Formel bearbeitet werden kann. Dabei wird die Bayes'sche Regel angewendet, mit

der die Formel $P(w|O)$ in drei andere Wahrscheinlichkeiten umgewandelt werden kann (Jurafsky 2000: 147):

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \frac{P(O|w)\,P(w)}{P(O)}$$

Diese Gleichung ist weitaus einfacher zu bearbeiten. $P(w)$ ist die Auftretenswahrscheinlichkeit des Wortes an sich und kann aus der Auftretenshäufigkeit in Textkorpora ermittelt werden. Der Fall für $P(O|w)$, der Auftretenswahrscheinlichkeit des Wortes in Abhängigkeit von der Beobachtung, ist etwas schwieriger und wird weiter unten behandelt. $P(O)$, die Wahrscheinlichkeit für die Beobachtung, ist nicht einzuschätzen. Sie kann aber ignoriert werden, da die Rechnung für alle möglichen Wörter bei einer bestimmten Beobachtung durchgeführt wird und $P(O)$ daher immer gleich ist (Jurafsky 2000: 149). $P(O)$ kann aus diesem Grund gleich eins gesetzt bzw. in der Gesamtrechnung gekürzt werden:

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} \frac{P(O|w)\,P(w)}{P(O)} = \underset{w \in V}{\operatorname{argmax}}\, P(O|w)\,P(w)$$

Also gilt als fundamentale Grundlage für die automatische Spracherkennung und Fehlerkorrektur, dass das gesuchte Wort $\hat{w}$ aus dem maximalen Wert der Werte $P(O|w)$ und $P(w)$ ermittelt wird. $P(O|w)$ wird auch *likelihood*, $P(w)$ auch *Prior probability* genannt (Jurafsky 2000: 149). Im folgenden Kapitel wird etwas praktischer betrachtet, was mit diesen Termen gemeint ist.

2.1.5 Beispiel zur Fehlerkorrektur isolierter Wörter

Der Vorgang der Fehlerkorrektur, wie er bei Kernighan, Church und Gale im Jahre 1990 beschrieben wird, soll im Folgenden am Beispiel des Wortes *acress* vorgeführt werden. Die Grundannahme ist, dass das falsch geschriebene Wort nur aufgrund eines einzigen Fehlers der in 2.1.2 eingeführten Typen Einfügung, Auslassung, Substitution und Transposition von der richtigen Schreibung abweicht. Der Algorith-

mus arbeitet nun so, dass zunächst alle möglichen korrekt geschriebenen Wörter, die in einem der vier Fehler abweichen, ermittelt werden. Anschließend erhalten diese eine Bewertung. Folgende Vorschläge werden nach dem Noisy-Channel-Modell bei Kernighan gemacht:

fehlerhafte Schreibung	korrigierte Schreibung	korrigierter Buchstabe	falscher Buchstabe	Buchstaben-position	Fehlertyp
acress	actress	t	-	2	Auslassung
acress	cress	-	a	0	Einfügung
acress	caress	ca	ac	0	Transposition
acress	access	c	r	2	Substitution
acress	across	o	e	3	Substitution
acress	acres	-	2	5	Einfügung
acress	acres	-	2	4	Einfügung

Tabelle 2.1: Korrektur von acress bei Kernighan et. al. 1990 (vgl. Kernighan et. al.: 205)

Die Nummerierung der Buchstabenpositionen beginnt bei null. Also haben wir im ersten Fall für *ac_tress* eine Auslassung an Position 2 der Zeichenfolge *acress*, im zweiten Fall *_cress* eine Einfügung an Position 0 usw. Im zweiten Schritt wird die Wahrscheinlichkeit für jede einzelne Worthypothese berechnet. Die geschieht mit der Gleichung, die oben bearbeitet wurde. Setzen wir *t* für die typografische Realität und *c* für die korrigierte Fassung, erhalten wir

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(t \mid c)\, P(c)$$

Der Wahrscheinlichkeitsterm im Nenner ist eliminiert, da *t* und somit auch *P(t)* konstant ist. *P(c)* kann ermittelt werden, indem die Häufigkeit der Wörter in einem Korpus ausgezählt wird und anschließend durch die Gesamtzahl der Wörter geteilt wird. So weit wurde das Vorgehen in 2.1.4 bereits beschrieben. In Tabelle 2.2 sind die Worthäufigkeiten für die Korrekturvorschläge angegeben, wie Kernighan und seine Mitarbeiter sie im AP-Newswire-Korpus von 1988 vorgefunden haben. Die erste Spalte gibt die Häufigkeit des Vorkommens an, die zweite bezeichnet den Wert

für die Auftretenswahrscheinlichkeit. Diese Zahl wird ermittelt, indem die Auftretenshäufigkeit eines Wortes durch die Gesamtzahl der Wörter (in diesem Fall 44 Mio.) geteilt wird.

Der Wert in der zweiten Spalte für *cress* ist nicht gleich null, da dies Berechnungsprobleme mit sich führen würde. Daher wird eine Operation durchgeführt, die »smoothing« genannt wird, um Nullwerte zu vermeiden. Wie man erkennen kann, ist der Wahrscheinlichkeitswert entsprechend der Häufigkeiten für *across* am größten und für *cress* am kleinsten.

c	Frequenz(c)	*P*(c)
actress	1343	0,0000315
cress	0	0,000000014
caress	4	0,0000001
access	2280	0,000058
across	8436	0,00019
acres	2879	0,000065

Tabelle 2.2: Worthäufigkeit und Auftretenswahrscheinlichkeit (vgl. Kernighan et. al.: 206)

Damit ist der Wert für *P*(c) ermittelt. Der Term *P*(t|c) stellt für die Forschung immer noch Probleme dar. Die Wahrscheinlichkeit, dass ein Wort falsch geschrieben wird, hängt von vielen Faktoren ab, wie z. B. wer der Schreiber ist, wie erfahren er im Maschineschreiben ist, wie müde er zum Zeitpunkt des Schreibens war, etc. Dennoch kann der Wert für *P*(t|c) relativ gut geschätzt werden. Die entscheidenden Faktoren hierfür liegen zum Teil auf der Tastatur selbst, nämlich die Nähe der Tasten zueinander. Die Nasale *n* und *m* werden beispielsweise häufig ausgetauscht. Dies liegt zum einen an der Nähe auf der Tastatur, zum anderen an der phonetischen Ähnlichkeit. Da sie ähnlich ausgesprochen werden, treten sie in ähnlichen Kontexten auf und werden häufig verwechselt.

Ein einfacher Weg, die Wahrscheinlichkeit einzuschätzen, wurde von Kernighan et. al. begangen. Sie ignorierten alle äußeren Faktoren und bewerteten nur die Wahrscheinlichkeit, dass mit dem falsch geschriebenen Wort *acress* eigentlich *across* gemeint war *P*(*acress*|*across*). Dazu benötigt man die Information, wie häufig in einem großen Fehlerkorpus *e* mit *o* verwechselt wurde. Dies geschieht mit einer

Verwechslungsmatrix, einer 26 x 26 Felder umfassenden Tabelle, in der für jedes Buchstabenpaar verzeichnet ist, wie häufig der eine Buchstabe mit dem anderen verwechselt wurde. Die Zelle [*t, s*] würde beispielsweise angeben, wie häufig ein *t* für ein *s* eingesetzt wurde (Jurafsky 2000: 152). Eine solche Tabelle ist 1983 erstmals von Grudin erstellt worden. Kernighan et. al. ermittelten beispielsweise in einer Liste von 25.000 Fehlern, dass der Vokal *a* 238-mal fehlerhaft für ein *e* eingesetzt wurde, während der Konsonant *s* 436-mal für ein *e* eingesetzt wurde (Kernighan et. al.: 206). Kernighan et. al. benutzten vier verschiedene Fehlermatrizen für die vier vorhandenen Fehlertypen Einfügung, Auslassung, Substitution und Transposition. Aus diesen Werten lassen sich ebenfalls Wahrscheinlichkeitswerte berechnen.

Abschließend werden die jeweiligen Wahrscheinlichkeitswerte für das Auftreten des Wortes an sich und das Auftretens des Fehlers miteinander multipliziert. Nimmt man die Werte für *P(c)* aus Tabelle 2.2 und berechnet man anhand der ermittelten Fehlertypen in Tabelle 2.1 die Wahrscheinlichkeiten eines Fehlers, erhält man abschließend die Werte in Tabelle 2.3:

c	Frequenz(c)	P(c)	P(t|c)	P(c) P(t|c)	%
actress	1343	0,0000315	0,000117	$3{,}69 \times 10^{-9}$	37 %
cress	0	0,000000014	0,00000144	$2{,}02 \times 10^{-14}$	0 %
caress	4	0,0000001	0,00000164	$1{,}64 \times 10^{-13}$	0 %
access	2280	0,000058	0,000000209	$1{,}21 \times 10^{-11}$	0 %
across	8436	0,00019	0,0000093	$1{,}77 \times 10^{-9}$	18 %
acres	2879	0,000065	0,0000321	$2{,}09 \times 10^{-9}$	21 %
acres	2879	0,000065	0,0000342	$2{,}22 \times 10^{-9}$	23 %

Tabelle 2.3: Ranking der Korrektur jedes Kandidaten für acress (Kernighan et. al.: 206)

In diesem Fall hätte *acres* mit ca. 45 % die höchste Wahrscheinlichkeit, da es auf zwei verschiedene Arten realisiert werden kann und sich die Wahrscheinlichkeiten daher addieren. Tatsächlich ist aber im Kontext *actress* gemeint. Am Ziel ist man an diesem Punkt also noch nicht. Um die richtige Worthypothese auszuwählen, benötigt man Informationen über das Umfeld, in dem das falsch geschriebene Wort auftritt.

2.1.6 Zuhilfenahme von Kontextinformation

Um zu dieser Lösung zu gelangen, muss Kontextinformation in Betracht gezogen werden. Analytische Ansätze wie Parsing haben auf diesem Gebiet bisher zu keinem brauchbaren Ergebnis geführt. Auch hier heißt also die Lösung Statistik.

Mithilfe von N-Gramm-Analyse-Techniken kann die Wahrscheinlichkeit eines Wortes in einer Wortkette bestimmt werden. Dies geschieht zumeist mit Dreiwortfolgen oder Trigrammen. Hierbei werden die letzten beiden Wörter betrachtet: *Der Zug verspätet sich um fünfzehn …* Diesen Satz könnte wohl jeder, der der deutschen Sprache mächtig ist, relativ eindeutig mit *Minuten* ergänzen, auch wenn die akustische Folge in der automatischen Spracherkennung oder das Schriftbild *Menuette, Minuskeln* oder *Ministranten* als wahrscheinlichste Kandidaten nahe legt. In der elektronischen Verarbeitung wird die Wortfolge *um fünfzehn* mit vorhandenen Trigrammen verglichen. Vorhanden sein könnten *um fünfzehn Uhr, um fünfzehn Minuten* und andere. Hat ein Trigramm eine ausreichende Auftretenswahrscheinlichkeit und passt die Beobachtung dazu, wird die Wortfolge ausgegeben.

Diese Technik kann auch auf Zeichenebene angewendet werden. Das einfachste aller N-Gramm-Modelle ist ein binäres Bigramm-Array. Es handelt sich dabei um ein zweidimensionales Array der Größe 26 x 26, dessen Elemente alle möglichen Zwei-Buchstaben-Kombinationen des Alphabets enthalten. Der Wert dieser Kombinationen ist entweder 0 oder 1, je nachdem ob diese Buchstabenfolge in wenigstens einem Wort des Referenzlexikons auftritt oder nicht. Ein entsprechendes Trigramm-Array hätte drei Dimensionen und deckte Drei-Buchstaben-Sequenzen ab oder die oben beschriebenen Dreiwortfolgen (Kukich 1992: 381).

Seit einiger Zeit werden Trigramm-Arrays auf Wortebene nicht mehr mit binären sondern mit probabilistischen Werten berechnet. Diese Statistiken wurden zunächst ausschließlich aufgrund großer Textkorpora berechnet. Es zeigte sich allerdings, dass Korrektursysteme nicht ausreichend hohe Fehlererkennungsraten lieferten.

Dieses Manko konnte korrigiert werden, indem die Trigrammtabelle direkt aus dem zu untersuchenden Text erstellt wurde. Auf diese Weise wurden das spezifische Vokabular und die thematische Struktur des Dokuments der orthografischen Korrektur zugänglich.

Morris und Cherry, die Mitte der siebziger Jahre mit probabilistischen Trigrammstatistiken experimentierten, entwickelten diese auf dem Dokument basierende Trigrammtabelle. Darüber hinaus berechneten sie für jedes Wort einen *Peculiarity Index* als Funktion der Trigrammhäufigkeiten dieses Wortes. Damit gelang es ihnen, nachzuweisen, dass 23 der 30 Wörter mit der größten »Seltsamkeit« tatsächlich falsch geschrieben waren (Kukich 1992: 382).

2.1.7 Korrektur multipler Fehler

Die von Kernighan et. al. benutzte Methode erlaubt nur die Korrektur von jeweils einem Fehler pro Wort. Um auch multiple Fehler in einem Wort berichtigen zu können, benötigt man Informationen über die *string distance*. Dieser Wert gibt an, wie ähnlich sich zwei Strings sind. Die hierfür verwendete Methode der dynamischen Programmierung wird unter anderem benutzt, um RNA- bzw. DNA-Stränge oder Gaschromatogramme für die Spektralanalyse zu vergleichen. In diesem Falle geht es um die Ähnlichkeit von Buchstabenfolgen. Im Falle eines fehlerhaft geschriebenen Wortes, das mehr als einen einfachen Fehler enthält, wählt man zur Korrektur den Kandidaten, der der beobachteten Zeichenfolge am nächsten kommt. Häufig wird nach den beiden Pionieren der Methode auf diese Modelle als *Damerau-Levenshtein-Metrik* referiert. Damerau implementierte erstmals 1964, Levenshtein ungefähr zur gleichen Zeit 1966 String-Distance-Modelle in der Rechtschreibkorrektur (Kukich 1992: 393).

Eine Variante dieses Modells namens *Minimum-Edit-Distance* hat sich durchgesetzt und findet sich bei Wagner und Fischer (Jurafsky 2000: 155). Die *Mini-*

mum-Edit-Distance ist die kleinste Zahl von Operationen, die nötig ist, um eine Zeichenfolge in eine andere zu transformieren. Der Abstand zwischen *Intention* und *Extension* ist beispielsweise drei, der Abstand zwischen *Intention* und *Inventur* vier Schritte groß:

Beispiel 1:		Beispiel 2:	
	intention		intention
Substitution t-s →		Substitution t-v →	
	intension		invention
Substitution i-e →		Löschen n →	
	entension		inventio
Substitution n-x →		Substitution o-r →	
	extension		inventir
		Substitution i-u →	
			inventur

Tabelle 2.4: Minimum-Edit-Distance

Indem jeder der drei möglichen Operationen Substitution, Auslassung und Einfügung der Wert 1 zugeordnet wird, ist eine einfache Gewichtung für die Distanz zweier Zeichenfolgen gegeben, auch Levenshtein-Distanz genannt nach ihrem Erfinder Levenshtein (vgl. Levenshtein, 1966). Die Levenshtein-Distanz ist also zwischen *Intention, Extension* 3 und zwischen *Invention, Inventur* 4. Levenshtein schlug eine alternative Methode vor, in der die Substitution nicht erlaubt ist, da sie durch Löschen und Einfügen erfolgen kann. Demnach wären die Levenshtein-Distanzen für *Intention/Extension* 6 und für *Intention/Inventur* 7. *Extension* würde gegenüber *Inventur* als Korrekturvariante bevorzugt werden.

Kombiniert mit den Fehlermatrizen aus Abschnitt 2.1.5 lässt sich nun ein »maximum probability alignment« berechnen, das die größte wahrscheinliche Ähnlichkeit zwischen der beobachteten Zeichenfolge und dem Korrekturkandidaten bestimmt, da die Werte für die einzelnen Operationen die Wahrscheinlichkeit angeben, mit denen sie auftreten. Sind nämlich die drei Substitutionen, die von *Intention* zu *Ex-*

tension führen, erheblich seltener als die vier Operationen, die *Intention* in *Inventur* überführen, kann auch Beispiel 2 die größere Ähnlichkeit aufweisen. In diesem Falle würde *Inventur* ausgegeben.

Die Levenshtein-Methode findet seit einiger Zeit Anwendung in Systemen, die mit automatischer Spracherkennung operieren. Sie ist nützlich bei der Ermittlung unterschiedlicher Schreibungen von Eigennamen in telefonischen Reservierungsautomaten und ähnlichen Anwendungen, in denen die genaue Orthografie eines Namens unbekannt ist.

Mit den beschriebenen Methoden steht bereits ein großes Inventar an statistischen Methoden zur Fehlerkorrektur bereit, die auf vielen verschiedenen Ebenen des Korrekturprozesses greifen und eine effektive Korrektur ermöglichen. Der grammatischen Realität der Sprache wurde jedoch dabei noch keine Rechnung getragen. Eine Annäherung an die linguistischen Probleme und Lösungsansätze im Bereich der Rechtschreibkorrektur erfolgt im Anschluss an dieses Kapitel.

2.2 Linguistische Anforderungen an die Fehlerkorrektur

Die oben beschriebenen Modelle sind seit Jahren im Einsatz und haben sich bewährt. Sie wurden allerdings für die englische Sprache erarbeitet und tragen den spezifischen Problemen des Deutschen wie Flexion oder Kompositabildung daher keine Rechnung. Im praktischen Gebrauch sind die Grenzen eines nur statistisch angelegten Ansatzes schnell offenkundig. Mag Fred Jelinek, ein Mitarbeiter des IBM-Sprachforschungsteams auch sagen: »Anytime a linguist leaves the group the recognition rate goes up« (Fred Jelinek, 1988, zit. nach Jurafsky 2000: 191). Für flektierende Sprachen und solche, die Wortbildung nach produktiven Mustern betreiben wie das Deutsche, ist ein linguistisch orientierter Ansatz unerlässlich, um zu

brauchbaren Ergebnissen zu gelangen, denn komplexe Wortbildungsprodukte wie *Donaudampfschifffahrtsgesellschaft* werden nie in einem Lexikon zu finden sein.

Das Englische ist im Gegensatz zum Deutschen eine weitgehend nicht flektierende Sprache. Stammflexion tritt kaum auf und die einzige Flexionsendung bei Verben und Substantiven ist das *-s*. Mit relativ wenigen Regeln lässt sich eine Morphologie des Englischen erfassen. Die deutsche Sprache flektiert ihre Wortklassen nach unterschiedlichen Kriterien und unterscheidet innerhalb dieser Klassen nach starker, schwacher und gemischter Flexion. Die Flexionstypen variieren je nach Wortklasse. So gibt es allein sechs verschiedene Flexionstypen der Substantive (Eisenberg [1], S. 158). Die Verben ordnen sich in zwei Hauptklassen, die starken und die schwachen. Die starken Verben sind aber in etwa 170 Paradigmen zu unterteilen (Eisenberg [1], S. 178), die kaum von einer computergestützten Morphologie zu verarbeiten sind. Sich dieses Wissen über das morphologische Verhalten zumindest teilweise zunutze zu machen, würde nicht nur das Lexikon entlasten und somit Speicherbedarf und Rechenzeit vermindern, sondern auch Möglichkeiten zur Korrektur von grammatikalischen Fehlern bieten.

Speicherzwänge waren auch die Motivation in den achtziger Jahren, die Speicherung morphologischer Varianten einzelner Wörter des Genus, Numerus, Kasus oder Person, Numerus, Tempus, Modus, Genus verbi zu vermeiden. Die einfachsten Ansätze ersetzten lediglich alternative Zeichenketten für die englische Sprache (*dictionary – dictionar**ies**, address*-θ – address***ed**, advise – advis**ed***). Um fehlerhafte Formengenerierung wie **adviseed* zu vermeiden, wurden Affixäquivalenzklassen entwickelt, die durch orthografische Ähnlichkeit motiviert waren (Kukich, 1992, S. 383). Später, mit der Entwicklung immer größerer Speicherkapazitäten, ging der Trend zur vollständigen Speicherung aller Flexionsformen.

2.2.1 Probleme der Modellierung von Flexion und Derivation

Die lexikalische Morphologie in Regeln zu fassen, mit denen die Wörter eines Lexikons automatisch generiert werden können, ist jedoch auch weiterhin äußerst reizvoll. Man bräuchte zur Formengenerierung nur noch ein Inventar an Grundmorphemen und Affixen sowie ein Regelwerk zu deren Kombination. Dabei darf man jedoch nicht außer Acht lassen, dass die lexikalische Morphologie von der Satzgrammatik verschieden ist. Die Formenbildung auf der Wortebene ist ungleich komplexer. »Neue Nominationseinheiten entstehen unter dem Druck gesellschaftlicher Bedürfnisse der Kognition und Kommunikation durch das Zusammenwirken semantischer und formativstruktureller Prozesse auf unterschiedliche Weise, z. B. durch Phraseologisierung und Terminologisierung wie auch andere Arten semantischer Umprägung, durch Entlehnung und durch Wiederbelebung untergegangener Ausdrücke« (Fleischer/Barz 1995: 2).

Die Wortbildung vollzieht sich also nur teilweise mithilfe des Morpheminventars und der dazugehörigen Regeln. Die semantischen und formativen Prozesse lassen sich nur sehr schwer in feste und vor allem einfache und überschaubare Regeln pressen. Nichtsdestotrotz kann das Lexikon eines sprachverarbeitenden Systems durch eine Formengenerierung im Bereich der Flexion bis hin zu sehr produktiven Wortbildungsmustern in seinem Speichervolumen stark eingeschränkt, in der Abdeckung der verarbeiteten Wortformen aber äußerst umfangreich sein. In der Praxis ist eine solche Umsetzung alles andere als leicht. Schon die Abgrenzung bereitet Schwierigkeiten, wie folgendes Beispiel verdeutlicht.

2.2.1.2 Affigierung

Die Homonymie bestimmter Affixe wie z. B. des Suffix *er,* das als Pluralsuffix (*Rind-er*) und als Derivationssuffix (*Läuf-er*) auftritt, stellt für die automatische Formengenerierung zwar kein unmittelbares Problem dar, da keine ungewollten, nicht wohlgeformten Wortformen entstehen. Bei der Erkennung der richtigen Form

spielt die Homonymie gleichwohl eine Rolle, da bei den Suffixen *er* oder *en* beispielsweise schwer zu entscheiden ist, welchen Zweck sie erfüllen. Hierbei können distributive Unterschiede helfen: Flexionssuffixe stehen immer am Ende, also hinter den Derivationssuffixen (Fleischer/Barz 1995: 5). Weiterhin ist die Beschaffenheit des Grundmorphems von Interesse. Im Beispiel fungiert das Suffix *-er* beim substantivischen Bestandteil als Pluralendung, beim verbalen als Derivationssuffix. Im Einzelfall lässt sich die Unterscheidung, welcher Formalismus vorliegt, relativ einfach bestimmen. Generalisiert man aber vom Pluralmorphem *-er* auf die Allomorphe des Pluralmorphems im Deutschen {*-e, -er, -s,* …}, vervielfacht sich die Zahl der Ambiguitäten.

2.2.1.3 Nichtkonkatenative Morphologie

Schwieriger wird die Modellierung, wenn man es nicht mehr mit der einfachen Verkettung von Morphemen zu tun hat (*Über-institut-ion-al-isier-ung-s-grad*), sondern mit nichtkonkatenativer Morphologie. Dabei kann es sich um Umlautung oder Ablautung handeln (*Vater* → *Väter; schieben* → *schob; stoßen* → *stieß*) oder um andere Phänomene wie subtraktive Morphologie oder Reduplikationen (*Abitur* → *Abi*).

Um sich der vorliegenden Fragestellung zu nähern, soll im Folgenden eine Formalisierung der Morphologie aufgrund von Finite-State-Transducern, basierend auf endlichen Automaten erklärt werden.

2.2.2 Modelle aus der generativen Linguistik

Die Sichtweise in der Einleitung zu diesem Kapitel war deskriptiv bzw. analytisch orientiert. In der generativen Morphologie ist die Herangehensweise genau umgekehrt. Es werden Modelle konstruiert, die aus einem festgelegten Regelinventar Wortformen generieren können. Diese Modelle müssen aufgrund der Fülle an möglichen Wortbildungen restriktiv bleiben, können aber für Problemfelder wie das vorliegende brauchbare Lösungen liefern.

Im Folgenden wird von wortbasierten, morphembasierten und realisierungsbasierten Ansätzen die Rede sein. In wortbasierten Ansätzen werden durch Regelanwendung aus vorhandenen Wörtern neue Wörter gebildet. Im morphembasierten Ansätzen werden segmentale Morpheme zu vollständigen Wortformen kombiniert. In realisierungsbasierten Ansätzen wird von der Bedeutung oder Funktion der Wortform ausgegangen. Entsprechende Regeln legen dann fest, wie die Wortform auszusehen hat (Carstensen, et. al. 2001: 180). Wie dies im Einzelnen aussieht, wird später behandelt.

Morphembasierte Ansätze sind formal sehr restriktiv, da sie nur eine Operation, die Verkettung von Morphemen zulassen, wenn auch genau festgelegt werden muss, welche Morpheme in welcher Reihenfolge verkettet werden können. Realisierungsbasierte Ansätze erlauben hingegen auch eine unmittelbare Umsetzung von nichtkonkatenativer Morphologie.

Weiterhin erlauben realisierungsbasierte Ansätze die Anwendung von Default-Regeln. Diese Regeln werden immer angewendet, wenn keine Regel mit höherer Priorität gegeben ist. Für die deutschen Pluralendungen könnte man folgende Regeln annehmen:

1.	(Plural – Nomen) X/*Kind, Ei, ...*	→	X + -er
2.	(Plural – Nomen) X	→	X + -e

Tabelle 2.5: Default-Regeln für die Plural-Bildung (Carstensen, et. al. 2001: 181)

X repräsentiert hier den Stamm, an den das Pluralmorphem *-er* bzw. *-e* angehängt wird. Für Regel 1 ist festgelegt, welche Substantive die Endung nehmen. Regel 2 ist für alle Substantive anwendbar und ist somit die Default-Regel. Damit es nicht zu Überschneidungen kommt, hat Regel 1 die höhere Priorität. Regel 2 wird, falls ein Lexem auftritt, für das eine andere Regel existiert (*Kind, Ei*), blockiert (Carstensen, et. al. 2001: 182). Ein anderer Einsatzbereich für Default-Regeln ist die Bildung des Infinitivs Präsens Aktiv auf *-en*. Indem man für jedes Flexionsparadigma die

Endung mit den meisten Synkretismen als Default definiert, spart man weitere Bildungsregeln ein, im konkreten Fall diejenigen für die 1. und die 3. Person Plural.

Weiter oben wurde bereits das Problem der Allomorphie angesprochen. In morphembasierten Ansätzen wird versucht, dieses Problem mithilfe von abstrakten Ausgangsrepräsentationen zu lösen. Im Falle von *Mutter, Mütter* kann man ein abstraktes Pluralsuffix *I* annehmen, dass der Wortform angefügt wird. Anschließend wird es durch eine entsprechende phonologische Pluralregel ersetzt. An der Oberfläche tauchen diese Symbole nicht auf.

Repräsentation	phonologische Regel
Mutter-I	
↓	u → Ü/_I
Mütter-I	
↓	I → ü/_ɪ
Mütter	

Tabelle 2.6: Pluralbildung mit Diakritika (Carstensen, et. al. 2001: 182)

Realisierungsbasierte Ansätze verzichten auf solche Symbole und setzen Alternationen wie die Pluralbildung mit diakritischen Symbolen unmittelbar anhand von morphologischen Merkmalen um. Es gibt noch mehr Bereiche, in denen die segmentale Phonologie zugunsten der autosegmentalen Morphologie in den Hintergrund geschoben werden kann. Prosodische Elemente wie die Silbe spielen beispielsweise bei der Kurzwortbildung die entscheidende Rolle.

[Ka] [tha] [ri] [na]	→	[Ka] [thi]
[A] [bi] [tur]	→	[A] [bi]
[Fun] [da] [men] [ta] [list]	→	[Fun] [di]

Tabelle 2.7: Kurzwortbildung auf –i (Carstensen, et. al. 2001: 182)

Die Regel könnte hier etwa lauten: Entferne so viele Segmente, dass unter Suffigierung von *–i* genau zwei Silben entstehen, wobei der Bezug auf die Silbe bei der Formbildung essentiell ist. Die Anzahl der abgetrennten Segmente ist dabei jedes Mal unterschiedlich.

2.2.3 Formale Sprachen und endliche Automaten

Der verbreitetste Formalismus bei der Formulierung morphologischer Regelsysteme sind die endlichen Automaten, in deren Rahmen auch die Finite-State-Transducer (FSTs) gehören, die schon mehrfach angesprochen wurden und in diesem Kapitel erklärt werden sollen. Die Automatentheorie gehört im Rahmen der theoretischen Informatik in den Bereich der formalen Sprachen. Für die maschinelle Verarbeitung natürlicher Sprachen haben diese sich als effiziente Mittel zur Beschreibung der jeweiligen Sprache oder des jeweiligen Sprachausschnitts erwiesen. Dazu benötigt man eine spezifische Grammatik.

Eine Grammatik *G* besteht zunächst aus zwei Alphabeten. Ein Alphabet enthält Terminalsymbole $\Sigma = \{a, b\}$, ein weiteres enthält Nichtterminalsymbole oder Variablen $\Phi = \{\alpha, \beta\}$. Nichtterminalsymbole dienen der Strukturierung und treten nicht an der Oberfläche auf, während die Terminalsymbole die Formelemente bezeichnen, aus denen der Text zusammengesetzt wird. Aus der Menge der Nichtterminalsymbole wird ein Startsymbol $S \in \Phi$ generiert. Als Letztes wird eine Regelmenge *R* benötigt, um die Terminalsymbole zu verketten (Carstensen, et. al. 2001: 60). Eine Grammatik $G = [\Phi, \Sigma, R, S]$ könnte folgendermaßen aussehen : G_1 = [{S, NP, EN, VP, V, Pron, N}, {Martin, Auftritt, seinen, inszeniert}, R, S] mit der Regelmenge R = {S → NP VP, NP → EN, NP → Pron N, VP → V NP, EN → Martin, N → Auftritt, V→ inszeniert, Pron → seinen} (Carstensen, et. al. 2001: 61). Eine solche Grammatik wird auch einfache Regelgrammatik genannt. Sie erzeugt die formale Sprache $L(G_1)$. Der Satz *Martin inszeniert seinen Auftritt* entstünde nach dieser Grammatik mit der Anwendung folgender Regeln:

S → NP VP	→ NP V NP
→ NP V Pron N	→ NP V Pron Auftritt
→ NP V seinen Auftritt	→ NP inszeniert seinen Auftritt
→ EN inszeniert seinen Auftritt	→ Martin inszeniert seinen Auftritt

Tabelle 2.8: Beispiel zur formalen Sprache mit der Grammatik G1 (Carstensen, et. al. 2001: 62)

Während Grammatiken Sprachen erzeugen, werden die Worte einer Sprache mithilfe von Automaten erkannt. Im einfachsten Fall wird nur mitgeteilt, ob der Ausdruck zur Sprache gehört oder nicht. Dies geschieht mit einseitig-linearen Grammatiken, die mithilfe von Ableitungsregeln und einer regulären Menge von Morphemen entsprechende Wörter generieren können. Solche Grammatiken sind entweder links-linear oder rechts-linear. Mit der folgenden rechts-linearen Grammatik G können die Wörter *unlehrbarkeit* und *unbelehrbarkeit* generiert werden. G: Φ = {S, A1, A2, A3, A4}, Σ = {un, be, lehr, bar, keit}, S, R = {S → un S, S → lehr A2, S → be A1, A1 → lehr A2, A2 → bar A3, A3 → keit A4, A3 → ε, A4 → ε }

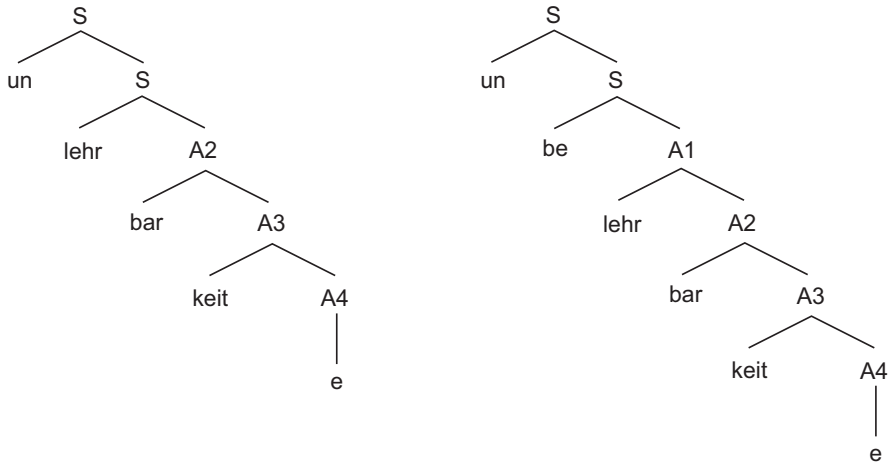

Abbildung 2.4: Rechts-linearer Ableitungsbaum für die Wörter Unlehrbarkeit, Unbelehrbarkeit (Carstensen, et. al. 2001: 65)

In einem Automaten werden Morpheme als mögliche Übergänge zwischen Knoten dargestellt. Mit solchen Modellen können gültige Derivationsprodukte erkannt und nicht wohlgeformte Ausdrücke ausgeschlossen werden. Eine etwas anschaulichere Darstellung findet sich in Abbildung 2.5 unten. Der Automat erhält das Wort als Eingabe und arbeitet es Morphem für Morphem ab. Jedes Wort fängt am Startknoten *S* an. In diesem Modell kann *bar* nur mit dem Verbstamm *lehr* verkettet werden. Nach *-bar* kann das Suffix *keit* kommen. *Un* ist nur als Präfix am linken Rand zugelassen und kann theoretisch beliebig oft wiederholt werden. Nur wenn das Wort

an einem der mit einem Doppelkreis markierten Endknoten ankommt, wird es als gültiges Wort akzeptiert.

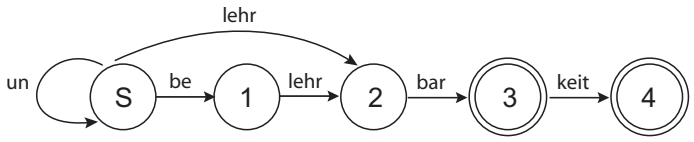

Abbildung 2.5: Akzeptanz aller mit den gegebenen Morphemen bildbaren Wörter (Carstensen, et. al. 2001: 66)

Im Folgenden wird eine einfache morphembasierte Modellierung von Flexion vorgestellt, die auch eine Abstimmung von phonologischen Regeln erlaubt. Als Beispiel wird ein Ausschnitt aus dem Paradigma des schwach flektierenden Verbs *waten* gewählt, der Flexion des Indikativ Präsens, und exemplarisch für ein Verfahren zur Formengenerierung der Verben dieser Flexionsklasse betrachtet. Um alle synthetischen Formen zu erfassen, müsste das Paradigma um den Konjunktiv Präsens sowie den Indikativ und den Konjunktiv Präteritum ergänzt werden. Das ist jedoch zum Verständnis des Verfahrens nicht notwendig.

In einem endlichen Automaten wie in Abbildung 2.6 lassen sich diese Formen wieder als Zustände und Zustandsübergänge darstellen. Die Kreise bezeichnen die Zustände 0, 1 und 2, zu deren Erreichung verschiedene Übergänge möglich sind.

Person	Sg	Pl
1	wate	waten
2	watest	watest
3	watet	waten

*Tabelle 2.9: Ausschnitt aus dem Verbparadigma von **waten***

Das Modell zur Flexion ist etwas einfacher als das zur Derivation. Die Suffixe werden an den Stamm angehängt. Daher gibt es nur einen möglichen Endzustand, zu dem aber vier verschiedene Übergänge führen. Zwei Übergänge werden eingespart, da zwischen der 1. und der 3. Person Plural und jeweils der 2. Person Singular und Plural Synkretismen bestehen. Der Stamm ist in diesem Modell für morphologische Funktionen nutzbar, aber nicht wortfähig, wenn keine Flexionsendung vorhanden ist.

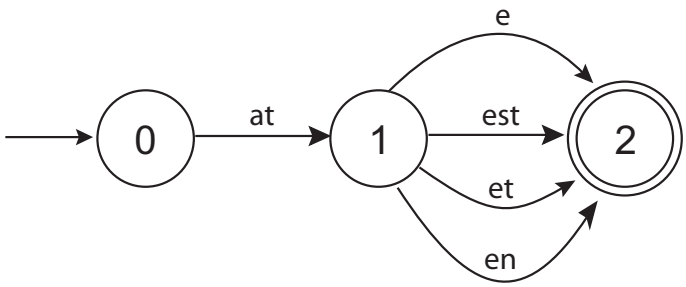

Abbildung 2.6 Morphologische Formen von waten in einem endlichen Automaten (Carstensen, et. al. 2001: 184)

Diese Darstellung segmentiert die Formen in die vorhandenen Morpheme. Ein Rechtschreibkorrektursystem erhält die Wortformen allerdings in der Praxis nicht vorsegmentiert, sondern als unstrukturierte Graphemketten. Um dies darstellen zu können, müssen neue Zustände eingefügt und vorhandene Übergänge durch eine Sequenz von Übergängen ersetzt werden:

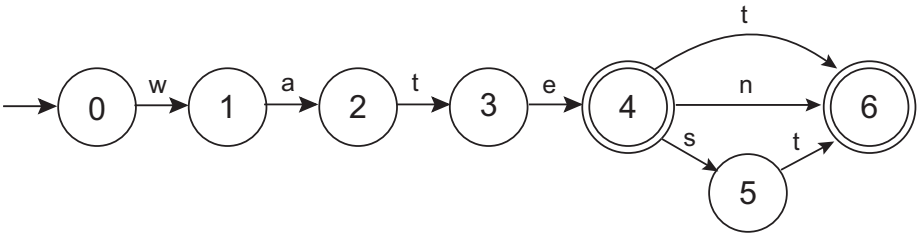

Abbildung 2.8: Graphemische Formen von waten in einem endlichen Automaten (Carstensen, et. al. 2001: 184)

Dieser Automat stellt in gewisser Weise nur eine ausführlichere Schreibweise des Automaten in Abbildung 2.6 dar. Da es sich hier nur um schematische Darstellungen handelt, ist es auch möglich, Automaten ganz ohne Zustände und Zustandsübergänge zu schreiben. Die Notation sieht dann wie folgt aus: wat (e | est | en | et | ε) (Carstensen, et. al. 2001: 184).

Die Anzahl der abgedeckten Formen lässt sich nun durch einfache Erweiterung steigern, etwa indem man weitere Verbstämme einfügt, die in das Paradigma, in diesem Falle der schwach flektierenden Verben, passen. Auch denkbar ist eine Ein-

fügung des Präteritumaffix *et* wie in Abbildung 2.8. Durch umfangreiche Listen von Verbstämmen kann durch solch ein einfaches Modell der Speicherbedarf des Lexikons beträchtlich reduziert werden. Für jedes Verb, das in die Klasse gehört, werden schon sieben Formen eingespart.

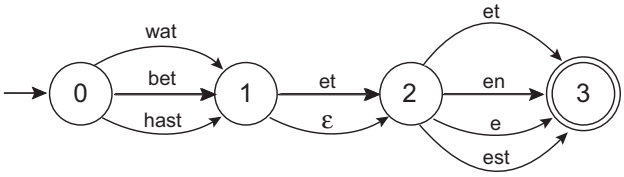

Abbildung 2.10: Verbformen von waten, beten, hasten (Carstensen, et. al. 2001: 185)

Mit den bisher beschriebenen Automaten kann man Wortformen auf ihre Wohlgeformtheit überprüfen, wie es für die automatische Rechtschreibkorrektur notwendig ist. Will man zusätzlich grammatische Korrekturen anbieten, benötigt man Informationen über die grammatischen Kategorien. Dies kann man durch Finite-State-Transducer erreichen. Diese speziellen Automaten bilden Ketten auf andere Ketten ab. Damit ermöglicht der Transducer sowohl Analyse (*beteten* → V Prät 1. Pl) als auch Formengenerierung (V Prät 1. Pl → *beteten*). So kann überprüft werden, ob eine gegebene Verbform im syntaktischen Kontext korrekt eingesetzt worden ist oder nicht.

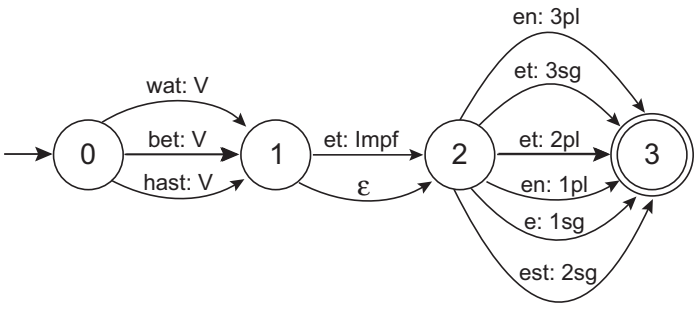

Abbildung 2.9: Verbformen von waten, beten, hasten in einem Finite-State-Transducer (Carstensen, et. al. 2001: 185)

Der Transducer besitzt mehr Übergänge als der Automat in Abbildung 2.8, da die Synkretismen, beispielsweise der 1. und 3. Person Plural, gesondert notiert werden müssen, um die grammatischen Kategorien abbilden zu können.

Eine offene Frage ist noch, wie verschiedene Bildungsmuster integriert werden können. Dazu ist man darauf angewiesen, phonologische Komponenten zu berücksichtigen. Die Endungen der Verbformen in der 2. Person Sg sind beispielsweise durch den letzten Laut des Stammes motiviert. Lautet der Stamm auf *s* aus, ist die Endung *t*. Endet er selbst auf *t*, ist die Endung *est*. Bei allen anderen Stämmen ist die Endung *st*.

Solche Fälle können ebenfalls in einem Transducer gelöst werden. Die Endbuchstaben werden eingelesen und lösen so die entsprechenden Übergänge aus. Selbst Fälle, in denen verschiedene Formen für eine Merkmalsstruktur zugelassen sind, lassen sich mit einem Transducer regeln. Die 2. Person Sg Präs lässt z. B. bei einigen Verben die Endungen *st* und *est* zu: *watest, watst; schriebest, schriebst*, etc. Solche Regeln werden in einer so genannten Zwei-Ebenen-Morphologie umgesetzt.

2.2.4 Kompositaanalyse

Die Literaturlage zur Kompositaanalyse ist sehr dünn, denn fast alle Erkenntnisse, die zur Zeit in der Rechtschreibkorrektur angewandt werden, wurden in den USA bzw. in englischsprachigen Untersuchungen ermittelt. In der englischen Sprache werden Komposita aber fast ausschließlich getrennt oder mit Bindestrich geschrieben. Durch das vorsegmentierte Auftreten ist also keine Strategie zur Dekomposition erforderlich. Für die deutsche Sprache ist die Komposition neben Flexion und Derivation sicher das wichtigste formbildende Element und bedarf eines Modells zur automatischen Analyse.

Komposita lassen sich nicht mit Automaten beschreiben, da ihre Anzahl und mögliche Form unendlich sind. Von dem System wird zunächst jede Zeichenkette

als Kompositum betrachtet, die nicht im Lexikon repräsentiert wird. Die Kompositaanalyse geschieht nun, indem alle Wortformen im Lexikon, die zur Bildung von Komposita zugelassen sind, also Substantive, Adjektive, Verbstämme und gegebenenfalls Konfixe etc. mit der zu untersuchenden Wortform verglichen werden. Wird eine Kette von Wortformen gefunden, die sich auf die Zeichenkette des Kompositums abbilden lässt, wird diese ausgegeben.

Die Auswahl analysierter Komposita in Tabelle 2.8 zeigt, dass es für viele Komposita unterschiedliche Möglichkeiten der Segmentierung gibt, die zu Ambiguitäten führen. Diese auszuräumen ist für die Rechtschreibkorrektur wichtig, da eine fehlerhafte Segmentierung beispielsweise keine Entscheidung über die richtige Form der Fuge mehr ermöglicht, da diese an der falschen Position angenommen wird.

Die richtige Variante erschließt sich in den meisten Fällen aus dem Kontext. Wenn in einem Geschäftsbericht oder -brief die Zeichenkette *Quartalstermine* auftritt, lässt sich mit dem Weltwissen, das jedem kompetenten Sprachnutzer zur Verfügung steht, erschließen, dass es sich um Termine bzw. die in der Wirtschaft gebräuchliche Zeiteinheit des Quartals handelt und nicht um eine Mine oder die Hamburger Alster. Von der Formseite lässt sich die Auswahl zugunsten von Variante 1 aber nur schwer begründen.

Mit einer Einteilung in semantische Klassen käme man hier weiter. *Quartal* und *Termin* sind eindeutig dem Bereich Wirtschaft zuzuordnen, während *Quart* in den Bereich der Musik (*Quartsextakkord*), des Fechtens oder der Maße gehört, *Alster* eine geografische oder Getränkebezeichnung sein kann und die *Mine* in die Bereiche Bergbau oder Schreibgeräte gehört. Schon dieses einfache Beispiel enthält Ambiguitäten für jeden Bestandteil und wird dadurch in der semantischen Analyse äußerst komplex. Ein Kategoriensystem wird niemals vollständig sein können und daher immer fehleranfällig bleiben und dem persönlichen Bias des Designers unterworfen sein.

Wortform	Variante 1	Variante 2
aufzugliedern	auf#gliedern	Aufzug#Liedern
Trampolin	Trampolin	Tram#Polin
dogmatisch	dogmatisch	Dogma#Tisch
Fischereibetrieb	Fischerei#Betrieb	Fische#Reibe#Trieb
Quartalstermine	Quartal#Termine	Quart#Alster#Mine
Literaturangabe	Literatur#Angabe	Literat#Uran#Gabe
Partnervermittlung	Partner#Vermittlung	Part#Nerv#Ermittlung
Musikerleben	Musiker#Leben	Musik#Erleben
Spezialgebieten	Spezial#Gebieten	Spezi#Alge#bieten
Treuhandvollmacht	Treuhand#Vollmacht	treu#Handvoll#Macht
Verbandsautonomen	Verband#Autonomen	Verband#Sau#Ton#Omen

Tabelle 2.10: Beispiele aus der Kompositaanalyse

Ähnlich kompliziert ist der Fall bei den Lexemen *Schreibtisch, Glastisch, Fototisch, dogmatisch.* Hier kommt die Problematik hinzu, dass die Konstituenten des Kompositums in unterschiedlichen semantischen Beziehungen zueinander stehen. Das Erstglied bezeichnet in den Beispielen entweder die Tätigkeit, zu der der Tisch dient, das Material, aus dem er besteht, oder Gegenstände, die auf ihm liegen. Selbst unter einem Dogmatisch lässt sich mit einiger Mühe ein Gegenstand vorstellen. Gemeint ist allerdings das Adjektiv mit der Bedeutung *streng an die Lehrsätze des Glaubens gebunden*. Die Fülle der semantischen Typen des Kompositums und deren Interpretierbarkeit sind also mannigfaltig. So sagt auch schon Wilmanns im Jahre 1896 zur Semantik des Kompositums, »dieselben in erschöpfender Weise in Gruppen einzuordnen ist kaum möglich.«[3] Heringer stellt 1984 allein 12 Interpretationen für das Kompositum *Fischfrau* vor.[4] Eine Nutzung vorhandener semantischer Kategorien ist also für die elektronische Analyse nicht möglich.

Mit Weltwissen ausgestattete Computersysteme sind seit einiger Zeit Gegenstand der Grundlagenforschung, aber noch längst nicht für praktische Anwendungen einsetzbar. Also muss zu einfacheren Mitteln gegriffen werden. Gerade für lange Kom-

posita werden wie im Beispiel häufig mehrere Varianten gebildet. Ein Werkzeug, um ungewollte Varianten wie *Dogma#Tisch* und *Tram#Polin* zu vermeiden, ist die Lexikalisierung des Adjektivs *dogmatisch* und des Substantivs *Trampolin*. Falls diese Lexeme bereits vorhanden sind, muss ihnen der Stellenwert als einzig gültige Variante eingeräumt werden. Dies kann z. B. wieder durch Default-Regeln für Varianten mit kleinerer Segmentzahl geschehen. Falsche Dekompositionen haben in der Regel mehr Bestandteile als die gewollten Varianten. Dies trifft auch auf fast alle Beispiele zu. In Einzelfällen ist die Fuge nicht klar zu bestimmen, wie in *Musiker#*Leben, bzw. Musik*#Erleben* oder *Druck#Erzeugnis* bzw. *Drucker#Zeugnis*. In solchen Fällen müssen beide Varianten erhalten bleiben, da eine Entscheidung zugunsten einer Variante nur mit kontextuellem Wissen möglich ist und dieses, wie oben gezeigt, nicht zu leisten ist.

Eine hinreichende Lösung für diesen Problembereich existiert bisher nicht und dürfte der Forschung auch in Zukunft noch Probleme bereiten. Mit den speziell angepassten Einzelfalllösungen lassen sich dennoch für die Praxis hinreichend brauchbare Ergebnisse erzielen. Langfristig wird man um ein semantisches Verfahren zur Kompositaanalyse jedoch kaum herumkommen.

3 KORPUSBASIERTER TESTLAUF – DITECT UND WORD 2000

In diesem Kapitel werden die Ergebnisse des korpusbasierten Testteils vorgestellt und anhand der in Kapitel 2 gewonnen Erkenntnisse analysiert. Wie bereits in der Einleitung erwähnt, wurde aus dem Bestand der *Neuen Presse* für die Untersuchung des Rechtschreibkorrektursystems DITECT ein Testkorpus zusammengestellt, das sich aus gleichmäßigen Teilen der fünf Ressorts *Seite Eins*, *Politik*, *Lokales*, *Wirtschaft* und *Sport* zusammensetzt. Die ausgewählten Artikel erschienen im Zeitraum vom 25. November 2002 bis zum 13. Januar 2003. Im Ganzen hat das Korpus eine Größe von über 50.000 Wörtern. Um einen Vergleichswert zur Leistung des Programms zu erhalten, wurde der Test neben DITECT mit dem Programm Word 2000 durchgeführt.

Die vorhandenen Rechtschreibfehler sowie Angaben zur Korpusgröße und zu den Beanstandungen der beiden Programme finden sich unten in Tabelle 3.1 (genauer Anhang A., S. 106). Dabei wurde zwischen der vollständigen Anzahl der Fehler und der Anzahl verschiedener Fehler unterschieden. DITECT findet insgesamt 1024 Fehler, Word sogar 1312. Insgesamt traten im Korpus 148 orthografische Fehler auf.

An den Daten fällt auf, dass die Anzahl der beanstandeten Schreibungen im krassen Gegensatz zur tatsächlichen Fehlerquote steht. Es werden in etwa zehnmal so viele Wörter beanstandet wie tatsächlich falsch geschrieben werden.

Ressort	Textgröße		Fehlerdetektion DITECT		Fehlerdetektion WORD		reale Fehler	
	Zeilen	Wörter	gesamt	verschiedene	gesamt	verschiedene	gesamt	verschiedene
EINS	2187	10062	114	90	144	103	35	22
POLI	2469	10134	126	104	110	86	39	32
LOKA	2443	10026	205	151	263	178	18	12
WIRT	2633	10080	148	109	204	141	39	36
SPOR	2451	10154	431	271	591	360	17	15
Gesamt:	**12183**	**50456**	**1024**	**725**	**1312**	**868**	**148**	**110**

Tabelle 3.1: Überblick über das Testkorpus

Für eine genauere Bewertung dieses Befunds müssen die Werte *Recall* und *Precision* eingeführt werden, die für die Evaluation von Suchmaschinen und ähnlichen Informationssystemen in den 50ern entwickelt wurden. Der *Recall* ist das Maß für die Vollständigkeit des Ergebnisses. Er ist wie folgt definiert:

$$\text{Recall} = \frac{\text{Anzahl der gefundenen orthografischen Fehler}}{\text{Anzahl der im Text vorhandenen orthografischen Fehler}}$$

Der Wertebereich liegt zwischen 0 und 1, wobei 0 das schlechteste und 1 das beste mögliche Ergebnis ist. Bei einem Recall von 1 sind also alle orthografischen Fehler gefunden worden, bei einem Wert von 0,75 sind es drei Viertel usw. In der Evaluation von Informationssystemen ist die *Recall*-Berechnung problematisch, da die Gesamtzahl der relevanten Dokumente in der Regel nicht bekannt ist, also abgeschätzt werden muss. In diesem Falle kann der Wert exakt bestimmt werden, da im vorhandenen Dokumentenbestand alle Fehler ausgezählt wurden.

Ist in einem großen Dokumentenbestand ein einziger Fehler vorhanden, es werden aber alle Wörter vom Programm als fehlerhaft markiert, so ist der Recall gleich 1. Dieser Wert reicht also alleine als Bewertungsmaßstab nicht aus. Um die Ballastquote mit in die Bewertung einzubeziehen, wird die *Precision* herangezogen. Mit diesem Wert können z.B. Eigennamen und nicht lexikalisierte Wörter, die korrekt

geschrieben sind, aber von dem Programm als fehlerhaft markiert werden, in die Bewertung mit einbezogen werden. Die Precision ist also ein Indikator für die Treffgenauigkeit der Fehlerdetektion und folgendermaßen definiert:

$$\text{Precision} = \frac{\text{Anzahl der gefundenen orthografischen Fehler}}{\text{Anzahl der im Text vorhandenen orthografischen Fehler}}$$

Auch die Precision hat einen Wertebereich von 0 bis 1. Ist die Precision hoch, stimmen Fehlerdetektion und vorhandene Fehler gut überein. Bei einem Precision-Wert von 0,3 sind 30 Prozent der beanstandeten Wörter falsch geschrieben. Es liegt nahe, die beiden vorgestellten Werte paarweise zu verwenden, da die Bewertung eines Ergebnisses nur so vollständig ist. Für die vorliegende Untersuchung erhalten wir folgendes Ergebnis:

	Recall verschiedene		Precision verschiedene		Recall gesamt		Precision gesamt	
DITECT	0,38	38 %	0,0580	5,8 %	0,297	30 %	0,0430	4,3 %
Word	0,44	44 %	0,0553	5,5 %	0,372	37 %	0,0419	4,2 %

Tabelle 3.2: Recall und Precision

Insgesamt traten bei einer Korpusgröße von 50.456 Wörtern 148 Fehler auf, von denen 110 Fehler verschieden waren. Von diesen 110 Fehlern erkannte DITECT 42 richtig, Word immerhin 48. Werden die Wiederholungen berücksichtigt, erkennt DITECT 44 korrekt und Word 55.

Die Erkennungsrate der tatsächlich vorhandenen Fehler ist relativ hoch. Sie beträgt nach den aktuellen Recall-Werten 38 Prozent bei DITECT und 44 Prozent bei Word. Dies liegt daran, dass falsch geschriebene Lexeme nicht im Lexikon zu finden sind und dadurch leicht aufgefunden werden können.

Es gibt allerdings noch anderes, was nicht im Lexikon zu finden ist, weshalb sich hier ein Blick auf die Precision-Werte lohnt. Es werden weitaus mehr Schreibungen beanstandet, als tatsächliche Fehler vorhanden sind. Von allen Lexemen, die DITECT als falsch erkennt, sind nur 5,8 Prozent wirkliche Fehler. Bei Word sind

es noch etwas weniger, nämlich 5.5 Prozent. Die Differenz fällt aber angesichts der Größenordnung kaum ins Gewicht. Führt man sich vor Augen, dass weniger als die Hälfte der Fehler gefunden wird, so wird die Genauigkeit, mit der die Programme arbeiten, äußerst gering.

	Recall, absolut	Recall, Prozent	Precision, absolut	Precision, Prozent
DITECT, EINS	0,45	45 %	0,111	11,1 %
DITECT, POLI	0,35	35 %	0,1154	11,5 %
DITECT, LOKA	0,31	31 %	0,0265	2,7 %
DITECT, WIRT	0,28	28 %	0,0917	9,2 %
DITECT, SPOR	0,57	57 %	0,0295	3,0 %
WORD, EINS	0,45	45 %	0,1262	12,6 %
WORD, POLI	0,31	31 %	0,1628	16,3 %
WORD, LOKA	0,38	38 %	0,0337	3,4 %
WORD, WIRT	0,22	22 %	0,0780	7,8 %
WORD, SPOR	0,54	54 %	0,0305	3,1 %

Tabelle 3.3: Recall und Precision in den einzelnen Ressorts

Tabelle 3.3 zeigt Recall- und Precision-Werte für die einzelnen untersuchten Ressorts. Die Werte für Recall sind hier einigermaßen gleichmäßig, während die Precision-Werte in den Ressorts *Lokales* und *Sport* erheblich einbrechen. In den Ressorts Seite Eins und Politik, die von der Struktur recht ähnlich sind, sind die Werte recht hoch. Die Gründe hierfür liegen vermutlich in dem Umfang, in dem Eigennamen auftreten. Die Fülle an Namen, die im Ressort Sport auftaucht (man denke allein an die 22 Spieler pro Fußballspiel plus Trainer), kann kein Rechtschreibkorrekturprogramm bewältigen. Noch dazu sind diese ständigen Änderungen unterworfen. Im Lokalressort ist die Erkennung regionaler Namen und Ortsbezeichnungen nicht von einem überregional orientierten Programm zu leisten. Man könnte allerdings erwarten, dass das Programm DITECT diesem Rahmen angepasst wäre. Dies ist nicht der Fall. Die Precision-Werte sind in beiden betrachteten Ressorts bei Word höher.

3.1 Bewertung der Fehlerdetektion

Auf Art und Gründe des Verhaltens der untersuchten Programme wird im Folgenden eingegangen. Tabelle 3.4 kategorisiert zunächst die beanstandeten Schreibungen nach einer Einteilung, die sich für das Verhalten der Programme als sinnvoll erwiesen hat. Im Ganzen ergeben sich die acht Bereiche einfache Lexeme, Wortschöpfungen (z.B. aus der Werbung), Komposita, Trennungsprodukte, Eigennamen, sonstige Zeichenketten außerhalb des Lexikons, grammatische Wendungen und Abkürzungen:

Problembereiche bei der Fehlerdetektion	DITECT						WORD 2000					
	EINS	POLI	LOKA	WIRT	SPOR	**GES.**	EINS	POLI	LOKA	WIRT	SPOR	**GES.**
einfache Lexeme	2	1	0	0	3	**7**	4	2	1	0	1	**8**
einfache Lexeme, großgeschrieben	5	3	4	0	6	**18**	0	0	0	0	1	**1**
Lexem mit Binnenmajuskel	0	0	0	2	0	**2**	1	0	0	2	0	**3**
Wortschöpfung	3	1	0	3	1	**9**	3	0	0	1	2	**7**
Derivationsprodukt	0	1	1	0	1	**2**	1	1	7	2	2	**12**
Komposita	12	15	18	26	8	**80**	1	4	4	6	8	**24**
Komposita mit Bindestrich	3	2	5	5	4	**19**	17	12	31	26	21	**107**
Komposita mit s-Fuge	4	12	9	8	4	**37**	1	2	0	0	1	**4**
Erweiterung mit Bindestrich	1	2	1	2	0	**6**	0	0	0	0	0	**0**
Trennung eines Wortes	1	5	1	1	1	**9**	1	5	2	0	2	**10**
geografische Namen	4	6	17	4	22	**53**	7	3	20	8	43	**81**
Namen von Institutionen, Firmen, etc.	4	1	4	8	3	**20**	5	2	11	28	14	**60**
Personennamen	33	36	72	28	208	**377**	41	35	69	38	249	**432**
Rechtschreibfehler	9	12	5	9	5	**40**	11	12	7	13	11	**54**
Fremdsprachliche Begriffe	3	3	7	3	0	**15**	2	5	18	6	1	**31**

Problembereiche bei der Fehlerdetektion	DITECT						WORD 2000					
	EINS	POLI	LOKA	WIRT	SPOR	**GES.**	EINS	POLI	LOKA	WIRT	SPOR	**GES.**
Internet-/E-Mail-Adresse	2	0	2	6	1	**11**	1	0	2	6	1	**10**
Gr. Wendung (Inkorporation von zu)	0	0	0	0	2	**1**	0	0	1	0	1	**2**
Wiederholung	3	1	2	2	0	**8**	3	1	2	2	0	**8**
Abkürzung	1	3	2	2	2	**10**	2		2	2	2	**8**
Akronym	0	0	1	0	0	**1**	2	2	1	1	0	**6**
	51	**55**	**90**	**50**	**217**	**725**	**62**	**55**	**101**	**68**	**265**	**868**

Tabelle 3.4: Überblick über die Fehlerkategorien

Auf diese Bereiche wird in den folgenden Unterkapiteln Bezug genommen. Die Tabelle zeigt, dass alle Problembereiche, die in Kapitel 2 beschrieben wurden, tatsächlich auftreten. Derivationsprodukte treten in der Fehlerdetektion recht selten auf, während Komposita einen großen Anteil an den beanstandeten Wortformen haben. Die Erweiterung mit Bindestrich und Worttrennungen stellen für automatische Korrekturprogramme die schon beschriebenen Probleme dar, da vor der Korrektur die Wortform wieder korrekt rekonstruiert werden muss:

Wirtschafts- und Arbeitsminister → Wirtschaftsminister und Arbeitsminister

Da in allen Textsorten die Eigennamen eine große Rolle spielen, stellt dies eine wichtige Kategorie der Fehlerdetektion dar. Die Zahlen zeigen eindrucksvoll den Umfang, den gerade die Kategorie der Personennamen unter den beanstandeten Wortformen hat. Ein ähnlicher Befund liegt bei der Beanstandung der fremdsprachlichen Begriffe und Internetadressen vor. Die Beanstandung dieser Begriffe geht darauf zurück, dass sie nicht im Lexikon angelegt sind und es keine Regel zu deren Generierung gibt. Daher werden sie als falsch geschriebene Wortformen

gekennzeichnet. Da die Rechtschreibfehler ähnlich wie Internet-Adressen und fremdsprachliche Bezeichnungen Zeichenketten darstellen, die nicht im Lexikon vorhanden sind und nicht mit formalen Regeln generiert werden können, wurden sie hier aufgenommen.

Zur Kategorie der grammatischen Wendungen zählen neben Wortauslassungen und Worteinfügungen sowie Kongruenz- und Topologiefehlern genau genommen auch die Bindestrichergänzungen. Ihnen wurde aber aufgrund der aus der Trennung resultierenden besonderen Probleme bereits ein Platz zugewiesen. In dieser Kategorie sollten sich die Fehler finden, die auf die Verletzung syntagmatischer Beziehungen der Kongruenz und Rektion zurückgehen. Dies ist jedoch nicht der Fall, da solche Fehler von den Programmen in der Regel nicht gefunden wurden. Sie gehören in den Bereich Fehlerkorrektur, der im Anschluss in Kapitel 3.2 behandelt wird. Die Kategorie umfasst daher nur Inkorporationen von *zu* wie in *hinterherzustiefeln*.

Die letzte Kategorie ist ein Sonderfall und behandelt die Abkürzungen und Akronyme. Ähnlich wie die Eigennamen können sich die Angehörigen dieser Kategorie nicht vollständig im Lexikon befinden und werden daher häufig beanstandet. Eine Möglichkeit, die häufig benutzt wird, besteht darin, sie zu ignorieren. Word ignoriert beispielsweise vollständig großgeschriebene Zeichenketten.

Auf das genaue Verhalten der Programme in den einzelnen Kategorien wird weiter unten eingegangen. Im Folgenden wird die quantitative Verteilung näher betrachtet.

3.1.1 Verteilung der beanstandeten Lexeme in Prozent

Die folgende Tabelle gibt den prozentualen Anteil aller beanstandeten Schreibungen grafisch wieder.

Diese Statistik zeigt das Missverhältnis, in dem tatsächliche Rechtschreibfehler und beanstandete Schreibungen im Ganzen stehen. Sehr dominant sind die Per-

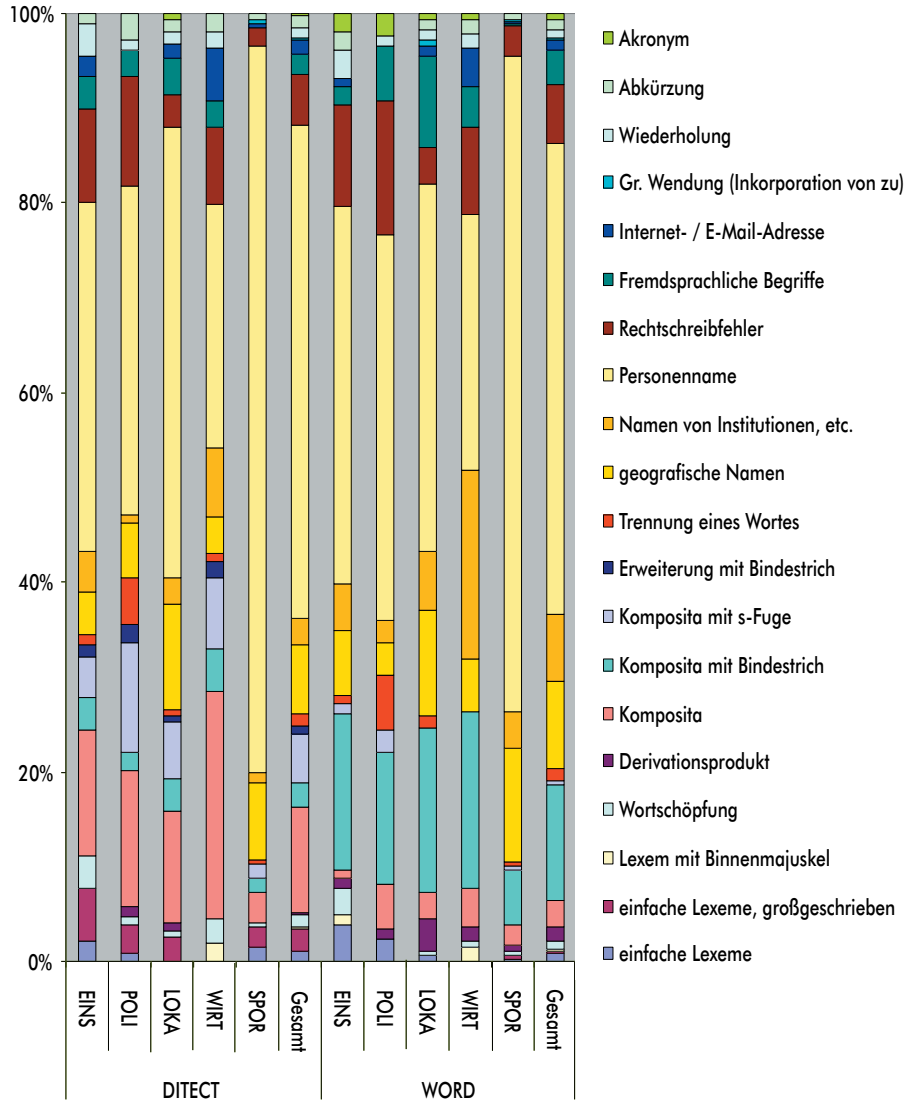

Tabelle 3.5: Statistik der als fehlerhaft erkannten Wörter in Prozent

sonennamen, die in jedem Ressort zwischen 25 und 77 Prozent der beanstandeten Wortformen ausmachen. Im Ganzen sind bei DITECT 52 %, bei Word 50 % der beanstandeten Wörter auf Personennamen zurückzuführen. Zusammen mit den geografischen Namen und Namen von Institutionen, Schiffen etc. kommen wir auf

einen Wert von 62 % bei DITECT und 66 % bei Word, die ausschließlich auf die Existenz von Namen zurückzuführen sind.

Im Bereich der Komposita zeigt sich, dass DITECT besonders viele einfache Komposita und in etwas geringerem Maße Komposita mit s-Fuge beanstandet. Komposita mit Bindestrich sind hier nur zu einem Bruchteil beanstandet worden. Word hingegen beanstandet kaum echte Komposita, es finden sich aber erstaunlich viele Komposita mit Bindestrich unter den als falsch markierten Wörtern. Dies ist ein interessanter Befund, der in Kapitel 3.1.5 näher beleuchtet wird.

Des Weiteren beanstandet DITECT mehr Substantivierungen und andere Großschreibungen sowie Abkürzungen, während Word die Beanstandungen schwerpunktmäßig bei Akronymen und fremdsprachlichen Ausdrücken durchführt. E-Mail- und Internet-Adressen sowie Wortwiederholungen und Abkürzungen werden von beiden Programmen gleichermaßen beanstandet.

3.1.2 Verteilung der beanstandeten Lexeme in absoluten Zahlen

Die prozentuale Darstellung oben ermöglicht einen guten Vergleich zwischen den einzelnen Kategorien. Diese Darstellung trägt allerdings der Tatsache keine Rechnung, dass Word insgesamt etwa 20 Prozent mehr Fehler gefunden hat als DITECT. Dieser Sachverhalt verzerrt das Ergebnis, weshalb in Tabelle 3.6 (siehe nächste Seite) die Werte in absoluten Zahlen angegeben sind. Hier wird deutlich, dass Word in allen Ressorts außer im Ressort Politik mehr Wörter zu beanstanden hat als DITECT.

Wenn es in der prozentualen Sichtweise zunächst so aussah, als würde DITECT mehr Namen beanstanden, so zeigt sich, dass DITECT zwar prozentual leicht höhere Werte im Bereich der Eigennamen hat. Diese erklären sich aber nur aus der geringeren Anzahl an insgesamt beanstandeten Fehlern. Absolut beanstandet Word mehr Eigennamen.

Die Verteilung der Beanstandungen ist nichtsdestotrotz relativ gleich. Es ergibt sich ein leichter Unterschied in den Ressorts *Seite Eins* und *Politik*. Diese Ressorts sind aber hinsichtlich des verwendeten Stils und der Terminologie sehr ähnlich. Ansonsten muss die Ähnlichkeit der Ergebnisse auf die Textsorte und das jeweils spezifische Vokabular zurückgeführt werden. In Artikeln aus dem Ressort Sport werden beispielsweise häufig am Ende die beteiligten Sportler wiedergegeben, was

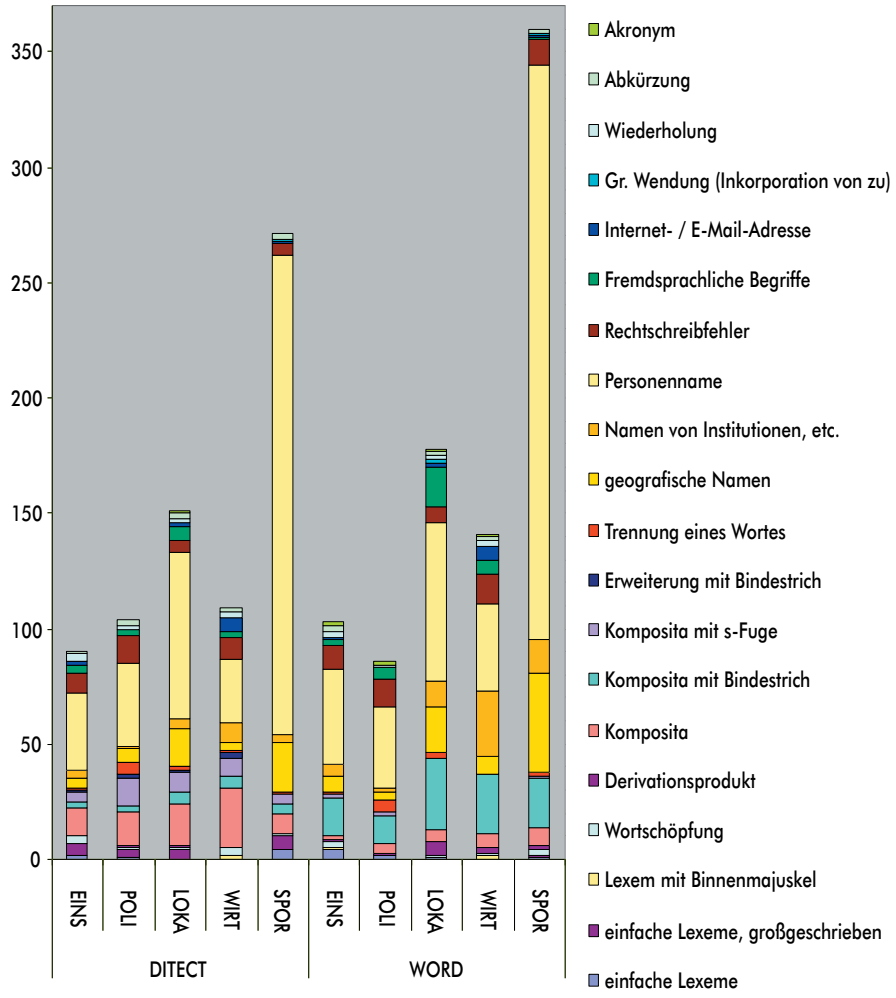

Tabelle 3.6: Statistik dre als fehlerhaft erkannten Wörter in absoluten Zahlen

in einem starken Anstieg der beanstandeten Wortformen, insbesondere der Eigennamen, resultiert.

Im Folgenden werden die einzelnen Fehlerkategorien, die oben aufgeführt wurden, eingehender untersucht und der Versuch unternommen, Gründe für das Verhalten der Programme zu finden.

3.1.3 Einfache Lexeme – Groß- und Kleinschreibung

Der erste Komplex betrifft nicht zusammengesetzte Lexeme des Grundwortschatzes. Da beide Programme zwischen Groß- und Kleinschreibung unterscheiden, wird diese Unterscheidung auch hier getroffen. So ergeben sich drei Unterkategorien. Die erste betrifft einfache Lexeme, bei DITECT vor allem Verbformen der Umgangssprache oder Konjunktive (*erfolge, entrutschte, daddelt*), bei Word Zahlwörter (*dutzende, hunderte, tausenden*). Das Korrektursystem von Word fordert hier Großschreibung der Zahlwörter.

Die zweite Unterkategorie betrifft substantivierte oder am Wortanfang großgeschriebene Lexeme, die aufgrund dieses Umstandes offenbar beanstandet wurden (s. Tab. 3.7). Die letzte Unterkategorie betrifft Lexeme mit Binnenmajuskel (*Auto-Future, BierRiese*), die aufgrund der offensichtlichen Schwierigkeit mit Majuskeln umzugehen hier mit behandelt werden. Im Falle von *BierRiese* muss sicherlich von einem Fehler ausgegangen werden, da es sich nicht wie bei *AutoFuture* um einen Namen handelt. Lexeme mit Binnenmajuskel werden von beiden Programmen immer als Fehler markiert.

Der Umgang mit Groß- und Kleinschreibung bei DITECT ist problematisch. Es treten einige Adjektive auf, die kleingeschrieben vom Programm korrekt erkannt werden, in der substantivierten Form allerdings als Fehler markiert werden. Hier wird offenbar bei großem und kleinem Buchstaben von zwei verschiedenen Zeichen ausgegangen. Die substantivierten Adjektive *Unsolides* und *Sportliches* werden

nicht gleich behandelt. Im ersten Fall schlägt DITECT Adjektive vor, im zweiten Fall sind es substantivische Personenbezeichnungen. Zusätzlich offenbart DITECT Schwächen bei der Erkennung von Flexionen. Auf diesen Umstand wird im Kapitel 3.2 näher eingegangen.

MICROSOFT WORD 2000			DITECT		
Lexem	Vorschlag 1	Vorschlag 2	Lexem	Vorschlag 1	Vorschlag 2
Hartz	Harzt	Harte	Freiheitlichen	freiheitlich	Freiheitlichkeit
Beust	Beulst	Beugst	Unsolides	unsolides	unsolide
ulrich	Ulrich	–	Sportliches	Sportliche	Sportlicher
Wittke	Wittre	–	–	–	–

Tabelle 3.7: Beispiele – Groß- und Kleinschreibung

Word trifft auch die Unterscheidung zwischen Groß- und Kleinschreibung, dies schlägt sich allerdings nicht in der Fehlerdetektion, sondern lediglich in der Wahl der Verbesserungsvorschläge wieder. Nach dem Verfahren, dass nur Vorschläge angegeben werden, die dem markierten Lexem am nächsten stehen, sind die Verbesserungsvorschläge bei Substantivierungen in der Regel auch Substantive oder Substantivierungen (siehe Tabelle 3.7).

3.1.4 Derivation und Wortneuschöpfungen

Der zweite Problembereich betrifft Derivationen und andere Wortschöpfungen, die nicht im Lexikon der Programme zu finden sind und nicht generiert werden können. Die Fälle, die bei DITECT auftreten, sind eher marginal. Es handelt sich um Diminutive (*Programmchen, Teestübchen*). Hinsichtlich der Akzeptabilität muss bemerkt werden, dass die Variante *Programmchen* gegenüber der Variante mit Umlautung des Stammvokals zumindest die markierte Version darstellt. Ob hier tatsächlich ein Fehler vorliegt, ist eine Stilfrage.

Bei Word treten vermehrt deverbale Substantivbildungen mit dem Suffix *er* auf (*Zögerer, Nasenstupser, Teppichknüpfer, Goldschürfer, Anscheinserwecker*). Zur Lösung wäre ein Modell zur Generierung des deverbalen Substantivs auf *er* denkbar. Zu-

mindest in den angegebenen Beispielen würde dies funktionieren. Wollte man diese Regel auf alle Verben erweitern, würden aber morphologische Probleme auftreten. Die Stammvokale einiger Verben müssen beispielsweise bei der Sustantivbildung umgelautet werden, andere nicht: *saufen – Säufer, rauchen – Raucher; zanken – Zänker, tanken – Tanker*. Andere Verben wechseln den Vokal: *singen – Sänger*. Zudem bilden generell nur solche Verben Substantive auf *er,* bei denen das Subjekt als Agens auftritt. Es gibt aber auch Ausnahmen: *Anlieger, Teilhaber* (Eisenberg [1]: 264). Andere Verben sind vollkommen von der Substantivbildung ausgeschlossen: **Erstauner, *Begegner*. Oder sie bilden diese in anderer Form, weshalb die Bildung auf *er* blockiert ist: **Glauber, Gäubiger*.

Solche Regularitäten der expliziten Derivation wie eine Wortbildungsregel für das Suffix *er* mit verbaler Basis lassen sich nur schwer formulieren. Man könnte beispielsweise die Verben kennzeichnen, die kein *er*-Suffix bilden, die Umlaute bilden, etc. Dazu benötigt man aber zuerst einmal Informationen über die Wortart, um die Verben zu lokalisieren. Solche Part-of-Speech-Tags gehören aber nicht zum Standard, da sie Speicherplatz benötigen und dieser bei der Lexikonerstellung kostbar ist. Zusätzlich braucht man einen Identifikator für den Verbstamm und die Informationen für die Derivation verschiedener Wortbildungsmuster, die bei den diversen produktiven Mustern des Deutschen recht zahlreich sind. Der Arbeitsaufwand ist erheblich geringer, wenn die Wortformen einfach ins Lexikon aufgenommen werden. Allerdings wird das Lexikon dadurch weniger flexibel, da kaum alle Wörter, die gebildet werden können, ins Lexikon aufgenommen würden. Dies führt wie im vorliegenden Fall zur Nichterkennung einiger Wortformen.

Verbneubildungen aus Adjektiven und Substantiven stellen für beide Programme ein Problem dar (*schwächelt, pfälzelt*). Dieses Verbbildungsmuster ist nur eingeschränkt produktiv, aber sowohl bei substantivischen, adjektivischen als auch bei verbalen Basen anwendbar: *Pfalz – pfälzeln, krank – kränkeln, Haufen/häufen – häu-*

feln, kochen – köcheln. Hier gilt das Gleiche wie für das Suffix *-er*. Man braucht eine umfassende Kategorisierung aller Lexeme, bevor eine solche Wortbildungsregel formuliert werden kann.

Die Rechtschreibprüfung von Word 2000 erkennt weiterhin nicht in allen Fällen das femininumanzeigende Suffix *in* (*Inhaberin, Besitzerin*). Diese Begriffe sollten, wenn sie auch nicht aus einem besitzanzeigenden männlichen Substantiv generiert werden können, zumindest im Lexikon verzeichnet sein, da ihre Verwendung recht häufig zu erwarten ist.

Weitere beanstandete Wortbildungen sind Wortschöpfungen aus der Werbung, die nicht ins Lexikon aufgenommen werden, wie z.B. Verbbildungen aus substantivischen Markenbezeichnungen (*gilden, gildet, flenst*), aktuelle Wortschöpfungen (*Teuro*) und Zusammenschreibungen (*niegelnagelneue*). Dies betrifft sowohl DITECT als auch Word, da Regionalität oder Aktualität dieser Wortbildungen von den Lexikographen nicht abgebildet werden kann.

3.1.5 Komposita

Der Bereich der Komposita ist ebenso wie die Derivation ein weites Feld und bedarf einer genauen Betrachtung. Zunächst wird eine grobe Einteilung in einfache Komposita, solche mit s-Fuge oder Tilgung des auslautenden *-e* im Erstglied und Komposita mit Bindestrich getroffen. Die Einteilung ist problematisch, da sie nicht eindeutig getroffen werden kann. Begriffe wie *burgenländische* oder *Migrantinnengruppe* können nicht eindeutig dem Gebiet Komposita oder Derivation zugewiesen werden, da sie Produkte beider Wortbildungstypen sind. In solchen Fällen wird die Kategorisierung nach der Hierarchie des Wortbildungsprozesses durchgeführt. Demnach ist *burgenländische* ein Derivationsprodukt von *Burgenland* und *Migrantinnengruppen* ein Kompositum aus *Migrantinnen* und *Gruppen*. Dass das Problem

bei der Erkennung von *Migrantinnengruppen* an der nicht erkannten Derivation und Flexion liegen könnte, spielt dabei zunächst keine Rolle.

3.1.5.1 Komposita ohne s-Fuge

Als Fugenelemente kommen in dieser Kategorie (e)n, es, e, er, ens und -θ in Frage, also alle Interfixe, abgesehen vom *s*, die mit deutschen Erstgliedern auftreten (Fleischer/Barz, 1995: 138). Die Kategorie umfasst neben dem Großteil rein substantivischer Komposita auch solche mit partizipialem oder verbalem Kopf (*fitgespritzt, kaputtreden, bloßstellen*) oder Modifikatoren anderer Wortarten wie Konfixen oder Partikeln (*Polarfüchse, umhergefahren, Hinserie*). Wie bereits im Unterkapitel zur Statistik erwähnt, treten Defizite bei der Erkennung von Komposita dieser Kategorie vorwiegend bei DITECT auf. Folgende Tabelle zeigt die nicht erkannten Komposita beider Programme sowie eine Auswahl nicht beanstandeter Komposita aus dem Testkorpus.

DITECT erkennt nicht		WORD erkennt nicht	DITECT und WORD erkennen
Absatzhit	Postsparbuchs	Bürgerämtern	Ansprechpartner
Abwehrkettenglied	Pulverschnees	Lichtergirlanden	Arbeitgeberverbänden
Aktienstrategin	Rodelspaß	Migrantinnengruppe	Atomwaffenmächte
Branchenausblick	Schmusekanzler	Toreschießen	Aufbruchstimmung
Chefkandidat	Schrotttankers		Chefredakteure
Dosenpfandes	Seezungen		Dosenpfandzeitalter
Farbenkarussell	Sorgenjahr	Sternsinger	Eliteeinheiten
Filmbasis	Spaltmaterialien	Christsozialen	Gazastreifens
Flutopferhilfe	Spritzenspezialist		Grundsatzfragen
Geheimdienst-erkenntnisse	Stahlschlitten	Fibrexnylon	Horrornachrichten
Goldzertifikat	Tankerwrack	Finanzdienstleisters	Interessengruppen
Guidomobil	Truppeneinsatz-versorger	Investmentbanking	Kinderbetreuung
Hochwasserflut	Türkeifrage	Parteivize	Kulturkreises
Kassentresen	umhergefahren		Lampenfieber
Kinogutschein	Vergabestopp	superklasse	Luftmatratzen
Koppelverträgen	Waffenbericht		Morgenmantel
Landesbischöfin	Waffendeklaration	fitgespritzt	Niedersachsenligisten

DITECT erkennt nicht		WORD erkennt nicht	DITECT und WORD erkennen
Mobilfunk-unternehmens	Waffendossier	kaputtreden	Parketthersteller
Modellstandort	Westgelände	wegmobben	Parteiausschluss-verfahren
Nachnutzer	Weststadtturnier		Plauschangriff
Nachtspringen	Winterhoch	Hinserie	Politikerleben
Narkosegas	Zugschiene		Politprominenz
Normalkameras	Zugtarife	dazugewonnen	Pulverschnee
Notdienst-vereinbarungen			Rentenreformkonzept
Notlichter	Postbox	Vorabbericht	Schnauzbartträger
Nullzinskrediten	Sportshirt		Sechsjahreshoch
Ölnebel	superklasse	Vielnutzern	Überflugspur
Originalberichts	drahtverstärkter		Umweltauflagen
Parteiriege	hingewerkelt		Unterschriftenaktion
Partyservices	nachgenutzt		Wahlauftaktrede
Polarfüchse	reingefeiert		Weltöffentlichkeit

Tabelle 3.8: Komposita

Die Fülle der Komposita, die von DITECT nicht erkannt werden und somit als Fehlerkandidaten gekennzeichnet werden, legt zunächst den Schluss nahe, dass keine Kompositaanalyse durchgeführt wird und somit alle nicht lexikalisierten Komposita als falsch geschrieben betrachtet werden. Die Liste der Komposita, die von beiden Programmen nicht beanstandet werden, widerspricht dieser These. Es fällt auf, dass bei DITECT vermehrt Flexionsformen beanstandet werden. *Dosenpfandes* wird beanstandet, *Dosenpfandzeitalter* allerdings nicht. *Pulverschnees* wird als falsch geschrieben vermutet, *Pulverschnee* stellt keinen Grund zur Sorge dar. Es scheint vor allem die Genitiv- oder Pluralmarkierung zu sein, die die Beanstandung zur Folge hat. Immerhin taucht das Genitiv- oder Plural-s in den 63 substantivischen Komposita bei DITECT neunmal auf, in den 40 Beispielen nicht beanstandeter Komposita nur zweimal.

Der Verdacht, dass das Problem der Kompositaerkennung durch Lexikalisierung aller gebräuchlichen Komposita gelöst wird, erhärtet sich durch die semantische

Identität der korrekt erkannten Komposita. Diese scheinen aus dem Bereich Politik zu stammen oder häufig verwandte Nominationseinheiten darzustellen (*Gazastreifen, Luftmatratze, Morgenmantel*). Ungewöhnliche Beispiele wie *Schauzbartträger* oder neue Wortbildungen wie *Dosenpfandzeitalter* widersprechen dem jedoch. Diese Frage ist anhand der Fehlerdetektion nicht eindeutig zu klären und wird später wieder aufgegriffen.

3.1.5.2 Komposita mit s-Fuge oder Tilgung des auslautenden -e

Komposita mit s-Fuge und solche mit Tilgung des auslautenden *e* wurden gesondert kategorisiert, da diese, wenn sie zerlegt werden, im Gegensatz zu Komposita mit anderen Fugen keine wohlgeformten Wortformen der deutschen Sprache mehr darstellen (**Meinungs Freiheit, *Elb Mündung*). Die s-Fuge tritt regelmäßig nach den Substantivierungssuffixen *keit, heit, igkeit, tum, schaft, sal* und *ling* sowie nach *en* (substantivierter Infinitiv) und den Fremdsuffixen *ion, ität* auf (Eisenberg [1]: 232). Es finden sich allerdings auch zahlreiche andere Erstglieder, darunter viele Grundmorpheme, bei denen das Fugen-s auftritt, so z.B. *Tag, Amt, Schiff, König*. In diesen Fällen ist der Ursprung der Fuge, die Kasusmarkierung des vorangestellten Genitivattribut noch deutlich sichtbar (Fleischer/Barz: 139).

Aufgrund der Überlegungen in Kapitel 3.1.4 zur Derivation kann man annehmen, dass die Schwierigkeit im Falle *Anscheinserwecker* nicht in der s-Fuge zu suchen ist, sondern in der deverbalen Substantivbildung auf *er*, die schon im Kapitel zur Derivation behandelt wurde. *Anscheinserwecken* wird von Word erkannt, so dass sich die Zahl der nicht erkannten Komposita mit s-Fuge auf zwei reduziert. Auch das Beispiel *Aufsichtsratsvize* offenbart keine Schwierigkeiten hinsichtlich des Fugen-s. Bereits in der vorigen Kategorie akzeptierte Word die Form *Parteivize* nicht, was darauf schließen lässt, dass keine Kompositabildung mit *Vize* erlaubt ist. Als einzelnes Lexem wird es nicht beanstandet.

DITECT erkennt nicht		DITECT und WORD erkennen
Anscheinserwecker	Höchstgeschwindigkeits-kursen	Altersversorgung
aufkommensstarker	Traditionsfleischerei	Arbeitsplatzverluste
Betriebsassistentin	Traditionsladen	Aufsichtsratsvorsitzender
Dienstagsdemonstration	vorgangsbezogen	Beratungsgesellschaft
Endbilanz	Einstiegsinteressenten	Besitzstandswahrung
Endredaktion	Einstiegsinvestitionen	Elbmündung
handlungsunwillig	Feiertagsziele	Festtagsreisen
Kanzleramtspapier	Produktionsalternativen	Finanzierungsgrundlage
Kommissionskonzeptes	Sauerstoffversorgungs-systeme	Führungskräften
Monatslos	Verkaufsseligen	Geschäftsaussichten
Regierungsmischung	Vorweihnachtssonnabend	Injektionsspritzen
Regierungsstart	Dreikönigsspringens	Kommunikationsbedarf
Rüstungsbericht	Lieblingsschanze	Regionspräsident
Satzungsfragen	Schnelligkeitsprobleme	Rücktrittsangebot
Schiffsvorderteil	Vollgesichtshelm	Stimmrechtsbe-schränkungen
Solidaritätslisten		Traditionsunternehmen
Steuervergünstigungs-abbaugesetz		Untersuchungshaft
Subventionsstaat		Verkehrsübungsplatz
Vergleichsfälle		Verpackungsordnung
Weltwirtschaftsarchiv	**Word erkennt nicht**	Vertragsverletzungs-verfahren
Hilfsbroschüren	Monatslos	Weihnachtspräsent
Migrationsausschuss	Aufsichtsratsvize	Wirtschaftsflaute
Regionspolitiker	Anscheinserwecker	Wirtschaftsforschungs-institut

Tabelle 3.9 Komposita mit s-Fuge

Bei DITECT stellen die Wortbildungskonstruktionen dieser Kategorie größere Schwierigkeiten an die elektronische Verarbeitung. Zwei- oder mehrgliedrige Komposita können, wie bereits oben erwähnt, nicht ohne weiteres in die Konstituenten zerlegt werden, da unweigerlich Bestandteile zurückbleiben, die nicht im Lexikon gespeichert sind. Tritt solch ein Fall auf, wie im Beispiel **End#Bilanz*, muss die fehlerhafte Konstituente korrekt auf das entsprechende Grundmorphem abgebildet

werden: *Ende#Bilanz*.In anderen Fällen ist die Konstituente mit einem zusätzlichen *s* versehen: *Dienstags#Demonstration*. Hier liegt nach dem Modell Kernighans ein Einfügungs-, im ersten Fall ein Tilgungsfehler vor, der wie folgt berichtigt wird: *Dienstag#Demonstration*.

Die vielen Beispiele, in denen Komposita mit s-Fuge korrekt erkannt wurden, sprechen dagegen, dass dieses Problem bei der Entwicklung des Systems vollkommen unbeachtet blieb. Die einfachste Möglichkeit, diesem falschen Verhalten der Fehlerkorrektur entgegenzuwirken, ist die vollständige Lexikalisierung aller in Frage kommenden Komposita, ein zeitaufwendiges und niemals vollkommen zu leistendes Unterfangen. Für die relativ gut einzuordnende Textsorte des Zeitungstextes könnte man allerdings mit einer begrenzten Zahl an Komposita auskommen.

Weitaus weniger ressourcenintensiv und viel eleganter ist es indes, eine Kompositaanalyse durchzuführen und für diejenigen Lexeme, die als Erstkonstituente ein Fugen-s nehmen, dies zu vermerken. Dies wären beispielsweise alle Lexeme, die die am Anfang des Abschnitts genannten Suffixe an der letzten Position tragen: *Schönheit, Erbschaft, Immunität* etc.

3.1.5.3 Komposita mit Bindestrich

Die letzte Unterkategorie in diesem Bereich stellen die Komposita mit Bindestrich dar (*Borussia-Trainer, Patriot-Raketen, Bild-Zeitung*). Seit der Rechtschreibreform ist es erlaubt, zur Verdeutlichung lange Komposita und solche mit drei aufeinander folgenden Vokalen mit Bindestrich zu schreiben (*Steuervergünstigungs-Abbaugesetz; Schnee-Eule*, aber *Schifffahrt*). Stilistisch lässt sich darüber gewiss streiten, es hat sich mittlerweile aber eingebürgert auch kürzere und sogar zweigliedrige Komposita mit Bindestrich zu schreiben (*Flaute-Jahr, Ehe-Gerücht*). Ein Großteil der Komposita im vorliegenden Testkorpus wurde mit Bindestrich geschrieben, was insbesondere von Word 2000 häufig beanstandet wurde.

Für die Beanstandungen bei DITECT lassen sich andere Gründe finden als die Existenz des Bindestrichs. Diese erklären sich zum einen aus der nicht erfolgten Lexikalisierung der Konstituenten. Auch als allein stehende Lexeme werden die Begriffe *gildet, Teuro, Acanthos, Rürup, Stajner, Winterabo* als falsch geschrieben beanstandet. Die Schreibung *ÜbersetzerInnen* enthält eine Binnenmajuskel, die für die Beanstandung verantwortlich ist, bei *US-Zeichtrickserie* liegt ein zweifacher Auslassungsfehler vor und im Falle von *Werks-klub* muss die Zweitkonstituente großgeschrieben werden (siehe Tab. 3.10).

Die sehr geringe Zahl der Bindestrichkomposita, die sich hier findet, lässt darauf schließen, dass DITECT Komposita mit Bindestrich nicht als Komposita behandelt. Ein Bindestrich kann als Fuge behandelt werden, in diesem Falle wird er durch eine Fugenmarkierung ersetzt: *Boulevard-Zeitung* → *Boulevard#Zeitung*. Der Bindestrich kann aber auch als Trennungszeichen behandelt werden. Dann entstehen bei der Verarbeitung aus einem Kompositum mehrere Einzelwörter *Boulevard-Zeitung* → *Boulevard Zeitung*. Die letztere Variante ist unpraktikabel, da sie die wertvolle Information, dass es sich bei der Zeichenfolge um ein Kompositum handelt, verwirft. Bindestrichkomposita sind einfach zu verarbeiten. Da die Fuge bereits vorgegeben ist, müssen die Komposita nicht analysiert werden, und es kann demzufolge nicht zu Erkennungsfehlern kommen.

Ein besonderer Umstand, der nicht als gewollte Design-Entscheidung gewertet werden kann, spricht gegen beide Varianten der Bindestrichbehandlung: Der Eigenname *Gerets* wird von beiden Programmen als fehlerhaft markiert. Es ist dabei nicht weiter verwunderlich, dass der aktuelle Trainer des Bundesligisten 1. FC Kaiserslautern den Programmen nicht bekannt ist. Es treten aber auch die Komposita *Gerets-Team* und *Gerets-Nachfolger* im Korpus auf. Diese werden von Word als falsch markiert. Bei DITECT tauchen sie nicht in der Bilanz auf. Dies spricht für eine

DITECT Komposita	Word Komposita	K. mit Eigennamen	K. mit Abk.	K. mit Fremdk.
Jetzt-gildet	Bundesliga-Hinrunde	Acanthos-Chef	Acht-Stimmen-GAU	Aldi-Airlines
Passarellen-Ebene	Deutsche-Bank-Chef	Alcantara-Dancers	CD-Roms	Austria-Adler
Eben-noch-Kanzlerkandidat	Eben-noch-Kanzlerkandidat	Baba-Berater	Dax-Werte	Call-Center
Teuro-Debatte	Im-Fall-des-Falles-Frage	Borussia-Trainer	DRC-Frauen	Cent-Münzen
Teuro-Gefühl	Ko-Trainer	Cobham-Gruppe	Duathlon-WM	Champ-Car
Werks-klub	Nachbarhaus-Besitzerin	Domotex-Austellerzahl	DVD-Player	Champ-Car-Rennen
	Passarellen-Ebene	Emnid-Umfrage	EM.TV-Aktie	Christmas-Party
K. mit Eigennamen	Premium-Zeitung	Fabricius-Brand	ING-BHF-Bank	Color-Line-Arena
Acanthos-Chef	Schalmeien-Klänge	Forsa-Umfrage	K.o.-Duelle	Event-Kultur
Baba-Berater	Sportfreunde-Trainer	Gerets-Team	M-Dax	Expo-Revival
Brauergilde-Aktionäre	Sternsinger-Spenden	Gerets-Nachfolger	Multimedia-BBS	Express-Airways
Rürup-Kommission		Gruner-und-Jahr-Chef	NP-Winterabo	Funky-Jam-Team
Stajner-Position	**K. mit Wortneubild.**	Haffa-Aussage	REWE-Touristik	Gin-Tonic-Bowle
	Teuro-Debatte	Hartz-Kommission	Spaß-FDP	Hannover-Concerts-Chef
K. mit Abkürzungen	Teuro-Gefühl	Interbrew-Angebot	Super-OB-Büro	Hannover-Open-Air
NP-Winterabo		Interbrew-Vorstandschef	Szene-Djs	mega-out
E-Netz	**K. mit Anf-zeichen**	Mölle-Supermann	UMTS-Lizenz	Prepaid-Karte
s!"-Buttons	„Bild"-Zeitung	Patriot-Luftabwehrrakete	UNMOVIK-Chef	Pre-Release-Party
UNMOVIK-Chef	„Patriot"-Raketen	Peppermint-Park-Studio	UNSCOM-Nachfolgeorganisation	Kunst-und-Design-Fachhochschulen
UN-ÜbersetzerInnen	„Skud"-Raketen	Rürup-Kommission	UN-ÜbersetzerInnen	Revival-Partys
US-Zeichtrickserie	„Spiegel"-Gründer	Stajner-Position	US-Zeichtrickserie	Sit-Ups
	„Jetzt-gildet's!"-Buttons	Transnet-Vizechef	VW-Bulli	Sledge-Eishockey
K mit Fremdkonst.	Use"-Kapazität	Transrapid-Variante	WOM-Verkäufer	Teenie-Idolen
Sit-Ups	„Focus"-Informationen	Yournightlife"-Chef	ZEW-Indikator	Top-Speed

Tabelle 3.10: Von DITECT und Word beanstandete Komposita mit Bindestrich

unabhängige Behandlung von Einzellexem (*Gerets*) und Kompositum (*Gerets-Nachfolger*).

Es ist möglich, dass der oben gezogene Schluss, Bindestrichkomposita würden nicht als Komposita betrachtet, dennoch zutrifft und das Verhalten im Fall *Gerets* andere Gründe hat. Lexika in professionellen Anwendungen bestehen zumeist aus mehreren Modulen: einem Hauptlexikon, das alle Grundmorpheme und weitere häufige Wörter der jeweiligen Sprache enthält, und einem oder mehreren zusätzlichen Lexika, in denen anwendungsspezifische Lexeme gespeichert oder aktuelle Eintragungen ergänzt werden können, wie z.B. die Namen der Bundesligatrainer.

Im Regelfall sollte solch eine Diskrepanz allerdings so aussehen, dass das Einzellexem nicht beanstandet wird, aber eventuell Probleme bei der Kompositaanalyse auftreten, wenn in einer Wortform Bestandteile auftreten, die in verschiedenen Lexika gespeichert sind, z.B. *Nachfolger* im Hauptlexikon und *Gerets* im Benutzerlexikon.

Word geht schärfer mit den Bindestrichkomposita ins Gericht: Das Programm beanstandet in dieser Unterkategorie fünf- bis sechsmal so viele Komposita wie DITECT. Darunter befinden sich im Wesentlichen drei große Gruppen: Komposita mit Eigennamen, Komposita, die Abkürzungen mit Lexemen verbinden, und Komposita mit fremdsprachlichen Bestandteilen (meistens englische).

Hierzu muss erwähnt werden, dass Word Fremdsprachen erkennt und dadurch rein englische (französische, spanische) Komposita korrekt erkannt werden können. Die Erkennungsleistung dieses Moduls ist allerdings nicht zuverlässig. Zudem können Begriffe, die deutsche mit fremden Stämmen verbinden, nicht zugeordnet werden.[5] Eine Beschäftigung mit diesem Thema führte zu weit. Da sich das Korpus lediglich auf deutschsprachige Texte bezieht, wird auch die Rechtschreibkorrektur nur in deutscher Sprache ausgeführt. Dabei wäre es wünschenswert, dass Fremdlexeme, die einen festen Platz im deutschen Sprachgebrauch haben wie *Event* oder *Revival*, lexikalisiert wären, um eine Kompositabildung mit diesen Bestandteilen zu ermöglichen.

Eine kleinere Gruppe betrifft Komposita, die nur aus deutschen Bestandteilen bestehen, oder solche, von denen angenommen werden muss, dass sie einen hohen Lexikalisierungsgrad besitzen (*Bundesliga-Hinrunde, Schalmeien-Klänge*). Daneben gibt es Komposita, die aufgrund von Wortneuschöpfungen nicht erkannt werden (*Teuro*).

Warum Word Komposita mit Bindestrich beanstandet ist nicht immer einsichtig. In vielen Fällen liegt der Grund in der fehlenden Lexikalisierung eines Bestandteils, was bei Word im Gegensatz zu DITECT zur Beanstandung des gesamten Kompo-

situms führt. Dies gilt zumindest für die Kategorie Komposita mit Eigennamen und fremdsprachlichen Konstituenten. Für die Kategorie Abkürzungen gilt dies zum Teil. Z.B. sind *Duathlon, ÜbersetzerInnen, Winterabo* und *Dax* nicht lexikalisiert. Ein Erkennungsproblem von Abkürzungen liegt nicht vor, da Word großgeschriebene Zeichenfolgen ignoriert.

K. mit dt. Konstituenten	Komposita mit Zahlen	K. mit Eigennamenskonst.	K. mit Abkürzungen
Abnick-Mehrheit	16-seitigen	Anzeiger-Hochhaus	ARD-Angaben
Anlage-Hafen	19 000-Einwohner-Stadt	Cartier-Damenuhr	AWACS-Aufklärungsflugzeuge
Bundesrats-Initiative	3,3-Millionen-Euro-Mann	Deutschland-Geschäftes	A-Waffenprogramms
Damen-Regionalliga	500-Euro-Scheinen	Dollar-Raum	AWD-Arena
Defizit-Grenze	53-Jährige	Dresdner-Bank-Chef	CDU-Bildungsexperte
Deutsch-Kanadiers	74,4-Prozent-Wahl	Euro-Bargelds	C-Waffenverbotsabkommen
Ehe-Gerüchte	Ex-96-Star	Fiat-Vorstandschefs	DB-Aufsichtratschef
Einweg-Dosen	Formel-1-Rennen	Gilde-Brauerei	DGB-Sozialexperte
Elektronik-Unternehmer	Noch-nie-96-Spieler	Heinrich-Böll-Stiftung	D-Netze
Ergebnis-Etappe	Pro-7-SAT-1-Chef	Irak-Konflikts	EU-Kommissar
Flaute-Jahr		Karl-Liebknecht-Hauses	Expo-Gelände
Flughafen-Trasse	**K. mit Fremdkonst.**	Kaufhof-Chef	Ifo-Geschäftsklima-Index
Fußball-Klubs	Bahn-Jobs	Konica-Foto-Niederlassung	IG-Metall-Chef
Interessen-Konflikt	Computer-Infrastruktur	Konica-Umsatzes	NRW-Amtskollege
Kassenärzte-Chef	Dow-Jones-Index	Krupp-Vorstand	S-Bahnstrecke
Klima-Fundis	Dumping-Angeboten	Metro-Aktie	SMS-Wettbewerbs
Klinikärzte-Verbandes	Handy-Hersteller	Minolta-Standort	SPD-Landesregierungen
Kosten-Nutzen-Verhältnis	High-Tech-Stützpunkten	Moskau-Erfahrung	T-Aktienkurs
Kufen-Duells	Open-Air-Arena	Pferdeturm-Idol	T-Mobile
Kurs-Höhenflug	Talk-Show-Präsenz	Schröder-Gerüchten	TUI-Konzern
Lockvogel-Angebot	Trainings-Dress	Simpson-Deal	U-Bahn-Station
Marine-Versorgungsschiff	Weltcup-Slalom	Stoiber-Auftritt	U-Boot
Medizintechnik-Firma		Toto-Lotto-Marketingmann	UN-Mandat
November-Rückgang	**K. mit Partikeln/Konf.**	Verdi-Chef	US-Dollar
Orientteppich-Importeure	anti-israelischer	Wall-Street-Firmen	X-City-Medien
Reform-Ideen	Ex-FDP-Spitzenpolitiker	Westerwelle-Stellvertreter	
Regime-Gegnern	Ex-Mannesmann-Chef		**Komposita mit Zahlwörtern**
Rot-Grün	Ex-Meister	**Namen mit Bindestrich**	Einschienen-Flitzer
Schlittschuh-Gehversuche	Partei-Vize	Baden-Württemberg	Neun-Milliarden-Finanzspritze
Schnee-Eulen	Solo-Album	Berlin-Schönefeld	Sechsjahres-Hochs
Standort-Aufwertung	Vize-Regierungssprecher	Hans-Dieter	Zehn-Kilometer-Freistilrennen
Tariftreue-Gesetz		Jörg-Dietrich	Zwei-Drittel-Mehrheit

K. mit dt. Konstituenten	Komposita mit Zahlen	K. mit Eigennamenskonst.	K. mit Abkürzungen
Umbau-Jahre	**K. mit s-Fuge**	Jürgens-Pieper	
Umsatz-Ausfälle	Präsidiums-Ultimatum	Kirch-Media	**Komposita mit NGr**
Vergabe-Stopp	Einzelhandels-Umsatzdaten	Richter-Reichhelm	Blut-Schweiß-und-Tränen-Rede
Weihnachtsmarkt-Idylle	Unions-Länder	schleswig-holsteinische	Drei-Pünktchen-Partei
Wetter-Verrücktheiten	Weihnachts-Persiflage	Techniker-Krankenkasse	Weiße-Tauben-Flugdienst

Tabelle 3.11: Nicht beanstandete Komposita mit Bindestrich

Tabelle 3.11 gibt darüber Aufschluss, dass Komposita mit Abkürzungen per se kein Problem darstellen. Die Erkennungsleistung bei Komposita mit Zahlen und solchen mit deutschen Bestandteilen ist auch zufrieden stellend.

Bei Word gilt in einigen Fällen der Bindestrich als Fehler. So werden *Nachbarhaus-Besitzerin* und *Sternsinger-Spenden* vom Programm akzeptiert, wenn der Bindestrich entfernt wird. Dies sind allerdings Einzelfälle. Die Fülle der nicht als fehlerhaft markierten Bindestrichkomposita in Tabelle 3.11 zeigt, dass eine konsequente Markierung von falsch gesetzten Bindestrichen nicht erfolgt. Nach der Neuregelung der Rechtschreibung müsste ein Großteil der Bindstrichkomposita als Fehler markiert werden, da dieser nur zur Verdeutlichung bei drei aufeinander folgenden Buchstaben oder extrem langen, unübersichtlichen Komposita erlaubt ist. Für *Defizit-Grenze, Einweg-Dosen, Umbau-Jahre* etc. gilt das eindeutig nicht. Offensichtlich gibt es eine erhebliche Diskrepanz zwischen Schreibusus und amtlicher Regelung. Das es einen erkennbaren Trend zur Bindstrichschreibung gibt, wird jedoch auf eine systematische Erfassung aller Bindestrichfehler im Korpus verzichtet.

Interessant an Tabelle 3.11 ist ein weiterer Einzelfall: Bindestrichkomposita mit der Konstituente *Vize* werden im Gegensatz zu echten Komposita mit diesem Bestandteil erkannt. Diese Konstituente hat offenbar einen eigenen Status. Sie wird als Zweitkonstituente nur mit Bindestrich zugelassen.

Die anderen Beispiele sind nicht beanstandet worden, weil ihre Konstituenten lexikalisiert sind. *Cartier* und *Fiat* sind offensichtlich gängiger als *Domotex* und *Rü-*

rup. Geht man davon aus, dass große Trainingskorpora bei der Lexikonerstellung benutzt wurden, verwundert dies nicht.

Warum die *Blut-Schweiß-und-Tränen-Rede*, die *Drei-Pünktchen-Partei* und der *Weiße-Tauben-Flugdienst* nicht beanstandet wird, wohl aber der *Eben-noch-Kanzlerkandidat* und die *Im-Fall-des-Falles-Frage*, ist eine weitere offene Frage. Werden Satzbestandteile (Syntagmen) in einem substantivischen Bindestrichkompositum verwendet, so muss dessen Erstkonstituente mit einer Majuskel anfangen. Alle anderen nichtsubstantivischen Konstituenten bleiben kleingeschrieben und werden mit Bindestrichen durchgekoppelt. Die Praxis der Korrekturprogramme läuft der Regel zuwider: So wird der *Eben-Noch-Kanzlerkandidat* nicht beanstandet, wenn jede Konstituente großgeschrieben wird.

Es ist sehr schwer, aufgrund dieser Listen auf das generelle Verhalten der Programme zu schließen, da davon ausgegangen werden muss, dass es mehrere Ebenen der Behandlung gibt. Neben einer generellen Strategie gibt es Ausnahmeregelungen und einzelne Lexikalisierungen, die nicht gefunden werden können, indem man betrachtet, was beanstandet wird und was nicht. Das Faktum, dass Word alle Komposita ähnlich behandelt und DITECT solche mit Bindestrich ignoriert ist jedoch ein wichtiges Indiz für die weitere Untersuchung in Kapitel 4.

3.1.6 Trennung am Zeilenende und Bindestrichergänzungen

Der nächste Problembereich behandelt Wörter, die durch einen Trennungsstrich getrennt geschrieben werden. Dies betrifft zunächst manuelle Worttrennungen am Zeilenende, die durch nachträgliche Bearbeitung wieder in den Fließtext zurückgekehrt sind (*Innensena-tor, schubs-ten*), und zum anderen Erweiterungen mit Bindestrich.

Erstere werden von beiden Programmen konsequent als fehlerhaft markiert. Anders als im Bereich der Bindestrichkomposita gibt es bei der Fehlerdetektion hier

keine Abweichungen. Formal kann man eine manuell durchgeführte Silbentrennung als Einfügungsfehler betrachten, der relativ einfach zu beheben sein sollte. Das Problem liegt in der Natur des Bindestrichs, dessen Verwendung als formatives Element Interpretationsspielraum zulässt. Solche Überlegungen gehören allerdings schon in den Bereich der Fehlerkorrektur, die im nächsten Unterabschnitt behandelt wird.

Während Bindestriche in der Silbentrennung relativ einfach erkannt werden können, ist der Fall bei Bindestrichergänzungen weitaus komplizierter. Hier liegt eine Koordinationsreduktion vor. Um solche Wendungen korrigieren zu können, muss die Information über die Syntax der Nominalgruppe vorliegen. Bindestrichergänzungen können vom derzeitigen Forschungsstand aus aufgrund der Komplexität dieses Phänomens elektronisch noch nicht zuverlässig wieder zusammengeführt werden.

Der häufigste Fall, die Koordination mit *und* bzw. *oder* bei gleichzeitiger Ersetzung der zweiten Konstituente im ersten Kompositum durch einen Bindestrich, ist der relativ einfache Regelfall (Bsp. 1 bis 4). Die Verwendung eines Bindestrichs im Kompositum erleichtert durch das Anzeigen der Fuge gegebenenfalls die Koordination für die elektronische Verarbeitung) (Bsp. 5 bis 8):

1. Bau- und Entwicklungspläne
2. Geschäfts- und Wirtschaftsklima
3. Volks- und Raiffeisenbanken
4. Gold- oder Rohölpreise

5. CD- oder DVD-Player
6. Energie- und Wasser-Riese
7. Europa- und Deutschland-Zentrale
8. Internet- und Telekom-Spekulationsblase

Die folgenden Beispiele zeigen die mögliche Komplexität dieses Phänomens. Die Bindestrichergänzung wird auch bei mehreren Komposita zugelassen, so dass Aufzählungen möglich sind (Bsp. 9). Sie kann zudem in dreigliedrigen Komposita beidseitig zur Tilgung der letzten Konstituente im ersten und der ersten Konstituente im zweiten Kompositum genutzt werden (Bsp. 10). Die Koordination muss nicht zwingend durch und, bzw. oder geleistet werden (Bsp. 11). Es muss sich des Weiteren nicht einmal um zwei Komposita handeln. In Beispiel 12 ist das Erstglied durch ein Adjektiv ersetzt worden.

9. Wirtschafts-, Arbeitsmarkt- und Sozialpolitik
10. Bundeswirtschafts- und -arbeitsminister

11. Schmuse- statt Basta-Kanzler
12. Technologie- und andere Werte
13. Einsatz als Fachbereichs- oder als Bereichsleiter
14. Gas- (tschechische Transgas) sowie Stromfirmen
15. Keine Abschmetterungs-, sondern eine Auffangstrategie
16. Wirtschafts-, Arbeitsmarkt- und Sozialpolitik

Darüber hinaus sind die Konstruktionen durch zusätzliche Attribute, Klammern, Abtönungspartikeln, Relativsätze usw. beliebig erweiterbar, so dass die Rekombination der Komponenten immer schwieriger wird (Bsp. 17 bis 19):

17. Fachbereichs- oder Bereichsleiter werden gesucht.
18. Sie arbeiten als Fachbereichs- oder eventuell als Bereichsleiter
19. Es wäre eine Einsetzung als Fachbereichs- oder bei besonders herausragenden Fähigkeiten eventuell auch die Position eines Bereichsleiters denkbar.

Die korrekte Rekombination der Komposita kann in diesem Bereich umgangen werden, da sie in erster Linie von semantischem Interesse ist, also in den Bereich des *Data Mining* und *Information Retrieval* gehört. Für die Rechtschreibprüfung ist gegenwärtig lediglich die Formseite von Interesse.

In der Diskussion der Komposita ist bereits deutlich geworden, dass für die Rechtschreibkorrektur die Fuge des Kompositums erkannt werden muss. Im Falle der Bindestrichergänzung reicht es also aus, wenn durch den Bindestrich die Existenz der Erstkonstituente eines Kompositums nachgewiesen wird. Liegt die Information vor, ob ein Fugen-s von dem Bestandteil genommen wird oder nicht, kann eine eindeutige Fehlerdetektion stattfinden, d. h. die Frage beantwortet werden, ob es sich um einen Bestandteil der deutschen Sprache handelt. Der andere Bestandteil der Bindestrichergänzung kann wie ein normales Kompositum behandelt werden.

3.1.7 Eigennamen

Auf das Problem der Eigennamen wurde bereits im Zusammenhang mir der Verteilung bei der Fehlerdetektion hingewiesen: Eine annähernd hinreichende Lexikalisierung der gebräuchlichen Eigennamen ist kaum leistbar, denn seit Beginn des Informations- und Globalisierungszeitalters hat sich die Gruppe fremdsprachlicher Namen, auf die referiert wird, stark vermehrt. Menschen und Daten bewegen sich immer schneller und in immer größerer Zahl um den Globus. Dadurch entsteht eine Internationalisierung von Orts- und Personennamen.

Aber nicht nur die Integration fremdsprachlicher Einflüsse in die deutsche Sprache bereitet Probleme bei der Fehlerdetektion. Die Wirtschaft braucht immer wieder neue, frische Bezeichnungen, und greift daher häufig zu in der Alltagssprache bisher unüblichen Schreibweisen aus der Internetsprache und dem IT-Bereich (*TransRapid, AutoFuture, ver.di, E.on*). Produktive Bildungsmuster zu formulieren, ist hier schwierig, da gerade durch die Originalität der Bildung eine Einprägung des Be-

griffs erreicht werden soll, das Neue also zum vitalen Kern der Wortproduktion gehören.

Namen stellen also nicht generell ein Problem für die Rechtschreibkorrektur dar. Gängige Ortsnamen, also solche von Städten, Staaten, Flüssen usw., lokal wie global, werden von beiden Programmen nicht beanstandet: *Angola, Tadschikistan; Baden-Württemberg, North Carolina; Lagos, Kuala Lumpur, Buenos Aires, Castrop-Rauxel*. Unbekannte und ungebräuchliche Ortsnamen, wie solche aus entlegenen Gegenden, auch wenn es sich um Großstädte mit mehr als einer Million Einwohnern handelt, sind jedoch von Word 2000 häufig beanstandet worden, wie *Porto Alegre, San Miguel de Tucuman, Hyderabad, Bangalore. Phnom Penh* gehört mit ähnlicher Größe allerdings nicht zu dieser Gruppe. Als Hauptstadt von Kambodscha kommt diesem Ort offenbar eine Bedeutung zu, der für eine Designentscheidung zugunsten der Lexikalisierung wichtig ist. *Vientiane*, die Hauptstadt von Laos, steht nicht im Lexikon. Eventuell wird auf Phnom Penh durch seine politische und historische Brisanz häufiger referiert und es taucht daher in Trainingskorpora als Größe auf, die dem Begriff einen Platz im Lexikon einräumt. Wir bewegen uns hier in den Grenzbereichen dessen, was vom Lexikon geleistet werden kann. Die getroffenen Designentscheidungen, was aufgenommen wird und was nicht, sind schwer nachzuvollziehen. Werden Städte über einer Million Einwohner lexikalisiert, diejenigen mit politischer und wirtschaftlicher Bedeutung oder ist die Häufigkeit in einem Trainingskorpus auschlaggebend?

Für Personennamen gilt Ähnliches: Native Vor- und Nachnamen wie die folgenden sind im Lexikon vorhanden: *Bärbel, Yvonne, Sebastian, Enno; Müller, Engler, Gruber, Neugebauer*. Auch Schreibvarianten werden weitgehend zugelassen: *Meier, Meyer, Maier, Mayer, Mayr, Meir* sind gültige Varianten des Namens, *Mair* wird von Word beanstandet. Ebenso folgendes Beispiel: *Katrin, Catrin, Kathrin* werden von Word 2000 nicht beanstandet, *Cathrin* wird allerdings nicht mehr akzeptiert,

obwohl diese Variante des Namens durchaus gebräuchlich ist. Dies spricht dafür, dass von dem System keine Schreibvariantennormalisierung im Sinne der Levenshtein-Methode angewendet wird. Sonst dürften Substitutionen, die die phonemische Struktur nicht verändern, wie *k–c* oder *t–th* bei Personennamen nicht zu einer Beanstandung des Wortes führen.

EINS (41)	POLI (43)	LOKA (92)	WIRT (39)	SPOR (232)	
Funtensee	Al-Mazroa	Albrecht-Schäffer-Weg	Büdelsdorf	Andertener	Godiva
Immensen	CORUÑA	Arnum	Garbsener	Andertens	Goulet
Rethen	Galiciens	Benthe	Senftleben	Bolzum	Groenewold
Uetze	Jammu-Kaschmir	Benther		Davenstedts	Gürek
	Laxe	Deisterstraße	Acanthos	Grasdorfs	Guzman
Ark	Traba	Döhrener	Bae	Havelser	Haletzki
Brauergilde	Nautile	Echternfeld	Barron's	Heiligenrode	Haseney
Calvann		Herrenstraße	Innogy	Hettich	Hintum
Kaida	Ambrozy	Laabs	Minoltas	Hiddestorf	Hipperling
	Annen	Laatzener	Sampo	Isla	Höhlig
Anda	Baradei	Loccumer	Stada	Kirchrödern	Hüper
Andas	Blix	Scholvinstraße	Transgas	Mitteltal-Obertal	Idrissou
Benke	Dehm	Südschnellweg		Mörsen	Ionannis
Bökel	Dückert	Uetze	Bolkestein	Nienstädt	Jäggis
Bouffier	Fücks	Usener	Borghoff	Pewsum	Janica
Duin	Geoff	Welfenplatz	Colaninno	Ramlingen	Jebok
Garrelt	Girin		Cromme	Ramlingens	Jescheniak
Hillus	Glos	Brauergilde	Dambrowski	Rehde	Jusufi
horst	Grüter	Parlophone	Deufel	Rethen	Kadrina
Karoff	Gusenbauer	Üstra	Grimmaer	Rethener	Kalex
Mojad	Hassam		Haffas	Stelingen	Kallabis
neufert	Heckl	Adil	Heckler	Tündern	Karakollukco
Nolting	Hoon	Arslan	Ipektchi		Kaufman
Pau	Kuhl	Begovic	Jakschies	Germanias	Kavapovic
Pellengahr	Lippelt	Bendig	Kannegiesser	Sachsenroß	Klingebiel
Raffelhüschen	Niessl	Bohnecke	Knorre		Kokott
Ranstorp	Petritsch	Bullerdieck	Kromphardt	Agac	Kollosky
Rongji	Rajoy	Cassan	Ladberg	Ahonen	Konstantinidis
Siddik	Rau	Dedek	Pellengahr	Alena	Kosgei

EINS (41)	POLI (43)	LOKA (92)	WIRT (39)	SPOR (232)	
Solbes	Salim	Dikau	Pokoj	Altin	Kostas
Staps-Finé	Schaich-Walch	Dlugosch	Rohner	Amvrossiadis	Kostelic
Steinbrück	Stiegler	Döll	Rösch	Andreina	Kotuljac
Terkingür	Tajes	Fellmann	Wennemer	Baacke	Kröckert
ulrich	Tischmann	Fiebelkorn	Yoshikatsu	Baba	Krupnikovic

Tabelle 3.12: Auswahl der beanstandeten Eigennamen bei DITECT

Wichtig ist aus Nutzerperspektive, ob die Inkonsistenz auffällig ist. Es sollte beispielsweise nicht Stuttgart erkannt, München aber als fehlerhaft markiert werden, da so der Nutzer verwirrt würde. Wenn allerdings ein Geo-Redakteur einen Artikel über Südamerika schreibt, wird dieser eher geneigt sein, Fehlermarkierungen von Ortsnamen in seinem Text zu akzeptieren. Aus dieser Perspektive ist die Leistung der beiden Systeme zufrieden stellend.

Tabelle 3.12 unten zeigt eine Auswahl der beanstandeten Eigennamen bei DITECT. Die Spalten sind unterteilt in einen ersten Abschnitt, der geografische Namen enthält, einen zweiten, der sich den Namen aus der Wirtschaft, anderen Institutionen und Schiffen widmet, und einen dritten Abschnitt, den größten, der Personennamen enthält. Hält man sich aber den Einsatzort des Programms DITECT vor Augen, also den begrenzten Kosmos einer Lokalzeitung, dann ist die Leistung als nicht befriedigend zu bewerten. Zumindest mit halbjährlichen Updates ist z.B. die Lexikalisierung der Spieler samt Trainer und Manager von Hannover 96 zu leisten. Damit wäre ein großer Teil der Beanstandungen von Personennamen getilgt.

Positiv ist zu bemerken, dass lokale Ortsnamen bei DITECT kaum auftreten. Die Lexikalisierung dieser Begriffe sollte also geleistet sein. Allerdings tritt das schon angesprochene Problem der Derivation wieder zutage. Bildungen aus Ortsnamen auf *er* oder mit dem besitzanzeigenden *s* werden nicht berücksichtigt: *Döhrener, Garbsener, Andertens.* Selbst wenn solche Terme nicht generiert werden können, weil dies zu unsicher im Hinblick auf mögliche Ausnahmen ist, liegt deren Bildung aus Nutzerperspektive doch nahe und sollte wenigstens durch Lexikalisierung der Begriffe

EINS (53)	POLI (40)	LOKA (100)	WIRT (74)	SPOR (306)	
Calenberger	Jammu-Kaschmir	Aegi	Garbsen	Anderten	Ahonen
Hasseröder	Kärntener	Albrecht-Schäffer-Weg	Garbsener	Andertener	Akyol
Maschsee	Traba	Arnum	Grimma	Andertens	Anschütz
Portsmouth	Dorint-Hotel	Benthe	Grimmaer	Bischofshofen	Baba
Rethen		Döhrener	Uphusen	Bolzum	Balakov
Sehnde	Nautile	Guantanamo		Borsum	Biskup
		Kröpcke	Acanthos	Davenstedts	Bodensiek
Forsa	Beust	Lindener	Barron's	Döhren	Bordeleau
Interbrew	Blix		Dax	Gerbrunn	Brdaric
Kaida	Bsirke	Arkadas	Dertour	Gleidingen	Briegel
Transrapid	Dückert	Media-Markt		Havelse	Bundt
	Glos	Parlophone	Hapag	Havelser	Burböck
Bökel	Grüter	Galeria	Hypo-Vereinsbank	Heiligenrode	Byrd
Hagenbecks	Gusenbauer	Sasi	Infineon	Kirchrode	Calmund
Hartz	Hassam	Scholvinstraße	Innogy	Kirchrödern	Cheruiyot
Hillu	Heckl	Sinn-Leffers		Borussia	Danijel
Nolting	Hoon	Usener	Mobilcom	Diayarbakirspor	Denboba
Osama	Jüttner	Üstra	Nemax	Fenerbahce	Depping
Pau	Künast	Vodafone	Transgas	Germanias	Friesinger
Raffelhüschen	Langguth		Transnet	Juventus	Gerets
Ranstorp	Mahamad	Adil	Vodafone	Mitteltal-Obertal	Giesel
Rongji	Schaich-Walch	Ayhan		Mörsen	Giesels
Schüssels	Schüssels	Ayse	Bolkestein	Pagelsdorf	Hintum
Solbes	Sigmar	Dikau	Dambrowski	Pewsum	Hoeneß
Staps-Finé	Tajes	Dlugosch	Dudenhöfer	Radisson-Hotel	Höhlig
Steinbrück	Weigelt	Döll	Haffa	Ramlingen	Höllwarth
Terkingür	Westenthaler	Eikemeier	Komatsu	Ramlingens	Hülsmann
Wittke	Wohlers	Fiebelkorn	Mehdorn	Ramlinger	Janne
Zhu	Zastrow	Fromberg	Pellengahr	Rethener	Kostelic

Tabelle 3.13: Auswahl der beanstandeten Eigennamen bei Word

berücksichtigt werden. Die nicht erkannten Personennamen sind, wie vermutet, zu einem sehr großen Teil fremdsprachlicher Herkunft, wie *Terkingür, Dlugosch, Kostelic*. Besonders im Ressort Sport sind solche Personennamen häufig.

Bei Word zeigt sich ein ähnlicher Befund hinsichtlich der Personennamen wie bei DITECT. Die Zahl der Namen ist sogar noch größer. Hier offenbart sich der Heimvorteil DITECTs, denn Begriffe von politischer Relevanz, wie sie für Zeitungstexte spezifisch sind, wie *Transrapid* oder *Dax* werden von Word als fehlerhaft markiert, von DITECT aber nicht. Ähnliches gilt für lokale Bezeichnungen wie *Aegi, Sigmar (Gabriel)*. Auch wenn die Zahl der beanstandeten Eigennamen bei Word in jedem Bereich größer ist als bei DITECT (siehe Abs. 3.1), hat auch Word einen Vorteil struktureller Natur. Es fällt auf, dass Ortsnamen entweder nicht auftreten oder in Begleitung der häufigsten Derivate: *Grimma, Grimmaer, Anderten, Andertener, Andertens*. Es ist klar, dass *Andertener* als falsch geschrieben beanstandet wird, wenn nicht einmal *Anderten* im Lexikon zu finden ist. Stichproben haben aber ergeben, dass Ortsnamen und Bildungen auf *er* bzw. *s* in fast allen Fällen entweder alle beanstandet oder alle als fehlerlos erkannt werden. Folgende Tabelle zeigt einige Stichproben von Namen kleinerer Orte aus Hannovers Umgebung und deren Derivaten, die nicht beanstandet wurden.

Hannover	Hannoveraner	Hannoveranerin	Hannovers
Celle	Celler	Cellerin	Celles
Büdelsdorf	Büdelsdorfer	Büdelsdorferin	Büdelsdorfs
Hemmingen	Hemminger	Hemmingerin	Hemmingens
Ronnenberg	Ronnenberger	Ronnenbergerin	Ronnenbergs
Wennigsen	Wennigsener	Wennigsenerin	Wennigsens
Laatzen	Laatzener	Laatzenerin	Laatzens

Tabelle 3.14: Derivate von Personennamen bei Word

Dieser Befund deutet darauf hin, dass in den Laboratorien von Microsoft bereits weit mehr hinsichtlich der Modellierung von Derivationsmustern getan worden ist, als ursprünglich vermutet wurde. Dennoch kann es sich immer noch um eine Lösung handeln, die auf der Lexikalisierung der Begriffe basiert. Sollten diese Begriffe aber durch das Programm gebildet worden sein, müssen mindestens Part-of-Speech-Tags vorhanden sein (siehe Abs. 2.1.3), um anzuzeigen, wann welche Affixe angefügt

werden können. Wahrscheinlich sind in diesem Fall zusätzliche semantische Kategorien, die beispielsweise Ortsnamen anzeigen. Dies ist schließlich nicht nur von akademischem Wert, sondern durchaus praxisrelevant, da Ortsnamen als Propria nicht wie normale Substantive flektieren (Eisenberg [2]: 160).

3.1.8 Fremdsprachliche Lexeme

Bestandteile aus Fremdsprachen spielen in den Textsorten, die sich in einer Tageszeitung finden, eine große Rolle. Der größte Teil der auftretenden Begriffe stammt aus dem Englischen, aber auch französische und spanische Bestandteile treten auf, meistens Bezug nehmend auf aktuelle Ereignisse in den Staaten mit der jeweiligen Landessprache (Tankerunglück vor der spanischen Küste an der *Playa de Traba*; Zinspolitik der französischen Regierung: *Credit Lyonnais, Assurance Generale de France*).

DITECT		Word						
Agricole	Internal	Red	Agricole	Events	Internal	Moody's	Playa	Shakedown
Assurance	Muerte	Rich	Assurance	Fire	Kitchen	Muerte	Racing	Shooting
cologne	Patriots	Shakedown	Concerts	Funky	Know	Net	Rating	Spice
Escapology	Patriots	start	dance	Horny	Let's	Patriots	Ratings	That
How	Pearls	Yournightlife	Escapology	How	Limited	Penalty	Rich	

Tabelle 3.15: Fremdsprachliche Begriffe

Die englischsprachigen Begriffe sind neben der Bezugnahme auf Nachrichten aus dem angloamerikanischen Raum zusätzlich in solchen Bereichen zu finden, die aus Marketing- und Werbungsgründen oder aufgrund einer starken kulturellen Affinität zu anglophonen Begriffen häufig auf Anglizismen oder ähnlich frisch klingende Nominationseinheiten zurückgreifen. Dies geschieht beispielsweise verstärkt in der DJ-Szene, wo sich Begriffe finden wie *Mousse T., Peppermint Park, Funky Kitchen* und *Shakedown* oder *Yournightlife*. Hier gilt ähnliches wie bei den Eigennamen im Bereich der Wirtschaft. Durch Wortneuschöpfungen soll ein Höchstmaß an Originalität erreicht werden. Um Wortschöpfungen wie *Shakedown* oder *Yournightlife*

in der automatischen Fehlerdetektion abzubilden, wäre zunächst ein umfassendes englischsprachiges Lexikon erforderlich. Dies ist zusätzlich zum deutschen Lexikon neben einem italienischen, einem französischen und einem spanischen Wortformenlexikon in den Installationsroutinen von Word 2000 und XP enthalten. Es fehlt aber eine Grammatik mit den entsprechenden Kombinationsregeln, die Wortbildungen wie *Yournightlife* möglich machen. Da diese Wortform keinem regulären Wortbildungsmuster folgt, hätte dies eine freie Kombination der Morpheme zur Folge. Damit wäre die Anzahl der möglichen Wortformen um ein Vielfaches erhöht und das Ziel, eine Abbildung der möglichen Wortformen der deutschen Sprache zu erreichen nicht realisierbar.

Die Zahl der fremdsprachlichen Begriffe ist sehr begrenzt. Eine zusätzliche Verwendung fremdsprachlicher Lexikoneinträge ist nicht ratsam, da viele Begriffe durch den Entlehnungsprozess eingedeutscht sind und daher formal, hinsichtlich der Flexionsmerkmale beispielsweise, nicht mehr den Äquivalenten aus der Ursprungssprache entsprechen. Eine Implementierung der Wortformen einer oder mehrerer Fremdsprachen würde zudem die Anzahl der möglichen Wortformen drastisch erhöhen. Durch die Vielzahl zusätzlicher Wortformen könnten einige Fehler nicht mehr als solche erkannt werden. Der Nutzen, den eine solche Lösung brächte, wäre durch den Schaden, der bei der Fehlerdetektion entstünde, wieder zunichte gemacht.

3.1.9 Grammatisch bedingte Fehler

Diese Kategorie behandelt zwei verschiedene Typen. Zum einen handelt es sich um grammatische Wendungen, die zwar orthografisch korrekt sind, von den Korrekturprogrammen allerdings dennoch als fehlerhaft markiert werden, da ihre Struktur offenbar nicht in der Morphologie der Programme vorgesehen ist. Beispiele:

1. Eine Mahnung an alle, seinen Mitmenschen nicht nach Äußerlichkeiten vorzuverurteilen.
2. Janne Ahonen hinterherzustiefeln bringt nichts.

Verben mit der Partikel *vor-* werden zusammengeschrieben, ebenso solche mit *hinterher.* Wird eine Infinitivkonstruktion mit *zu* gebildet, erfolgt die Inkorporation von *zu* zwischen der Partikel und dem verbalen Bestandteil. Aus diesem Grund wird der zu-Infinitiv bei Eisenberg als Bestandteil des verbalen Paradigmas betrachtet und *zu* nicht etwa als Konjunktion angesehen (Eisenberg [2]: 345). Die Wendung wird zusammengeschrieben, wenn auch Partikel und Verb sonst zusammengeschrieben werden. Beide Programme erkennen diese Strukturen nicht oder zumindest nur unzureichend. Das Verb *hinterherstiefeln* ist nicht im Lexikon enthalten, es wird von beiden Programmen beanstandet. *Vorverurteilen* wird allerdings erkannt, die Variante *vorzuverurteilen* jedoch nicht.

Dieser Befund stellt ein weiteres Indiz für die Vermutung dar, dass eine Morphologie zur Flexionsformengenerierung allenfalls rudimentär vorhanden ist, denn das hier untersuchte Phänomen ist nicht so selten, dass es ignoriert werden könnte. Die zur Erkennung des inkorporierten *zu* erforderliche Segmentierung der Verben in Partikel und Verbbestandteil hätte über den Bereich des Infinitivs mit *zu* hinaus den großen Vorteil, dass Wortbildungen wie *hinterherstiefeln* problemlos von dem Korrekturprogramm generiert und somit auch erkannt würden.

Der zweite Bereich in dieser Kategorie betrifft syntaktische Fehler, die nicht mehr in einem einzelnen Wort lokalisiert werden können. Es handelt sich dabei häufig um ausgelassene oder verdoppelte Wortformen (Beispiele 1, 2, 5) oder um Fehler in den Rektionsbeziehungen zwischen Verb und Argument oder innerhalb einer Nominalgruppe (Beispiele 3, 4, 5). Beispiele:

1. Zudem *will* RWE1 *will* aus Randsparten wie Bau (Hochtief) und Druckmaschinen (Heidelberger) aussteigen.
2. Droht dadurch ein Preisverfall kommen?
3. Die den Zukäufen zu verdankenden Steigerungen im Kerngeschäft hatten die noch höheren Einbußen in den nicht zum Kerngeschäften gehörenden Aktivitäten nicht ausgeglichen.
4. Der Vorschlag der SPD-Politiker verunsichern Goldanleger
5. Experten fordern, dass die Zentralbanken in Europa ein Teil der Goldreserven zu verkaufen.

Mit den bisher beschriebenen und untersuchten Methoden kommt man hier nicht weiter, da diese lediglich auf dem Finden von Schreibvarianten basieren, die durch einen der vier beschriebenen Fehlertypen von der lexikalisierten Form abweichen. Hier haben wir es mit den schon besprochenen Real-Word-Errors zu tun (vgl. Kap. 2.1.2), einem Fehlertyp, der nicht auf motorischen Fehlleistungen oder mangelnder orthografischer Kenntnis beruht, sondern seine Ursache in der Vermischung verschiedener syntaktischer Varianten hat.

Die mittelbare Entstehung der Fehler ist allerdings einsichtig. Im ersten Beispiel unten steht das Subjekt im Vorfeld und die Präpositionalgruppe, hier Adverbial zum Satz, im Mittelfeld innerhalb der Verbalklammer. Durch Einfügung der Konjunktion *zudem* wird das Subjekt ins Mittelfeld gerückt:

6. RWE1 will aus Randsparten wie Bau (Hochtief) und Druckmaschinen (Heidelberger) aussteigen.
7. Zudem will RWE1 aus Randsparten wie Bau (Hochtief) und Druckmaschinen (Heidelberger) aussteigen.

Im Fehlerbeispiel ist die Verbform in beiden Stellungen vorhanden geblieben. Ähnliches geschieht in den Beispielen drei und vier, wo die Rektionsbeziehungen nicht vollständig eingehalten wurden. Die Nominalgruppe *zum Kerngeschäften* aus dem dritten Beispiel steht zwar im Dativ, die Kongruenz hinsichtlich des Kasus ist also eingehalten worden. Die Nominale kongruieren allerdings nicht hinsichtlich des Numerus. Der Artikel steht im Singular, das Substantiv im Plural. Auch hier ist wahrscheinlich der Ursprungstext unvollständig korrigiert worden. Korrekt müsste die gesamte Nominalgruppe analog zum Numerus des entsprechenden Begriffs im Subjekt des Satzes im Singular stehen.

Die Rechtschreibprüfung von Word besitzt zwar schon seit der 97er-Version das Feature der Grammatikprüfung. Diese ist aber bisher nicht in der Lage, einen der bisherigen Fehler zu erkennen. Der Fehler aus Beispiel vier ist der einzige der fünf Beispielfehler, der von dem Programm erkannt und korrekt auf die fehlerhafte Subjekt-Verb-Kongruenz zurückgeführt wird. Für *Der Vorschlag der SPD-Politiker verunsichern Goldanleger* bietet Word *Der Vorschlag der SPD-Politiker verunsichere Goldanleger* an. Es ist sicher kein Zufall, dass die erfolgreiche Korrektur eines Grammatikfehlers in dem kürzesten der Beispiele erfolgte. Bei einem einfachen flektierten Verb, das als Argumente ein Subjekt und ein direktes Objekt im Akkusativ nimmt, haben wir es trotz des etwas komplexeren Subjekts mit einem einfachen Aussagesatz, also mit der unmarkierten Form des Satzes zu tun.

3.1.10 Abkürzungen und Akronyme

Abkürzungen werden von Word kaum beanstandet, da die meisten Abkürzungen substantivischer Art sind und diese vollständig großgeschrieben werden (*SPD, USA, HTML*). Zeichenketten, die komplett aus Majuskeln bestehen, sind von der Rechtschreibkorrektur ausgenommen. Ausnahmen sind silbische Abkürzungen (*allg., Werbespr., jap.*) oder Abkürzungen von Adjektiven, Adverbien etc. (*ugs., bzw.*),

aber auch Abkürzungen, die neben substantivischen auch andere Bestandteile enthalten (*GmbH, UdSSR*).

Für die Orthografie einer substantivischen Abkürzung ist nur der semantische Gehalt von Bedeutung. D.h. es gibt keine Wohlgeformtheitskriterien, deren Einhaltung oder Missachtung auf einen Fehler schließen ließen. Die Masse der Abkürzungen, die in den verschiedenen Fachsprachen gebräuchlich sind, ist darüber hinaus nicht lexikalisierbar. Es existiert vielmehr eine Varietät, die häufig zu Überschneidungen oder nur geringfügigen Abweichungen führt. Ein Versuch, die Abkürzungen in einem juristischen Text zu korrigieren, ist nur mit tiefer Fachkenntnis zu leisten, da zu jedem Gesetzesakronym geprüft werden müsste, ob das jeweilige Gesetz in dem gegebenen Zusammenhang semantisch korrekt angewendet wurde. Dieses Problem verlässt also den Bereich der Orthografie und betrifft den Bereich der semantischen Analyse, die derzeitig von einem elektronischen System noch nicht zufrieden stellend geleistet werden kann. Die Entscheidung, großgeschriebene Zeichenketten von der Rechtschreibkorrektur auszunehmen, muss also als vernünftig bewertet werden, obwohl dadurch ein weiterer Bereich der Korrektur verschlossen bleiben muss.

Akronyme werden im Gegensatz zu Abkürzungen häufig nur teilweise oder am Anfang großgeschrieben (*BUND, Unicef, Juso, BamS*), also sind sie Gegenstand der Rechtschreibkorrektur. Falls Ambiguitäten auftreten wie im Falle von *BUND* wird der Deutlichkeit halber großgeschrieben. Im Prinzip gilt für Akronyme das Gleiche wie für Abkürzungen, mit dem Unterschied, dass Akronyme in dem Sinne wohlgeformt sein müssen, dass ihre Aussprache möglich ist. Eine Korrektur nach Aussprachekriterien verbietet sich allerdings schon aufgrund der verschiedenen möglichen Schreibweisen für einzelne Phoneme oder die Längenmarkierung. Folgende Abkürzungen und Akronyme wurden von den Programmen beanstandet:

Die Anzahl der beanstandeten Abkürzungen ist in etwa gleich, ihre Natur aber unterschiedlich. Beide Programme beanstanden vornehmlich Abkürzungen, die

DITECT			Word		
BamS	imm	Dehoga	T.s	Winterabo	Unicef
UNMOVIK	OSVer		dpa-AFX	winterabo	BamS
ca.	St.		imm	DJs	Kitas
tel.	Winterabo		OSVer		Dehoga
EM.TV	winterabo		SpVg		E.on

Tabelle 3.16: Abkürzungen und Akronyme

Flexions- oder Derivationssuffixe oder andere Bestandteile wie Satzzeichen enthalten (*OSVer, DJs, Kitas, E.on, dpa-AFX, EM.TV*). Ob DITECT im Falle von Großschreibungen ähnlich vorgeht wie Word, kann anhand dieser wenigen Beispiele nicht beantwortet werden. Denkbar wäre, dass lediglich Abkürzungen bis zu drei oder vier Zeichen ignoriert werden. Die einzige rein substantivische und durchgehend großgeschriebene Abkürzung wird von DITECT beanstandet. Es handelt sich dabei um das Akronym UNMOVIK, einen neueren tagespolitischen Term, der beim letzten Update des Programms noch nicht vorliegen konnte.

Bei DITECT fällt vor allem auf, dass allgemeinsprachliche Abkürzungen offenbar nicht im Lexikon enthalten sind (*ca., St., tel.*). Diese Begriffe sind zu häufig, als dass sie bei der Erstellung des Lexikons hätten übersehen werden können. Dieser Befund legt den Schluss nahe, dass auch andere Abkürzungen nicht im Lexikon vorhanden sind, sondern ähnlich behandelt werden wie bei Word.

3.1.11 Zusammenfassung

Im Bereich Fehlerdetektion liegt eines der Hauptprobleme beider Korrektursysteme. Mit einem Anteil an falschen Fehlermeldungen von ca. 95 Prozent ist hier noch einiges an Forschung und Entwicklungsarbeit zu wünschen. Beide Programme haben Schwächen in der Erkennung von Eigennamen, fremdsprachlichen Begriffen und Akronymen. Dies führt dazu, dass Fehler innerhalb solcher Lexeme effektiv nicht berichtigt werden können.

Word zeigt zusätzlich Schwächen in der Erkennung von Komposita mit Bindestrich, DITECT gibt bei Komposita ohne Bindestrich zu viele Fehlermeldungen aus. Treten Bindestriche auf, wird von DITECT nur der Teil hinter dem Bindestrich korrigiert. Dadurch kann kein plausibler Korrekturvorschlag erstellt werden. Zudem bleibt ein großer Teil der Fehler unerkannt. Word hat in den Bereichen Bindestrichkomposita und Korrektur von manueller Silbentrennung eindeutig bessere Ansätze.

Beide Programme leisten zu wenig im morphologischen Bereich. Word ist in der Lage, Flexionsformen abzubilden, zeigt aber Schwächen, Wortbildungskonstruktionen produktiver Muster nachzuvollziehen. DITECT zeigt keine Lösungen zur Erkennung von Derivationsprodukten gespeicherter Grundmorpheme. Flexionsformen werden häufig als fehlerhafte Wortformen gekennzeichnet. Beide Programme sind zudem nicht in der Lage, grammatische Fehler zu erkennen und zu korrigieren.

Die Stärken der Fehlerdetektion liegen in der Kennzeichnung einfacher Rechtschreibfehler. Wird durch einen typografischen Fehler aber eine wohlgeformte Wortform erzeugt (*Seite* → *Seide*), ist die Erkennung des Fehlers unwahrscheinlich.

3.2 Bewertung der Fehlerkorrektur

Die bisherige Betrachtung, die der Fehlerdetektion vorbehalten war, konzentrierte sich auf die Natur der beanstandeten Wortform, also zu einem nicht geringen Teil auf die Beschaffenheit des Lexikons. Die nun folgende Untersuchung betrachtet die Korrekturmethodik der Programme. Dabei muss neu kategorisiert werden. Als Gliederungspunkte dienen die einfachen Fehlertypen, die im theoretischen Kapitel beschrieben wurden: Auslassung, Einfügung, Substitution und Transposition. Es bleibt allerdings nicht aus, dass auf Begriffe aus einer der acht Kategorien aus Kapitel

3.1 verwiesen wird, da diese nach wie vor den Daten-Pool darstellen, der für die Untersuchung verwendet wird. Zunächst wird der Anteil betrachtet, den die einzelnen Fehlertypen am Gesamtbestand der Fehler haben.

3.2.1 Verteilung der typografischen Fehlertypen

Die orthografischen Fehler im Testkorpus wurden nach der Fehlertypendefinition aus 2.1.2 (S. 13) kategorisiert. Die Zuordnung war nicht immer eindeutig zu leisten, da in wenigen Fällen multiple Fehler, d. h. mehrere Fehler in einer Wortform, auftraten. Die Zuordnung geschieht dabei rein formal auf der Zeichenebene. Für das Lexem *Mikrofon* existieren beispielsweise die beiden Schreibweisen *Mikrofon* und *Mikrophon*. Letztere wurde als Fehler betrachtet, da die empfohlene und im Verlag übliche Schreibweise die Variante der neuen Rechtschreibung *Mikrofon* ist. Aus phonetischer Sichtweise handelt es sich hier um die Substitution der beiden im Deutschen möglichen Graphem-Phonem-Korrespondenzen für den Laut [f], also um einen einfachen Fehler ([f] → <ph>, [f] → <f>). Die Minimum-Edit-Distance beträgt aber 2, auf der Zeichenebene handelt es sich also um zwei Fehler, da ein Buchstabe substituiert und der andere eingefügt werden muss, um von der Zeichenkette *Mikrofon* zu *Mikrophon* zu gelangen. In diesem Fall ist ein solches Vorgehen befremdlich, da die Ursache des Fehlers nicht eine motorische Fehlleistung, sondern die Wahl einer anderen Schreibvariante, also das Resultat eines kognitiven Prozesses ist.

Ähnlich schwierig gestaltet sich die Kategorisierung nach diesem System, wenn ein gesamtes Wort ausgelassen, zusätzlich eingefügt ist oder an der falschen Position steht. Um nicht mit zweierlei Maß zu messen, wurde das System der typografischen Fehlertypen jedoch beibehalten und auch im Falle einer Worteinfügung jedes zusätzliche Zeichen im Text als Einfügungsfehler bewertet. Die Kategorisierung von

Wortfehlern erfolgte allerdings gesondert am unteren Ende der Tabelle. Im Folgenden wird die Übersicht über die vorhandenen Fehler dargestellt:

Fehlertyp	Fehleranzahl		ø MED	DITECT						Word					
				gefunden		Pos. 1		Pos. 2		gefunden		Pos. 1		Pos. 2	
Auslassung	48	44 %	1,10	25	52 %	5	10 %	8	17 %	30	63 %	23	48 %	0	0 %
Einfügung	17	15 %	1,24	5	29 %	3	18 %	0	0 %	7	41 %	6	35 %	0	0 %
Substitution	34	31 %	1,15	9	26 %	3	9 %	1	3 %	8	24 %	2	6 %	2	6 %
Transposition	3	3 %	1,00	2	67 %	2	67 %	0	0 %	2	67 %	2	67 %	0	0 %
Einfügung, Substitution	2	2 %	2,00	0	0 %	0	0 %	0	0 %	0	0 %	0	0 %	0	0 %
Wortauslassung	2	2 %	3,00	0	0 %	0	0 %	0	0 %	0	0 %	0	0 %	0	0 %
Worteinfügung	4	4 %	4,50	1	25 %	0	0 %	0	0 %	1	25 %	0	0 %	0	0 %

Tabelle 3.17: Typografische Fehler bei DITECT und Word

Fast die Hälfte aller Fehler geht auf Auslassungen von Lettern oder Leerzeichen zurück. Innerhalb dieser Kategorie ist auch die Erkennungsrate am höchsten: DITECT erkennt die Hälfte aller Auslassungsfehler, Word etwas mehr. Die Substitutionsfehler bilden mit etwas mehr als 30% die zweithäufigste Fehlergruppe, gefolgt von Einfügungsfehlern mit 15%. In diesen beiden Fehlerkategorien ist die Erkennungsleistung bei beiden Programmen deutlich geringer als beim Typ Auslassung. Transposition und multiple Fehler treten nur selten auf. Auf die Natur der Fehler wird in den entsprechenden Unterkapiteln näher eingegangen.

Anhand der Tabelle wird deutlich, dass die Annahme Dameraus, es handele sich in der Mehrzahl der Tippfehler um einfache Fehler (vgl. Damerau 1964: 175), eine durchaus veritable Arbeitshypothese darstellt. Lässt man Worteinfügungen und -auslassungen außen vor, bewegt sich die Minimum-Edit-Distance in einem Rahmen, der deutlich unter 1,3, also bei etwa einem Editierungsschritt zur Berichtigung pro falsch geschriebener Wortform liegt. Es treten zwar multiple Fehler wie im oben

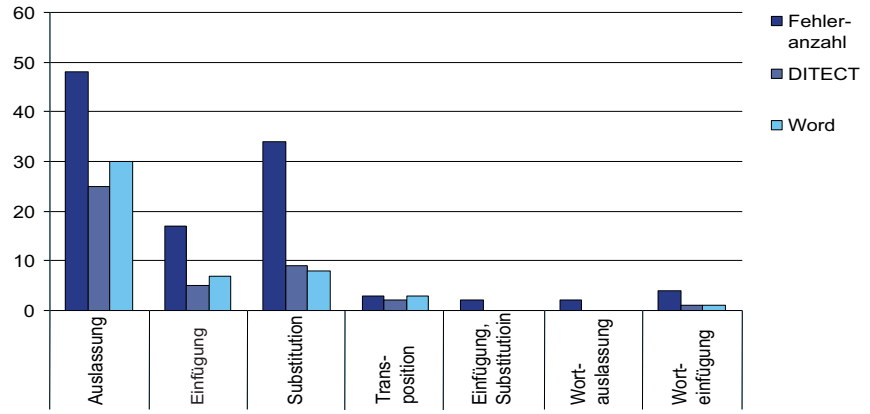

Tabelle 3.18: Fehlerdetection bei DITECT und Word

beschriebenen Beispiel auf, das Gros lässt sich aber auf einfache Fehlleistungen zurückführen.

Tabelle 3.19 zeigt, wie das Verhältnis zwischen Word und DITECT bei der Fehlerkorrektur aussieht. Die Balken zeigen die Korrekturleistung an, wobei der blau markierte Bereich für richtige Korrekturvorschläge an erster Position und der

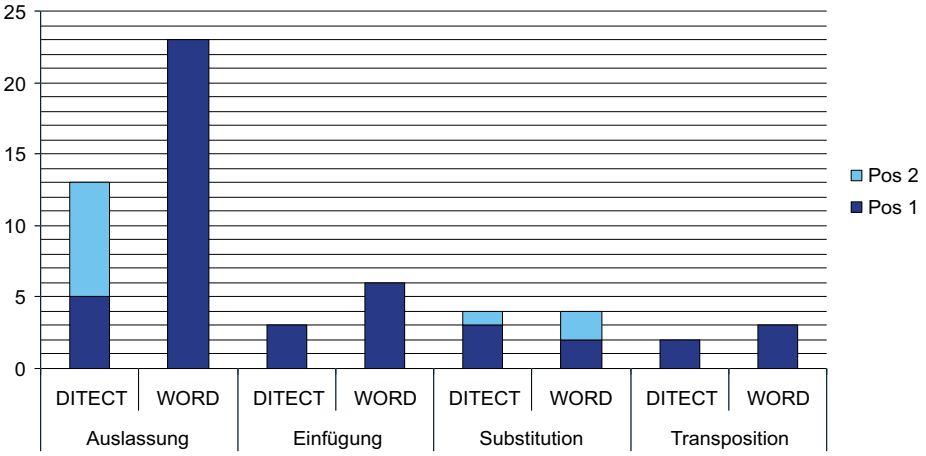

Tabelle 3.19: Fehlerkorrektur bei DITECT und Word

rot markierte Bereich für richtige Korrekturvarianten an zweiter Position steht. Word zeigt insgesamt bessere Korrekturleistungen als DITECT. Dies ist besonders deutlich bei den Auslassungsfehlern. Word korrigiert doppelt so viele Fehler wie

DITECT und diese alle mit dem ersten Korrekturvorschlag, während sich die richtige Wortform bei DITECT häufig erst im zweiten Vorschlag findet. Auch bei den Einfügungsfehlern liegt die Korrektur von Word deutlich vorn.

Beide Programme korrigieren etwa gleich viele Substitutionsfehler. DITECT korrigiert von neun gefundenen Fehlern vier. Word findet nur acht Fehler und korrigiert ebenfalls vier. Bei insgesamt 34 vorhandenen Substitutionsfehlern ist die Korrekturleistung damit bei beiden Programmen sehr schwach.

3.2.2 Auslassungsfehler

Auslassungsfehler sind im untersuchten Testkorpus die häufigsten aller Fehlertypen. Dabei handelt es sich um typografische (*fortwärendenkampfes, in de Nacht*) sowie kognitive Fehler (*andersfalls, Wachstumprognose*). Die Korrekturleistung ist von der Beschaffenheit des einzelnen Fehlers abhängig. Die Zeichenkette *fortwärendenkampfes* konnte beispielsweise nicht berichtigt werden. Es besteht ein multipler Fehler und die korrekte Schreibung besteht in zwei Lexemen, was die Korrektur erschwert. Gleichwohl wurde der Fehler gefunden. Dies war beim Auslassungsfehler im Artikel der Nominalgruppe *in de Nacht* nicht möglich, da die Zeichenfolge *de* in einigen Kontexten eine akzeptable Wortform darstellt. Dies betrifft lediglich Fremdsprachen, weshalb der Fehler hier einfach zu finden sein sollte.

Die Wortformen *andersfalls* und *Wachstumprognose* können von Word effektiv berichtigt werden. Das Programm liegt in beiden Fällen mit dem ersten Vorschlag richtig. DITECT liefert in beiden Fällen ein schlechteres Resultat. Im ersten Beispiel bietet das Programm zuerst die großgeschriebene Variante von *andernfalls* an, im zweiten Beispiel findet es den Fehler überhaupt nicht. Auch im Beispiel *Katasrophe* bestätigt sich die in Tabelle 3.19 getroffene Aussage: Word kann den Fehler mit dem ersten Korrekturvorschlag berichtigen, während DITECT dies erst im zweiten

Anlauf bewerkstelligt. Der erste Vorschlag stellt keine gültige Wortform dar. Offenbar ist hier das Lexikon schlecht gepflegt worden (siehe Tab. 3.20).

Besonders häufig ist die Auslassung des Apostrophs, der für weggelassene Laute am Wortanfang eingefügt wird (Duden, Rechtschreibung: 21). Bei den auftretenden Klitika handelt es sich um eine Enklise der phonologisch schwachen Form *es* an die finite Verbform (*stehts, gings, gabs, wenns*). Hier kann noch nicht von Grammatikalisierungsvorgängen gesprochen werden, weshalb die Auslassung des Apostrophs als Fehler angesehen werden muss. Beide Programme sind in der Lage, diese Fehler aufzuspüren und zu berichtigen. DITECT findet den Fehler jedoch nur in etwa drei Viertel der Fälle. Die Rechtschreibprüfung von Word funktioniert hier konsequenter. Jeder fehlende Apostroph einer Enklise wird gefunden und korrigiert.

Viele andere Fehler sind auf die Neuregelung der Getrennt- und Zusammenschreibung zurückzuführen. Die Zusammenschreibung *nochmal* ist umgangssprachlich wahrscheinlich analog gebildet zu *einmal, zweimal, dreimal,* etc. Sie besteht aus den beiden getrennten Wortformen *noch* und *einmal* bzw. umgangssprachlich *mal*. Nach der Neuregelung ist diese Schreibung erlaubt. Word findet hier jedoch einen Fehler.

Zeichenkette	korrekte Schreibung	DITECT		Word	
		Vorschlag 1	Vorschlag 2	Vorschlag 1	Vorschlag 2
Chistian	Christian	Christina	Christian	Christian	
andersfalls	andernfalls	Andernfalls	andernfalls	andernfalls	Andersfalls
Katasrophe	Katastrophe	Katastophe	Katastrophe	Katastrophe	
Wachstumprognose	Wachstumsprognose			Wachstumsprognose	
spanende	spannende			spannende	sparende
tafen	trafen	tafeln	taften	taufen	taften
sowas	so was	soaps	Sofas	so was	
vollgepackte	voll gepackte	voll gepackt	vollbepackte		

Tabelle 3.20: Korrektur von Auslassungsfehlern

Des Weiteren werden streng nach Regelwerk viele Verbindungen mit Verben oder Partizipien falsch zusammengeschrieben (*zurückkönnen, tiefgreifende*). Eine genaue Betrachtung des Phänomens Getrennt- und Zusammenschreibung findet sich in

Kapitel 4, wo systematisch überprüft wird, inwieweit die Programme in der Lage sind, dieses Problem zu behandeln. Andere Einzelfälle sind kein Produkt der Neuregelung, sondern Zusammenschreibungen von Wörtern, die häufig nebeneinander stehen (*Darüberhinaus, nichtständigen (Mitglieder des Sicherheitsrats)*). Häufige Fehler wie *desweiteren* oder *darüberhinaus* werden von beiden Programmen berichtigt.

Tabelle 3.20 stellt einige der aufgetretenen Auslassungsfehler und deren Korrekturvorschläge dar. Von den acht Beispielen konnte Word sechs mit dem ersten Vorschlag berichtigen. DITECT schaffte drei im zweiten Anlauf, brachte also ein deutlich schlechteres Ergebnis. Fraglich ist, warum die richtigen Vorschläge in den ersten beiden Beispielen an zweiter Position auftreten. In beiden Fällen ist der erste Vorschlag um je einen Editierungsschritt weiter von der beobachteten Zeichenfolge entfernt als der zweite, also weniger plausibel.

Offenkundig wird die Schwäche DITECTs, mit Flexionsformen umzugehen, so auch im letzten Beispiel in der Tabelle. Die Getrenntschreibung der beiden Konstituenten wird zunächst richtig vorgeschlagen, die Flexionsendung jedoch nicht korrekt wiedergegeben. In der zweiten Variante stimmt die Endung, das Lexem enthält aber eine andere Konstituente als die Beobachtung und ist zusammengeschrieben. Eine mögliche Erklärung ist, dass Flexionsformen nicht im Lexikon enthalten sind und daher der richtige Vorschlag nicht gemacht werden kann. Dagegen spricht der Vorschlag 2, der eine flektierte Form enthält, die der falschen Schreibung entspricht.

Das Ergebnis besteht in uneinheitlichen Korrekturvorschlägen hinsichtlich der Flexionsendung und hinsichtlich der Entscheidung zugunsten von Getrennt- oder Zusammenschreibung. Beides sollte im Idealfall gleich behandelt werden. Die Korrektur wirkt hier wie ein einfacher Lexikonabgleich, ohne dass dem eine kohärente Struktur zugrunde läge. Ein praktischer Nutzen ist hier nicht erkennbar.

3.2.3 Substitutionsfehler

Die zweithäufigste Fehlergruppe bilden die Substitutionsfehler. Eine Hauptursache ist die Neuregelung der Groß- und Kleinschreibung (**im einzelnen, *abend, *Blauer Brief*) bzw. die Großschreibung von Substantivierungen (**der vornehmste, *zum Schweigen*). Abgesehen vom Lexem *Abend*, das in der Grundform nur großgeschrieben auftritt, ist keiner dieser Fehler von einem der Programme erkannt worden (detaillierte Auflistung aller Fehler in Anhang B). Andere auf Groß- und Kleinschreibung zurückzuführende Fehler wie die Binnenmajuskel im Wort *Bier-Riese* wurden von beiden Programmen erkannt, aber nicht korrigiert. Die gewollte Kleinschreibung im Namen der Dienstleistungsgewerkschaft *ver.di* konnte ebenso nicht erkannt werden, da die Homophonie mit dem Namen des Komponisten *Verdi* zur Wahl dieser Schreibvariante geführt hat. Da die Zeichenfolge *Verdi* eine gültige Wortform darstellt, wird sie nicht beanstandet. Es handelt sich hier um einen Real-Word-Error.

Zeichenkette	korrekte Schreibung	DITECT		Word	
		Vorschlag 1	Vorschlag 2	Vorschlag 1	Vorschlag 2
Repiblik	Republik	Republik	Republic	Republik	
Megabites	Megabytes	Megabits	Megabytes	Megabytes	
Know How	Know-how				
Know				–	
How		how	Howie	Hob	Hof
Elvtalaue	Elbtalaue	Elbtalaue		–	
verkaufes	verkaufen	Verkaufes	verkauftes	Verkauftes	verkaufest
Werks-klub	Werks-Klub	Klub	Klaub	Werks-klug	Werks-Klub

Tabelle 3.21: Korrektur von Substitutionsfehlern

Einfache orthografische Substitutionsfehler treten in dieser Kategorie weniger häufig auf, werden aber in der Mehrzahl gefunden und es wird der richtige Korrekturvorschlag gemacht. Die Auswahl in Tabelle 3.21 enthält Beispiele mit einfachen Fehlern, wie *Elvtalaue*, in dem zwei auf der Tastatur benachbarte Buchstaben vertauscht wurden. Word konnte diesen Fehler nicht berichtigen, vermutlich aufgrund der Komplexität des Kompositums. DITECT machte den richtigen Korrekturvor-

schlag. Die Zeichenkette *verkaufes* vermochte keines der Programme zu berichtigen, da zu viele Verbformen in Frage kommen. Ohne weitere Informationen über die syntaktischen Einheiten des Satzes sind die Vorschläge arbiträr und relativ unbrauchbar.

DITECT zeigt sich in der Auswahl der Korrekturvorschläge flexibler. Word präsentiert fast ausschließlich Vorschläge, die einen einfachen Fehler zur Grundlage haben. DITECT lässt auch Varianten zu, die mehrere Editierungsschritte erfordern, wie von *Repiblik* zu *Republic*. Dennoch erreicht Word eine bessere Korrekturrate. In Tabelle 3.16 hat sich gezeigt, dass die meisten Fehler auf einfache Fehlleistungen zurückgehen. Daher ist es vernünftig, auch bei der Korrektur nur einfache Fehler zuzulassen, da sonst die Abweichung zum korrekten Lexem zu groß wird.

Die Bindestrichauslassung bei *Know-how* ist in dieser Form nicht von den Programmen erkannt worden. Dieser Fehler könnte theoretisch auch als Auslassungsfehler betrachtet werden, er erfordert jedoch die Einfügung eines Leerzeichens. Ohne dieses wäre der Fehler relativ unproblematisch zu berichtigen. So reicht er jedoch über die Wortgrenze hinaus, was die Erkennung erschwert. Daher werden nur die einzelnen Konstituenten getrennt voneinander behandelt. Etwas augenfälliger ist das Bindestrichproblem im Beispiel *Werks-klub*. DITECT ist nicht in der Lage, diesen Fehler zu berichtigen, da es den Bindestrich vermutlich als Wortgrenze annimmt. Das Programm berichtigt nur den fehlerhaften Teil des Wortes. Word hingegen betrachtet das Bindestrichkompositum als Ganzes und kann den richtigen Korrekturvorschlag machen.

Im Beispiel *Megabites* sind zwei Varianten innerhalb von einem Schritt zu erreichen, die auch semantisch sehr ähnlich sind: *Megabits* und *Megabytes*. Typografisch wäre die Variante *Megabits* plausibler, da der Fehler durch das gemeinsame Treffen der Tasten *E* und *S* entstanden sein könnte, während *I* und *Y* sehr weit voneinander entfernt sind. Phonologisch gesehen ist der Variante *Megabytes* der Vorzug zu geben,

da diese der Aussprache der fehlerhaften Zeichenkette entspricht. Diese Variante ist tatsächlich die richtige. Bei DITECT treten beide Varianten auf, die richtige jedoch erst an Position 2. Word macht nur einen Vorschlag und erzielt damit die korrekte Prognose.

Eine weitere Fehlerquelle in dieser Kategorie, die zu einem Großteil unerkannt und unkorrigiert blieb, ist im Bereich der syntaktischen Relationen und der Einheit der grammatischen Kategorien innerhalb von Nominalgruppen zu finden. Häufig sind diese Beziehungen nicht eingehalten (*mit gebremsten Verve, dem Grundsatzfragen, in wahrsten Sinne*), ein Problem, das auch im Rahmen der Einfügungsfehler häufig auftritt (*Der Preisanstiegs, zum Kerngeschäften*). Detektion und Korrektur dieser Fehler gestalten sich so schwierig, da es sich bei den Kandidaten ausnahmslos um gültige Wortformen handelt. Zur Korrektur ist der Kontext entscheidend. Trigrammstatistiken können dabei hilfreich sein (vgl. Kap. 2.1.6). Bei der Fehlerdetektion sind sie jedoch nicht tauglich, da der Text zur Überprüfung vollständig auf Trigrammketten abbildbar sein muss. Jedes nicht gespeicherte Trigramm würde zu einem Abreißen der Kette und somit zu einer Beanstandung des aktuellen syntaktischen Zusammenhangs führen.

Zu einer effektiven Grammatikprüfung ist vielmehr ein Parser notwendig, der die syntaktische Struktur analysiert und direkt auf Fehler in syntagmatischen Beziehungen hinweisen kann. Der nötige Rechenaufwand bei einer Grammatikprüfung ist allerdings immens, da jede Wortform auf die Korrektheit ihrer syntaktischen Umgebung geprüft werden muss. Word bietet dies mit der Grammatikprüfung an. Diese ist jedoch äußerst unzuverlässig und wird aus diesem Grund nicht standardmäßig mit der Rechtschreibkorrektur ausgeführt.

3.2.4 Einfügungsfehler

In der drittgrößten Gruppe, der Kategorie der Einfügungsfehler, treten hauptsächlich typografische Fehler auf (*Ministerprsäsident, vorausssichtlich, beigetreiten*). In diesen drei Fällen konnten beide Programme im ersten Anlauf die richtige Korrekturhypothese erstellen. Alle drei Beispiele enthalten einfache Einfügungsfehler bei Wörtern in der Grundform.

Die Wortform *weltpolistische* konnte nur von Word berichtigt werden. DITECT bietet stattdessen zunächst das korrekte Lexem, aber statt der flektierten Form nur die Grundform *weltpolitisch*. Als Vorschlag 2 wird das substantivische Kompositum *Weltpolizisten* angeboten. Man muss mindestens einen fünffachen Fehler annehmen, um diesen Vorschlag zu rechtfertigen, was nahezu ausgeschlossen ist. Das Beispiel legt den Schluss nahe, dass in dem von DITECT verwendeten Lexikon lediglich die Grundform *weltpolitisch* gespeichert ist, da sich ansonsten andere Flexionsformen in den Korrekturvorschlägen finden müssten.

Zeichenkette	korrekte Schreibung	DITECT		Word	
		Vorschlag 1	Vorschlag 2	Vorschlag 1	Vorschlag 2
Vermögenssteuer	Vermögensteuer				
Ministerprsäsident	Ministerpräsident	Ministerpräsident	Ministerpräsidenten	Ministerpräsident	–
vorausssichtlich	voraussichtlich	voraussichtlich	Voraussichtlichkeit	voraussichtlich	–
UN-Waffeninspektektionskommission	UN-Waffeninspektionskommission	–		–	
beigetreiten	beigetreten	beigetreten	beigetretene	beigetreten	beigebreiten
weltpolistische	weltpolitische	weltpolitisch	Weltpolizisten	weltpolitische	weltsolistische
auf's	aufs			aufs	
Meier's	Meiers			Meiers	

Tabelle 3.22: Korrektur von Einfügungsfehlern

Word bietet in diesem Fall die richtige Wortform an und als Alternative den Begriff *weltsolistische*, der ebenfalls nur einen Editierungsschritt erfordert, um zur beobachteten Zeichenkette zu gelangen. Nicht nur die Flexionsform wird berücksichtigt, es wird auch ein passendes Kompositum aus *Welt* und *solistisch* gebildet.

Die Zeichenkette *UN-Waffeninspektektionskommission* enthält einen multiplen Fehler, der von beiden Programmen gefunden, aber offenbar aufgrund der Länge des Lexems nicht korrigiert werden konnte.

Im Falle von *Vermögenssteuer* ist entgegen der Duden-Norm ein Fugen-s eingefügt worden. Da diese Schreibweise in den betreffenden Artikeln häufiger auftritt, wird von einem kognitiven Fehler ausgegangen. Keines der beiden Programme hat diesen Fehler erkannt.

Ebenfalls zu den typografischen Einfügungsfehlern muss der überflüssige Apostroph zählen, wie bei der enklitischen Verschmelzung *auf's* oder dem besitzanzeigenden *s* in *Meier's.* Die Apostrophfehler konnten wiederum von Word berichtigt werden und von DITECT nicht. Es muss eingeräumt werden, dass die Grammatikalität der Verschmelzung *auf's* nicht hundertprozentig eindeutig ist. Analog zu anderen Verschmelzungen aus Präposition und Artikel wird das Apostroph hier aber als Fehler angesehen.

Die falsche Auseinanderschreibung von *weiter gearbeitet, weiter machen* stellt ein größeres Problem dar. Sie ist von beiden Programmen in keinem der Fälle entdeckt worden. Dass Getrennt- und Zusammenschreibung ein breites Problemfeld darstellt, ist bereits angesprochen worden. Das Thema wird in Abschnitt 4.3 eingehend behandelt.

3.2.5 Transpositionsfehler

Transpositionsfehler beruhen auf dem Vertauschen zweier Buchstaben. Diese Fehlerquelle trat im Testbestand eher selten auf (*Bretange, wordne, entsprechnede*). In allen drei Fällen ist das *n* vertauscht worden, zweimal mit dem Vokal *e*, einmal mit dem Konsonanten *g.* Aufgrund der französischren Aussprache von *Bretagne* bzw. des Schwalauts *en* kann vermutet werden, dass der Fehler nicht rein typografischer

Natur ist. Tritt die Vertauschung von *en* durch *ne* häufiger auf, müsste sich dies in der entsprechenden Verwechslungsmatrix niederschlagen.

Zeichenkette	korrekte Schreibung	DITECT		Word	
		Vorschlag 1	Vorschlag 2	Vorschlag 1	Vorschlag 2
Bretange	Bretagne	Bretagne	Brenntage	Bretagne	
wordne	worden	worden	wofern	worden	
entsprechnede	entsprechende	entsprechend	Entsprechendes	entsprechende	

Tabelle 3.23: Korrektur von Transpositionsfehlern

Wieder zeigt sich bei der Korrektur ein ähnliches Bild wie in den anderen Kategorien. Word liefert jeweils nur einen Korrekturvorschlag und erstellt damit in jedem Fall die richtige Hypothese. DITECT tut sich schwerer. In den ersten beiden Fällen wird die richtige Hypothese erstellt, es erscheinen aber noch weitere Vorschläge, die von dem ursprünglichen Lexem zu weit abweichen, als dass sie einen plausiblen Korrekturvorschlag darstellten. Im letzten Beispiel ist von DITECT die Flexionsendung nicht berücksichtigt worden. Als erste Variante wird die Grundform des Adjektivs, als zweite Variante eine substantivierte Flexionsform angegeben, deren Endungen nicht mit der beobachteten Wortform übereinstimmen.

Was bereits im Rahmen der Flexion und der Getrennt- und Zusammenschreibung beobachtet wurde, wiederholt sich im Bereich Groß- und Kleinschreibung. Die Korrektur bei DITECT wirkt inkohärent hinsichtlich der grammatischen Markierungen Substantivierung und Flexion. Der Fehler wird zwar korrigiert, die übrigen Markierungsverhältnisse innerhalb der Wortform werden allerdings stark verändert.

Trotz dieser Schwächen hat sich gezeigt, dass Transpositionsfehler den Fehlertypus darstellen, den DITECT am häufigsten berichtigt, während die Leistung bei Auslassungs- und Einfügungsfehlern schlechter ausfällt. Die Neuordnung der Buchstaben bereitet offenbar weniger Probleme als die Veränderung der Wortlänge mit der sich daraus ergebenden Komplexität.

3.2.6 Sonstige Fehlerquellen

Multiple Fehler, die mehr als einem Fehlertyp zugrunde liegen, treten bis auf den Fall *Mikrophon,* der einleitend bereits diskutiert wurde, nicht auf. Dieser Fehler wurde von beiden Programmen jedoch nicht berichtigt, da beiden Schreibvarianten gültig sind.

Besonderes Augenmerk liegt hier auf der Einfügung oder Auslassung ganzer Wörter. Wortauslassungen werden von beiden Programmen nicht erkannt. Beide Programme beanstanden jedoch Wörter, die in direkter Folge wiederholt werden, wobei es sich nicht in allen Fällen um einen Fehler handelt. Häufig sind beispielsweise ein einleitendes Relativpronomen und ein darauf folgender Artikel. Hier bietet keines der Programme Vorschläge zur Korrektur an.

3.2.7 Besonderheiten der Fehlerkorrektur bei DITECT und Word

In Tabelle 3.24 wird ein fundamentaler Unterschied in der Strategie der beiden Programme deutlich. Die Vorschläge, die Word unterbreitet, sind viel näher an der beobachteten Zeichenfolge als die Vorschläge bei DITECT. Keiner der Korrekturvorschläge bei Word weicht in mehr als einem Editierungsschritt von dem zu korrigierenden Wort ab. Das Prinzip des einfachen Fehlers ist also konsequent durchgehalten.

Für die Auswahl der Korrekturvariante ist dieses Vorgehen sehr förderlich. Alle Vorschläge sind höchst plausibel. Die Schattenseite ist, dass multiple Fehler ausnahmslos zwar gefunden, aber nicht berichtigt worden sind (*US-Zeichtrickserie, fortwärendenkampfes*). Dennoch gewährleistet dieses enge Korsett eine effektive Korrektur einfacher orthografischer Fehler.

DITECT scheint keine Beschränkung hinsichtlich der Minimum-Edit-Distance zu besitzen, zieht daraus aber bei der Korrektur multipler Fehler keinen Vorteil. DITECT war ebenso wenig wie Word nicht in der Lage, einen einzigen multiplen

Zeichenkette	DITECT		Word	
	Vorschlag 1	Vorschlag 2	Vorschlag 1	Vorschlag 2
Innensena-tor	to	Tor	Innensenator	
Integra-tionskraft	Timokrat	Tischkarte	Integrationskraft	
Borghoff	Bohrkopf	Bohrgut	Borgzoff	Borghaff
Bullerdieck	Bullterrier	Buchdrucke	Bullerdick	Bullerdeck
Fiebelkorn	Fibelkinder	fieberkrank	Fieberkorn	Fiedelkorn
Gerets	Gerste	Gerster	Geerts	Genets
Hintum	hintue	hintum	Hintut	Hintun
Knorre	Knorren	Knorzer	Knurre	Knarre
Pellengahr	Pelagianer	Pelzgegner	Pellenwahr	Pellenjahr
Raffelhüschen	Raffaelisch	Randgeschehen	Raffelhübschen	Raffelhülschen
Schröcksnadel	Schmuckhandel	Schmucknadel	Schrecksnadel	
Funtensee	Fugenteile	Fundstelle	Funkensee	Furtensee
Heiligenrode	Heiligenberg	Heiligenfeld	Heiligenrüde	Heiligenrohe
Senftleben	Setzfehlern	Sendefolgen	Senfleben	Senftreben
Anscheinserwecker	Anästhesiezwecke	Anästhesiezwecken	Anscheinserweckter	Anscheinserwecken
Kettenkarrussel	Kettenkarussell	Kettenkarussells	Kettenkarrdussel	Kettenkarrfussel
Nasenstupser	Nasenstübers	Nasenblutens	Nasenstupses	Nasenstupse
Winterabo	Winterbau	Winterobst	Winterambo	

Tabelle 3.24: Allgemeine Korrektur bei DITECT und Word

Fehler zu korrigieren. Die Korrekturvorschläge weichen hier äußerst stark von der Beobachtung ab. Teilweise ist noch eine Verwandtschaft der Lexeme zu erkennen (*Heiligenrode → Heiligenberg*). Die gemeinsame Erstkonstituente resultiert aber wahrscheinlich nur aus der allgemeinen Ähnlichkeit der beiden Begriffe. In anderen Fällen ist diese Ähnlichkeit nicht gegeben (*Anscheinserwecker → Anästhesiezwecke*). Höchstwahrscheinlich sind alle Korrekturvorschläge als volle Wortformen lexikalisiert.

Die Annahme, DITECT sei nicht zur Wortbildung oder Kompositaanalyse fähig, ergibt sich aus der Beobachtung bei Word. Hier findet sich eine Vielzahl an Korrekturvorschlägen, die von dem Programm selbst generiert sein müssen, da sie nicht lexikonfähig sind (*Raffelhüschen → Raffelhübschen, Heiligenrode → Heiligen-*

rüde). Dadurch steigt die Zahl der möglichen Wortformen um ein Vielfaches, was dazu führt, dass sich Word die Annahme des einfachen Fehlers leisten kann. Bei DITECT blieben, würde man dieses Konzept konsequent durchsetzen, in dieser Auswahl nur drei, streng genommen nur zwei Begriffe übrig[6] (*hintue, hintum, Knorren*).

Weitere Probleme liegen in der Behandlung von Bindestrichen. Die Korrektur, die Word problemlos leistet, ist bei DITECT nicht möglich. So wird auch die geringe Zahl der beanstandeten Komposita mit Bindestrich plausibel. DITECT beschränkt sich vermutlich in der Korrektur auf die Zweitkonstituente. Daher tauchten die Komposita *Gerets-Team* und *Gerets-Nachfolger* nicht in der Fehlerstatistik bei DITECT auf, wohl aber der Name *Gerets*.

Flexion wird von DITECT in keiner Form abgebildet. Ähnlich wie im Beispiel zur Korrektur von *Heiligenrode* ähneln die Korrekturvorschläge einander, diese Ähnlichkeit ist aber nicht konsequent genug durchgehalten, um sich auch auf die Flexionsendungen zu erstrecken. Die Inkonsistenz der Korrekturvorschläge DITECTs ist in mehreren Beispielen offenkundig geworden. Kernprobleme der Orthografie wie Getrennt- und Zusammenschreibung und Groß- und Kleinschreibung sowie Flexion werden bei der Korrektur von einfachen Fehlern widersprüchlich behandelt. Die Vorschläge wirken dadurch arbiträr und stellen für den Nutzer keine Hilfe dar. Ein übergeordnetes Konzept zur Korrektur ist bei DITECT nicht erkennbar.

4 ANALYTISCHE TESTPHASE

Im letzten Abschnitt der Untersuchung werden noch einmal einige zentrale Probleme herausgegriffen, die in deren Rahmen auftraten, deren Klärung aber anhand der Korpusdaten nicht vollständig möglich war. DITECT zeigt Schwächen bei der Abbildung von Flexionsformen. Das hat der Test deutlich ergeben. Dieser Komplex wird den ersten Teil des analytischen Tests bilden. Anhand von exemplarischen Ausschnitten aus den Flexionsparadigmen von Verben und Substantiven wird überprüft werden, ob die Korrekturprogramme den Flexionsendungen Rechnung tragen und ob Variationen des Stammvokals bei der Flexion die Korrektur beeinflussen.

Der nächste Testteil wird das Verhalten DITECTs bei der Korrektur von Bindestrichkomposita untersuchen. Hier traten im ersten Testlauf einige Unstimmigkeiten auf, in Fällen, wo z.B. der Korrekturvorschlag nur auf die zweite Konstituente des Kompositums zielte. Solche Probleme traten bei Word nicht auf, weshalb sich dieser Test ausschließlich auf das bei der Madsack Verlagsgesellschaft eingesetzte Programm bezieht.

Die letzten beiden Problemkomplexe beziehen sich auf die Neuregelung der deutschen Rechtschreibung, genauer auf die Getrennt- und Zusammenschreibung und die Groß- und Kleinschreibung. Diese Bereiche sind aus zwei Gründen interessant. Zum einen wurde um die Rechtschreibreform eine erhitzte Debatte geführt, ohne dass ein nennenswerter Kompromiss zwischen Befürwortern und Gegnern erreicht wurde. Dies führte dazu, dass sich die »Arbeitsgruppe der deutschsprachigen Nachrichtenagenturen« zusammensetzte und eigene Vorschläge zu deren Umsetzung

erarbeitete, die in den genannten Bereichen Getrennt- und Zusammenschreibung und Groß- und Kleinschreibung grobe Vereinfachungen vorsahen. Welchem Konzept die Korrekturprogramme folgen, soll in diesem Test geklärt werden. Der andere Grund für diesen Teil der Untersuchung ist die Komplexität des Problems. Die Erkennung von Fehlern, die mit der Getrennt- und Zusammenschreibung zusammenhängen, erfordert zumindest bei falscher Getrenntschreibung die Kenntnis des Kontextes, in den das zu untersuchende Wort eingebettet ist. Dies ist aus schon beschriebenen Gründen kompliziert. Im Falle der Groß- und Kleinschreibung liegt der Fall ähnlich, da bei Substantivierungen immer der Kontext entscheidend ist.

4.1 Flexion

Um die Fähigkeit zur korrekten Modellierung von Flexionsendungen zu überprüfen, wurde zu jedem Flexionstyp, wie er bei Eisenberg angenommen wird (Eisenberg [1]: 144ff.), mindestens ein Beispiel ausgewählt (siehe Anhang C). Diese Beispiele wurden im Stamm mit je einem einfachen Fehler versehen, der in der Korrektur zu der entsprechenden Wortform führte. Die unterbreiteten Korrekturvorschläge sind hier nicht von Interesse. Wenn andere Stammformen als die ursprünglich intendierte mit einem einfachen Fehler angenommen werden können, ist dies nicht von Bedeutung für den Erfolg dieses Tests (*lehgen* → *legen, lehren, lehnen*). Entscheidend ist lediglich, ob die grammatischen Kategorien korrekt repräsentiert sind, also Person, Numerus, Tempus und gegebenenfalls Modus beim Verb bzw. Kasus und Numerus beim Substantiv.

4.1.1 Verben

Bei der Verbflexion wurde lediglich zwischen schwacher und starker Flexion unterschieden. Aus diesen Bereichen wurden jeweils drei Beispiele ausgewählt, die als exemplarische Größe für diesen Test ausreichen sollten. Eine Einschätzung der

Flexion schwacher Verben bedarf keiner großen Menge an Beispielen, da die Endungen im Großen und Ganzen gleich sind und der Stamm nicht abgelautet wird. Als Beispiele wurden jeweils ein Verb auf *en*, eins auf *ln* und eines mit der Endung *rn* gewählt.

Bei den stark flektierenden Verben eröffnet sich eine Vielzahl an Kategorien, die häufig nur ein Verb enthalten. Eine umfassende Untersuchung aller Flexionsklassen beim Verb ist aber weder möglich noch nötig. Der Umstand, dass dieses Feld so unüberschaubar ist, führt dazu, dass keine Formengenerierung durchgeführt wird. Die Verbformen werden als ganze Wortformen im Lexikon gespeichert. Es reicht also eine Auswahl aus, um eine Idee vom Verhalten des Programms zu bekommen.

Die Flexionsformen, die für die einzelnen Verben ausgewählt wurden, beschränken sich auf die Präsens- und Präteritumformen des Aktiv. Auf den Konjunktiv wurde verzichtet. Die Indikativformen reichen für eine Einschätzung des Verhaltens der Korrekturprogramme aus. Analytische Formen wurden ebenfalls nicht berücksichtigt, da sich ihre Form aus den Partizipien und den Hilfsverben ergibt. Dennoch wurden infinite Formen im Test berücksichtigt. Die Partizipien des Präsens und des Perfekt sowie der Infinitiv Präsens sind mit aufgenommen worden. Das Partizip Präsens wird bei Eisenberg als Adjektiv betrachtet und somit nicht als zum verbalen Paradigma zugehörig gesehen (Eisenberg [1]: 192). Da es jedoch regelmäßig aus dem Verbstamm bildbar ist, wurde es in diesen Test aufgenommen.

Präsens				Präteritum			
		Sg	Pl			Sg	Pl
1.	leg	e	en	1.	legt	e	en
2.		st	t	2.		est	et
3.		t	en	3.		e	en

Tabelle 4.1: Flexion schwacher Verben am Beispiel legen

Im Präsens treten regelmäßig Synkretismen in der 3. Person Singular und der 2. Person Plural sowie in der 1. und 3. Person Plural auf. Im Präteritum fallen in den

beiden Numeri jeweils die 1. und die 3. Person zusammen. Insgesamt bestehen damit für die Ausschnitte des Paradigmas jeweils vier Formen, die zusammen mit den beiden Partizipien für die getesteten Verben im Anhang ab Seite 114 zusammen mit den Ergebnissen des Tests aufgelistet wurden.

Das Ergebnis ist eindeutig: Word beherrscht die Formengenerierung ausgezeichnet. Die Endung des Korrekturvorschlags stimmt in jedem Fall mit derjenigen der falsch geschriebenen Verbform überein. Selbst die Schwa-Epenthese in der zweiten Person Singular Präsens Aktiv (*segele, segle; rudere, rudre*) wird korrekt abgebildet.

DITECT hat größere Schwierigkeiten auf diesem Gebiet, die schon allein dadurch bedingt sind, dass die Korrektur nicht auf der Grundlage einfacher Fehler erfolgt. So wurde bei den schwach flektierenden Verben nur etwa knapp die Hälfte aller fehlerhaften Wortformen mit der korrekten Endung zurückgegeben. Darunter jeweils der Infinitiv und einige Partizipien. Des Weiteren finden sich vermehrt die Verbformen mit dem höchsten Markierungsgrad (*legst, segelst, segeltest*).

Bei den stark flektierenden Verbformen fand sich nur etwa ein Fünftel mit den richtigen Flexionsendungen. Hier traten keine Infinitive auf, dafür wieder Partizipien und eher markierte Formen. Der Grund dafür ist offensichtlich. Bei den Verbformen mit dem höchsten Markierungsgrad ist die Verwechslung mit anderen Formen weit weniger wahrscheinlich. Allerdings ist grundsätzlich eine Unterscheidung der Formen in jedem Falle möglich. Die Annahme, DITECT weise in der Abbildung von Flexionsformen Schwächen auf, hat sich voll bestätigt.

4.1.2 Substantive

Flexionsparadigmen deutscher Substantive enthalten weit weniger Formen als diejenigen der Verben, allerdings gibt es mehr Flexionsklassen, von denen zwischen vier und über 20 angesetzt werden (Eisenberg [2]: 135). Wir halten uns an die Darstellung bei Eisenberg, die aus dem Duden von 1973 übernommen wurde (Eisenberg

[1]: 152). In der Regel enthält ein Paradigma wie bei der starken Flexion der Maskulina und Neutra nicht mehr als fünf Formen, was die Formengenerierung weitaus einfacher macht.

Allerdings gibt es auch hier Unregelmäßigkeiten, die berücksichtigt werden müssen. Beispielsweise wird der Plural von *Angehöriger* nicht durch weitere Verkettung gebildet, sondern durch Tilgung des Konsonanten *r*. In diesem Fall kann das Problem übergangen werden, indem *Angehöriger* als Bestandteil des Paradigmas von *angehörig* betrachtet wird. So kann mit dem Bestand an substantivischen Flexionssuffixen (*e, (e)r, (e)s, (e)n, (e)rn*) jede Flexionsform dieses Lexems gebildet werden. Das Problem der nichtkonkatenativen Morphologie ist damit aber noch nicht vollständig gelöst, da Fremdlexeme wie *Thesaurus, Medium, Paradigma* etc. unregelmäßige Pluralformen bilden, bei denen sich die gesamte Endsilbe ändert. Solche Formen können nicht mehr automatisch generiert werden, sondern müssen ähnlich wie die Formen der starken Verbflexion im Einzelnen gespeichert werden.

Maskulina und Neutra, stark

		Sg	Pl
Nom	Berg	-	e
Gen		es	e
Dat		(e)	en
Akk		-	e

Neutra, stark

		Sg	Pl	
Nom	Kind	-		er
Gen		es	er	
Dat		(e)	ern	
Akk		-	er	

Tabelle 4.2: Flexion starker Maskulina und Neutra (vgl. Eisenberg [1]: 152)

Die Auswahl der Testbegriffe erfolgte nach dem gleichen Schema wie im ersten Teil des Tests. Zu jedem Flexionstyp wurde mindestens ein Vertreter ausgewählt, dessen Flexionsformen, Synkretismen ausgenommen, mit einem einfachen Fehler versehen, dem Test unterzogen wurden, um die Abbildung der Flexionsendung zu untersuchen. Auf die s-Flexion, die sich gegenwärtig im Deutschen schnell ausbreitet, wurde verzichtet, da davon ausgegangen werden kann, dass das Anhängen von *s* beherrscht wird, wenn die übrigen Paradigmen dargestellt werden können. Das

komplexeste Paradigma der Substantivflexion bildet die starke Flexion der Maskulina und Neutra, die jeweils einen Pluralmarker und das Genitiv-s im Singular sowie eine Markierung für den Dativ Plural enthalten (siehe Tabelle 4.2).

Maskulina, schwach

		Sg	Pl
Nom	Buchstabe	-	e
Gen		es	e
Dat		(e)	en
Akk		-	e

Neutra, gemischt

		Sg	Pl
Nom	Ende	-	er
Gen		es	er
Dat		(e)	ern
Akk		-	er

Tabelle 4.3: Flexion schwacher Maskulina und gemischter Neutra

Daneben wurde die schwache Flexion der Maskulina angesetzt, die als Markierung nur den Schwa-Auslaut auf *(e)n* kennt, und die gemischte Flexion der Neutra, die einen zusätzlichen Genitiv-Marker enthält (siehe Tabelle 4.3).

Feminina mit n-Plural

		Sg	Pl	
Nom	Entnahme		-	n
Gen		-	n	
Dat		-	n	
Akk		-	n	

Feminina mit e-Plural

		Sg		Pl
Nom	Kunst	-	Künst	e
Gen		es		e
Dat		(e)		en
Akk		-		e

Tabelle 4.4: Flexionstypen der Feminina

In den ausgewählten Beispielen war die Zuordnung teilweise schwierig, da das Lexem *Urheber*[WP] keinen Pluralmarker enthält oder das Paradigma von *Staatsangehöriger*[WP] den Plural unter Tilgung des auslautenden *r* bildet. Daher wird für die Substantive als unterscheidendes Kriterium für die Zuordnung der Genitivmarker angesehen. Für die ausgewählten Neutra gilt Ähnliches. Teilweise trat ein Nullplural auf wie im Paradigma von *Verfahren*[WP] oder ein überhaupt nicht berücksichtigter Plural wie derjenige im unregelmäßigen *Medium*[WP]. Die Zuordnung zu starker oder gemischter Flexion wurde dabei nach dem Kriterium entschieden, ob für den Plural ein Dativmarker vorhanden war oder nicht.

Die Kategorisierung der Feminina war schließlich am eindeutigsten. Hier muss entschieden werden, ob der Plural auf *(e)n* gebildet wird und keine Alternation des Stammes erfolgt oder ob der Plural auf *e* auslautet und eine Umlautung erfolgt. In diesem Falle existiert eine Dativmarkierung im Plural (siehe Tabelle 4.4).

Insgesamt wurden die Beispiele danach kategorisiert, ob die Formen lediglich mit Hilfe von konkatenativer Morphologie erfasst werden können oder ob andere Formmerkmale auftraten wie Alternation der auslautenden Silbe oder Umlautung des Stammvokals, wie sie im zweiten Flexionstyp von Tabelle 4.2 obligatorisch ist, wenn der Stammvokal umgelautet werden kann (*Haus, Buch*). Eine komplette Auflistung der im Test verwendeten Begriffe mit Angaben, ob die Flexionsendung von den Programmen abgebildet wurde, findet sich im Anhang C ab Seite 116.

Das Ergebnis fällt ähnlich aus wie im ersten Testteil. Word war in der Lage, jede Flexionsendung korrekt auf die fehlerhafte Zeichenfolge abzubilden. Bereits in den vorhergegangenen Tests hatte sich gezeigt, dass die Rechtschreibkorrektur von Word mit dem Konzept arbeitet, nicht mehr als notwendig zu berichtigen, da die Wahrscheinlichkeit, dass nur ein einfacher Fehler auftritt, sehr hoch ist. Diese Annahme hat sich ein weiteres Mal bestätigt. Zusätzlich scheinen alle nötigen Flexionsformen generierbar oder im Lexikon gespeichert zu sein, da kein einziger Ausfall auftrat. Um dies verifizieren zu können, wäre allerdings eine Datenmenge nötig, die den Rahmen dieser Untersuchung sprengen würde.

Die Rechtschreibkorrektur von DITECT hat wiederum Schwierigkeiten, die richtige Hypothese zu generieren. Nur etwa die Hälfte der Fehler konnten berichtigt werden. Dabei trat kein erkennbares Muster auf. Im Beispiel *Drittland* wurde jede Flexionsform inklusive der umgelauteten korrekt abgebildet, während dies im Beispiel *Vertrag* nicht möglich war. Ähnlich verhielt es sich in den Beispielen *Kunst*, deren Korrektur fehlerfrei war, und *Übereinkunft*, dessen Formen nicht korrigiert wurden. Die Schwierigkeiten hingen also nicht primär an der Beschaffenheit der

Flexionsform. Vielmehr traten häufig Korrekturvorschläge auf, die in der typografischen Form weit von der intendierten Wortform abwichen. So bestätigt sich auch hier die Annahme, dass DITECT zur Flexionsformenabbildung nicht in der Lage ist. Die einzelnen Flexionsformen weichen voneinander meistens nur in einem und höchstens in drei Editierungsschritten ab. Vielfach ist die Fehlerkorrektur von DITECT zu ungenau, um solche Feinarbeit zu leisten.

4.2 Komposition

Die Ziel dieses Tests war, nachzuweisen, dass DITECT bei der Korrektur von Komposita mit Bindestrich den Fehler macht, nur die Zweitkonstituente zu berücksichtigen. Zu diesem Zweck wurden dem System einige Zeichenketten eingegeben, die der Struktur nach einem substantivischen Kompositum ähneln. Dabei wurden in einigen Fällen bewusst Fehler eingebaut. In anderen Fällen wurden einzelne frei erfundene Konstituenten erdacht, wie sie als Bestandteile der deutschen Sprache vorkommen könnten (*Droms, Schmopf*).

Die Annahme, DITECT würde nur die Zweitkonstituente berichtigen, schien sich zunächst zu bestätigen. In den Beispielen in Tabelle 4.5 finden sich Fehler in der Zweitkonstituente der Komposita. Die Korrekturvorschläge bei DITECT beziehen sich nur auf die Zweitkonstituente, während Word Vorschläge macht, die auf die gesamte Wortform zielen.

Testterm	DITECT	Word
Schmauch-Berzen	Betern	Schmauch-erzen
Zeisig-Drone	Drogen	Zeisig-Krone
Kurden-Staart	Staat	Kurden-Start
Tisch-Fuußball	Fußball	Tisch-Fußball

Tabelle 4.5: Korrekturvarianten bei Bindestrichkomposita – Fehler in Konst. 2

In Tabelle 4.6 befinden sich Fehler in beiden Konstituenten. Zunächst wurden zwei Reihen gebildet, in denen jeweils die Erstkonstituente verändert wurde (*Sturden* → *Strurden* → *Strqurden*). Diese Fantasiebegriffe wurden als Erstkonstituenten von Bindestrichkomposita verwendet, in einem

Fall mit einem Fantasiewort (*Kaat*), im anderen Fall mit einer realen Konstituente (*Staat*). Wie sich in der ersten Zeile zeigt, hat DITECT für die Wortform ebenso einen Korrekturvorschlag anzubieten wie Word. Wird diese Zeichenkette jedoch als Erstkonstituente eines Bindestrichkompositums verwendet, in dem beide Konstituenten Fehler aufweisen (*Sturden-Kaat*), bezieht sich der Korrekturvorschlag nur auf die Zweitkonstituente. Befindet sich kein Fehler in der Zweitkonstituente, wird kein Fehler gefunden (*Sturden-Staat*), obwohl die Erstkonstituente für sich genommen korrigiert werden müsste.

Wird die Erstkonstituente nun so verändert, dass sie als unakzeptabel gelten muss, bleibt der Korrekturvorschlag für die Zweitkontituente aus, wie in den Fällen *Strqurden-Kaat* und *Struhrden-Staat*. Das Verhalten ist dabei bei den beiden Zweikonstituenten *Staat* bzw. *Kaat* analog.

Interessant hieran ist, dass die Änderung in der Vorgehensweise erfolgt, wenn die Konstituente nicht mehr wohlgeformt ist. Es ist vorstellbar, dass das Programm mit einem Inventar an Vollsilben oder alternativ auch Initial- und Terminalsilben ausgestattet ist und ein Fehler angenommen wird, wenn die Silbenränder in der untersuchten Wortform nicht in diesem Inventar vorhanden sind. Dies wäre bei dem Silbenrand *Strqu-* bzw der Silbe *Strqur* der Fall. Abgesehen vom genauen Vorgehen des Programms, das an der Oberfläche schwer einzuschätzen ist, ist die

Testterm	DITECT	Word
Sturden	strudeln	Stunden
Sturden-Kaat	Kata	Sturdehn-Kat
Strurden-Kaat	Kata	(Fehler gefunden, keine Vorschläge)
Strqurden-Kaat	(Fehler gefunden, keine Vorschläge)	(Fehler gefunden, keine Vorschläge)
Sturden-Staat	(kein Fehler)	Sturen-Staat
Strurden-Staat	(kein Fehler)	(Fehler gefunden, keine Vorschläge)
Struhrden-Staat	(kein Fehler)	(Fehler gefunden, keine Vorschläge)

Tabelle 4.6: Korrekturvarianten bei Bindestrichkomposita – Fehler in beiden Konst.

Erstkonstituente in diesen Fällen offensichtlich maßgeblich für die Fehlerdetektion. Die Strategie, die ganze Zeichenfolge in der Fehlerdetektion zu berücksichtigen, allerdings nur Vorschläge für die Zweitkonstituente zurückzugeben, ist aus Nutzerperspektive allerdings schwer nachvollziehbar.

In einer zweiten Reihe ist das Verhalten an einem anderen Beispiel erneut überprüft worden (Tabelle 4.7). Der Ergebnis in der Fehlerdetektion ist ähnlich. Für die Erstkonstituente wird kein Fehler angenommen, solange die Lautkombinationen wohlgeformten Einheiten der deutschen Sprache entsprechen. Erst bei einem Verstoß gegen diese Regeln wird bei DITECT ein Fehler angenommen, wie in der Silbe *Drkoehms.* Eine Ursache für die Markierung als Fehler könnte darin liegen, dass hier im Onset der Silbe zwischen Sonorant und Vokal ein Plosiv eingefügt wurde. Verschlusslaute nach Frikativen treten im Deutschen aber nur in den Kombinationen [ʃt] und [ʃp] auf (eine Ausnahme ist die Lautkombination [ʃk] bei Integrationen von Fremdwörtern und Fremdnamen wie in *Schkopau, Schkeuditz*) (Maas, 1999: 186). Damit wäre das Wohlgeformtheitskriterium nicht erfüllt. Nach dem allgemeinen Silbenbaugesetz werden fünf Hauptklassen der Sonoritätshierarchie unterschieden. Demnach können die Laute in der Silbe in folgender Reihenfolge realisiert sein (Eisenberg[1]: 106):

Obstruent, stl. Obstruent, sth. Nasal Liquid Vokal
Liquid Nasal Obstr, sth. Obstr, stl.

Ein Verschlusslaut ist also nach einem Liquid generell ausgeschlossen. Ob ein derartiger Silbenbau der Rechtschreibkorrektur bei DITECT zugrunde liegt, ist schwer zu sagen. Nach den bisherigen Ergebnissen scheint ein solches Vorgehen eher unwahrscheinlich, eine andere Erklärung lässt sich aber für das Verhalten des Programms nicht finden.

In Tabelle 4.7 zeigt sich eine Ungereimtheit im Verhalten DITECTs, die während des Tests auftrat. Nicht in jedem Fall bezieht sich der Korrekturvorschlag bei DITECT lediglich auf die Zweitkonstituente.

Testterm	DITECT	Word
Droms-Tasche	(kein Fehler)	Drops-Tasche
Drohms-Tasche	(kein Fehler)	Drohäs-Tasche
Droehms-Tasche	(kein Fehler)	(Fehler gefunden, keine Vorschläge)
Drkoehms-Tasche	Drechselmaschine	(Fehler gefunden, keine Vorschläge)

Tabelle 4.7: Korrekturvarianten bei Bindestrichkomposita

Im Unterschied zu den vorigen Beispielen wurde von DITECT in diesem Fall ein Korrekturvorschlag für die gesamte Zeichenkette gemacht. Das Verhalten des Programms wird dadurch undurchsichtig und folgt keinem einheitlichen Konzept. Die Fehlerkorrektur von Word hingegen bietet bei zunehmender Abweichung von möglichen Wortformen keine Korrekturvorschläge mehr an.

4.3 Getrennt- und Zusammenschreibung

4.3.1 Vorbemerkungen

In diesem Abschnitt werden Probleme der Getrennt- und Zusammenschreibung nach der neuen Rechtschreibung besprochen. Als korrekte Schreibvariante gelten die Vorschläge der Zwischenstaatlichen Kommission für deutsche Rechtschreibung von 1996, die das amtliche Regelwerk bilden (vgl. Heller, 1996).

Der Testdatenbestand ist eng an das amtliche Regelwerk angelehnt und in drei Hauptgruppen untergliedert. Die Einteilung erfolgt anhand der Wortart der Zweitkonstituente. Es handelt sich dabei um Verbindungen mit Verben, Verbindungen mit Partizipien und Verbindungen mit Adjektiven. Diese drei Hauptgruppen sind weiter untergliedert nach der Beschaffenheit der Erstkonstituente: Es treten in Ver-

bindungen mit Verben als Zweitglied beispielsweise Verben, Partizipien, Adjektive, Adverbien, Substantive und Partikeln als Erstglied auf. Weitere Verbindungen vor allem mit Partizipien werden im Anschluss daran aufgeführt. Eine komplette Übersicht der Testdaten mit Kommentaren zu den einzelnen Kategorien findet sich in Anhang E.1. Im zweiten Teil von Anhang E wurde aus den Bestandteilen aller Kategorien ein kleines Korpus an Beispielsätzen erstellt. So kann geprüft werden, ob der unmittelbare Kontext entscheidend ist für eine wirkungsvolle Korrektur oder nicht.

In den meisten Fällen ist die Entscheidung jedoch eindeutig, da die Regelung, ob getrennt oder zusammengeschrieben wird unabhängig von direkter oder übertragener Bedeutung ist. Bei Verbindungen aus Adjektiv und Verb oder Partizip wird beispielsweise getrennt geschrieben, wenn das Adjektiv steiger- oder erweiterbar ist (Bsp.: *dicht behaart* wegen *dichter behaart; freihalten* wegen **freier halten*). Die Schreibung ist unabhängig davon, ob die Bedeutung direkt (*einen Stab freihalten*) oder übertragen ist (*Einfahrt freihalten*; aber: *eine Rede frei halten*).

Dennoch ist für die elektronische Korrektur der Kontext entscheidend. Bei einer falschen Getrenntschreibung von *freihalten* muss der Kontext von *frei* analysiert werden. Tritt darauf folgend *halten* auf, handelt es sich um einen Fehler. Prinzipiell muss also jedes Adjektiv in der Kurzform (und ebenso jedes Verb im Infinitiv, jedes Partizip, Adverb und eine Reihe von Partikeln) auf ein eventuell folgendes Verb überprüft werden. Das kann zu Performance-Problemen führen, weshalb Design-Entscheidungen denkbar sind, dieses Problem in der Korrektur auszuklammern.

4.3.2 Testergebnis

Beide Programme sind in der Lage, einen Teil der fehlerhaften Schreibungen zu korrigieren. Die Leistung der Programme ist erstaunlich ähnlich. In der tabellari-

schen Fassung konnte von 161 Beispielen DITECT 23 Fälle erkennen und Word 24, was jeweils etwa 15 Prozent entspricht.

Die Verteilung der Fälle offenbart die Schwächen bei der Korrektur dieser Phänomene. Ein Drittel der Beispielfehler resultiert aus falscher Zusammenschreibung, zwei Drittel sind falsch getrennt geschrieben. Von den falschen Zusammenschreibungen haben beide Programme etwa 35 bis 40 Prozent richtig erkannt. Von den über 100 falschen Getrenntschreibungen erkennt DITECT keine, Word nur eine (**ab holen*). Einige der Fälle beruhen auf der fehlenden Lexikalisierung eines Bestandteils und treten deshalb gesondert auf (**Elfenbein farben*, **Nitrofen belastet*). Streng genommen müssen diese aber zu den nicht erkannten Getrenntschreibungen gezählt werden.

	Gesamt	DITECT	Word
Falsche Zusammenschreibung	54	20	19
Falsche Getrenntschreibung	107	0	1
Wort nicht lexikalisiert		3	4

Tabelle 4.8: Fehlerkorrektur – Getrennt- und Zusammenschreibung

Falsche Zusammenschreibung zu erkennen ist viel einfacher als die Erkennung von falscher Getrenntschreibung. Die Wortgrenze muss nicht überschritten werden. Die Korrektur falscher Zusammenschreibung ist also kontextunabhängig. Im anderen Fall müssen die unmittelbar angrenzenden Wortformen betrachtet werden, was offenbar bei beiden Programmen nicht geschieht.

Aus Sicht des Programmierers kann die Zusammenschreibung also als der markierte Fall angesehen werden. Die relativ geringe Erkennungsrate bei den Zusammenschreibungen erklärt sich aus der hohen Produktivität der Komposition im Deutschen. Wie sich in Kapitel 3 zeigte, haben beide Programme Probleme bei der Erkennung einiger Kompositagruppen. In jenem Bereich ist also ein hoher Grad an Kombinatorik gefragt. Im vorliegenden Fall hingegen muss stark differenziert werden, was zusammengeschrieben werden darf und was nicht. Die Vermutung liegt nahe, dass die Programme ihre Leistung hinsichtlich Getrennt- und Zusam-

menschreibung auf einzelne Bereiche beschränken, die nicht mit den produktiveren Kompositionsmustern kollidieren (z. B. rein substantivische Komposita).

Beide Programme erkennen falsche Zusammenschreibung bei Verbindungen mit Partizipien, Adjektiven oder Adverbien im Erstglied und Verben im Zweitglied. Die Leistungen von Word sind dabei im Bereich der Adjektive und Partizipien besser, während DITECT eine höhere Erkennungsrate im Bereich Adverbien zeigt. Reine Verbverbindungen werden von DITECT nicht berichtigt. Word zeigt hier bessere Leistungen. Bei Verbindungen mit Partizipien im Zweitglied und Adjektiven oder Partizipien im Erstglied zeigt DITECT die besseren Ergebnisse, während Word hier kaum ein Beispiel berichtigt. In allen genannten Kategorien ist das Bild lückenhaft. Eine grundlegende Bemühung, diese Problembereiche abzudecken, ist jedoch erkennbar.

Anders sieht es bei Verbindungen mit Partikeln im Erstglied aus. Hier ist die Erkennungsrate gleich null. Ein Ansatz zur Korrektur ist offenbar in beiden Programmen nicht enthalten, was an dem Umstand liegt, dass Verbindungen mit den genannten Partikeln (Anhang E. 1.9) konsequent zusammengeschrieben werden. Eine Erkennung falsch getrennt geschriebener Verbindungen wäre aufgrund der Häufigkeit der Partikeln zu aufwendig.

Falsch zusammengeschriebene Verbindungen aus Substantiven und Verben, Adjektiven oder Partizipien werden ebenfalls von beiden Programmen nicht erkannt. Von Einzelfällen abgesehen, in denen Bestandteile nicht als freie Morpheme auftreten (*Elfenbein farben*) oder in denen flektierte Formen auftreten, die nicht als Bestandteil eines Kompositums zugelassen sind (*gewinnebringende*) ist die Erkennungsrate gleich null. Das lässt darauf schließen, dass für diesen Bereich bewusst keine Problemlösungsstrategie implementiert wurde. Die Gründe dafür sind zum einen die schon angesprochene mögliche Kollision mit der Erkennung von Kompo-

sita. Der andere Grund liegt in der starken Kontextabhängigkeit der Getrennt- und Zusammenschreibung dieser Verbindungen. Beispiel:

Die Verarbeitung erfolgt computergesteuert.
Der Prozess ist vom Computer gesteuert.

Wie schon hinlänglich gezeigt wurde, beschränkt sich die Rechtschreibprüfung jedoch bei beiden Programmen auf Einzelwortkorrektur. Der zweite Teil des Tests, in dem die gleichen Beispiele in syntaktischen Kontext eingebettet sind, untermauert den Befund: Es wurden in den Beispielsätzen nur Fehler gefunden, die auch in der tabellarischen Form erkannt wurden.

4.4 Groß- und Kleinschreibung

4.4.1 Vorbemerkungen

Der Problembereich Groß- und Kleinschreibung ist recht vielfältig. Die Großschreibung von Substantiven und substantivischen Wendungen lässt sich über das Lexikon regeln (*Cordon bleu, Conditio sine qua non, Alma Mater, Ultima Ratio*) (IDS, Regeln und Wörterverzeichnis: 54, § 55). Dieser Bereich beruht weitgehend auf Einzelregelungen und wird im Test nicht weiter betrachtet. Die Ergebnisse würden in der Substanz mit denen aus Kapitel 3 korrespondieren. Des weiteren wird Großschreibung, die auf bestimmten Textsorten und -elementen beruht, nicht behandelt, wie die Großschreibung von Überschriften und Werktiteln, Gesetzesbezeichnungen, Anschriften (IDS, Regeln und Wörterverzeichnis: 52, § 53) oder Konventionen wie die Großschreibung am Zeilenanfang im Vers.

Interessanter ist die Großschreibung von Substantivierungen. Hier wird konsequenter großgeschrieben, als dies früher der Fall war. Lt. dem amtlichen Regelwerk erkennt man Substantivierungen unter anderem an einem vorausgehenden Artikel oder unbestimmten Zahlwort, an einem vorangestellten adjektivischen Attribut oder an ihrer Funktion als kasusbestimmtes Satzglied (IDS, Regeln und Wörterverzeichnis: 58, § 57). Die Großschreibung einer Substantivierung ist also ausschließlich kontextbestimmt. Ohne einen Parser, der bestimmt, ob sich beispielsweise ein Adjektiv auf ein Substantiv in einer Nominalgruppe bezieht oder ob es selbst ein Satzglied darstellt, ist die Korrektur einer kleingeschriebenen Substantivierung nicht zu leisten.

Die Großschreibung hingegen kann generell als der markierte Fall angesehen werden. Tritt ein großgeschriebenes Adjektiv auf, kann ein Parser aktiviert werden, der den Satz prüft. Beispiel: *Die Grünen Abweichler entschieden sich anders.* Hier befindet

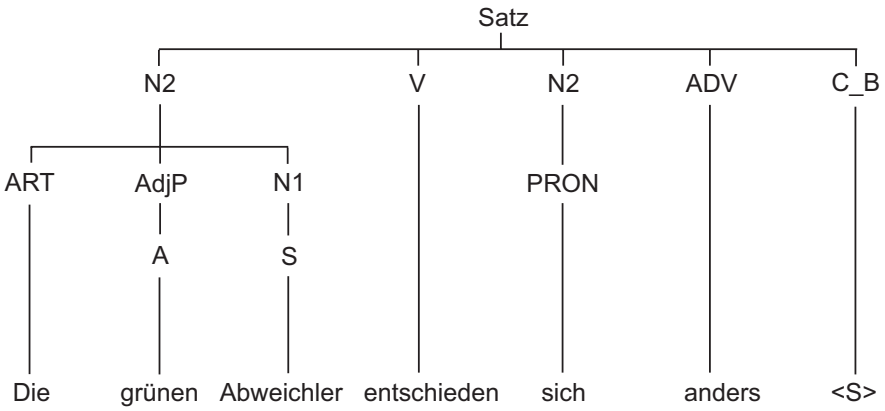

Tabelle 4.9: Analyse eines Satzes zur Groß- und Kleinschreibung des UIS-Parsers

sich ein Fehler in der ersten Nominalgruppe. Das adjektivische Attribut bezieht sich auf das Kernsubstantiv der Nominalgruppe und müsste demzufolge kleingeschrieben werden. In einem solchen Fall kann zunächst geprüft werden, ob ein Artikel, ein unbestimmtes Zahlwort etc. vorausgeht. Dies ist der Fall. Anschließend muss mithilfe des Parsers geprüft werden, ob dieses Adjektiv die Funktion eines Satzglie-

des übernimmt (siehe Tabelle 4.9). Dies ist nicht der Fall. Der verwendete Parser[7] erkennt das großgeschriebene Adjektiv als Teil der Nominalgruppe, die in Subjektrelation zum Verb steht. Das Adjektiv kongruiert mit dem Substantiv hinsichtlich Kasus und Numerus und wird hinsichtlich des Genus vom Substantiv regiert.

Die korrekte Ausgabe erfolgt jedoch nur, wenn unmarkierte fehlerfreie Sätze eingegeben werden. Eine Modifikation zur Korrektur von Fehlern in der Groß- und Kleinschreibung ist sicher recht einfach. Die Analyse von Sätzen, die mehrere Konstituentensätze oder diskontinuierliche Konstituenten enthalten, ist jedoch derzeit kaum vorstellbar.

Der Bereich der Substantivierungen umfasst den größten Teil des Tests. Genauer untersucht werden Substantivierungen von Verben und Adjektiven. Substantivierungen im Verbbereich beschränken sich auf die Infinitive. Diese werden in Anhang F in Abschnitt 1.1 betrachtet. Die folgenden Unterabschnitte behandeln andere grammatische Formen, in denen der Infinitiv oft fälschlich großgeschrieben wird: den Infinitiv mit *zu* (Anh. F, Abs. 1.2) und den Infinitiv in Verbindung mit Hilfsverben, Modalverben oder den Verben *gehen* und *sehen* (Abs. 1.3 und 1.4). Substantivierte Adjektive werden in Abschnitt 2 behandelt. Die großzuschreibenden Substantivierungen finden sich in Abschnitt 2.1. Die Abschnitte 2.2 und 2.3 behandeln Adjektive, die fehlerhaft großgeschrieben wurden. Hier gibt es viele Fehlerquellen, wie z. B. die falsche Annahme eines substantivierten Adjektivs bei Koordinationsreduktion innerhalb einer diskontinuierlichen Nominalgruppe (*Das neue Auto ist nicht viel besser als das alte.*). Andere Fehlerquellen sind die Großschreibung von Superlativen und Nationen- oder geografischen Bezeichnungen innerhalb von Nominalgruppen. Auch Sprach-, Farb- und Zahladjektive werden häufig großgeschrieben, Numeralia allerdings nicht. Dieses Problem wird in Abschnitt 2.4 behandelt. Abschließend werden feste adjektivische Verbindungen geprüft, die nach der Neuregelung großgeschrieben werden.

Weitere Testfälle beziehen sich auf die Großschreibung nach einem Doppelpunkt (Abschnitt 3). Hier wird nur großgeschrieben, wenn ein vollständiger Satz folgt. Nach Ellipsen oder Nominalgruppen ohne Prädikat wird kleingeschrieben. Schließlich wird der Umgang mit Akronymen (Abschnitt 4) und festen Wendungen aus Substantiv und Adjektiv betrachtet (Abschnitt 5). Letztere müssen von Fall zu Fall betrachtet werden, da eine formale Regel nicht existiert. Die Regelungen sind häufig semantisch motiviert, wie die Großschreibung von botanischen und zoologischen Bezeichnungen (*Roter Milan*) oder historischen Ereignissen (*Dreißigjähriger Krieg*, *Zweiter Weltkrieg*).

4.4.2 Testergebnis

Wie im korpusbasierten Testteil deutlich wurde, berücksichtigt Word bei der Korrektur, ob das als fehlerhaft erkannte Lexem groß- oder kleingeschrieben ist. Bei großgeschriebenen Adjektiven oder Verben, die einen Fehler enthalten, ist der Korrekturvorschlag ebenfalls großgeschrieben. Die Tatsache, dass die Groß- und Kleinschreibung bei der Korrektur erhalten bleibt, offenbart, dass dem Problem eine gewisse Bedeutung beigemessen wird. Das lässt darauf schließen, dass das Phänomen auch in der Fehlerdetektion Beachtung findet. Das Testergebnis widerlegt diese Vermutung.

	Anzahl der Beispiele	DITECT	Word
Substantivierungen – Verben	19	1	0
Substantivierungen – Adjektive	30	2	0
Schreibung nach Doppelpunkt	4	0	0
Akronyme	3	0	0
Feste Wendungen aus Adj. und Sbst.	32	0	0
Summe	88	3	0

Tabelle 4.10: Korrektur aus dem Bereich Groß- und Kleinschreibung

Bei 88 Beispielen aus den verschiedenen Kategorien entdeckte DITECT eine falsche Kleinschreibung und in zwei Fällen falsche Großschreibung. Damit ist zwar

erwiesen, dass DITECT dem Problem der Groß- und Kleinschreibung grundsätzlich Rechnung trägt. Allerdings beschränkt sich die Detektion auf einige wenige Fälle, die im Lexikon gespeichert sind. Da nicht einmal fünf Prozent erreicht werden, kann von einzelnen Glückstreffern ausgegangen werden. Eine systematische Behandlung des Phänomens findet offensichtlich nicht statt.

Word fand keinen einzigen der Fehler, dabei finden sich unter den Testdaten viele prominente Beispiele, die in der Diskussion um die Rechtschreibreform immer wieder zitiert wurden (*Im Großen und Ganzen, gleich und gleich, schwarz auf weiß*). Anhand des Ergebnisses drängt sich der Verdacht auf, dass das Problem bewusst nicht behandelt wird, um falsche Ratschläge zu vermeiden. Die starke Kontextabhängigkeit des Phänomens Groß- und Kleinschreibung lässt sich nicht mit der von den Programmen durchgeführten Einzelwortkorrektur vereinbaren.

5 FAZIT

Vor nur zwanzig Jahren besaß nur eine kleine Gruppe von Interessierten einen eigenen PC. Das Internet war dem Militär und einem spezialisierten Insiderkreis vorbehalten. Seitdem hat sich der Prozess der Globalisierung vollzogen, die Gesellschaft ist vom Industrie- ins Informationszeitalter katapultiert worden. Die Rechenleistung der Heimrechner verdoppelt sich dabei fast im Jahrestakt, und der Trend ist weiterhin ungebrochen. Durch die Vielzahl der immer neuen Möglichkeiten von Homebanking über digitale Fotografie bis hin zur globalen Positionsbestimmung auf wenige Meter wird eine Welt grenzenloser Möglichkeiten des technologischen Fortschritts suggeriert.

Auch die elektronische Rechtschreibkorrektur hat dabei große Fortschritte zu verzeichnen, die sich nicht nur in der Erweiterung des Lexikons erschöpfen. Von einem zuverlässigen System, das eine automatische Korrektur ermöglicht, kann aber keine Rede sein. Zu groß sind die Lücken, die die Programme offen lassen. Die Grenzen der verglichenen Programme liegen dabei erstaunlich eng beieinander, obwohl sie aus vollkommen verschiedenen Kontexten stammen. Diese Grenzen resultieren schließlich auch nicht nur im Unvermögen der Programmierer oder der fehlenden visionären Kraft der Systementwickler. Die unendlichen Ausdrucksmöglichkeiten, die die Sprache bietet, die Bildung neuer Nominationseinheiten oder die fließenden Wortgrenzen, können nur zu einem bestimmten Grad vollständig abgebildet werden. Das System betrachtet aber jede nicht gespeicherte Zeichenkette als Fehler, was zur Folge hat, dass auf einen Fehler etwa zwanzig Begriffe kommen,

die nicht falsch geschrieben sind, aber dennoch von den Programmen beanstandet werden.

Offenbar scheint dieses recht starre Konzept an die Grenzen seiner Leistungsfähigkeit gestoßen zu sein. Ein Paradigmenwechsel ist auf diesem Gebiet nötig, um den gestiegenen Ansprüchen an die Fehlerkorrektur gerecht zu werden. Werkzeuge, die in der Lage sind, auch unbekannte Zeichenketten als orthografisch korrekte Ausdrücke zu erkennen, dürften dabei genügend vorhanden sein.

Das Problem der Fehlerdetektion ist im ersten Teil von Kapitel 3 hinreichend beschrieben worden. Es konnte festgestellt werden, dass etwa zwei Drittel der Fehleinschätzungen auf nicht gespeicherten Namen beruhen. Diese Information ließe sich für eine verbesserte Fehlerdetektion nutzen. Ein denkbares Werkzeug dazu stellt die Minimum-Edit-Distance dar, die von Word nur im Bereich der Fehlerkorrektur eingesetzt wird. Falls eine unbekannte Wortform nicht innerhalb von zwei Editierungsschritten auf eine gespeicherte Wortform abbildbar ist, kann davon ausgegangen werden, dass es sich nicht um einen Tippfehler handelt. In solchen Fällen macht Word beispielsweise keine Korrekturvorschläge mehr. Konsequenterweise müsste dann geprüft werden, ob die Wortform die Kriterien eines Namens besitzt. Der formale Begriff des Proper Name lässt sich dabei am Flexionsverhalten und am Artikelgebrauch festmachen (Eisenberg[2]: 160). Wenn die Wortform z. B. häufiger im Text auftritt, eventuell zusätzlich mit einem besitzanzeigenden *s,* aber ohne weitere Flexionsendungen, kann die Existenz eines typografischen Fehlers bereits ausgeschlossen werden. Selbst in kürzeren Texten des Testkorpus könnte ein solches Verfahren einen Teil der überflüssigen Beanstandungen zuverlässig und ohne komplizierte Operationen entfernen.

Andere Textelemente könnten durch ihre Form von der Fehlerdetektion ausgeschlossen werden, wie z. B. URLs oder E-Mail-Adressen. Diese werden von beiden Programmen beanstandet, obwohl ihre feste Syntax (*www.*) oder das Sonderzei-

chen @ eine zweifelsfreie Identifikation ermöglichen. Word versucht mehr oder weniger erfolgreich die Erkennung von Fremdsprachen. Die Einordnung von Internet-Adressen findet jedoch nicht statt. Ebenso wäre eine Erkennung von Straßen-, Orts- oder anderen Namen denkbar, die unbekannte Bestandteile enthalten, aber anhand prototypischer Bestandteile identifiziert werden könnten (*X-straße, X-hausen, X-bach*, etc).

Ein überhaupt nicht angesprochener Bereich ist die Trainierung des Lexikons. Häufig benutzte Begriffe können dem Lexikon hinzugefügt werden. Die Flexionsformen werden jedoch weder automatisch ergänzt, noch wird darauf hingewiesen, dass dies nicht geschieht. Der Nutzer findet es vielmehr selbst heraus, wenn das nächste Mal eine flektierte Form eines solchen Lexems auftritt. Diese kann dann nämlich nicht vom Korrekturprogramm erkannt werden. Dem Problem könnte leicht Abhilfe geschaffen werden, vorausgesetzt die Formengenerierung erfolgt mit Finite-State-Transducern. In diesem Falle müssten in einem Dialog einige Beispielparadigmen angegeben werden, die angeben, wie die Flexion des Lexems aussehen könnte. Der Nutzer müsste nur noch die richtige Variante auswählen.

Die Möglichkeiten zur Fehlerkorrektur sind ähnlich eingeschränkt wie diejenigen zur Fehlerdetektion. Wie die Tests gezeigt haben, beschränkt sich die Fehlerkorrektur prinzipiell auf Non-Word-Errors, also Zeichenketten, die keine gültige Wortform darstellen. Wortformen im falschen Kontext können nicht automatisch erkannt werden. Dies schließt große Bereiche mit erheblichem Fehlerpotenzial ein, wie die untersuchten Gebiete Getrennt- und Zusammenschreibung, Groß- und Kleinschreibung, Kongruenz- und Rektionsfehler aller Art, falsche Auseinanderschreibung von Komposita etc. All diese Fehler zu berichtigen, erfordert höhere Anstrengungen und erheblich mehr Rechenleistung als die Vorschläge zur Verbesserung der Fehlerdetektion. Ohne eine Analyse des Kontextes ist die Korrektur nicht möglich. Die Leistungsfähigkeit der Parser ist zudem begrenzt. Komplexe Satz-

strukturen können bisher kaum entschlüsselt werden. Dennoch könnte ein Lexikon, in dem die grammatischen Kategorien zu den einzelnen Wortformen mit Hilfe von Finite-State-Transducern gespeichert sind, kombiniert mit einem Parser, der Aufschluss über die Gliederung der syntaktischen Konstituenten gibt, ein nützliches Werkzeug zur Korrektur grammatischer Probleme darstellen. Die grammatische Korrektur von Windows XP leistet hier bereits einiges, die Praxis zeigt aber, dass hier noch einige Jahre Forschungsarbeit zu leisten sind.

Für die Probleme Groß- und Kleinschreibung und Getrennt- und Zusammenschreibung kann ein Parser nur das Grundwerkzeug darstellen. Zur effektiven Korrektur sind weitere Werkzeuge nötig, die beispielsweise im Falle des ersteren Problems analysieren, ob ein Adjektiv zusammen mit einem Artikel, mit Präposition und Artikel oder anderen eine Substantivierung anzeigenden Wörtern steht (*alles, allerlei, etwas, genug, nichts, viel, wenig*). So ein Programmsegment müsste ausgelöst werden, wenn der Parser eine Nominalgruppe findet, der kein Substantiv zugeordnet werden kann (*alles Übrige* im Gegensatz zu *alle übrigen Dinge*). Solche Methoden sind allerdings extrem fehleranfällig.

Für die Getrennt- und Zusammenschreibung sind ähnliche Segmente denkbar. Diese müssten entsprechende Verbindungen suchen (VB-VB, VB-Part, ADJ-Part, etc.) und die Regeln anwenden. Auch dies ist nicht unproblematisch, da das System schwer entscheiden kann, ob ein Adjektiv in einer Verbindung mit einem Partizip steigerbar ist oder nicht. Selbst für einen Muttersprachler sind solche Entscheidungen manchmal schwer zu treffen. Damit diese Entscheidungen elektronisch getroffen werden können, müssten entsprechende Merkmalspaare für jedes Adjektiv existieren, die solche Informationen angeben (*falsch* = - steigerbar, *dicht* = + steigerbar). Doch auch diese Lösung kann allenfalls näherungsweise funktionieren, da diese Frage nicht lediglich vom beteiligten Adjektiv abhängt.

Als Fazit des Tests muss festgehalten werden, dass automatische Korrektur bisher nicht möglich ist. Die Leistung eines Rechtschreibkorrekturprogramms erschöpft sich vielmehr in der Assistenz beim Auffinden von einfachen Tippfehlern und in der Suche der korrekten Schreibweise. Das Lektorat kann dadurch effizienter gestaltet werden, ersetzbar wird es aber noch lange nicht. In jedem Falle ist der menschliche Faktor bei der Entscheidung über richtige oder falsche Schreibung weiterhin unersetzlich.

ANMERKUNGEN

1 Christoffel Walther: Bericht von vnterscheid der Biblien vnd anderer des Ehrnwirdigen vnd seligen Herrn Doct. Martini Lutheri Bücher / so zu Wittemberg vnd an andern enden gedruckt werden / dem Christlichen leser zu nutz. Wittemberg 1563, zit. nach Klaeren 1997
2 Unternehmensberatung Dieckmann: ↦ http://www.ub-dieck.com
3 W. Wilmanns: Deutsche Grammatik. Gotisch, Alt-, Mittel- und Neuhochdeutsch. Zweite Abteilung: Wortbildung. Straßburg 1896. Zit. Nach Eisenberg [1], S. 221
4 H. J. Heringer: Wortbildung: Sinn aus den Chaos. DS 12, 1984. S. 1-13. Zit. Nach Eisenberg [1], S. 221
5 Hannover-Concerts-Chef ist lt. Word Französisch, Haffa-Aussage wird als Englisch ausgewiesen.
6 Bezieht man Groß- und Kleinschreibung in die Betrachtung ein, ist hintue zwei Schritte von Hintum entfernt.
7 Es handelt sich um den UIS-Parser (Universitäts-Informations-System) der Universität Zürich: Volk, Martin et. al., Rev.: 16.03.2000, ↦ http://www.ifi.unizh.ch/CL/UIS/parser.html

A ÜBERBLICK ÜBER DAS TESTKORPUS

Diese Tabelle gibt einen Überblick über das Testkorpus. Jeder verwendete Artikel mit allen vorhandenen Fehlern und Daten zur Fehlerdetektion ist hier aufgelistet. Die Informationen, die direkt aus der Madsack-Datenbank übernommen wurden, sind Datum und Uhrzeit der letzten Änderung, Arbeitstitel, Ressort und Zeilenanzahl. Für die Analyse sind diese Angaben nur von mittelbarem Interesse. Die Anzahl der Wörter entspricht der Zählung von Microsoft Word. Geringe Differenzen sind hier möglich.

Im Folgenden finden sich zwei Spalten, in denen Angaben zur Fehlerdetektion gemacht werden. Es finden sich hier nur Angaben für die Gesamtzahl der Fehler, da Wiederholungsfehler für das gesamte Ressort betrachtet wurden. Eine ausführliche Betrachtung der Wiederholungsfehler wurde nur im Rahmen der tatsächlich aufgetretenen Fehler durchgeführt.

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
24.11.02 21:21 Ausländer	LOKA	48	236	26		22		0	0
24.11.02 19:44 cart	LOKA	64	345	6		19		0	0
24.11.02 17:55 Gabriel1	LOKA	53	254	2		5		0	0
25.11.02 19:04 Ballons	LOKA	36	154	6		6		0	0
25.11.02 19:22 cart-info	LOKA	45	174	1		7		0	0
25.11.02 19:21 cart1gol	LOKA	70	347	14		15		0	0
25.11.02 19:24 cart2gol	LOKA	70	335	3		13		0	0
25.11.02 22:10 dehoga	LOKA	41	154	9		8		0	0

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
25.11.02 20:48 Diebstahl	LOKA	31	135	1		1		0	0
25.11.02 20:48 EikemeierHaupttext	LOKA	100	415	7		14		0	0
26.11.02 18:57 AufmGerichtzp	LOKA	87	357	1		1		0	0
26.11.02 20:32 Aufm.Mio.-loch	LOKA	71	281	1		1		1	1
26.11.02 19:54 Aufm.Mousse-T	LOKA	90	365	15		35		1	1
26.11.02 19:13 Handy-Aufmacher	LOKA	89	406	4		7		0	0
26.11.02 18:11 Gilde-Chronik	LOKA	78	280	6		8		0	0
01.01.03 17:43 aufm.-PROJEKT	LOKA	76	251	2		0		0	0
01.01.03 19:04 Dosen	LOKA	81	316	6		6		0	0
01.01.03 19:04 Fundbüro	LOKA	88	373	12		7		0	0
02.01.03 20:50 Dosenhaupt	LOKA	102	490	11		9		2	2
02.01.03 19:35 Härke1	LOKA	97	416	9		9		1	1
02.01.03 16:34 Robbie	LOKA	124	475	7		15		1	1
02.01.03 19:41 strom-feature	LOKA	102	398	11		13		2	2
05.01.03 17:51 postagentur	LOKA	119	482	8		7		4	2
05.01.03 17:51 gerichtunterhalt	LOKA	89	358	1		2		1	0
29.12.02 13:13 glosserückerl	LOKA	89	295	3		2		2	2
06.01.03 18:54 Baustraße	LOKA	79	318	7		6		0	0
06.01.03 20:01 Gericht	LOKA	81	345	3		0		0	0
06.01.03 19:58 jederzeithilfsbereit	LOKA	77	323	1		3		0	0
06.01.03 19:58 Wasserrettungen	LOKA	66	247	6		6		0	0
06.01.03 18:16 Streikallgemein	LOKA	66	235	4		4		3	0
06.01.03 18:16 Loccum/dl	LOKA	70	237	7		7		0	0
06.01.03 19:25 Sternsingen-GABRIEL	LOKA	64	229	5		5		0	0
Summe:	**LOKA**	**2443**	**10026**	**205**	**151**	**263**	**178**	**18**	**12**

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
24.11.02 23:21 AufsetzerÖsterreich	EINS	34	190	1		1		0	0
24.11.02 20:23 Brandanschlag	EINS	35	149	4		3		0	0
24.11.02 18:28 KommetarSeite11	EINS	49	216	3		1		0	0
24.11.02 16:49 Mölle	EINS	38	136	2		2		2	2
24.11.02 20:24 TextAufmacher	EINS	54	233	1		6		0	0
24.11.02 20:23 Winterabo	EINS	33	105	8		7		1	1
25.11.02 20:40 Aufm.Text1	EINS	22	107	0		0		0	0
25.11.02 22:13 Aufm.Text2	EINS	22	103	1		0		0	0
25.11.02 22:12 Kommi	EINS	54	246	0		1		2	1
25.11.02 20:44 Vermögensteuer	EINS	37	185	3		4		4	0
25.11.02 19:50 Winterabo	EINS	36	145	6		7		0	0
26.11.02 21:02 Bildung	EINS	34	173	0		5		0	0
26.11.02 20:59 GildeText	EINS	43	238	4		6		1	1
26.11.02 20:42 Kommi	EINS	62	267	3		4		2	0
26.11.02 20:57 Vermögensst	EINS	30	188	0		1		3	0
26.11.02 20:03 Mölli	EINS	15	94	0		0		0	0
26.11.02 20:11 patriot	EINS	15	101	1		3		0	0
09.12.02 21:24 ÄrzteUZ&Text	EINS	60	236	0		0		0	0
09.12.02 19:51 Kommi	EINS	56	222	2		0		0	0
09.12.02 21:24 US-Manöver	EINS	30	142	4		5		0	0
09.12.02 21:26 Awacs	EINS	26	138	2		1		0	0
10.12.02 18:45 Aufmachi	EINS	50	184	4		5		3	3
10.12.02 20:31 kommi	EINS	52	235	2		2		0	0
10.12.02 22:01 Irak	EINS	47	177	5		2		0	0
01.01.03 19:22 Kommi	EINS	51	228	2		7		0	0
01.01.03 19:22 Transrapid	EINS	38	202	3		4		0	0
01.01.03 19:22 RürupUZ&Text	EINS	36	178	5		5		0	0
02.01.03 20:18 Kranke	EINS	39	151	2		2		0	0
03.01.03 20:39 Kommi	EINS	58	258	1		0		0	0
03.01.03 20:39 Hochwasser	EINS	34	169	3		6		0	0
03.01.03 20:43 Konsumflaute	EINS	34	186	0		4		0	0
05.01.03 19:11 kommi	EINS	52	225	1		2		2	2
05.01.03 19:11 Hillu	EINS	49	293	5		6		2	2

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
06.01.03 20:44 Kommi	EINS	47	201	0		2		0	0
06.01.03 20:44 GrünerPunkt	EINS	44	156	2		1		2	2
07.01.03 19:43 Aufm.Text	EINS	49	196	2		1		0	0
07.01.03 18:39 Einzelhandel	EINS	49	183	2		1		0	0
07.01.03 18:39 Kommi	EINS	49	230	3		1		0	0
08.01.03 19:47	EINS	42	236	1		5		0	0
08.01.03 20:46 frosttext	EINS	46	238	4		4		1	1
08.01.03 19:50 Kommi	EINS	44	204	1		0		0	0
08.01.03 19:50 MittelEU	EINS	44	164	2		2		0	0
09.01.03 20:06 Arbeitsmarkt	EINS	42	188	1		0		0	0
09.01.03 23:51 Aufma.text	EINS	35	183	0		0		1	1
09.01.03 23:51 Kommentar	EINS	50	183	3		4		3	3
09.01.03 21:07 Kommi	EINS	35	236	2		2		2	0
10.01.03 20:52 Aufm.Verdi	EINS	36	202	1		2		1	1
10.01.03 20:38 Aufs.ElKaida	EINS	34	176	3		6		0	0
10.01.03 20:38 Kommi	EINS	50	194	1		3		1	0
12.01.03 20:50 Auif.Text	EINS	40	227	5		3		0	0
12.01.03 20:53 Kommi	EINS	57	260	1		1		1	1
12.01.03 21:12 Tarifflucht	EINS	42	169	1		2		1	1
2.01.03 21:12 Irak	EINS	27	136	1		2		0	0
Summe:	**EINS**	**2187**	**10062**	**114**	**90**	**144**	**103**	**35**	**22**

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
24.11.02 19:26 AufmacherParteitag	POLI	150	580	3		4		1	1
24.11.02 19:29 Bayern	POLI	70	214	2		3		0	0
24.11.02 18:19 Haiser	POLI	50	187	1		2		1	1
24.11.02 18:19 Hessen	POLI	70	266	2		1		0	0
24.11.02 21:00 Kompakt	POLI	64	420	5		4		2	1
25.11.02 19:48 Aufm.Schüssel	POLI	118	433	9		9		0	0
25.11.02 19:58 AufmacherAugstein	POLI	79	333	2		4		0	0
25.11.02 21:22 Aufsetzer	POLI	74	312	1		1		0	0
25.11.02 21:56 Irak	POLI	120	456	2		5		0	0
25.11.02 21:56 ThemadesTages	POLI	115	502	4		3		0	0
08.12.02 14:46 UN	POLI	210	748	18		16		10	8
08.12.02 14:46 tanker	POLI	141	613	25		13		4	4
08.12.02 17:06 interharms	POLI	191	830	5		8		10	9
08.12.02 20:34 textAufm	POLI	141	545	8		8		0	0
07.01.03 18:43 SPDAufmacher	POLI	109	499	11		9		1	1
07.01.03 20:46 AufmacherThema	POLI	106	517	2		4		4	2
07.01.03 22:25 Aufsetzer	POLI	104	377	7		4		0	0
08.01.03 21:00 AufmacherThema	POLI	116	517	6		7		1	1
08.01.03 19:49 FDP1	POLI	117	472	2		2		0	0
08.01.03 19:26 Interview	POLI	138	479	3		0		2	2
08.01.03 19:39 Kommi1	POLI	118	464	3		1		3	3
08.01.03 19:26 Bagdadmenschelnd	POLI	68	370	5		2		0	0
Summe:	**POLI**	**2469**	**10134**	**126**	**104**	**110**	**86**	**39**	**32**
24.11.02 20:04 Aufmacher	WIRT	77	286	4		2		1	1
24.11.02 20:10 börse	WIRT	102	351	3		14		4	4
24.11.02 20:20 Kompakt	WIRT	69	479	5		10		0	0
24.11.02 20:02 Mobilcom	WIRT	79	216	4		6		0	0
24.11.02 20:07 M1lyonais	WIRT	20	101	3		3		0	0
24.11.02 20:08 M2	WIRT	20	98	0		0		1	1
25.11.02 20:39	WIRT	53	186	0		4		0	0
25.11.02 21:52 AufsetzerBahn	WIRT	66	237	2		4		2	2
25.11.02 21:05 HAFFA	WIRT	66	241	10		13		2	2
25.11.02 21:58 MittelBosch	WIRT	86	288	5		4		0	0
25.11.02 21:59 NP-Umfrage	WIRT	84	322	1		3		0	0
08.12.02 14:18 grimma	WIRT	117	467	10		10		2	2

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
08.12.02 14:30 Banken	WIRT	109	396	1		4		0	0
08.12.02 13:32 Auto	WIRT	109	411	7		9		4	4
08.12.02 19:04 Reisen	WIRT	80	276	1		3		1	1
08.12.02 19:06 AufsetzerBörse	WIRT	78	283	5		12		0	0
01.01.03 19:52 aufmacher	WIRT	108	381	6		6		0	0
01.01.03 16:10 börse	WIRT	96	375	3		1		1	1
03.01.03 21:53 aufsetzer	WIRT	125	419	21		8		2	1
06.01.03 15:16 aufm.börseKopie	WIRT	90	291	3		11		1	1
06.01.03 19:12 aufs. niedersachsenKopie	WIRT	77	307	2		14		1	1
06.01.03 20:21 niedersachsenkompakt	WIRT	94	441	4		3		1	1
07.01.03 19:48 aufsetzerwirt1	WIRT	45	220	1		0		0	0
07.01.03 19:48 Minolta	WIRT	87	321	7		4		1	1
07.01.03 18:56 Aufmacher	WIRT	57	248	5		9		0	0
07.01.03 18:54 Aufsetzer	WIRT	58	261	5		4		2	1
07.01.03 18:53 TextSF	WIRT	43	73	2		1		0	0
07.01.03 20:50 BöeseAufmacher	WIRT	47	174	3		3		0	0
07.01.03 20:55 Börsegestern	WIRT	46	146	1		9		1	0
07.01.03 20:55 Börsemittel	WIRT	46	136	3		7		0	0
09.01.03 17:38 Gold	WIRT	138	521	6		5		8	8
09.01.03 15:55 Grundig	WIRT	62	220	6		0		0	0
09.01.03 21:51 kompakt	WIRT	54	291	3		5		0	0
08.01.03 19:39 RWE	WIRT	81	292	3		6		2	2
08.01.03 19:39 TUI	WIRT	64	325	3		7		2	2
Summe:	**WIRT**	**2633**	**10080**	**148**	**109**	**204**	**141**	**39**	**36**

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
24.11.02 20:03 lautern	SPOR	86	376	6		13		0	0
24.11.02 18:49 Reaktion	SPOR	53	276	7		10		0	0
24.11.02 19:44 Rugby	SPOR	50	144	2		1		0	0
24.11.02 18:24 stajner	SPOR	92	367	18		18		0	0
24.11.02 19:22 textzuhabensiesgewusst	SPOR	66	230	5		4		0	0
25.11.02 21:58 Dortmund	SPOR	93	321	7		11		0	0
25.11.02 17:52 Biskup/Giesel	SPOR	43	226	9		15		2	2
25.11.02 17:41 Lautern	SPOR	98	395	4		17		0	0
25.11.02 18:57 Sportlerwahl	SPOR	83	297	5		14		0	0
25.11.02 11:39 tt	SPOR	53	213	17		22		0	0
08.12.02 19:56 stef-195657	SPOR	66	277	13		12		0	0
08.12.02 20:57 sportkompakt	SPOR	56	230	9		14		0	0
08.12.02 12:31 spielderwoche	SPOR	79	299	32		38		0	0
08.12.02 19:42 niedersachsenliga	SPOR	58	183	32		37		0	0
08.12.02 19:45 Meldungrechts	SPOR	66	302	14		20		1	1
01.01.03 17:25 03-AUFMACHER	SPOR	100	362	11		16		0	0
01.01.03 16:30 03-inzahlen	SPOR	139	419	50		47		1	1
05.01.03 20:41 01-96-test	SPOR	45	227	33		35		0	0
05.01.03 20:54 01-aufmacher	SPOR	54	317	8		9		0	0
05.01.03 20:54 01-aufsetzer	SPOR	64	303	21		23		0	0
05.01.03 20:54 01-Baba	SPOR	28	134	10		15		3	2
05.01.03 18:34 02-anstoss	SPOR	45	208	4		4		0	0
05.01.03 19:36 02-aufmacher	SPOR	69	308	5		24		0	0
05.01.03 19:36 02-aufsetzer	SPOR	47	246	14		22		2	1
05.01.03 19:37 02-HSV	SPOR	81	315	9		17		1	1
05.01.03 19:01 04-aufmacher	SPOR	76	346	13		21		2	2
05.01.03 19:27 02-	SPOR	55	236	13		19		1	1
05.01.03 19:27 02-wolf	SPOR	50	216	1		3		0	0
05.01.03 19:27 friesinger	SPOR	55	253	3		10		0	0
05.01.03 19:27 scorps	SPOR	88	342	13		17		1	1
05.01.03 19:27 04-aufsetzer	SPOR	68	370	14		22		1	1
05.01.03 18:14 04-reformen	SPOR	46	209	4		4		0	0
07.01.03 16:32 anstoss	SPOR	62	276	1		2		0	0
07.01.03 16:32 02-aufmacher	SPOR	83	339	13		16		1	1

Datenbankreferenz		Textgröße		DITECT		WORD		reale Fehler	
Datum, Titel des Artikels	Ressort	Zeilen	Wörter	ges.	vers.	ges.	vers.	ges.	vers.
07.01.03 18:24 02-oddsetcup	SPOR	83	269	2		2		0	0
07.01.03 19:36 02-sportkompakt	SPOR	71	323	9		17		1	1
Summe:	**SPOR**	**2451**	**10154**	**431**	**271**	**591**	**360**	**17**	**15**
Summe gesamt:		**12183**	**50456**	**1024**	**725**	**1312**	**868**	**148**	**117**

B ÜBERBLICK ÜBER DIE VORHANDENEN FEHLER

Die folgenden Tabellen geben alle orthografischen Fehler wieder, die im Testkorpus aufgetreten sind. Die Fehler sind nach ihrem Fehlertyp sortiert und soweit möglich nach der Art des Fehlers weiter kategorisiert. Zusätzlich werden das Ressort angegeben, in dem sich der Fehler befand, die Häufigkeit, in der der Fehler auftrat (*Frequenz*), und die Minimum-Edit-Distance , in der die Schreibung von der korrekten Zeichenfolge abweicht (*MED*). Nur in seltenen Fällen tritt ein multipler Fehler auf.

Die Spalten zu den einzelnen Programmen beziehen sich auf Fehlerdetektion und Fehlerkorrektur. In der Spalte *Det.* (Detektion) sind die Fehler mit einem Häkchen markiert, die von dem jeweiligen System gefunden wurden. Die Spalte *Kor.* (Korrektur) gibt an, ob sich der richtige Korrekturvorschlag an Position 1 oder an Position 2 befand bzw. nicht vorhanden war (X).

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
Fehlertyp: Auslassung								
gibts	EINS		Apostroph	1	✓	2	✓	1
wars	EINS		Apostroph	1			✓	1
gabs	EINS		Apostroph	1			✓	1
wirds	EINS		Apostroph	1	✓	1	✓	1
habs	POLI		Apostroph	1	✓	1	✓	1
gibts	LOKA	3	Apostroph	1	✓	2	✓	1
solls	LOKA		Apostroph	1	✓	2	✓	1

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
ists	LOKA		Apostroph	1	✓	2	✓	1
gabs	LOKA	2	Apostroph	1			✓	1
gibts	WIRT	2	Apostroph	1	✓	2	✓	1
stehts	WIRT		Apostroph	1	✓	2	✓	1
Wenns	WIRT		Apostroph	1	✓	x	✓	1
wirds	WIRT		Apostroph	1	✓	1	✓	1
gabs	SPOR	2	Apostroph	1			✓	1
gings	SPOR		Apostroph	1	✓	1	✓	1
hats	SPOR		Apostroph	1	✓	x	✓	1
siehts	SPOR		Apostroph	1			✓	1
gehts	SPOR		Apostroph	1	✓	x	✓	1
Darüberhinaus	POLI		G. u. Z.-schreibung	1	✓	1	✓	1
nichtständigen	POLI	3	G. u. Z.-schreibung	1				
auseinandergebrochenen	POLI		G. u. Z.-schreibung	1	✓	1		
sowas	POLI		G. u. Z.-schreibung	1	✓	x	✓	1
Ex-Unionswirtschafts politischerSprecher	LOKA		G. u. Z.-schreibung	1	✓	x	✓	x
nochmal	WIRT		G. u. Z.-schreibung	1			✓	1
zurückkönnen	WIRT		G. u. Z.-schreibung	1				
vollgepackte	WIRT		G. u. Z.-schreibung	1	✓	x		
hinterherzustiefeln	SPOR		G. u. Z.-schreibung	2	✓	x	✓	x
Westerwelleauf	EINS		Zusammenschr.	1	✓	x	✓	x
die offizielle anerkannten Atomw.-Mächte	POLI		Deklinationstyp	1				
Die deutsche Automobilzulieferer	WIRT		Kongruenz, Num.	1				
ihr schlechteste Jahr	WIRT		Kongruenz, Gen.	1				
Experten fordern, ein Teil zu verkaufen	WIRT		Kongruenz, Kas.	2				

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
sein Minolta-Kollegen	WIRT		Kongruenz, Num.	1				
Alle ist schwierig	SPOR		Kongruenz, Num.	1				
Sie bekräftigen gestern	EINS		Tempus, Bezug	1				
top	SPOR		Neue Rechtschr.	1			✓	1
Parte	POLI		Rechtschreibung	1				
andersfalls	EINS		Rechtschreibung	1	✓	2	✓	1
Chistian	EINS		Rechtschreibung	1	✓	2	✓	1
Wachstumprognose	EINS	2	Rechtschr., Fuge	1			✓	1
Katasrophe	POLI		Rechtschreibung	1				
die Beschäftigen	POLI		Rechtschreibung	1				
Harmut	WIRT		Rechtschreibung	1	✓	2	✓	1
US-Zeichtrickserie	WIRT		Rechtschreibung	2	✓	x	✓	x
fortwärendenkampfes	WIRT		Rechtschreibung	2	✓	x	✓	x
spanende	WIRT		Rechtschreibung	1			✓	1
in de Nacht	WIRT		Rechtschreibung	1				
tafen	SPOR		Rechtschreibung	1	✓	x	✓	x
immer düster werdenden	WIRT		Semantik	2				
Bremsen und Elektronikherstellers	WIRT		Zeichensetzung	1				
Wer wird-Millionär-Moderator	SPOR		Zeichensetzung	1	✓	x	✓	x
Fehlertyp: Substitution								
der vornehmste	EINS		Groß- u. Kleinschreibung	1				
zum Schweigen	EINS		Groß- u. Kleinschreibung	1				
sinn	EINS		Groß- u. Kleinschreibung	1				
Im einzelnen	POLI		Groß- u. Kleinschreibung	1				
abend	POLI		Groß- u. Kleinschreibung	1	✓	1	✓	2
Blauen Brief	POLI		Groß- u. Kleinschreibung	1				
Gelben Sack	LOKA		Groß- u. Kleinschreibung	1				

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
Gelbe Riese	LOKA		Groß- u. Kleinschreibung	1				
die gleiche	LOKA		Groß- u. Kleinschreibung	1				
hier zu lande	WIRT		Groß- u. Kleinschreibung	1				
Neuer Markt	WIRT	3	Groß- u. Kleinschreibung	1				
Werks-klub	SPOR		Groß- u. Kleinschreibung	1	✓	x	✓	2
[...] Linie: seit Abzug [...]	POLI		Groß- u. Kleinschreibung	1				
zieht es sie nach Berlin?	POLI		Groß- u. Kleinschreibung	1				
Der Vorschlag verunsichern	WIRT		Kongruenz, Numerus	1				
gebremsten Verve	POLI		Kongruenz, Genus	1				
dem Grundsatzfragen	POLI		Kongruenz, Kasus	1				
in wahrsten Sinne	POLI		Kongruenz	1				
BierRiese	EINS		Rechtschr., Binnenmajuskel	1	✓	x	✓	x
Repiblik	EINS		Rechtschreibung	1	✓	1	✓	1
Megabites	POLI		Rechtschreibung	1	✓	2	✓	1
BBK (Bundesdelegiertenk.)	POLI		Rechtschreibung	1				
Elvtalaue	POLI		Rechtschreibung	1	✓	1	✓	x
verkaufes	WIRT		Rechtschreibung	1	✓	x	✓	x
Verdi	EINS	5	Wortschöpfung	2				
Verdi-Chef	EINS		Wortschöpfung	2				
Verdi-Chef	POLI		Wortschöpfung	2				
Verdi	POLI	3	Wortschöpfung	2				
Verdi-Vorstandsmitglied	POLI		Wortschöpfung	2				
Verdi	LOKA	3	Wortschöpfung	2				
Verdi-Regionsvorsitzende	LOKA		Wortschöpfung	2				
Focus-Online	EINS		Zeichensetzung	1				
Know How	POLI		Zeichensetzung	1	✓	x	✓	x

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
Hapag Lloyd	WIRT		Zeichensetzung	1				
Hapag Lloyd Flug	WIRT		Zeichensetzung	2				

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
FEHLERTYP: EINFÜGUNG								
auf's	POLI		Apostroph	1			✓	1
Meier's	WIRT		Apostroph	1			✓	1
weiter gearbeitet	LOKA		G. u. Z.-schreibung	1				
offen gelassen	WIRT		G. u. Z.-schreibung	1				
Axel- Springer-Verlags	WIRT		G. u. Z.-schreibung	1				
weiter machen	WIRT		G. u. Z.-schreibung	1				
Der Preisanstiegs	WIRT		Kongruenz, Kasus	1				
zum Kerngeschäften	WIRT		Kongruenz	2				
Ein Niedersachsen brachte [...]	EINS		Kongruenz, Kasus	1				
Ministerprsäsident	EINS		Rechtschreibung	1	✓	1	✓	1
vorausssichtlich	EINS		Rechtschreibung	1	✓	1	✓	1
Kärntener	POLI		Rechtschreibung	1			✓	1
UN-Waffeninspek-tektionskommission	POLI		Rechtschreibung	3	✓	x	✓	x
beigetreiten	POLI		Rechtschreibung	1	✓	1	✓	1
weltpolistische	WIRT		Rechtschreibung	1	✓	x	✓	1
Vermögenssteuer	EINS	7	Rechtschr., Fuge	1				
"SZ"	WIRT		Zeichensetzung	2				
FEHLERTYP: TRANSPOSITION								
Bretange	POLI	1	Rechtschreibung	1	✓	1	✓	1
wordne	POLI	1	Rechtschreibung	1	✓	1	✓	1
entsprechnede	WIRT	1	Rechtschreibung	1			✓	1

					DITECT		WORD	
Zeichenfolge	Ressort	Frequenz	Art des Fehlers	MED	Det.	Kor.	Det.	Kor.
FEHLERTYP: EINFÜGUNG, SUBSTITUTION								
Journalistenmikrophone	POLI	1	Neue Rechtschr.	2				
Mikrophon	LOKA	1	Neue Rechtschr.	2				
FEHLERTYP: WORTAUSLASSUNG								
Droht ein Preisverfall [ ] kommen?	WIRT	1	Syntax	3				
[ ] waren gewesen	SPOR	1	Syntax	3				
FEHLERTYP: WORTEINFÜGUNG								
und	EINS	1	Syntax	4	✓	x	✓	x
[...] den Atlantik für zu verseuchen.	POLI	1	Syntax	4				
[...] fordern auf, dass [...] zu verkaufen	WIRT	1	Syntax	5				
Zudem will RWE will [...]	WIRT	1	Syntax	5				

C FLEXION

Dieser Testlauf untersucht die Fähigkeit der Programme, die korrekten Flexionsendungen in den Vorschlägen beizubehalten. Die Markierung besagt, ob die Programme den Fehler korrekt im ersten Versuch berichtigen konnten

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
VERBEN: SCHWACH FLEKTIEREND			
lehgen	legen	✓	✓
lehge	lege		✓
lehgst	legst	✓	✓
lehgt	legt		✓
lehgte	legte	✓	✓
lehgtest	legtest		✓
lehgtet	legtet		✓
lehgten	legten		✓
lehgend	legend		✓
gelehgt	gelegt		✓
sehgeln	segeln	✓	✓
sehgele	segele		✓
sehgle	segle		✓
sehgelst	segelst	✓	✓
sehgelt	segelt		✓
sehgelte	segelte	✓	✓
sehgeltest	segeltest	✓	✓
sehgeltet	segeltet		✓
sehgelnd	segelnd	✓	✓

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
gesehgelt	gesegelt	✓	✓
rzudern	rudern	✓	✓
rzudere	rudere		✓
rzudre	rudre		✓
rzuderst	ruderst	✓	✓
rzudert	rudert		✓
rzuderte	ruderte	✓	✓
rzudertest	rudertest	✓	✓
rzudertet	rudertet		✓
rzudernd	rudernd	✓	✓
gerzudert	gerudert		✓
VERBEN: STARK FLEKTIEREND			
schraiten	schreiten		✓
schraitest	schreitest		✓
schraite	schreite		✓
schriett	schritt		✓
schriette	schritte		✓
schriettst	schrittst		✓
schriettest	schrittest		✓
schriettet	schrittet		✓
schraitend	schreitend	✓	✓
geschrietten	geschritten	✓	✓
behrgen	bergen		✓
behrge	berge		✓
bihrgst	birgst	✓	✓
bihrgt	birgt		✓
behrgt	bergt		✓
bahrg	barg	✓	✓
bahrgst	bargst		✓
bahrgen	bargen	✓	✓
bahrgt	bargt		✓
behrgend	bergend		✓

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
behrgende	bergende		✓
gebohrgen	geborgen		✓
wqerfen	werfen		✓
wqerfe	werfe		✓
wqirfst	wirfst		✓
wqirft	wirft		✓
wqerft	werft		✓
wqarf	warf		✓
wqarfst	warfst		✓
wqarfen	warfen		✓
wqarft	warft		✓
wqerfend	werfend		✓
gewqorfen	geworfen	✓	✓
wqürfe	würfe	✓	✓
SUBSTANTIVE: GRUNDFORMFLEXION			
Maskulina, stark			
Urehber	Urheber	✓	✓
Urehbern	Urhebern	✓	✓
Urehbers	Urhebers	✓	✓
Maskulina, schwach			
Buchstane	Buchstabe		✓
Buchstanen	Buchstaben	✓	✓
Maskulina, gemischt			
Staatsangehoriger	Staatsangehöriger		✓
Staatsangehorige	Staatsangehörige		✓
Staatsangehorigen	Staatsangehörigen		✓
Neutra, stark			
Hoheitsgebit	Hoheitsgebiet	✓	✓
Hoheitsgebite	Hoheitsgebiete		✓
Hoheitsgebiten	Hoheitsgebieten		✓
Hoheitsgebits	Hoheitsgebiets	✓	✓
Neutra, gemischt			
Verfaren	Verfahren	✓	✓

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
Verfarens	Verfahrens		✓
Feminina mit n-Plural			
Entnahne	Entnahme	✓	✓
Entnahnen	Entnahmen	✓	✓
Ubermittlung	Übermittlung		✓
Ubermittlungen	Übermittlungen		✓
SUBSTANTIVE: FLEXION MIT UMLAUTUNG ODER UNREGELMÄSSIGER STAMMÄNDERUNG			
Maskulina, stark			
Gruntsatz	Grundsatz	✓	✓
Gruntsätze	Grundsätze	✓	✓
Gruntsätzen	Grundsätzen	✓	✓
Gruntsatzes	Grundsatzes	✓	✓
Verrtrag	Vertrag		✓
Verrträge	Verträge		✓
Verrträgen	Verträgen		✓
Verrtrages	Vertrages		✓
Verrtrags	Vertrags		✓
Neutra, stark			
Dritttland	Drittland	✓	✓
Dritttlandes	Drittlandes	✓	✓
Dritttländer	Drittländer	✓	✓
Dritttländern	Drittländern	✓	✓
Dritttlande	Drittlande	✓	✓
Haüs	Haus		✓
Haüses	Hauses		✓
Haüse	Hause		✓
Häüser	Häuser		✓
Häüsern	Häusern		✓
Neutra, gemischt			
Mehdium	Medium	✓	✓
Mehdiums	Mediums	✓	✓
Mehdien	Medien		✓

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
Feminina mit e-Plural			
Cunst	Kunst		✓
Cünste	Künste		✓
Cünsten	Künsten		✓
Ubereinkunft	Übereinkunft	✓	✓
Ubereinkünfte	Übereinkünfte	✓	✓
Ubereinkünften	Übereinkünften	✓	✓

D KOMPOSITION

Testterm	DITECT	Word
Schmopf-Kerzen	Schrumpfnieren	Schopfkerzen
Speizeit-Fraß	Spielzeitstart	(kein Fehler)
Schmauch-Berzen	Betern	Schmauch-erzen
Zeisig-Drone	Drogen	Zeisig-Krone
Kurden-Staart	Staat	Kurden-Start
Tisch-Fuußball	Fußball	Tisch-Fußball
Sturden	strudeln	Stunden
Sturden-Kaat	Kata	Sturdehn-Kat
Strurden-Kaat	Kata	(Keine Vorschläge)
Strqurden-Kaat	(keine Vorschläge)	(Keine Vorschläge)
Sturden-Staat	(kein Fehler)	Sturen-Staat
Strurden-Staat	(kein Fehler)	(Keine Vorschläge)
Struhrden-Staat	(Keine Vorschläge)	(Keine Vorschläge)
Fleiß-Schmiege	(kein Fehler)	Fleiß-Schmiegen
Fleiß-Schmiegel	Schmirgel	Fleiß-Schmirgel
Droms-Tasche	(kein Fehler)	Drops-Tasche
Drohms-Tasche	(kein Fehler)	Drohäs-Tasche
Droehms-Tasche	(kein Fehler)	(Keine Vorschläge)
Drkoehms-Tasche	Drechselmaschine	(Keine Vorschläge)
Tropf-Kerze	(kein Fehler)	(kein Fehler)
Kropf-Terze	(kein Fehler)	Kropf-Sterze
Kopf-Trerze	Treber	
Topf-Krerze	Kreder	Topf-kreuze
Krerzen-Topf	(kein Fehler)	Kreuzen-Topf
Trerzen-Kopf	(kein Fehler)	Terzen-Kopf
Krarzen-Topf	(kein Fehler)	Kratzen-Topf
Trärzen-Kopf	(kein Fehler)	(Keine Vorschläge)
Trärzen-Kopf	(kein Fehler)	(Keine Vorschläge)
Krrerzen-Topf	(Keine Vorschläge)	Karrerzen-Topf
Trrerzen-Kopf	(Keine Vorschläge)	Türerzen-Kopf
Krartzen-Topf	(Keine Vorschläge)	Kratzen-Topf
Trärtzen-Kopf	(Keine Vorschläge)	(Keine Vorschläge)

E GETRENNT- UND ZUSAMMENSCHREIBUNG

1 GZS – Tabellarisch

Dieser Abschnitt zeigt alle Beispiele, die im analytischen Testteil zum Thema Getrennt- und Zusammenschreibung an den beiden Suchmaschinen getestet wurden. Die Auswahl beinhaltet Beispiele aus allen möglichen Bereichen, in denen die Neuregelung zur Getrennt- und Zusammenschreibung wirkt. Dies betrifft weitestgehend Verben bzw. Partizipien in Verbindung mit einer verbalen, adjektivischen, substantivischen oder adverbialen Erstkonstituente.

Die Gliederung der Beispiele erfolgt nach der Zweitkonstituente. D. h. zunächst werden alle Verben mit ihren möglichen Erstkonstituenten behandelt, dann die Partizipien, sofern es nicht schon Überschneidungen gab, dann die Adjektive und schließlich sonstige Gruppen. Substantivische Verbindungen sind von der Regelung nicht betroffen, da die Komposition relativ eindeutig geregelt ist.

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
1.1 Verb (im Infinitiv Präsens Aktiv) und Verb Ohne Unterscheidung zwischen eigentlicher und übertragener Bedeutung werden Verbindungen aus Verben im Zuge der Rechtschreibreform getrennt geschrieben.			
kennenlernen	kennen lernen		✓
sitzenbleiben (in der Schule)	sitzen bleiben		✓

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
1.2 Partizip und Verb			
verlorengehen	verloren gehen	✓	✓
gefangennehmen	gefangen nehmen		✓
bekanntgeben*	bekannt geben	✓	✓
1.3 Adjektiv und Verb Diese Verbindungen schreibt man getrennt, wenn das Adjektiv gesteigert oder die Wortgruppe erweitert werden kann.			
genaunehmen	genau nehmen (sehr genau nehmen)		✓
gutgehen	gut gehen (besser gehen)		✓
zufriedenstellen	zufrieden stellen (ganz zufrieden stellen)	✓	✓
aufrecht erhalten	aufrechterhalten		
bereit halten	bereithalten		
bloß stellen	bloßstellen		
brach liegen	brachliegen		
fest nehmen	festnehmen		
gerade biegen	geradebiegen		
groß geschrieben	großgeschrieben		
groß schreiben	großschreiben		
gut schreiben	gutschreiben		
hoch jubeln	hochjubeln		
hoch zufrieden	hochzufrieden		
kaputt machen	kaputtmachen		
kaputt machen	kaputtmachen		
schwarz arbeiten	schwarzarbeiten		
zum wahnsinnig werden	zum Wahnsinnigwerden		
1.4 Verbindungen mit dem Verb *sein*			
dasein	da sein (aber: das Dasein)		✓
zusammensein	zusammen sein		
1.5 Verbindungen aus Adjektiv und Verb mit Ableitungen auf *ig, isch, lich*			
heiligsprechen	heilig sprechen	✓	✓
übrigbleiben	übrig bleiben	✓	

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
heimlichtun	heimlich tun		✓

1.6 Verbindungen aus zusammengesetztem Adverb und Verb
Verbindungen aus aneinander, aufeinander, auseinander oder abwärts, aufwärts, vorwärts etc. und Verb schreibt man getrennt.

aneinanderfügen	aneinander fügen	✓	
aufwärtsgehen (besser werden)	aufwärts gehen	✓	✓
aneinanderreihen	aneinander reihen	✓	✓
auseinandersetzen	auseinander setzen	✓	

1.7 Verbindungen von Substantiv und Verb
Wenn Substantiv und Verb eine untrennbare Zusammensetzung bilden, wird zusammengeschrieben. Ansonsten schreibt man getrennt.

radfahren	Rad fahren		✓
das Rad fahren	das Radfahren		
eislaufen	Eis laufen		✓
das Eis laufen	das Eislaufen		
das Fußball spielen	das Fußballspielen		
das Ski laufen	das Skilaufen		
Berg steigen	bergsteigen		
Hand haben	handhaben		
Maß regeln	maßregeln		

1.8 Sonstige Verbindungen von Substantiv und Verb
In der Regel schreibt man Verbindungen aus Substantiven und Verben getrennt, es sei denn, die Bedeutung verblasst, wie bei *heim-, irre-,Preis-, stand-, statt-, teil-, wett- und wunder-*.

Heim gehen	heimgehen		
irre führen	irreführen		
Stand halten	standhalten		
statt geben	stattgeben		

1.9 Verbindungen ausPartikel und Verb
Weiterhin zusammengeschrieben werden auch Verbindungen ausPartikel und Verb unter anderem mit *ab-, an-, aus-, dabei-, daneben-, davon-, entgegen-, entlang-, gegenüber-, heraus-, herunter-, hervor-, hinein-, hinweg-, hinzu-, los-, nieder-, über-, umher-, vorbei-, vorher-, weg-, weiter-, wieder-, zurecht-, zurück-, zusammen-, zuvor-* und *zwischen-*.

ab holen	abholen		✓
bevor stehend	bevorstehend		
dagegen halten	dagegenhalten		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
davon kommen	davonkommen		
dazu gehören	dazugehören		
dran gehängt	drangehängt		
entgegen gekommen	entgegengekommen		
entgegen treten	entgegentreten		
heraus kommen	herauskommen		
herunter laufen	herunterlaufen		
hinein bringen	hineinbringen		
hinein gehen	hineingehen		
hinein treten	hineintreten		
hinterher kommen	hinterherkommen		
hinweg kommen	hinwegkommen		
hinzu gekommen	hinzugekommen		
meist gejagt	meistgejagt		
rüber kommen	rüberkommen		
vorbei gehen	vorbeigehen		
weiter bilden	weiterbilden		
weiter spielen	weiterspielen		
wieder beleben	wiederbeleben		
wieder kommen	wiederkommen		
zurück gefallen	zurückgefallen		
zurück gehen	zurückgehen		
zusammen gewohnt	zusammengewohnt		
zusammen halten	zusammenhalten		

1.10 Verbindungen mitPartizipien oder Adjektiven

1.10.1 Allgemein

Bei Verbindungen mitPartizipien als zweitem Bestandteil schreibt man getrennt, wenn auch das demPartizip zugrunde liegende Verb vom ersten Bestandteil getrennt geschrieben wird.

kennengelernt	kennen gelernt	✓	✓
obengenannt	oben genannt	✓	
genaugenommen	genau genommen	✓	
verlorengegangen	verloren gegangen	✓	
Radfahrend	Rad fahrend	✓	

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
vorwärtsblickend	vorwärts blickend		
zufriedenstellend	zufrieden stellend		
Siedendheiß	siedend heiß		
Bläulichgrün	bläulich grün		

1.10.2 Erster Bestandteil ist erweiter- oder steigerbar

besserverdienend	besser verdienend		
blankgeputzt	blank geputzt		
dichtbehaart	dicht behaart (dichter behaart)	✓	
hellstrahlend	hell strahlend	✓	
hochqualifiziert	hoch qualifiziert (höher qualifiziert)		
Klarmachen	klar machen		
neugegründet	neu gegründet		
Vollgepumpt	voll gepumpt	✓	
alt bewährt	altbewährt		
alt eingesessen	alteingesessen		
anders lautend	anderslautend		
fest gemacht	festgemacht		
frei gelassen	freigelassen		
hoch geschlossen	hochgeschlossen		
hoch intelligent	hochintelligent		
Leck/leck geschlagen	leckgeschlagen		

1.10.3 Verbindungen, die selbst erweiter- oder steigerbar sind
Verbindungen aus adjektivischemPartizip und Adjektiv oder aus Adjektiv undPartizip schreibt man getrennt, wenn der erste Bestandteil gesteigert oder erweitert werden kann.

die furchteinflößende Gestalt	die Furcht einflößende Gestalt		
die große furchteinflößende Gestalt	die große Furcht einflößende Gestalt	✓	
eine äußerst Furcht einflößende Gestalt	eine äußerst furchteinflößende Gestalt		
die gewinnbringende Idee	die Gewinn bringende Idee		
die gewinnebringende Idee	die Gewinne bringende Idee		✓
eine noch Gewinn bringendere Idee	eine noch gewinnbringendere Idee		
Den abendfüllend	Den Abend füllend		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
Die Abend füllende Veranstaltung	Die abendfüllende Veranstaltung		
Lippen schürzend	lippenschürzend		
Die roten lippenschürzend	Die roten Lippen schürzend		
Beispiel gebend	beispielgebend		
Beitrag steigernd	beitragsteigernd		
kapitalbildend*	Kapital bildend		
Kräfte fördernd	kräftefördernd		
Skandal umwittert	skandalumwittert		
Markt gerecht	marktgerecht		
Maß geschneidert	maßgeschneidert		
Schweiß gebadet	schweißgebadet		
Sitten gefährdend	sittengefährdend		
Weg weisend	wegweisend		
Zähne knirschend	zähneknirschend		

1.10.4 Erstbestandteil steht für eine Wortgruppe
Wenn die Verbindung selbst erweiter- oder steigerbar ist, kann auch weiterhin zusammengeschrieben werden. Liegt eine solche Erweiterung oder Steigerung vor, ist nur die Zusammenschreibung möglich

ein Herz erfrischendes Lachen	ein herzerfrischendes Lachen		
ein das herzerfrischendes Lachen	ein das Herz erfrischendes Lachen		
Sagen umwoben	sagenumwoben		
von sagenumwoben	von Sagen umwoben		
Ball führend	ballführend		
den ballführend	den Ball führend		
Computer gesteuert	computergesteuert		
vom computergesteuert	vom Computer gesteuert		
Monate lang	monatelang		
mehrere monatelang	mehrere Monate lang		
Gold schimmernd	goldschimmernd		
wie goldschimmernd	wie Gold schimmernd		
Achsel zuckend	achselzuckend		
Elfenbein farben	elfenbeinfarben	✓	✓
Form vollendet	formvollendet		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
Hass erfüllt	hasserfüllt		
Hersteller abhängig	herstellerabhängig		
Herz zerreißend	herzzerreißend		
Kilometer weit	kilometerweit		
Millionen schwer	millionenschwer		
Nitrofen belastet	nitrofenbelastet	✓	✓
Polizei bekannt	polizeibekannt		
Potenz steigernd	potenzsteigernd		
Schmerz stillend	schmerzstillend		
Sonnen beschienen	sonnenbeschienen		
Sport orientiert	sportorientiert		
Tage lang	tagelang		
Tränen überströmt	tränenüberströmt		

1.11 Verbindungen von so, wie und zu mit Adjektiv

Verbindungen von so, wie und zu mit Adjektiv werden getrennt geschrieben. Bei Gebrauch von soviel, solange oder soweit als Konjunktion wird aber weiterhin zusammengeschrieben, z.B. in Soweit ich weiß, ...

zuwenig	zu wenig	✓	
soviel	so viel (als Numeraladjektiv)		
wieviel	wie viel	✓	✓
all zu	allzu		

1.12 Sonstige

bei Zeiten	beizeiten		
dort hin	dorthin		
hinten an	hintenan		
sogenannt*	so genannt		
zum soundso vielten Mal	zum soundsovielten Mal		✓
drei Mal	dreimal		
jedesmal*	jedes Mal		✓
frühest möglich	frühestmöglich		
lila farben	lilafarben	✓	✓
unzulässiger Weise	unzulässigerweise		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
Dreiviertel*	drei Viertel		
Zweidrittel*	zwei Drittel		
ein Viertel Gramm	ein Viertelgramm		

2 GETRENNT- UND ZUSAMMENSCHREIBUNG IN SYNTAKTISCHEM KONTEXT

Die Beispiele in dieser Tabelle sind denen aus der tabellarischen Auflistung oben entnommen. Da einige der Beispiele kontextabhängig sind, soll dieser Test zeigen, ob die Programme den syntaktischen Kontext zur Disambiguierung der Schreibungen berücksichtigen können.

	Testsatz	DITECT	Word
Falsch	Der Abend füllende Film begeistert die Massen.		
Korrekt	Der abendfüllende Film begeistert die Massen.		
Falsch	Achsel zuckend drehte sie sich um.		
Korrekt	Achselzuckend drehte sie sich um.		
Falsch	Sie fand die Aufgabe nicht all zu schwierig.		
Korrekt	Sie fand die Aufgabe nicht allzu schwierig.		
Falsch	Der alt eingesessene Betrieb beschäftigt über 200 Menschen.		
Korrekt	Der alteingesessene Betrieb beschäftigt über 200 Menschen.		
Falsch	Du darfst die Wörter nicht sinnlos aneinanderreihen.	✓	✓
Korrekt	Du darfst die Wörter nicht sinnlos aneinander reihen.		
Falsch	Kann er die Geschwindigkeit aufrecht erhalten?		

	Testsatz	DITECT	Word
Korrekt	Kann er die Geschwindigkeit aufrechterhalten?		
Falsch	Ich werde mich damit auseinandersetzen!	✓	
Korrekt	Ich werde mich damit auseinander setzen!		
Falsch	Beitrag steigernde Maßnahmen verunsichern die Kassenpatienten.		
Korrekt	Beitragsteigernde Maßnahmen verunsichern die Kassenpatienten.		
Falsch	Der Verein muss seine Finanzen bei Zeiten in Ordnung bringen.		
Korrekt	Der Verein muss seine Finanzen beizeiten in Ordnung bringen.		
	Sie hat die Ergebnisse bekannt gegeben.		
Korrekt	Sie hat die Ergebnisse bekannt gegeben.		
Falsch	Zum Zeitpunkt der Invasion sollen 250.000 Soldaten bereit stehen.		
Korrekt	Zum Zeitpunkt der Invasion sollen 250.000 Soldaten bereitstehen.		
Falsch	Für besserverdienende Steuerzahler weht ein anderer Wind.		
Korrekt	Für besser verdienende Steuerzahler weht ein anderer Wind.		
Falsch	Die bevor stehende Reform fällt wahrscheinlich mager aus.		
Korrekt	Die bevorstehende Reform fällt wahrscheinlich mager aus.		
Falsch	Die Schuhe sind blankgeputzt.		
Korrekt	Die Schuhe sind blank geputzt.		
Falsch	Sie konnte keine gewichtigen Argumente dagegen halten.		
Korrekt	Sie konnte keine gewichtigen Argumente dagegenhalten.		
Falsch	Es war schwer, gewichtige Argumente dagegen zu halten.		

	Testsatz	DITECT	Word
Korrekt	Es war schwer, gewichtige Argumente dagegenzuhalten.		
Falsch	Er sollte dazu gehören.		
Korrekt	Er sollte dazugehören.		
Falschx	Dort hin führt nur eine verstaubte Schotterpiste.		
Korrekt	Dorthin führt nur eine verstaubte Schotterpiste.		
Falsch	Sie hat sich dran gehängt.		
Korrekt	Sie hat sich drangehängt.		
Falsch	Du wirst mich drei Mal verraten.		
Korrekt	Du wirst mich dreimal verraten.		
Falsch	Etwa Dreiviertel der Deutschen sind mit ihrem Job nicht zufrieden.		
Korrekt	Etwa drei Viertel der Deutschen sind mit ihrem Job nicht zufrieden.		
Falsch	Ihre Elfenbein farbene Haut faszinierte ihn.	✓	✓
Korrekt	Ihre elfenbeinfarbene Haut faszinierte ihn.		
Falsch	Ich werde Ihnen entgegen kommen.		
Korrekt	Ich werde Ihnen entgegenkommen.		
Falsch	DiePolizei hat die Verbrecher fest genommen.		
Korrekt	DiePolizei hat die Verbrecher festgenommen.		
Falsch	Auf der Automobilmesse wurden Form vollendete Fahrzeuge vorgestellt.		
Korrekt	Auf der Automobilmesse wurden formvollendete Fahrzeuge vorgestellt.		
Falsch	Nach drei Wochen wurde er wieder frei gelassen.		

	Testsatz	DITECT	Word
Korrekt	Nach drei Wochen wurde er wieder freigelassen.		
Falsch	Er reichte die Unterlagen nicht eben frühest möglich ein.		
Korrekt	Er reichte die Unterlagen nicht eben frühestmöglich ein.		
Falsch	Das kann man wieder gerade biegen.		
Korrekt	Das kann man wieder geradebiegen.		
Falsch	Dafür sind gewaltverherrlichende Videospiele verantwortlich.		
Korrekt	Dafür sind Gewalt verherrlichende Videospiele verantwortlich.		
Falsch	Dieses Geschäft ist Gewinn trächtig.		
Korrekt	Dieses Geschäft ist gewinnträchtig.		
Falsch	Substantivierte Verben werden groß geschrieben.		
Korrekt	Substantivierte Verben werden großgeschrieben.		
Falsch	Den Betrag kann ich Ihnen gut schreiben.		
Korrekt	Den Betrag kann ich Ihnen gutschreiben.		
Falsch	Er blickte in Hass erfüllte Augen.		
Korrekt	Er blickte in hasserfüllte Augen.		
Falsch	Morgen werde ich heim kommen.		
Korrekt	Morgen werde ich heimkommen.		
Falsch	Die Funktionen des Geräts sind Hersteller abhängig.		
Korrekt	Die Funktionen des Geräts sind herstellerabhängig.		
Falsch	Sie schluchzte Herz zerreißend.		
Korrekt	Sie schluchzte herzzerreißend.		

	Testsatz	DITECT	Word
Falsch	Du kannst die Sachen schon hinein bringen.		
Korrekt	Du kannst die Sachen schon hineinbringen.		
Falsch	Es ist schwer, über den Verlust hinweg zu kommen		
Korrekt	Es ist schwer, über den Verlust hinwegzukommen.		
Falsch	Es sind noch zweiPersonen hinzu gekommen.		
Korrekt	Es sind noch zweiPersonen hinzugekommen.		
Falsch	Dieses Kleid ist hoch geschlossen.		
Korrekt	Dieses Kleid ist hochgeschlossen.		
Falsch	Mit der Arbeit war er hoch zufrieden.		
Korrekt	Mit der Arbeit war er hochzufrieden.		
Falsch	DasPassiert mir jedesmal.	✓	
Korrekt	DasPassiert mir jedes Mal.		
Falsch	Wenn das so weitergeht, wird er es noch kaputt machen.		
Korrekt	Wenn das so weitergeht, wird er es noch kaputtmachen.		
Falsch	Das Land erstreckte sich Kilometer weit vor ihren Augen.		
Korrekt	Das Land erstreckte sich kilometerweit vor ihren Augen.		
Falsch	Das solltest du dir noch einmal klarmachen.		
Korrekt	Das solltest du dir noch einmal klar machen.		
Falsch	Schwarzbrot wirkt Kräfte fördernd.		
Korrekt	Schwarzbrot wirkt kräftefördernd.		
Falsch	Das Schiff ist vor der Küste Spaniens leck geschlagen.		
Falsch	Das Schiff ist vor der Küste Spaniens Leck geschlagen.		

	Testsatz	DITECT	Word
Korrekt	Das Schiff ist vor der Küste Spaniens leckgeschlagen.		
Falsch	Die lila farbene Kuh steht für eine besondere Marke.	✓	✓
Korrekt	Die lilafarbene Kuh steht für eine besondere Marke.		
Falsch	DasProdukt ist in seiner derzeitigen Form nicht Markt gerecht.		
Korrekt	DasProdukt ist in seiner derzeitigen Form nicht marktgerecht.		
Falsch	Maß geschneiderte Anzüge sind teuer.		
Korrekt	Maßgeschneiderte Anzüge sind teuer.		
Falsch	Er ist der meist gejagte Mann der Welt.		
Korrekt	Er ist der meistgejagte Mann der Welt.		
Falsch	Der Millionen schwere Zigarettenfabrikant wurde entführt.		
Korrekt	Der millionenschwere Zigarettenfabrikant wurde entführt.		
Falsch	Er wartete Monate lang auf ihre Rückkehr.		
Korrekt	Er wartete monatelang auf ihre Rückkehr.		
Falsch	Die Nahrungsmittel sind Nitrofen belastet.	✓	✓
Korrekt	Die Nahrungsmittel sind nitrofenbelastet.	✓	✓
Falsch	DasPotenz steigernde Mittel ist umstritten.		
Korrekt	DasPotenzsteigernde Mittel ist umstritten.		
Falsch	Sie war Schweiß gebadet.		
Korrekt	Sie war schweißgebadet.		
Falsch	Das Ski laufen im Harz ist um diese Zeit noch nicht möglich.		
Korrekt	Das Skilaufen im Harz ist um diese Zeit noch nicht möglich.		

	Testsatz	DITECT	Word
Falsch	Sie hat sich beim Schlittschuh laufen verletzt.		
Korrekt	Sie hat sich beim Schlittschuhlaufen verletzt.		
Falsch	Es handelt sich um das sogenannte Stockholmsyndrom.		
Korrekt	Es handelt sich um das so genannte Stockholmsyndrom.		
Falsch	Er hat das zum soundso vielten Mal getan.	✓	
Korrekt	Er hat das zum soundsovielten Mal getan.		
Falsch	Ich werde eine Weile spazierengehen.	✓	✓
Korrekt	Ich werde eine Weile spazieren gehen.		
Falsch	Ein Viertel Gramm Marihuana reicht nicht für eine Verurteilung.		
Korrekt	Ein Viertelgramm Marihuana reicht nicht für eine Verurteilung.		
Korrekt	Ein viertel Gramm Marihuana reicht nicht für eine Verurteilung.		
Falsch	Er ist mit Dopingmitteln vollgepumpt.	✓	
Korrekt	Er ist mit Dopingmitteln voll gepumpt.		
Falsch	Er kann an keinem Schnellimbiss vorbei gehen.		
Korrekt	Er kann an keinem Schnellimbiss vorbeigehen.		
Falsch	Der Élysée-Vertrag war Weg weisend für die deutsch-französischen Bez.		
Korrekt	Der Élysée-Vertrag war wegweisend für die deutsch-französischen Bez.		
Falsch	Wer sich nicht weiter bildet, verliert den Anschluss.		
Korrekt	Wer sich nicht weiterbildet, verliert den Anschluss.		
Falsch	Der Versuch, sie wieder zu beleben, misslang.		
Korrekt	Der Versuch, sie wiederzubeleben, misslang.		

	Testsatz	DITECT	Word
Falsch	Sie antwortete Zähne knirschend.		
Korrekt	Sie antwortete zähneknirschend.		
Falsch	Er ist auf den drittenPlatz zurück gefallen.		
Korrekt	Er ist auf den drittenPlatz zurückgefallen.		
Falsch	Sie haben über drei Jahre zusammen gewohnt .		
Korrekt	Sie haben über drei Jahre zusammengewohnt.		
Falsch	Zweidrittel der Stimmen sind erforderlich.		
Korrekt	Zwei Drittel der Stimmen sind erforderlich.		
Falsch	Das Fußball spielen macht mir Spaß.		
Korrekt	Das Fußballspielen macht mir Spaß.		
Falsch	Es ist zum wahnsinnig werden!		
Korrekt	Es ist zum Wahnsinnigwerden!		

F GROSS- UND KLEINSCHREIBUNG

	Testsatz	DITECT	Word
1.1 Substantivierte Infinitive			
Falsch	Es ist zum verzweifeln.		
Korrekt	Es ist zum Verzweifeln.		
Falsch	DasPrinzip: Klassik zum kennenlernen, reinschnuppern, genießen.	✓	
Korrekt	DasPrinzip: Klassik zum Kennenlernen, Reinschnuppern, Genießen.		
Falsch	Denn im Qualm fand der Experte eine Reihe von Kohlenwasserstoffen, die durch einatmen gesundheitsgefährdend sein können.		
Korrekt	Denn im Qualm fand der Experte eine Reihe von Kohlenwasserstoffen, die durch Einatmen gesundheitsgefährdend sein können.		
Falsch	Da hilft kein beten, Südkorea!		
Korrekt	Da hilft kein Beten, Südkorea!		
Falsch	Dieses Buch ist ein muss für jeden Germanisten.		
Korrekt	Dieses Buch ist ein Muss für jeden Germanisten.		
Falsch	Das lesen fällt ihm immer noch schwer.		
Korrekt	Das Lesen fällt ihm immer noch schwer.		
Falsch	Dabei blieb es, am Ende hatten die Langenhagener gut Lachen.		
Korrekt	Dabei blieb es, am Ende hatten die Langenhagener gut lachen.		
1.2 Verben im zu-Infinitiv			
Falsch	Der Einbrecher hatte vor Gericht nichts mehr zu Lachen.		
Korrekt	Der Einbrecher hatte vor Gericht nichts mehr zu lachen.		
Falsch	Viele Menschen auf der Welt haben nicht genug zu Essen.		
Korrekt	Viele Menschen auf der Welt haben nicht genug zu essen.		

	Testsatz	DITECT	Word
Falsch	Außerdem habe der Gemeinderat beschlossen, nur bis 30Prozent der Betreuungskosten durch Elternbeiträge zu Decken.		
Korrekt	Außerdem habe der Gemeinderat beschlossen, nur bis 30Prozent der Betreuungskosten durch Elternbeiträge zu decken.		

1.3 Verben in Verbindung mit Modal- oder modifizierenden Verben

Falsch	Hier kann man auch Schreien.		
Korrekt	Hier kann man auch schreien.		
Falsch	Er will Einkaufen.		
Korrekt	Er will einkaufen.		
Falsch	Er selbst habe nicht Ausweichen können.		
Korrekt	Er selbst habe nicht ausweichen können.		
Falsch	Die Abwehr offenbarte ungeahnte Lücken, was Seidensticker alles andere als Schwärmen ließ.		
Korrekt	Die Abwehr offenbarte ungeahnte Lücken, was Seidensticker alles andere als schwärmen ließ.		

1.4 Verben in Verbindung mit den Verben gehen, sehen, sein (Absentiv)

Falsch	Er will Einkaufen gehen.		
Korrekt	Er will einkaufen gehen.		
Falsch	Beide unternahmen Ausflüge, gingen Wandern, fuhren Fahrrad.		
Korrekt	Beide unternahmen Ausflüge, gingen wandern, fuhren Fahrrad.		
Falsch	Dann habe er aus dem Haus gewollt, aber Wilhelm hatte ihn Kommen gesehen.		
Korrekt	Dann habe er aus dem Haus gewollt, aber Wilhelm hatte ihn kommen gesehen.		
Falsch	Sie ist Einkaufen.		
Korrekt	Sie ist einkaufen.		

2.1 Substantivierte Adjektive

Falsch	EtwasPersönliches stecke dahinter.		
Korrekt	EtwasPersönliches stecke dahinter.		
Falsch	Und sein Verriss eines Bob-Dylan-Auftritts gehört zum garstigsten und spaßigsten, was man derzeit auf Deutsch lesen kann: …		
Korrekt	Und sein Verriss eines Bob-Dylan-Auftritts gehört zum Garstigsten und Spaßigsten, was man derzeit auf Deutsch lesen kann: …		

	Testsatz	DITECT	Word
Falsch	Sie ähneln sich nicht im entferntesten.		
Korrekt	Sie ähneln sich nicht im Entferntesten.		
Falsch	Es ist das beste, wenn du jetzt gehst.		
Korrekt	Es ist das Beste, wenn du jetzt gehst.		
Falsch	Sie tappt im dunkeln.		
Korrekt	Sie tappt im Dunkeln		
Falsch	Das kostet das dreifache		
Korrekt	Das kostet das Dreifache		
Falsch	Sie tut immer das richtige		
Korrekt	Sie tut immer das Richtige		
Falsch	Er zog den kürzeren.		
Korrekt	Er zog den Kürzeren.		
Falsch	Ich tue mein möglichstes.		
Korrekt	Ich tue mein Möglichstes.		
2.2 Nicht substantivierte Adjektive			
Falsch	Ich verkaufe nur gegen Bar.		
Korrekt	Ich verkaufe nur gegen bar.		
Falsch	Sie versucht es schon seit Längerem.	✓	
Korrekt	Sie versucht es schon seit längerem.		
Falsch	Das neue Auto ist nicht viel besser als das Alte.		
Korrekt	Das neue Auto ist nicht viel besser als das alte.		
Falsch	Dieser Berg ist am Steilsten.		
Korrekt	Dieser Berg ist am steilsten.		
2.3 Adjektivische Nationen- oder andere geografische Bezeichnungen			
Falsch	die Deutschen Schlagersänger		
Korrekt	die deutschen Schlagersänger		
Falsch	die Litauische Mafia		
Korrekt	die litauische Mafia		
Falsch	Mitarbeiter von 40 Hannoverschen Betrieben		
Korrekt	Mitarbeiter von 40 hannoverschen Betrieben		

	Testsatz	DITECT	Word
2.4 Adjektivische Numeralia, Sprach-, Farb- und Zahladjektive			
Falsch	Er ist der Erste von acht Atlasspinnern.		
Korrekt	Es ist der erste von acht Atlasspinnern.		
Falsch	Der Schulausschuss ist der Einzige, in dem die Grünen die Sitzungsleitung stellen.		
Korrekt	Der Schulausschuss ist der einzige, in dem die Grünen die Sitzungsleitung stellen.		
Falsch	Dies ist der dritte Dundee-Film, 13 Jahre nach dem Zweiten.		
Korrekt	Dies ist der dritte Dundee-Film, 13 Jahre nach dem zweiten.		
Falsch	Das Buch ist auf deutsch und auf französisch erhältlich.		
Korrekt	Das Buch ist auf Deutsch und auf Französisch erhältlich.		
Falsch	Die Ampel hat auf rot geschaltet.		
Korrekt	Die Ampel hat auf Rot geschaltet.		
Falsch	Sie hatte mit ihrer Vermutung ins schwarze getroffen.		
Korrekt	Sie hatte mit ihrer Vermutung ins Schwarze getroffen.		
Falsch	Er bekam für die Klassenarbeit eine sechs.		
Korrekt	Er bekam für die Klassenarbeit eine Sechs.		
Falsch	Die Grünen Abweichler entschieden sich anders.		
Korrekt	Die grünen Abweichler entschieden sich anders.		
2.5 Feste adjektivische Verbindungen			
Falsch	Im großen und ganzen ist das richtig.		
Korrekt	Im Großen und Ganzen ist das richtig.		
Falsch	Ein Fest für groß und klein.		
Korrekt	Ein Fest für Groß und Klein.		
Falsch	Gleich und gleich gesellt sich gern.		
Korrekt	Gleich und Gleich gesellt sich gern.		
Falsch	Er mag süßes und salziges.		
Korrekt	Er mag Süßes und Salziges.		
Falsch	Gemeinsam gehen wir durch Dick und Dünn.	✓	
Korrekt	Gemeinsam gehen wir durch dick und dünn.		
Falsch	Ich habe es Schwarz auf Weiß.		
Korrekt	Ich habe es schwarz auf weiß.		

	Testsatz	DITECT	Word
3.Schreibung nach Doppelpunkt und innerhalb von Klammern			
Falsch	Kampf gegen den internationalen Terrorismus: Ja.		
Korrekt	Kampf gegen den internationalen Terrorismus: ja.		
Falsch	An einer Seite des idyllischen Weihnachtsmarktes hat der Hamburger Lion's Club eine Ansammlung von Blechcontainern aufgestellt:Provisorische Verkaufsräume für mehr als 20 Händler.		
Korrekt	An einer Seite des idyllischen Weihnachtsmarktes hat der Hamburger Lion's Club eine Ansammlung von Blechcontainern aufgestellt:Provisorische Verkaufsräume für mehr als 20 Händler.		
Falsch	Eine erweiterte Ladenöffnung scheint auf der Zielgeraden zu sein: offenbar		
Korrekt	Eine erweiterte Ladenöffnung scheint auf der Zielgeraden zu sein: Offenbar sind sich fast alle Bundesländer schon einig.		
Falsch	Eine Untersuchung (Vollständige Ergebnisse in Finanztest, Heft Dezember 2001) vor allem für die, die in eine Krankenversicherung wechseln können …		
Korrekt	Eine Untersuchung (vollständige Ergebnisse in „Finanztest", Heft Dezember 2001) vor allem für die, die in eine [andere] Krankenversicherung wechseln können …		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
4 AKRONYME			
AIDS	Aids		
EXPO	Expo		
UNICEF	Unicef		
5 FESTE WENDUNGEN AUS ADJEKTIV UND SUBSTANTIV			
der technische Direktor	der Technische Direktor		
der parlamentarische Staatssekretär	der Parlamentarische Staatssekretär		
der heilige Vater	der Heilige Vater		
der regierende Bürgermeister	der Regierender Bürgermeister		
die französische Revolution	die Französische Revolution		
der westfälische Friede	der Westfälische Friede		
der dreißigjährige Krieg	der Dreißigjährige Krieg		
das viktorianische Zeitalter	das Viktorianische Zeitalter		
der Viktorianische Stil	der viktorianische Stil		
die Goldenen Zwanziger	die goldenen Zwanziger		
der Kalte Krieg	der kalte Krieg		
der zweite Weltkrieg	der Zweite Weltkrieg		
der stille Ozean	der Stille Ozean		
der gelbe Fluss	der Gelbe Fluss		

Falsche Schreibung	Korrekte Schreibung	DITECT	Word
der Blaue Planet	der blaue Planet		
die Dritte Welt	die dritte Welt		
der Große Teich	der große Teich		
der Rote Planet	der rote Planet		
der weiße Tod	der Weiße Tod		
die deutsche Bundesbahn	die Deutsche Bundesbahn		
Das Schwarze Brett	das schwarze Brett		
Erste Hilfe	erste Hilfe		
das fleißige Lieschen (Botanik)	das Fleißige Lieschen (Botanik)		
der schwarze Holunder	der Schwarze Holunder		
der rote Milan	der Rote Milan		
der Deutsche Schäferhund	der deutsche Schäferhund		
die Schwarze Johannisbeere	die schwarze Johannisbeere		
die gelbe Karte	die Gelbe Karte		
der Gelbe Sack	der gelbe Sack		
der Blaue Brief	der blaue Brief		
die Goetheschen Gedichte	die goetheschen Gedichte		
das Ohmsche Gesetz	das ohmsche Gesetz		

LITERATURVERZEICHNIS

Grammatik, Phonetik

Dudenredaktion: DUDEN. Grammatik der deutschen Gegenwartssprache. Mannheim, Leipzig, Wien, Zürich [4]1998.

Eisenberg, Peter [1]: Grundriss der deutschen Grammatik. Band 1: Das Wort. Stuttgart, Weimar 1998

Eisenberg, Peter [2]: Grundriss der deutschen Grammatik. Band 2: Der Satz. Stuttgart, Weimar 1999

Fleischer, Wolfgang, Barz, Irmhild: Wortbildung der deutschen Gegenwartssprache. Leipzig 1995[2]

Maas, Utz: Phonologie. Einführung in die funktionale Phonetik des Deutschen. Opladen 1999

Computerlinguistik

Beardon, Colin, Lumsden; David, Holmes, Geoff: Natural language and computational linguistics: An introduction. New York 1991.

Carstensen, Kai-Uwe, Ebert, Christian, Endriss, Cornelia, Jekat, Susanne, Klabunde, Ralf, Langer, Hagen (Hrsg.): Computerlinguistik und Sprachtechnologie. Eine Einführung. Heidelberg, Berlin 2001.

Damerau, Fred J.: A technique for computer detection and correction of spelling errors. In: Communications of the ACM, 7(3) 1964, 171-176. ↦ http://delivery.acm.org/10.1145/370000/363994/p171-damerau.pdf?key1=363994&key2=6069500401&coll=portal&dl=ACM&CFID=6294689&CFTOKEN=53876703

Emele, Martin Christoph: Der TFS-Repraesentationsformalismus und seine Anwendung in der maschinellen Sprachverarbeitung. Stuttgart 1997.

Jurafsky, Daniel, Martin, James H..: Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall 2000.

Kernighan, M. D., Church, K. W., Gale, W. A.: A spelling correction program based on a noisy channel model. In: COLING-90, Helsinki 1990, Vol. II, S. 205-211.

Kukich, Karen: Techniques for automatically correcting words in text. In: Communications of the ACM (Association for Computing Machinery), 24(4) 1992, 377-439. ↦ http://delivery.acm.org/10.1145/150000/146380/p377-kukich.pdf?key1=146380&key2=8478500401&coll=portal&dl=ACM&CFID=6294495&CFTOKEN=23029417

Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions and reversals. In: Cybernetics and Control Theory, 10(8), 1966, S. 707-710. Original in: Doklady Akademii Nauk SSSR 143(4), 1965, S. 845-848.).

Pennington, Martha C. (Hrsg.): Computers in applied linguistics: an international perspective.

Cleverdon 1992.
Peterson, James, L.: A note on undetected typing errors. In: Communications of the ACM 29(7) 1986, S. 633-637. ↦ http://delivery.acm.org/10.1145/10000/6146/p633-peterson.pdf?key1=6146&key2=0938500401&coll=portal&dl=ACM&CFID=6294495&CFTOKEN=23029417

Rechtschreibung

Arbeitsgruppe der deutschsprachigen Nachrichtenagenturen: Rechtschreibreform. Beschluss zur Umsetzung der Rechtschreibreform. Hamburg 21.06.1999. ↦ http://www.dpa.de/info/rechtschr/rs_index.htm
Dudenredaktion: DUDEN. Die Rechtschreibung. Mannheim, Leipzig, Wien, Zürich [4]1998.
Duden - Die neue Rechtschreibung. Mannheim. ↦ http://www.duden.de/
Heller, Klaus: Rechtschreibreform. Eine Zusammenfassung von Dr. Klaus Heller. IDS Sprachreport. Extraausgabe 1.07.1996: ↦ http://www.ids-mannheim.de/pub/sprachreport/reform/reform.html
Institut für deutsche Sprache, Mannheim: Deutsche Rechtschreibung. Regeln und Wörterverzeichnis. Amtliche Regelung. 14.09.2000. ↦ http://www.ids-mannheim.de/grammis/reform/regel.pdf
Neue Zürcher Zeitung: Rechtschreibreform in der NZZ. 15.05.2000. ↦ http://www.nzz.ch/dossiers/dossiers2000/rechtschreibung/

Korrekturprogramme

ARD Ratgeber Technik: Rechtschreibprüfungen - Was taugen Rechtschreibprüfungen am Computer? Rev.: 07.06.2001. ↦ http://www.ndrtv.de/ratgebertechnik/themen/rechtschreibung.html
Brinkmann, Sebastian: OpenOffice 1.0: Die Alternative zu MS Office.ZDNet Deutschland. Rev.: 24.05.2002. ↦ http://produkte.zdnet.de/test/77/14/1493.html
Greve, Georg C. F.: Brave GNU World. Linux Magazin Rev.: 13.11.2002. ↦ http://www.linux-magazin.de/Artikel/ausgabe/2002/12/bgw/bgw.html
Hexaglot Deutsch Korrekt 2000. Produktinformation: ↦ http://www.hexaglot.de/
Klaeren, Herbert: Stenokontorist-parteilos. Wenn Morgenstern ein Textprogramm gehabt hätte. Vortrag in der Marathon-Vorlesung. Tübingen 11.12.1997. ↦ http://www-pu.informatik.uni-tuebingen.de/users/klaeren/skp/vortrag.htm
Kortstock, Ulf: Rechtschreibkontrolle und Autokorrektur. Rev.: 12.4.2001 ↦ http://www.kortstock.de/Word/rechtsch.htm
Primus Korrektur-Manager. Produktinformation (SoftEx) Rev.: 05.07.2002: ↦ http://www.medienwerkstatt-online.de/products/primus/primus.html
Projekt Wörterbuch. TU Graz 15.03.2001. ↦ http://www.freedict-project.de/de/konzept.html
R.O.M. logicware Soft- und Hardware GmbH: Neuerungen in papyrus 8 “HyperOFFICE”. Berlin 18.10.2002. ↦ http://www.rom-logicware.com/neu_8.htm
Schlick, Andreas: AmigaWriter / Update 1.2 12.1998. ↦ http://www.amzeiger.de/sw/9812aw.htm

ALLE NETWORX-ARBEITEN IM ÜBERBLICK

Networx Einführung

Jens Runkehl, Peter Schlobinski & Torsten Siever
Sprache und Kommunikation im Internet (Hannover, 1998)
web**sprache** • medien**analyse**

Networx Nr. 1

Lena Falkenhagen & Svenja Landje
Newsgroups im Internet (Hannover: 1998)
web**sprache**

Networx Nr. 2

Gisela Hinrichs
Gesprächsanalyse Chatten (Hannover, 1997)
web**sprache** • medien**analyse**

Networx Nr. 3

Julian Hohmann
Web-Radios (Hannover, 1998)
web**sprache**

Networx Nr. 4

Silke Santer
Literatur im Internet (Hannover, 1998)
web**sprache**

Networx Nr. 5

Peter Schlobinski
Pseudonyme und Nicknames (Hannover, 1998)
web**sprache** • medien**analyse**

Networx Nr. 6

Jannis K. Androutsopoulos
Der Name @ (Heidelberg, 1999)
web**sprache**

Networx Nr. 7

Laszlo Farkas & Kitty Molnár
Gäste und ihre sprachlichen Spuren im Internet (Hannover, 1999)
web**sprache**

Networx Nr. 8

Peter Schlobinski & Michael Tewes
Graphentheoretisch fundierte Analyse von Hypertexten (Hannover, 1999)
web**sprache** • medien**analyse**

Networx Nr. 9

Barbara Tomczak & Cláudia Paulino
E-Zines (Hannover, 1999)
web**sprache**

Networx Nr. 10

Katja Eggers et al.
Wissenstransfer im Internet – drei Beispiele für neue wissenschaftliche Arbeitsmethoden (Hannover, 1999)
web**sprache** • medien**analyse**

Networx Nr. 11

Harald Buck
Kommunikation in elektronischen Diskussionsgruppen (Saarbrücken, 1999)
web**sprache**

Networx Nr. 12

Uwe Kalinowsky
Emotionstransport in textuellen Chats (Braunschweig, 1999)
web**sprache**

Networx Nr. 13

Christian Bachmann
Hyperfictions – Literatur der Zukunft? (Zürich, 1997)
web**sprache**

Networx Nr. 14

Peter Schlobinski
Anglizismen im Internet (Hannover, 2000)
web**sprache** • medien**analyse**

Networx Nr. 15

Marijana Soldo
Kommunikationstheorie und Internet (Hannover, 2000)
web**sprache** • medien**analyse**

Networx Nr. 16

Agnieszka Skrzypek
Werbung im Internet (Hannover, 2000)
web**sprache** • werbe**sprache**

Networx Nr. 17

Markus Kluba
Der Mensch im Netz. Auswirkungen und Stellenwert computervermittelter Kommunikation (Hannover, 2000)
web**sprache**

Networx Nr. 18

Heinz Rosenau
Die Interaktionswirklichkeit des IRC (Potsdam, 2001)
web**sprache**

Networx Nr. 19

Tim Schönefeld
Bedeutungskonstitution im Hypertext (Hamburg, 2001)
web**sprache** • medien**analyse**

Networx Nr. 20

Matthias Thome
Semiotische Aspekte computergebundener Kommunikation (Saarbrücken, 2001)
web**sprache** • medien**analyse**

Networx Nr. 21

Sabine Polotzek
Kommunikationssysteme Telefonat & Chat: Eine vergleichende Untersuchung (Dortmund, 2001)
web**sprache**

Networx Nr. 22

Peter Schlobinski et al.
Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten in der SMS-Kommunikation (Hannover, 2001)
web**sprache** • handy**sprache**

Networx Nr. 23

Andreas Herde
www.du-bist.net. Internetadressen im werblichen Wandel (Düsseldorf, 2001)
web**sprache** • werbe**sprache**

Networx Nr. 24

Brigitte Aschwanden
›Wär wot chätä?‹ Zum Sprachverhalten deutschschweizerischer Chatter (Zürich, 2001)
web**sprache** • medien**analyse**

Networx Nr. 25

Michaela Storp
Chatbots. Möglichkeiten und Grenzen der maschinellen Verarbeitung natürlicher Sprache (Hannover, 2002)
web**sprache** • werbe**sprache** • medien**analyse**

Networx Nr. 26

Markus Kluba
Massenmedien und Internet – eine systemtheoretische Perspektive (Hannover, 2002)
web**sprache** • medien**analyse**

Networx Nr. 27

Melanie Krause & Diana Schwitters
SMS-Kommunikation – Inhaltsanalyse eines kommunikativen Phänomens (Hannover, 2002)
handy**sprache**

Networx Nr. 28

Christa Dürscheid
SMS-Schreiben als Gegenstand der Sprachreflexion (Zürich, 2002)
handy**sprache**

Networx Nr. 29

Jennifer Bader
Schriftlichkeit & Mündlichkeit in der Chat-Kommunikation (Zürich, 2002)
web**sprache** • medien**analyse**

Networx Nr. 30

Olaf Krause
Fehleranalyse für das Hannoversche Tageblatt (Hannover, 2003)
medien**analyse**

Networx Nr. 31

Peter Schlobinski & Manabu Watanabe
SMS-Kommunikation – Deutsch/Japanisch kontrastiv. Eine explorative Studie (Hannover/Tokyo, 2003)
handy**sprache**

Networx Nr. 32

Matthias Wabner
Kreativer Umgang mit Sprache in der Werbung. Eine Analyse der Anzeigen- und Plakatwerbung von McDonald's (Regensburg, 2003)
werbe**sprache**

Networx Nr. 33

Steffen Ritter
Kohärenz in moderner, interaktiver und handlungsbasierter Unterhaltung. Die Textwelten von Adventures (Mannheim, 2003)
werbe**sprache**

Networx Nr. 34

Peter Schlobinski
Sprache und Denken ex machina? (Hannover, 2003)
werbe**sprache**

Networx Nr. 35

André Kramer
Rechtschreibkorrektursysteme im Vergleich. DITECT versus Microsoft Word (Hannover, 2003)
werbe**sprache** • medien**analyse**