

Funktionale Belastung des kontrastiven Wortakzents im Deutschen

Studienarbeit Nr. 61 im Fach Computerlinguistik

von

Steven Nehls

**Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Azenbergstraße 12
D-70174 Stuttgart**

Prüfer u. Betreuer: apl. Prof. Dr. phil. Bernd Möbius

Anmeldung der Arbeit: 14. Juni 2007
Abgabe der Arbeit: 10. September 2007

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Stuttgart, den 10. September 2007

(Steven Nehls)

Inhaltsverzeichnis

1	Einleitung	1
2	Akzent und Wortakzent	3
2.1	Allgemeine Definition	3
2.2	Akustische Korrelate des Wortakzents	3
2.3	Typologische Einteilung nach Wortakzent	4
2.4	Wortakzent im Deutschen	5
3	Funktionale Belastung	7
3.1	Das Konzept	7
3.2	Bedeutung der funktionalen Belastung	7
3.3	Unterschiedliche Methoden der Ermittlung des Wertes	8
4	Das Modell von Niyogi und Surendran	10
4.1	Objekte und Typen	10
4.2	Die Beschreibung von Kontrasten	11
4.3	Entropie	12
4.4	Die funktionale Belastung	13
5	Vorgehen	16
5.1	Implementierung des Modells	16
5.2	Erstellen einer Minimalpaarliste	18
5.3	Verwendete Korpora: Auswahl und Bearbeitung	19
5.3.1	CELEX	19
5.3.2	SmartKom-Unitselection-Korpus	20
5.3.3	Die CQP-Korpora	22
5.3.4	Abweichungen von Idealen	24
6	Ergebnisse und Evaluation	26
6.1	Ermittelte Werte	26
6.2	Theoretische Schlüsse	29
A	Appendix	31
A.1	Skript zur Ermittlung der funktionalen Belastung	31
A.2	Liste der Wortakzentminimalpaare aus CELEX	33
A.2.1	Minimalpaare aus dem Lemmalexikon	33
A.2.2	Ergänzende Minimalpaare aus dem Vollwortlexikon	37

1 Einleitung

Das Ziel dieser Arbeit ist die Bestimmung der funktionalen Belastung von kontrastivem Wortakzent im Deutschen. Ausgangspunkt ist dabei die Arbeit von Niyogi und Surendran, die ein Modell für die Berechnung der funktionalen Belastung von Phonemkontrasten vorstellt (Niyogi und Surendran, 2003). Die funktionale Belastung von Wortakzent in Sprachen wie dem Deutschen gilt im Allgemeinen als gering. Diese Arbeit soll dies auf ein statistisches Fundament stellen, indem die funktionale Belastung des Wortakzents im Deutschen mit anderen kontrastiven Phänomenen wie der Vokallänge oder Stimmhaftigkeit von Konsonanten verglichen wird. Als Basismaterial für die Berechnung dienten einige, am IMS zur Verfügung stehende Korpora. Diese sind vor allem CELEX (Baayen et al., 1995) und der SmartKom-Unitselection-Korpus (Schweitzer et al., 2003). Zusätzlich dazu sind auch zwei CQP-Korpora, die am IMS zur Verfügung stehen, benutzt worden.

Die Arbeit kann in zwei Teile unterteilt werden: Der erste Teil besteht aus den Abschnitten zwei und drei, in denen die theoretischen Grundlagen, nämlich die Beschreibung der Begriffe des Wortakzents und der funktionalen Belastung, vorgestellt werden. Zuerst wird der Begriff des Akzents bzw. Wortakzents näher beschrieben.¹ Dabei wird auf verschiedene sprachtypologische Einteilungen von Verschiedenen Sprachen nach der Funktion von Wortakzent eingegangen. Besonders soll dabei die Stellung des Deutschen in dieser Einteilung und die Eigenschaften des deutschen Wortakzents beschreiben werden.

Danach folgt eine Beschreibung des Konzepts der funktionalen Belastung von der ersten Vorstellung im Prager Linguistenzirkel bis zu der Arbeit von Niyogi und Surendran. Dabei werden auch einige der Ziele beschrieben, die sich die Vertreter des Konzepts von der Ermittlung der funktionalen Belastung erhoffen.

Der Zweite Teil besteht aus den Abschnitten vier, fünf und sechs. Abschnitt vier stellt die technischen Grundlagen, das Modell von Niyogi und Surendran genauer vor. In diesem Abschnitt wird, unter dem Punkt 4.3, auch der Begriff der Entropie, so wie er bei Shannon (Shannon, 1948) in der Informationstheorie beschrieben wird, vorgestellt, da dieser Begriff für das Modell von Niyogi und Surendran wichtig ist. Der fünfte Abschnitt beschreibt die Arbeitsschritte, die unternommen worden sind, um das Modell von Niyogi und Surendran zu implementieren. Für die Umsetzung des Modells wurde Perl verwendet. Die wichtigen Formeln aus dem Modell werden mit Code-Auszügen belegt und die Arbeitsweise des Perlskripts näher erklärt. Einige wenige Grundkenntnisse über das Programmieren sind dabei für das Verständnis der Erklärungen vorausgesetzt. Ein Teil der Aufgaben war auch das Finden

¹Der Wortakzent ist ein Unterbegriff des allgemeinen Begriffs Akzent, neben dem Satzakzent. Um Verwechslungen vorzubeugen, soll hier schon angemerkt sein dass Akzent eher im Sinne von Betonung gemeint ist, als z.B. so etwas wie Akzent wie er bei nicht Muttersprachlern einer Sprache zu hören ist.

von Minimalpaaren aus CELEX. Die Funktionsweise dieses Skripts wird im Unterabschnitt 5.2 beschrieben. Zuletzt wird noch eine Beschreibung der Korpora, die für die Berechnung benutzt worden sind, gegeben. Dieser Unterabschnitt ist relativ ausführlich, da die Formatierung der Daten, in eine, für das Berechnungsskript, benutzbare Form den Hauptarbeitsaufwand ausgemacht hat. Im letzten Abschnitt werden die Ergebnisse dieser Arbeit, das sind die Werte der funktionalen Belastungen des kontrastiven Wortakzents im Deutschen, im Vergleich mit einigen anderen kontrastiven Features der deutschen Phonologie, vorgestellt und diskutiert. Der Anhang enthält das Skript, das zur Berechnung der funktionalen Belastung geschrieben wurde und eine Liste von Wortakzentminimalpaaren, die aus der Phonologielemmaliste aus CELEX ermittelt wurde.

Allgemein werden in dieser Arbeit Umschriften in IPA-Notation gegeben. Bei Auszügen aus den Korpora, speziell CELEX, wird die in dem jeweiligen Korpus verwendete Transkription verwendet. Beispiele für Wortakzentminimalpaare werden in der üblichen Orthographie gegeben, wobei die betonte, d.h. die akzentuierte Silbe fett gedruckt ist. Phoneme bzw. Umschriften in IPA- oder SAMPA-Notation außerhalb von Korpusauszügen als Beispielen stehen zwischen zwei Schrägstrichen (z.B. /p/). Orthographische Elemente sind in eckige Klammern (z.B. <p>) gesetzt.

2 Akzent und Wortakzent

2.1 Allgemeine Definition

Der Akzent bezeichnet eine suprasegmentale Eigenschaft in der Phonologie und beschreibt die Hervorhebung einer linguistischen Einheit, gegenüber den ihr benachbarten Einheiten. In der englischen Literatur wird der Akzent auch mit „stress“ oder „accent“ bezeichnet. Im Deutschen ist statt „Akzent“ auch oft „Betonung“ zu lesen. In dieser Arbeit, wird vorwiegend der Begriff „Akzent“ benutzt. Beide Begriffe sollen hier austauschbar sein. Wiese (Wiese, 1996, S. 272) versteht unter Akzentuierung (stress):

... patterns of speech [...] such that some parts of a linguistic unit are judged to be more prominent than other parts in the same unit.

Es handelt sich also um eine Prominenzrelation bei der ein Teil einer Einheit akustisch gegenüber den anderen hervorgehoben wird. Die kleinste Einheit, die einen Akzent tragen kann, ist die Silbe. Damit aber diese Prominenzrelation innerhalb einer dieser linguistischen Einheiten bestehen kann, muss diese Einheit aus mindestens zwei Silben bestehen (siehe Lehiste, 1970, Kap. 4.5.2). Im Allgemeinen gibt es zwei Domänen für Akzentuierung: das Wort und den Satz. Ist eine Silbe in einem Wort hervorgehoben, spricht man von Wortakzent. Ist ein Wort in einem Satz hervorgehoben spricht man von Satzakzent. Die vorliegende Arbeit wird sich im Weiteren allerdings nur mit dem Wortakzent beschäftigen. Wenn also die Rede von Akzentposition ist, ist damit immer die akzentuierte Silbe im Wort gemeint. Eigentlich ist die Domäne des Wortakzents das phonologische Wort, das nicht immer mit orthographischen Wörtern übereinstimmt. So können längere Komposita aus mehreren phonologischen Wörtern bestehen (vgl. hierzu Wiese, 1996).

Da es sich beim Akzent um eine Prominenzrelation handelt, gibt es verschiedene Grade der Akzentuierung, meistens Haupt-, Neben- und gar keinen Akzent. Die einzelnen Akzentuierungsgrade sollen hier keine so wichtige Rolle spielen, es soll im Weiteren also nur zwischen Hauptakzent und keinem bzw. allen anderen Graden unterschieden werden. In der phonetischen Umschrift IPA, wird die hauptakzentuierte Silbe mit einem *ˈ* gekennzeichnet, welches am Anfang der akzentuierten Silbe oder vor dem akzentuierten Vokal steht.

2.2 Akustische Korrelate des Wortakzents

Nach Ladefoged können Definitionen von Betonung (stress) aus zwei verschiedenen Perspektiven gegeben werden: der des Sprechers und der des Hörers der Äußerung. Am einfachsten ist die Definition aus der Perspektive des Sprechers, nämlich

dass eine betonte Silbe mit größerem respiratorischem Aufwand produziert wird (Ladefoged, 2001, S. 112):

A stressed syllable is produced by pushing more air out of the lungs in one syllable relative to others. A stressed syllable thus has greater respiratory energy than neighbouring unstressed syllables. [...] Stress can always be defined in terms of what a speaker does in one part of an utterance relative to another.

Nach Lehiste (Lehiste, 1970), machen sich akzentuierte Silben dem Hörer dadurch bemerkbar, dass sie vom Sprecher mit einem höheren Aufwand produziert werden. Es können drei akustische Merkmale für betonte Silben angegeben werden:

- Die Veränderung der Grundfrequenz gegenüber den benachbarten Silben. Meistens wird die Grundfrequenz erhöht.
- Erhöhung der Intensität gegenüber den benachbarten Silben.
- Erhöhung der Vokaldauer in der akzentuierten Silbe.

Als wichtigstes Merkmal gilt im Allgemeinen die Erhöhung der Grundfrequenz. Nach Lehiste gehen erhöhte Grundfrequenz und erhöhte Intensität miteinander einher, da sowohl die Erhöhung der Lautstärke als auch die Erhöhung der Frequenz durch eine Erhöhung des subglottalen Drucks erreicht werden. Die Erhöhung der Intensität scheint also ein untergeordneter Cue für die Erkennung von akzentuierten Silben zu sein. Im Deutschen ist die Erhöhung der Vokallänge allerdings kein akzentgebendes Mittel, da unterschiedliche Vokallänge, bei einem ansonsten gleichen Wort schon bedeutungsunterscheidend sein kann (Wängler, 1974). Als Viertes kann noch die Vokalqualität angegeben werden, da es Vokale gibt, die nur in unakzentuierten Silben zu finden sind, wie z.B. /ə/ und /ɐ/ im Deutschen. In den wenigsten Sprachen scheint es dabei der Fall zu sein, dass nur eines dieser Merkmale, für einen Hörer den Ausschlag gibt, eine Silbe als akzentuiert zu erkennen. Es ist viel mehr eine Mischung der oben genannten Merkmale, die dabei auch Sprachabhängig zu sein scheint.

2.3 Typologische Einteilung nach Wortakzent

Allgemein werden Sprachen je nach Positionierung des Wortakzents in zwei Gruppen geteilt: in Sprachen mit freiem und Sprachen mit gebundenem Akzent. Meistens gilt, dass Sprachen, bei denen der Akzent immer auf eine bestimmte Silbe im Wort fällt, Sprachen mit gebundenem, festem oder delimitativem Akzent genannt

werden. Die Akzentposition ist bei ihnen abhängig von der Wortgrenze. Der Wortakzent dient in diesen Sprachen dem Hörer als Anhaltspunkt für Wortgrenzen im Redestrom. Seine Funktion ist delimitativ (vgl. dazu Trubetzkoy, 1939; Wängler, 1974; Clark et al., 2007). Zu den Sprachen mit gebundenem Wortakzent gehört z.B. Polnisch, wo der Akzent immer auf der vorletzten Silbe liegt oder Französisch, wo der Akzent auf die letzte Silbe fällt (vgl. dazu Lehiste, 1970). Im Finnischen, Lettischen und Tschechischen ist ebenfalls die letzte Silbe akzentuiert (Wängler, 1974). In diesen Sprachen gibt es keine Wortpaare, die sich nur durch die Akzentposition im Wort unterscheiden.

Sprachen der zweiten Gruppe werden im Allgemeinen als Sprachen mit freiem oder phonemischem Wortakzent bezeichnet (Kohler, 1977; Lehiste, 1970). In ihnen fällt der Wortakzent nicht auf eine fest abzählbare Silbe und kann somit wortunterscheidend sein. Als Beispiel für eine dieser Sprachen wird vor allem Russisch genannt (Kohler, 1977; Lehiste, 1970), wo der Akzent grammatikalische Unterschiede in verschiedenen Formen des selben Wortes kennzeichnen kann (Wängler, 1974). Lehiste bemerkt aber dazu, dass der Wortakzent in diesen Sprachen trotzdem nur eine geringe funktionelle Belastung hat, da es nur wenige Wortpaare gibt, die sich nur durch die Akzentposition unterscheiden (Lehiste, 1970).

2.4 Wortakzent im Deutschen

Bei Clark et al. (2007) wird Deutsch zu den Sprachen mit freiem Akzent gezählt, da es im Deutschen Wortpaare gibt, die sich nur durch die Position der akzentuierten Silbe unterscheiden. Nach Lehiste und Wängler dagegen zählt das Deutsche zu einem Zwischentypen, zwischen freiem und gebundenem Akzent. Es gibt zwar Wortpaare, die sich nur durch die Akzentposition unterscheiden, doch ist der Akzent fest an bestimmte Morpheme gebunden, weswegen das Deutsche als Sprache mit morphologischem Akzent bezeichnet wird (Lehiste, 1970; Wängler, 1974). In Komposita, die aus mehreren Nomen zusammengesetzt sind, gibt die Akzentposition Aufschluss über die Struktur des Kompositums. Es gibt also prinzipiell Minimalpaare bei längeren Komposita, die unterschiedliche Bedeutung haben, je nachdem wo der Akzent liegt.

Es gibt außerdem im Deutschen eine ganze Reihe von Verben, die sich nur durch die Akzentposition im Wort unterscheiden. Die Akzentposition entscheidet dabei ob es sich bei dem betreffenden Verb um ein Partikelverb oder ein Präfixverb handelt. Liegt der Hauptakzent auf der Partikel, ist diese abtrennbar und es handelt sich um ein Partikelverb. Liegt der Hauptakzent dagegen auf dem Stamm, ist die Partikel nicht abtrennbar und verhält sich wie ein Präfix. Es handelt sich dann um ein Präfix- oder Partikelpräfixverb. Die betreffenden Partikel sind „durch-“, „hinter-“, „um-“, „über-“, „unter-“ und „wider-“. Die Akzentposition entscheidet also bei diesen Verben darüber, ob die Partikel abtrennbar ist oder nicht.

Einige Beispiele dafür sind **umbauen** vs. **umbauen**, **übersetzen** vs. **übersetzen** oder **übergehen** vs. **übergehen**. Es gibt aber auch Paare, bei denen der Wortakzent darüber entscheidet, ob es sich um ein Verb oder Nomen handelt, wie z.B. bei **Widerstand** vs. **widerstand**. Nach Wiese können Partikelverben mit abtrennbarer Partikel, bei denen der Hauptakzent also auf der Partikel liegt, als „minimale Phrasen irgendeiner Art“ (engl. „minimal phrases of some sort“) (Wiese, 1996, S. 295 f.) betrachtet werden, weswegen sich die Akzentuierung von den Partikelverben mit unabtrennbarer Partikel unterscheidet. Wiese gibt daher die Funktion des Akzents im Deutschen als Unterscheidungsmerkmal für verschiedene Ebenen an, denen eine geäußerte Konstituente angehört (vgl. Wiese, 1996, S. 311), also ob es sich bei der Äußerung nun um ein Wort, ein Kompositum oder eine Phrase handelt. Als Beispiel gibt Wiese dafür „die kranken **Schwestern**“ vs. „die **Krankenschwestern**“ an. Im ersten Fall, mit dem Hauptakzent auf „**Schwestern**“, handelt es sich um Phrasenakzent, also das Vorhandensein einer Phrase. Im zweiten Fall, mit dem Hauptakzent auf „**kranken**“, handelt es sich um ein Kompositum. Der Akzent zeigt also, nach Wiese, die Grenzen von Konstituenten an und hat somit auch eine demarkative Funktion im Deutschen, ähnlich der in den Sprachen mit gebundenem Akzent.

3 Funktionale Belastung

3.1 Das Konzept

Das Konzept der funktionalen Belastung ist unter Phonologen oft (unter den unterschiedlichsten Bezeichnungen) diskutiert und beschrieben worden. Es geht auf den Prager Linguistenzirkel in den dreißiger Jahren, zurück. Eine exakte Definition des Begriffes, scheint es bei ihnen allerdings nicht zu geben. Mathesius nennt sie z.B. nur „*Grad der Ausnützung*“ von phonologischen Einheiten und macht keine weiteren Beschreibungen (Mathesius, 1931). Trubetzkoy beschreibt die funktionale Belastung ebenfalls nicht näher (Trubetzkoy, 1939). Es geht nur aus dem Kontext hervor, dass es sich um einen statistischen Wert handeln soll. Martinet beschreibt sie als „*Wichtigkeit einer Opposition für die kommunikativen Bedürfnisse*“ (Martinet, 1981, Kap. 2.21). Eine etwas anschaulichere Beschreibung, worum es bei funktionaler Belastung geht, gibt (Hockett, 1966, S.300):

The function of a phonemic System is to keep utterances of a language apart. Some contrasts between phonemes in a system apparently do more of this job than others.

Auch wenn es keine exakte Definition bei den unterschiedlichen Autoren gibt, so ist die Idee, die dahinter steht, bei allen gleich. Oft, vor allem anfänglich, wurde dabei nur an binäre Phonemoppositionen gedacht. Aber das Konzept ist vielleicht am sinnvollsten, wenn man es eher auf phonetische Merkmale, die wortunterscheidend sein können, ausdehnt.

3.2 Bedeutung der funktionalen Belastung

Für den Prager Zirkel gehört die funktionale Belastung zur vollständigen strukturellen Beschreibung der Phonologie einer Sprache. Durch die funktionale Belastung soll, neben der qualitativen Beschreibung einer Sprache, auch die quantitative Beschreibung ihren Platz finden (Mathesius, 1931). Die funktionale Belastung soll also zur Vollständigkeit der strukturellen Beschreibung einer Sprache dienen. Hockett spekuliert, neben der strukturellen Vollständigkeit der Beschreibung, auch darüber, ob die funktionale Belastung bestimmte Phänomene voraussagen bzw. erklären kann, wie z.B. das Nichtvorhandensein mancher Sprachlaute, die von Sprechern einer Sprache als notwendig erachtet werden (Hockett, 1955).

Hockett war der Auffassung, dass die funktionale Belastung später in der automatischen Spracherkennung, Anwendung finden könnte. Meyerstein führt noch viele andere, vermutete Anwendungsgebiete für das Konzept der funktionalen Belastung an: von Anwendungen bei der Analyse von Gedichten über Textklassifikation bis zur Autorenidentifikation (Meyerstein, 1970). Diese ganzen Vorschläge

zeigen die weite Beschäftigung in der Phonologie bzw. in der gesamten Linguistik, mit diesem Thema.

Martinets Annahme, dass die funktionale Belastung ein wichtiger Faktor im Sprachwandel ist, genauer eigentlich im Lautwandel, soll hier besonders Erwähnung finden, da sie oft in der Literatur wieder aufgegriffen wurde. Die Vermutung Martinets war, dass Phonemoppositionen, die eine hohe funktionale Belastung aufweisen, dem Lautwandel weniger ausgesetzt sind, als Phonemoppositionen, die eine niedrigere funktionale Belastung haben (vgl. dazu Martinet, 1981). Damit ist gemeint, dass Phonemoppositionen mit einer ausreichend hohen Belastung nicht im Laufe der Zeit wegfallen. Die funktionale Belastung würde damit zu den internen Faktoren im Sprachwandel zählen.²

Diese Vermutung wurde stark diskutiert und King, der seine Arbeit (King, 1967) sogar nur dieser Hypothese gewidmet hat, hat eher Hinweise für das Gegenteil gefunden und bezweifelt daher, dass die funktionale Belastung ein wichtiger Faktor für den Sprachwandel ist.³ Seine Arbeit ist dabei aber auch nicht ohne Kritik geblieben. Martinet und Hockett bezweifeln, dass Kings Ergebnisse denn überhaupt zwangsläufig auch zu diesen Schlussfolgerungen führen müssten, da ohnehin so gut wie nichts über Lautwandel bekannt sei und Kings dafür herangezogene Texte, etwa 20.000 Phoneme lang, zu klein sind (vgl. dazu Martinet, 1981; Hockett, 1966).

Auf diese Debatte soll hier aber nicht weiter eingegangen werden, da das Gebiet des Sprachwandels einfach zu komplex ist und über die Gründe von Lautwandel bislang nur spekuliert werden kann. Ob funktionale Belastung beim Sprachwandel eine Rolle spielt und welche, können eigentlich nur ausführliche und weitreichende Studien zeigen. Am sinnvollsten scheint die funktionale Belastung als vollständige Beschreibung der phonologischen Struktur einer Sprache. Wenn funktionale Belastung keinen Faktor im Sprachwandel darstellen sollte, so sollte sich doch wenigstens das Fakt des Wandels überhaupt in der phonologischen Statistik (=der funktionalen Belastungsverteilung) bemerkbar machen oder ankündigen.

3.3 Unterschiedliche Methoden der Ermittlung des Wertes

Es gibt, in der Geschichte des Begriffs der funktionalen Belastung, verschiedene Vorschläge für dessen Ermittlung, jedoch gibt es bis heute keine allgemein gültige Methode. Schon Trubetzkoy weist auf die bestehende Problematik hin (Trubetzkoy, 1939, Kap. 7, S. 230 f.). Er schlägt vor, die entsprechenden statistischen Häufigkeiten aus einem Wörterbuch einfach zu zählen oder sie mit den Zahlen, die aufgrund von Kombinationsregeln erwartet werden, zu vergleichen. Die Methode der ein-

²Im Gegensatz zu externen Faktoren wie Einflüsse anderer Sprachen oder Modeerscheinungen.

³Die Arbeit von Niyogi und Surendran scheint dabei auf ein ähnliches Ergebnis hinzuweisen (vgl. Niyogi und Surendran 2003 Kap.9).

fachen Zählung der Minimalpaare findet oft Verwendung in der Literatur (z.B. in Lehiste, 1970, Kap.4). Diese Methode ist zwar einfach, doch wurde vermehrt darauf hingewiesen, dass sie den einzelnen Häufigkeiten der Wörter in einer Sprache nicht Rechnung trägt.

Die Vokallänge im Deutschen, macht eine nur doppelt so hohe Anzahl an Minimalpaaren (etwa 227) wie der Wortakzent (etwa 110) aus, aber die einzelnen Worte haben eine erheblich höhere Vorkommenshäufigkeit. Einige Autoren haben, aus oben genannten Gründen, entropiebasierte Methoden, die auf Shannons Informationstheorie beruhen, vorgeschlagen, da die einfache Minimalpaarzählung die Worthäufigkeiten der einzelnen Worte nicht berücksichtigt. Der wichtigste Vorschlag stammt von Hockett (Hockett, 1955, 1966), auf den auch Niyogi und Surendran zurückgreifen. Hockett hat allerdings keinerlei Berechnung selbst durchgeführt. Ein weiterer Artikel ist dabei noch sehr wichtig: (Wang, 1967). Er vergleicht darin mehrere Ansätze für die Berechnung. Hocketts Methode schneidet dabei am besten ab.

Hier soll auch die Arbeit von Carter (Carter, 1987) erwähnt werden, obwohl sie aus einem anderen Gebiet, der automatischen Spracherkennung, stammt. Sie soll aber trotzdem erwähnt werden, da auch sie eine der Grundlagen für das Modell von Niyogi und Surendran darstellt. Carters Arbeit beschäftigt sich mit der Evaluierung von Spracherkennung, die nur unvollständige Daten wie z.B. nur Artikulationsart oder -ort zur Verfügung hat. Er hat dafür ein ganz ähnliches Modell wie das Hocketts, seiner Arbeit zugrunde gelegt.

Seit den siebziger Jahren allerdings findet der Begriff der funktionalen Belastung in der Literatur kaum noch Erwähnung. Erst Niyogi und Surendran beschäftigen sich nach dieser Zeit wieder damit und stellen somit die neueste nennenswerte Arbeit, die sich ausschließlich mit diesem Thema beschäftigt dar. Sie beschreibt sehr genau eine Methode zur Berechnung und wird im folgenden Abschnitt näher erklärt, da diese Arbeit der Fokuspunkt dieser vorliegenden Arbeit darstellt. Außerdem beschäftigt sich ihre Arbeit mit der Berechnung einiger phonologischer Merkmale und der Robustheit der so ermittelten Werte, und versucht einige Anwendungen in der Phonologie. Niyogis und Surendrans Arbeit stellt somit eine Möglichkeit zur Schaffung empirischer Grundlagen in der phonologischen Statistik zur Verfügung.

4 Das Modell von Niyogi und Surendran

In diesem Teil der Arbeit wird das Modell von Niyogi und Surendran für die Berechnung der funktionalen Belastung vorgestellt. Das Modell ermöglicht es, genauere Aussagen über die Wichtigkeit von Kontrasten bzw. phonemischen Merkmale für eine Sprache zu machen. Es ist somit ein wichtiges und nützliches Werkzeug für die statistische Phonologie.

Die Erläuterungen geben nur einen Überblick über die Funktionsweise des Modells und des danach geschriebenen Perlskriptes, das im Abschnitt 5.1 erklärt wird. Die Details können in Niyogis und Surendrans Arbeit (Niyogi und Surendran, 2003, Abschnitt 3) nachgeschlagen werden. Alle in diesem Abschnitt gegebenen Erläuterungen sind, sofern keine anderen Quellen angegeben sind, in der Arbeit unter Abschnitt 3 zu finden. Der Unterabschnitt über Entropie ist hauptsächlich aus Shannon (1948), sofern keine anderen Quellen angegeben sind.

4.1 Objekte und Typen

Sehr wichtig für das Modell, ist das Konzept von Objekttypen oder einfach Objekten (Niyogi und Surendran 2003 Abschnitt 3.1). Eine Sprache, L , wird als eine Folge, L_T , von Objekten des Typs T verstanden. Da es sich um eine Methode für die Untersuchung von phonologischen Phänomenen handelt, d.h. meistens kontrastiver, phonetischer Merkmale, können diese Typen eben Phoneme, Silben oder Wörter sein. Phoneme stellen einen atomaren Typen dar.

Jedes Objekt hat einen Wert. Zwei Objekte sind genau dann gleich, wenn sie den gleichen Wert haben. Also im Falle eines Phonems ist ein Objekt mit dem Wert /p/ dann gleich einem anderen Objekt, wenn es auch den Wert /p/ hat. Komplexere Objekte sind dabei aus atomaren Objekten zusammengesetzt. So sind z.B. Silben Objekte, die aus zwei Komponenten bestehen: die erste ist eine Folge von Objekten des Typs „Phonem“ und die zweite besteht aus dem Typ Akzentuierung (engl. stress), welche hier nur die Werte „akzentuiert“ und „nicht akzentuiert“ haben kann. Silben haben dann genau den gleichen Wert, wenn die Folge von Phonemen gleich ist und sie den gleichen Akzentuierungsgrad haben. Wörter letztlich bestehen aus zwei Komponenten, nämlich der ersten, einer Folge von Silben, und der zweiten, nämlich einer Bedeutung (engl. meaning), welche aber vernachlässigt werden kann, da es sich um eine phonologische Berechnungsmethode handelt. Es wären auch andere Typen, die weniger phonologisch begründet sind, vorstellbar.

Für die vorliegende Arbeit wurde lediglich der Typ „Wort“ benutzt, in einer etwas vereinfachten Form. D.h. dass die Wörter nicht, wie es die formale Beschreibung verlangt aus zweikomponentigen Silben mit einer Folge von Phonemen als erster und einer zweiten Komponente „Betonung“ zusammengesetzt ist. Vielmehr wird das Wort einfach durch seine SAMPA-Transkription wiedergegeben, also nur

mit einer Komponente. Das Zeichen /' / steht am Silbenanfang, wenn die Silbe betont ist, wie am Beispiel für das Wort „umdrehen“ /'ʊm-dre:ɪ-ən/ zu sehen ist.

4.2 Die Beschreibung von Kontrasten

Ein Kontrast c in der Sprache L wird in dem Modell als die Äquivalenzrelation⁴ „gleich in Abwesenheit von c “⁵ beschrieben. Je nach Objekttyp und seinen möglichen Werten, beschreibt ein Kontrast eine Partition⁶ der Menge der verschiedenen möglichen Werte des Objekttyps. Die einzigen Äquivalenzklassen⁷ darin, die mehr als ein Element enthalten, sind die, deren Werte gleich wären, wenn es den Kontrast nicht gäbe. Diese Partition bzw. nur die Mengen, die mehr als ein Element enthalten, also in denen die kontrastierenden Objekte enthalten sind, werden mit θ beschrieben.

Ein Beispiel dafür wäre der Kontrast zwischen stimmhaften und nicht stimmhaften Phonemen, also $\theta_{\text{Stimmhaftigkeit}}$. In jeder Teilmenge würde nur ein Phonem stehen, wie etwa $\{a\}$ oder $\{m\}$. Die einzigen Äquivalenzklassen mit mehr als einem Element wären die, deren Elemente sich nur durch Stimmhaftigkeit unterscheiden, wie z.B. $\{p, b\}$, $\{t, d\}$, $\{k, g\}$, $\{s, z\}$ usw. . .

Wenn ein Kontrast aus einer Sprache L_T verschwindet, dann wird daraus die Sprache L_{T_θ} , was wiederum eine Folge von Objekten des Typs T_θ ist. T_θ ist im Wesentlichen der gleiche Objekttyp, nur enthalten seine möglichen Werte den Kontrast θ nicht mehr. T_θ hat genau die gleiche Komponentenstruktur, seine Werte aber sind Äquivalenzklassen in θ . θ wird mit einer Funktion $g_{T,\theta} : \Phi_T \rightarrow \theta$ gleichgesetzt, die Werte von T auf die Äquivalenzklassen von θ abbildet. Das heißt, dass die Werte der Objekte die gleichen wie bei „ T “ sind, mit dem Unterschied, dass die Äquivalenzklassen, die in T mehr als ein Element enthalten, in „ T_θ “ nur noch ein Element enthalten. Die Funktion, die also aus L_T L_{T_θ} macht, arbeitet Objekt für Objekt ab und ersetzt jeden Wert von Φ_T , durch den entsprechenden von Φ_{T_θ} . Das heißt, dass im Falle eines Kontrastpaares beide Werte nun auf einen Wert abgebildet werden, so dass das Paar nicht mehr zu unterscheiden ist. Im Beispiel oben würde das heißen, dass /a/ auf /a/ abgebildet würde, aber z.B. /b/ von /p/ nicht mehr zu unterscheiden wäre. Beide würden auf ein anderes Zeichen abgebildet werden.⁸

Wichtig ist, dass Kontraste zwischen atomaren Typen sich auf komplexere Ty-

⁴Ein Äquivalenzrelation ist symmetrisch, transitiv und reflexiv. Zwei Elemente stehen in einer Äquivalenzrelation, wenn sie gleichwertig sind.

⁵engl.: „Equal in the absence of c “ (Niyogi and Surendran, 2003, Abschnitt 3.2).

⁶Eine Partition ist die Aufteilung einer Menge in Untermengen, so dass kein Element in mehr als einer dieser Teilmengen enthalten ist.

⁷Das sind die Teilmengen der Partition.

⁸Da der Zeichenvorrat der noch nicht in SAMPA benutzten Symbole begrenzt ist, ist es effizienter /b/ und /p/ beide entweder auf /b/ oder auf /p/ abzubilden.

pen, die aus den entsprechenden atomaren zusammengesetzt sind, vererben. Es ist logisch, dass Kontraste zwischen Phonemen eben nicht nur diese Phoneme, sondern eben auch Silben und und Wörter unterscheiden können. Daher ist es nicht unbedingt notwendig, bei einem Phonemkontrast die Sprache als Folge von „ $T = \text{Phonem}$ “ zu repräsentieren. Sie kann statt dessen auch als Folge von „ $T = \text{Wort}$ “ oder „ $T = \text{Silbe}$ “ repräsentiert werden.

4.3 Entropie

Da der Begriff der Entropie für das Modell von Niyogi und Surendran, sehr wichtig ist, soll das Konzept der Entropie hier kurz näher beschrieben werden. Der Begriff Entropie stammt aus der Informationstheorie, die Shannon begründet hat (Shannon, 1948, 1951).

Eine Sprache kann, nach Shannon, als ein statistischer Prozess repräsentiert werden, der Nachrichten aus einer endlichen Menge von „Zeichen“ erzeugt. Jedes der „Zeichen“, wird einzeln nacheinander nach einer bestimmten Wahrscheinlichkeit⁹ erzeugt. Die Wahrscheinlichkeit, für die Erzeugung eines Zeichens, hängt nicht nur von dem Zeichen selbst ab, sondern auch von seinen $n - 1$ Vorgängern. Shannon nennt das die Annäherung der n -ten Ordnung an die Sprache (Shannon, 1948, S.5 ff). Annäherung deshalb, da n unendlich groß sein müsste, um exakt der Sprache zu entsprechen. Ein Prozess mit $n = 0$ generiert dabei alle Zeichen mit der gleichen Wahrscheinlichkeit, also unabhängig von der Sprache, ein Prozess mit $n = 1$ dann die Zeichen in der Wahrscheinlichkeit, in der sie in der Sprache auftreten. $n = 2$ ist die Wahrscheinlichkeit, mit der zwei Zeichen hintereinander auftreten, also das Zeichen a dem Zeichen b folgt. Das wird dann mit der Wahrscheinlichkeit von Bigrammen dargestellt. Für $n = 3$ ergibt das dann Trigramme.

Entropie stellt ein Maß für die Information dar, die ein Zeichen¹⁰ durchschnittlich in einer Sprache produziert (Shannon, 1951). Eigentlich ist Entropie ein Maß für die durchschnittliche Unsicherheit darüber, welches Zeichen als nächstes folgt, wenn man die vorausgehenden $n - 1$ Zeichen bereits kennt. Diese Unsicherheit wiederum stellt ein Maß für den Grad an Auswahl, welches Zeichen als nächstes produziert werden kann, dar. Information entsteht durch die Beseitigung dieser Unsicherheit. Dieses Maß wird allgemein mit H abgekürzt. Die Gleichung zur Errechnung von H ist:

⁹Das bedeutet, dass jedes Zeichen eine feste, ihm zu eigene Auftretenswahrscheinlichkeit hat. Die Wahrscheinlichkeit dieses Zeichens ändert sich in der Sprache bzw. dem Prozess nicht. Das bedeutet, die Wahrscheinlichkeitsverteilung ist statisch.

¹⁰Diese Zeichen entsprechen in Niyogi und Surendrans Modell den T-Objekten.

$$H = \frac{1}{n} \sum_{i \in A} p_i \log_2 p_i \quad (1)$$

Das n aus $\frac{1}{n}$ steht dabei für den Grad der Annäherung an die Sprache. In dieser Arbeit spielt allerdings nur die Annäherung der ersten Ordnung an die Sprache eine Rolle. Das heißt, dass nur die Wahrscheinlichkeitsverteilung einzelner Zeichen ohne Kenntnis oder Beeinflussung der Wahrscheinlichkeit von vorangehenden Zeichen herangezogen wird. Es gilt also $n = 1$ und die etwas kürzere Formel:

$$H = \sum_{i \in A} p_i \log_2 p_i \quad (2)$$

p_i ist die Wahrscheinlichkeit eines Zeichens. A ist ein Alphabet oder eine Menge von Zeichen. Für $n > 1$ ist p_i dann die Wahrscheinlichkeit einer bestimmten n -Gramm-Folge für alle n -Gramme, die in der Nachricht erscheinen. Die Basis des Logarithmus gibt die Einheit der Information an. Normalerweise ist das die 2, was dann Bit an Information ergibt. H ergibt dann die durchschnittliche Anzahl an Bits pro Zeichen, die mindestens erforderlich ist um es in der Sprache zu verschlüsseln. Die Basis ist allerdings nicht ausschlaggebend für die Formel.¹¹ Die Formel summiert also nur die Wahrscheinlichkeit mal des binären Logarithmus der Wahrscheinlichkeit jedes Zeichens bzw. n -Gramms von Zeichen. Man kann sagen, dass so jedes Zeichen seinen Beitrag zu der Gesamtentropie leistet.

Die wichtigsten Eigenschaften der Entropie H im Bezug auf die Berechnungen, die in dieser Arbeit gemacht werden, sind, dass H genau dann 0 ist, wenn alle p_i (die Wahrscheinlichkeit eines der Symbole) bis auf eines, die Wahrscheinlichkeit 0 haben und eines die Wahrscheinlichkeit 1. Die Unsicherheit über das Zeichen, das als nächstes folgt, wenn die $n - 1$ vorangegangenen Zeichen bekannt sind, ist gleich null, wenn es nur ein Zeichen gibt, das übertragen werden kann. H ist am höchsten wenn alle Symbole gleich wahrscheinlich auftreten. Also bei N verschiedenen Zeichen jedes genau die Wahrscheinlichkeit $\frac{1}{N}$ hat. Außerdem erhöht jede Veränderung in Richtung Angleichung der Wahrscheinlichkeiten H .

4.4 Die funktionale Belastung

Niyogi und Surendrans Modell lehnt sich bei der Berechnung der funktionalen Belastung an die Vorschläge von Hockett (Hockett, 1955, 1966) und an die Arbeit von

¹¹Die Basis stellt nur einen konstanten Faktor k dar: $H = k \sum_{i \in A} p_i \log_2 p_i$. Will man die Basis 2 nach 10 umrechnen ist k etwa $3\frac{1}{3}$ (vgl. Shannon, 1948, S.1). Für das gesamte Modell ist die Basis noch weniger von Bedeutung, da sich k jeweils im Zähler und Nenner von Formel 3 kürzt.

Carter an (Carter, 1987). Carters Arbeit hat sich nicht direkt mit der Berechnung von funktionaler Belastung beschäftigt, sondern mit der Berechnung der Effizienz von Spracherkennern bei unvollständigen Daten, was dem Problem der funktionalen Belastung sehr ähnlich ist. Er schlägt dafür einen informationstheoretischen Ansatz vor, wie Hockett für die Berechnung von funktionaler Belastung.¹² Niyogi und Surendrans Formel für die funktionale Belastung (FL) ist, davon abgeleitet, nun:

$$FL_T(\theta) = \frac{H(L_T) - H(L_{T_\theta})}{H(L_T)} \quad (3)$$

Im Zähler steht nur der Entropieverlust der Sprache, im Falle des Verschwindens des Kontrastes θ . $H(L_T) - H(L_{T_\theta})$ ist das, was an Entropie verloren geht wenn der Kontrast θ aus der Sprache L verschwindet. L_{T_θ} stellt also die Sprache ohne den zu untersuchenden Kontrast dar. Die funktionale Belastung ist nun also das Verhältnis zwischen der Entropie, die verloren geht, wenn der Kontrast aus der Sprache verschwindet, und der Entropie der unveränderten Sprache (siehe Hockett, 1955, S.217).

Eine Sprache ist, wie weiter oben beschrieben, eine Folge L_T von Objekten des Typs T , die auch als statistischer Prozess aufgefasst werden kann, der diese Sprache erzeugt. Dieser Prozess hat eine Wahrscheinlichkeitsverteilung, in der jedes Objekt bzw. jedes n -Gramm von Objekten eine eigene Vorkommenswahrscheinlichkeit hat. Somit hat dieser Prozess auch eine Entropie $H(L_T)$. Einige weitere Parameter kommen für die endgültige Fassung der Formel noch hinzu. Diese beschreiben im wesentlichen $L_{T,n}$. Mit allen Parametern hat die Formel in ihrer endgültigen Fassung, die folgende Form:

$$FL_{T,n}(\theta; S) = \frac{H(L_{T,n}; S) - H(L_{T_{\theta,n}}; g_{T,\theta}(S))}{H(L_{T,n}; S)} \quad (4)$$

Diese Formel stellt laut Niyogi und Surendran (Niyogi und Surendran, 2003 S.8) die Annäherung n -ter Ordnung an die funktionale Belastung eines Kontrastes θ dar, entsprechend der Größe n weiter oben unter 4.3 beschrieben. Die Bedeutung der verschiedenen Parameter für das Modell und die Wahl ihrer Werte für alle Berechnungen in dieser Arbeit, ist im Folgenden:

- L ist die Sprache, die untersucht werden soll, also eine Folge von T -Objekten. Da nur das Deutsche Thema dieser Arbeit ist steht L hier immer für Deutsch.

¹²Eine detailliertere Beschreibung davon ist in Niyogi und Surendran(2003 Abschnitt 2.2) nachlesbar.

- n steht für die Ordnung der Annäherung an die Sprache, wie in Abschnitt 4.3 weiter oben schon erklärt wurde. Für die gesamte vorliegende Arbeit wurde $n = 1$ gewählt, d.h. dass nur die Wahrscheinlichkeitsverteilung von Unigrammen für die Entropieberechnung herangezogen wurde.
- T ist der Typ von T -Objekten, aus denen sich der Korpus bzw. die Häufigkeitsliste zusammensetzt, wie in Abschnitt 4.1 bereits erklärt, wurde. T ist hier für alle Berechnung als $T = Wort$ gewählt worden.
- S ist der Korpus, der für die Berechnung der funktionalen Belastung benutzt wurde und stellvertretend für die Sprache L steht. Wenn z.B. mit CELEX gearbeitet wurde, gilt $S = CELEX$.
- θ ist der Kontrast bzw. die Partition der Menge aller möglichen Werte Φ_T , die den Kontrast repräsentiert, und dessen funktionale Belastung errechnet werden soll. In Abschnitt 4.2 wurde das bereits genauer beschrieben.
- $g_{T,\theta}(S)$ ist die Funktion, die wie oben unter 4.2 erläutert, die Objekte des Typs T der Sprache L in die Objekte des Typs T_θ umwandelt und somit aus L die Sprache L_θ macht. In dieser Arbeit wurde diese Funktion als eine Reihe von regulären Ausdrücken umgesetzt.

Die Parameter, die das Ergebnis am meisten beeinflussen sind n , S und T . n und T verändern direkt die Wahrscheinlichkeitsverteilung und damit die Entropie, da sich durch größere n die Menge an Wahrscheinlichkeiten in der Wahrscheinlichkeitsverteilung erhöht. Größere n fangen mehr Information ein als niedrige. Von T kann das selbe gesagt werden: $T = Phonem$ hat weniger mögliche Werte und somit weniger Ereignisse in der Wahrscheinlichkeitsverteilung, als $T = Wort$. Die funktionalen Belastungswerte sind höher bei $T = Phonem$ als bei $T = Wort$. Unterschiedliche Korpora verhalten sich nicht eindeutig voraussagbar, was von ihrem Aufbau abhängt. Allerdings zeigen Niyogi und Surendran, dass eine recht hohe Korrelation zwischen den funktionalen Belastungswerten der unterschiedlichen Korpora besteht. Auch bei unterschiedlichen n ist die Korrelation sehr hoch.¹³ Um die Vergleichbarkeit der Werte in dieser Arbeit zu bewahren, wurde n und T bei allen Berechnungen gleich gewählt. Es wurden lediglich verschiedene Korpora für die Berechnung herangezogen.

¹³Für Genaueres über die Bestimmung der Korrelation (siehe Niyogi und Surendran, 2003 Abschnitt 6)

5 Vorgehen

Das Vorgehen in der Arbeit bestand aus zwei Teilen: Der erste, kürzere Teil bestand darin, eine Liste von Minimalpaaren zu erstellen, die sich nur durch die Position des Wortakzents unterscheiden. Der zweite, längere Teil bestand darin, das Modell, das Niyogi und Surendran vorschlagen, zu implementieren und die funktionale Belastung des deutschen Wortakzents festzustellen. Das soll alles in dem folgenden Abschnitt genauer erläutert werden. Außerdem wird die Auswahl und Bearbeitung der für die Berechnung benutzten Korpora genauer beschrieben sowie die Arbeitsschritte, um die Korpora in eine für die Skripte gut verwendbare Form zu bringen.

5.1 Implementierung des Modells

Die Skriptumsetzung des Modells von Niyogi und Surendran wurde in Perl geschrieben und ist im Anhang A.1 nachschlagbar. Die Formeln aus Abschnitt 4 werden mit Auszügen aus dem Programm belegt. Das Skript erfordert eine Liste mit Wörtern und ihrer jeweiligen Häufigkeit¹⁴ als Eingabe, die als Übergabeparameter beim Starten des Programms mit angegeben wird. Der Programmaufruf sieht dann schematisch etwa so aus:

```
$perl comp_FL [WORTHÄUFIGKEITSLISTE]
```

Die Worthäufigkeitsliste muss dabei folgendes Format haben:

```
[Wort1]\[Häufigkeit1]
[Wort2]\[Häufigkeit2]
usw...
```

Jedes Wort muss in einer eigenen Zeile stehen, gefolgt von seiner Häufigkeit in dem Korpus, beide durch einen Backslash (\) getrennt. Die Wortliste wird Zeile für Zeile eingelesen. Die Worte und Häufigkeiten werden dabei in zwei Hashes¹⁵ gespeichert, mit dem Wort als Schlüssel und der Häufigkeit als Wert. In den ersten Hash wird das unveränderte Wort als Schlüssel gespeichert. Danach wird der Kontrast, dessen funktionale Belastung untersucht werden soll, mit einem oder mehreren regulären Ausdrücken aus dem jeweiligen Wort entfernt:

```
$fldr [0] =~ s / ' //;
```

¹⁴Die Worthäufigkeitsliste entspricht der Wahrscheinlichkeitsverteilung unter 4.4. Wenn der gewählte Typ nicht $T = Wort$ ist und/oder $n \neq 1$ muss die Liste anstelle der Worte eben n -Gramme des Typs T mit ihren jeweiligen Häufigkeiten enthalten.

¹⁵Perl stellt schon eine fertige Datenstruktur für den Hash zur Verfügung, so dass der Datentyp lediglich als Hash deklariert werden muss und keine eigene Hashfunktion erstellt werden muss, wie in anderen Programmiersprachen.

„*\$fldr[0]*“ enthält der Reihe nach jedes Wort aus der Worthäufigkeitsliste. Das „s“ im Ausdruck „s’/“ vor den Slashes zeigt an, dass der reguläre Ausdruck zwischen den zwei ersten Slashes mit dem Ausdruck zwischen dem zweiten und letzten Slash ersetzt werden soll. In diesem Fall wird das Akzentzeichen /’/ gelöscht, da zwischen dem zweiten und dritten Slash nichts steht. Die regulären Ausdrücke entsprechen dabei dem Ausdruck „*g_{T,θ}(S)*“ aus Formel 4, der die Funktion darstellt, die die Sprache L_T in L_{T_θ} umwandelt. Für die anderen Kontraste, die mit dem Wortakzent verglichen werden sollten, waren teilweise mehrere hintereinander gestellte Ausdrücke nötig. Das so veränderte Wort wird dabei direkt wieder in „*\$fldr[0]*“ gespeichert und dann als Schlüssel für den zweiten Hash, der die veränderte Sprache enthält, genutzt. Bei der Zuweisung von Schlüssel und Wert wird dabei genau darauf geachtet, dass im Falle, dass ein Schlüssel schon vorhanden ist, der Wert lediglich addiert wird und nicht neu zugewiesen.

Die Entropie wird danach sowohl für die Liste mit dem Kontrast als auch für die Liste ohne den Kontrast berechnet. Die betreffenden Zeilen im Programm für Formel (1) $H = \sum_{i \in A} p_i \log_2 p_i$, die Entropieberechnung, stehen in einer eigenen Subroutine und sehen so aus:

```
foreach $wert ( values %{ $shlst } ) {
    if ( $wert != 0 ) {
        $zwischen = ( $wert / $gesamtwrd ) * ( log ( $wert
            / $gesamtwrd ) / log ( 2 ) );
        $sum = $sum + $zwischen ; }
};
```

„*#{ \$shlst }*“ ist die Dereferenzierung des Hashes, da der Hash an eine Subroutine übergeben wird und Perl die Übergabe von Hashes nur als Referenz erlaubt, die später wieder dereferenziert werden muss. „*values*“ weist der Variablen „*\$wert*“ jeweils einen Wert aus dem übergebenen Hash, also der Worthäufigkeitsliste, zu. „*\$shlst*“ entspricht also dem Subskript A unter dem Summenzeichen. „*\$wert*“ ist dabei die jeweilige Häufigkeit des Wortes und entspricht dem i in den Subskripten unter dem Summenzeichen und p . p , die Vorkommenswahrscheinlichkeit, setzt sich dabei aus der Häufigkeit des jeweiligen Wortes, geteilt durch die Anzahl aller Wörter, „*\$gesamtwrd*“. „ $\log(\$wert/\$gesamtwrd)/\log(2)$ “ ist eine etwas umständliche Umrechnung der Basis des Logarithmus, da Perl nur die Basis e ¹⁶ für den Logarithmus zur Verfügung stellt. Die Basis 2 erhält man indem man den $\log_e p_i$ nochmal durch $\log_e 2$ teilt.

Die Formel $FL_{T,n}(\theta) = \frac{H(L_{T,n}) - H(L_{T_\theta,n})}{H(L_{T,n})}$ hat in Perl dann folgende Form:

```
$f1 = ( $lang1 - $lang2 ) / $lang1 ;
```

¹⁶Die Eulersche Zahl. Dieser Logarithmus wird dann natürlicher Logarithmus genannt. e hat etwa den Wert 2,718.

wobei “\$lang1“ $H(L_{T,n}; S)$ die Entropie der untersuchten Sprache mit dem Kontrast, ist und “\$lang2“ die Entropie der untersuchten Sprache ohne Kontrast ist. Das Ergebnis der Division wird in der Variablen \$fl gespeichert und später als Ergebnis ausgegeben.

Die Ausführung des Skriptes, die Berechnung der funktionalen Belastung, dauert in der Regel selbst bei größeren Korpora wie CELEX nicht länger als einige Sekunden.

5.2 Erstellen einer Minimalpaarliste

Für das das möglichst restlose Auffinden der Wortminimalpaare, wurde CELEX als Korpus herangezogen. Eine Liste mit Minimalpaaren für kontrastiven Wortakzent, aus dem Lemmalexikon von CELEX, ist in A.2 angehängt. Das Skript, für das Finden der Minimalpaare ist sehr kurz und müsste aus der folgenden Beschreibung einfach nachzuschreiben sein. Deshalb werden hier keine Auszüge aus dem Skript gezeigt, um die einzelnen Schritte zu veranschaulichen.

CELEX enthält verschiedene Informationen über die Phonologie jedes Wortes, wie z.B. Silbenstruktur und SAMPA-Transkription. Eine Beschreibung von CELEX mit der genauen Informationsstruktur für jeden Eintrag ist im Abschnitt 5.3.1 zu finden. Die Spalten, mit den für das Erstellen der Liste uninteressanten Informationen wie etwa Informationen über die Silbenstruktur, werden gelöscht und nur die Spalten mit den wichtigen Informationen, also nur Orthographie, SAMPA-Transkription und Vorkommenshäufigkeit, werden beibehalten. Sie werden in eine neue Liste gespeichert und sortiert. Das SAMPA-Feld eines jeden Wortes wird dabei kopiert und mittels eines regulären Ausdrucks der entsprechende zu untersuchende Kontrast entfernt. Im Falle des Wortakzents wird dann einfach das Akzentsymbol entfernt. Die daraus resultierende Liste hat dann etwa folgende Gestalt:

```
ap-bIl-d@n\ 'ap-bIl-d@n\abbilden\14
ap-bIlt\ 'ap-bIlt\Abbild\21
ap-bIl-dUN\ 'ap-bIl-dUN\Abbildung\85
usw...
```

Wichtig für die richtige Sortierung ist, dass sich in dieser Liste die Spalte mit der veränderten SAMPA-Transkription ganz vorne befindet, gefolgt von der unveränderten SAMPA-Spalte. Diese Liste wird dann sortiert und danach werden die jeweils direkt benachbarten Zeilen miteinander verglichen. Wenn die jeweiligen Spalten, die den Kontrast nicht enthalten, gleich sind, aber die jeweiligen Spalten, die den Kontrast noch enthalten, ungleich sind, werden beide Zeilen auf die Konsole ausgegeben. Das Finden der Minimalpaare dauert, selbst bei einer langen Liste, für gewöhnlich nur einige Sekunden.

Die Anzahl der Wortpaare, die sich nur durch die Akzentposition unterscheiden und die aus dem Lemmalexikon gewonnen wurden, liegt bei 171. Für das Vollwortlexikon liegt die Anzahl bei 2223 Wortpaaren. Wobei zu sagen bleibt, dass viele Aussprachevarianten ohne tatsächliche Bedeutungsunterscheidungen in der Liste enthalten sind. Die reale Anzahl liegt dadurch natürlich niedriger, wobei auch zu beachten ist, dass Wortpaare wie z.B. **Tenor** vs. **Tenor**, **Erlangen** vs. **erlangen** in CELEX nicht enthalten sind. Auf Wortpaare wie **damit** und **damit** oder **dazu** und **dazu** wurde hier nicht eingegangen, da das den Rahmen der Arbeit gesprengt hätte und CELEX keine unterschiedlichen Einträge für diese Wörter hat.

5.3 Verwendete Korpora: Auswahl und Bearbeitung

5.3.1 CELEX

CELEX (Baayen et al., 1995) ist ein dreisprachiger Korpus, der aus Englisch, Niederländisch und Deutsch besteht. Hier wurde die CD-ROM-Version benutzt. CELEX wurde als Referenz herangezogen, weil es nicht nur die einzelnen Wörter mit Häufigkeitsangaben enthält, sondern zusätzlich noch phonologische Informationen wie z.B. die Silbenstruktur und eine eigene phonologische Umschrift. CELEX ist in zwei Teile in den jeweiligen Sprachen unterteilt, den Lemma- und den Vollwortteil. Diese sind wiederum in Phonologie, Morphologie, Syntax, Orthographie und Frequenz unterteilt. In dieser Arbeit wurde nur der jeweilige Phonologieteil verwendet.

Der Lemmateil enthält allgemeine Informationen über die Lemmata in dem Korpus. Die Liste mit den Wortakzentminimalpaaren im Anhang wurde aus dem Lemmateil extrahiert. Der Vollwortteil, der zur Berechnung der funktionalen Belastung benutzt wurde, enthält die Vollformen der Wörter. Die Struktur eines Eintrages in diesem Teil ist etwas anders als die im Lemmateil, da es noch ein Extrafeld gibt, das die Lemmanummer enthält, unter der das Wort im Lemmateil gefunden werden kann.

Jedes Wort hat einen Eintrag in einer eigenen Zeile, mit verschiedenen Informationen zu diesem Wort, in verschiedenen Feldern angehängt. Die Felder enthalten Informationen über das Lemma, dem das Wort angehört, eine SAMPA-Transkription des Wortes sowie Informationen über die Vorkommenshäufigkeit des Wortes im Korpus. Die einzelnen Informationsfelder sind dabei in jeweils durch einen Backslash (\) getrennt. Der deutsche Vollwortteil des Korpus enthält etwa 365.000 Einträge. Hier ist ein Beispieleintrag aus dem Vollwortteil:

```
11462\dankt\35\8654\'daNkt\[daNkt]\[CVCCC]
11463\dankte\104\8654\'daNk-t@[daNk][t@]\[CVCC][CV]
11464\danktet\0\8654\'daNk-t@[daNk][t@t]\[CVCC][CVC]
usw...
```

Die Felder sind folgendermaßen zusammengestellt: ganz links steht die „Id-Num“, was der Zeilennummer entspricht. Danach folgt das Wort in seiner orthographischen Umschrift, dann die Häufigkeit, die „IdNum“ des Lemmas im entsprechenden Lemmateil und die SAMPA-Transkription des Wortes. Danach das Wort in CPA-Transkription, das ist eine CELEX-eigene Umschrift, die SAMPA recht ähnlich ist, hier aber ignoriert wurde, da schon ein Feld mit einer SAMPA-Transkription, die vor allem Wortakzentinformation enthält, da ist. In der CPA-Transkription steht jede Silbe in eckigen Klammern. Danach folgt, ganz rechts, eine Transkription der Silbenstruktur, bei der jeder Vokal mit „V“ und jeder Konsonant mit „C“, bezeichnet wird.

Da CELEX schon phonologische Informationen enthält, musste dieser Korpus kaum bearbeitet werden. Die Perl-Skripte müssen lediglich die nicht gebrauchten Informationen löschen und eine Liste erstellen, die dann aus obigem Beispiel etwa folgendes erstellt:

```
'daNkt\35
'daNk-t@\104
'daNk-t@t\0
usw. . .
```

Es bleiben also nur die SAMPA-Transkription gefolgt von der Häufigkeit des Wortes in CELEX übrig. „[WORT][HÄUFIGKEIT]“, ist die Standardform, in die auch die anderen Korpora gebracht wurden, da die Skripte¹⁷ eine solche Struktur als Standardformat für die Eingabe von Worthäufigkeitslisten erwarten.

5.3.2 SmartKom-Unitselection-Korpus

Der Unitselection-Korpus (Schweitzer et al., 2003) besteht aus 17.379 Wörtern, aus 2.586 Äußerungen. Ein Teil davon stammt aus dem TAZ-Zeitungskorpus. Der Rest besteht aus SmartKom-domänenspezifischen Äußerungen. Alle Äußerungen wurden von einem professionellen Sprecher vorgelesen und aufgenommen.

Zu jeder Äußerung, die in einer Sounddatei enthalten ist, gibt es eine Reihe verschiedener Labelfiles, die etwa Informationen über Tonhöhenverlauf, die Phoneme oder die Wörter enthalten, die in der Aufnahme der Reihe nach auftreten. Für die Zusammenstellung einer Worthäufigkeitsliste aus diesem Korpus wurden nur die Labelfiles mit den orthographischen Wörtern (.words) und den Phonemen in SAMPA-Transkription (.sylstructure) benutzt. Die Struktur einer Labeldatei hat beispielsweise folgende Form:

¹⁷Sowohl das Berechnungsskript für die funktionale Belastung als auch das Skript für das Finden von Minimalpaaren.

```
0.720000 121 e: " |
0.760000 121 l
0.800000 121 @|
0.920000 121 f
usw...
```

Das obige Beispiel ist ein Auszug aus einer „sylstructure“-Datei. In jeder Zeile stehen die Informationen für ein Label. Die Zeile kann in drei Spalten, mit jeweils unterschiedlicher Information geteilt werden. Jede Spalte ist durch ein Leerzeichen getrennt. In der letzten Spalte steht das Label, in diesem Falle das Phonem, das geäußert wurde. In der ersten Spalte steht der Zeitpunkt, an dem das Phonem in der Aufnahme erscheint. Die mittlere Spalte enthält lediglich einen Farbcode für die Farbe, in der das Label angezeigt wird.

Das Perlskript, das die Worthäufigkeitsliste aus diesen Labelfiles zusammensetzt, arbeitet in folgenden Schritten: Alle „words“-Dateien werden nacheinander geöffnet. Die Zeitindizes für jedes Wort, bei dem der Index vor dem Wort für den Endzeitpunkt des Wortes steht und der Index des Wortes der Zeile darüber, für den Startzeitpunkt, werden genutzt, um die entsprechende „sylstructure“-Datei nach den Labels, die zu dem entsprechenden Wort gehören, zu durchsuchen. Die Labels, die dann zwischen Start- und Endindex stehen, werden an das dazugehörige Wort unter Abtrennung mit einem Backslash (\) angefügt. So erhält jedes Wort in dem Korpus eine SAMPA-Transkription.

Alle Satzzeichen sowie spezielle Annotationen für Pausen in der Rede, wurden weggelassen. Andere Konventionen mussten umgewandelt werden. So wurde z.B. das SAMPA-Symbol für den Hauptakzent /‘/ durch das in CELEX benutzte /’¹⁸ ersetzt, das Zeichen, das das Ende einer Silbe markiert, //, durch /-/ ersetzt und das // am Wortende gelöscht. Das Ergebnis sieht folgendermaßen aus:

```
den\ 'de:n
Mythos\ 'my:-tOs
theatralisch\te-at-'Ra:-lIS
usw...
```

Eigentlich ist die Information über die Orthographie der Wörter für das Skript, das die funktionale Belastung errechnet, nicht nötig. Sie wurde aber benötigt für den letzten Arbeitsschritt, in dem Abweichungen des realisierten Akzents von der Standardakzentuierung im Deutschen „korrigiert“ und vereinheitlicht wurden, da sonst die Anzahl an Wortakzentminimalpaaren nach oben verfälscht worden wäre. Als Beispiel wurde die Realisierung **Zukunft** nach **Zukunft** korrigiert und bloße

¹⁸Diese Änderung wurde lediglich Vorgenommen, um die Notation etwas an CELEX anzugleichen, da in CELEX mit /‘/ der Nebenakzent markiert wird.

Aussprachevarianten, wie **dazu** und **dazu**, vereinheitlicht. Wörter, die keinen Akzent aufweisen, bekamen einen Standardakzent. Wörter, bei denen die Veränderung der Akzentuierung die Bedeutung des Wortes ändert, wurden dabei ausgelassen.¹⁹

Danach wurde aus dem Korpus eine Worthäufigkeitsliste erstellt. Da die Arbeit unter einem Linuxsystem gemacht wurde, genügte dafür ein Befehl der Form „sort [KORPUS] | uniq -c > [WORTHÄUFIGKEITSLISTE]“. Da dieser Befehl zuerst die Häufigkeit schreibt und dann, durch einen Tab getrennt, das Wort, wurde noch ein Skript benutzt, welches das Wort mit der Häufigkeit vertauscht und beide durch einen Backslash, wie in CELEX, trennt. Das Orthographiefeld wurde dabei weggelassen, so dass die Worthäufigkeitsliste nur die SAMPA-Transkription enthält.

5.3.3 Die CQP-Korpora

Um den Anforderungen, die Trubetzkoy (vgl. Trubetzkoy, 1939, S. 230) schon an Korpora gestellt hatte, gerecht zu werden, nämlich möglichst repräsentativ für die gesprochene Sprache zu sein, wurden aus CQP der Stuttgarter-Zeitung-Korpus und der Bundestagsdebattenkorpus ausgewählt.

Die Stuttgarter Zeitung wurde deswegen ausgewählt, weil eine Zeitung von dem Themenspektrum, das sie abdeckt, sehr vielseitig ist und somit der gesprochenen Sprache, in diesem Aspekt, relativ nahe kommt. Für gewöhnlich halten sich Zeitungsartikel allerdings an einen besonderen Schreibstil, wie z.B. möglichst objektiv zu sein, oder aber die Benutzung nur bestimmter Tempusformen, welche in der gesprochenen Sprache eher selten zu finden sind. Auch enthalten Zeitungen oft wortwörtliche Wiederholungen von ganzen Sätzen, was ebenfalls in der gesprochenen Sprache eher selten auftritt.

Als zweites wurde der Bundestagsdebattenkorpus ausgewählt, da es sich dabei um wirklich gesprochene Sprache handelt. Die Unterschiede zur normalen Umgangssprache sind aber dennoch groß, da man bei solchen Debatten für gewöhnlich davon ausgehen kann, dass die Redner rhetorisch gebildet sind (oder zumindest ausgebildete Redenschreiber haben), was nicht auf den normalen Durchschnittssprecher einer Sprache gilt. Außerdem ist das Themenspektrum stark begrenzt — es handelt sich nur um politisch relevante Themen und keineswegs alltägliche Themen (wie z.B. das Wetter).

Die Arbeitsschritte an den beiden Korpora waren die selben, da sie im selben Format vorlagen. CQP erlaubt es, Suchergebnisse in Dateien zu speichern. Es wurde einfach nach jedem Eintrag gesucht²⁰ und das Ergebnis in eine Datei gespeichert. Die Einstellungen in CQP für die Ergebnisanzeige wurden so geändert, dass

¹⁹Es gibt in diesem Korpus eigentlich nur zwei Wortpaare die sich durch Akzentposition unterscheiden: „Kaffe“ vs. „Cafe“ und „Erlangen“ vs. „erlangen“

²⁰Der entsprechende Suchausdruck dafür ist „*.*“.

nur das entsprechende gefundene Wort pro Zeile ohne Kontext²¹ und nur mit seinem entsprechenden POS-Tag angezeigt wurde. In jeder Zeile der so gespeicherten Korpusdatei stand nur ein Wort und dahinter sein entsprechendes POS-Tag. Der gesamte Eintrag ist dabei in spitze Klammern (<>) eingeschlossen. Das Wort und das POS-Tag sind durch einen Slash (/) voneinander getrennt.

```
<für/APPR>
<großes/ADJA>
<Aufsehen/NN>
<./IPNORM>
usw...
```

Wie man dem obigen Ausschnitt sehen kann, erhalten Satz- und Sonderzeichen ebenfalls eigene Einträge.

Da die CQP-Korpora keine phonologischen, oder gar phonetischen Informationen enthalten, wurden die einzelnen Wörter mit Hilfe eines Perlskripts mit phonologischen Informationen aus CELEX versehen. Dabei wurde in dem Skript aus CELEX jede Orthographieinformation als Schlüssel in einen Hash gespeichert, mit der dazugehörigen SAMPA-Transkription als Wert. Danach wurde der extrahierte CQP Korpus eingelesen. Alle Wörter die nicht als „NN“ oder „NE“ getaggt waren, wurden kleingeschrieben, um Wörter, die zufällig nach einem Punkt stehen und daher großgeschrieben sind, nicht für die CELEX-Schlüssel in dem Hash mit einem Eintrag für Nomen, die in CELEX großgeschrieben sind, verwechselbar zu machen. <ß>, <ä>, <ö> und <ü> wurden durch <ss>, <ae>, <oe> und <ue> ersetzt um die Wörter der Schreibweise in CELEX anzugleichen. Danach wurde jedes Wort durch seine SAMPA-Transkription aus dem zuvor erzeugten Hash ersetzt. Da nicht jedes Wort, das in den CQP-Korpora erscheint, auch einen Eintrag in CELEX hat, blieben dadurch einige Wörter in der Orthographieform bestehen. Die POS-Tags und die spitzen Klammern wurden danach gelöscht. Hier folgt Beispielauszug aus einem der Korpora, nach den oben erläuterten Arbeitsschritten:

```
'de:r
kommunalpolitiker
'kan
'an
'de:n
usw...
```

Wie man sehen kann, steht in jeder Zeile nur noch ein Wort, in den meisten Fällen nur in SAMPA-Transkription. Die einzigen Wörter, die nicht transkribiert

²¹D.h. Der Kontext wurde auf 0 gestellt. Standardvoreinstellung ist ein Kontext von fünf Wörtern vor dem Treffer und nach dem Treffer.

werden konnten, waren meistens längere Komposita, Zahlen, Namen und Abkürzungen. Um die Zahl der Komposita gering zu halten, wurden längere Komposita mit Bindestrich (-) an dem Bindestrich getrennt und jeder Teil in eine neue Zeile geschrieben, so dass die einzelnen Bestandteile, in den meisten Fällen, durch ihre SAMPA-Transkription ersetzt werden konnten.

Der aufwändigste Arbeitsschritt war nun, die Wortpaare, die sich nur durch die Position ihres Wortakzents unterscheiden, aus dem Kontext in dem sie stehen, eindeutig einer Aussprache zuzuordnen und das Akzentzeichen /' / an der entsprechenden Stelle zu setzen. Dazu wurde die Minimalpaarliste, die, wie unter 5.2 beschrieben, aus CELEX gewonnen wurde, benutzt.

Um den Aufwand möglichst gering zu halten, wurden die Korpora auf jeweils etwa zwei Millionen Einträge beschränkt und es sind verallgemeinernde Annahmen über die Betonung der Wortpaare aus der Liste gemacht worden. Bei Wortpaaren, die in der Auflistung im Anhang A.2 gesternt sind, wurde pauschal nur eine Variante gewählt. Bei dem Wortpaar „**durch**suchen“ vs. „durch**s**uchen“ wurde, sofern das Wort Polizei im selben Satz als Agens erschien, immer die zweite Variante gewählt. Das Paar „**um**fassen“ vs. „um**f**assen“ z.B. wurde immer als „umfassen“ interpretiert, da das einzige sinngebende Vorkommen von „umfassen“, nur auf einen Ring oder etwas anderes, das eine Fassung aufweist, bezogen werden konnte und ein solches Wort nie im Kontext des Wortes erschien.

Danach wurde daraus, wie oben bei dem Unitselectionkorpus, eine Worthäufigkeitsliste erstellt. Nachdem alle Satzzeichen aus der Liste gelöscht wurden, hatte der Stuttgarter-Zeitung-Korpus einen Umfang von etwa 1,8 Millionen Wörtern und der Bundestagsdebattenkorpus einen Umfang von etwa 1,6 Millionen Wörtern.

5.3.4 Abweichungen von Idealen

Alle Korpora und die Veränderungen, die an ihnen vorgenommen wurden, sind nur Kompromisse zwischen schaffbarer Arbeit und möglichst genauen Daten und Informationen in den Korpora. Von seinem Format her, ist der SmartKom-Korpus am nächsten an einem Ideal. Er besteht aus gesprochenen und vorgelesenen Aufnahmen. Das heißt, dass man Daten hat, die wirklich so geäußert worden sind. Allerdings ist der eigentliche Verwendungszweck, nämlich die Sprachsynthese, auch seine größte Abweichung von einem Idealkorpus, für statistische Informationen über eine Sprache: er ist viel zu wenig umfangreich mit etwa siebzehntausend Wörtern, so dass schon keine Minimalpaare für Partikelverben, die sich nur durch Akzentposition unterscheiden, zu finden waren. Die einzigen Minimalpaare waren „Kaffee“ vs. „Café“ und „Erlangen“ vs. „erlangen“.

CELEX ist auch sehr nahe an einem idealen Korpus. Allerdings ist er, gerade was die Partikelverben angeht, zu sehr verallgemeinert. Es erscheinen viele Minimalpaare, die es so im Deutschen nicht gibt bzw. nicht benutzt werden. Dazu

wurden die Häufigkeiten der Wörter einfach fünfzig zu fünfzig auf beide Teile des Paares verteilt. Außerdem liegt CELEX schon als Wortliste vor, was heißt, dass z.B. bei Berechnungen der funktionalen Belastung keine $n > 1$ gewählt werden können, also der Kontext von Wörtern nicht berücksichtigt werden kann.

In Sachen Umfang waren die CQP-Korpora, gegenüber dem Unitselection-Korpus, besser. Nachteilig ist bei ihnen, dass sie ohne phonetische oder phonologische Information vorliegen. Die CQP-Korpora mussten erst um phonologische Informationen aus CELEX ergänzt werden und nicht für jedes Wort, in den CQP-Korpora gab es in CELEX eine SAMPA-Transkription.²² Das heißt, dass die Wortlisten am Ende nicht vollständige, adäquate, phonologische Informationen enthalten und so die Werte der funktionalen Belastungen auch nur Annäherungen an die tatsächlichen Werte sind.

Ein idealer Korpus für die Berechnungen, wäre wohl ein Korpus aus Sprachaufnahmen, vielleicht Mitschnitten von alltäglichen Gesprächen, der vollständig mit allen phonetisch-phonologischen Informationen, auf allen Ebenen, versehen ist, so wie der Unitselection-Korpus, nur mit einem größeren Umfang von etwa 1-2 Millionen Wörtern.

²²Das waren im Besonderen Zahlen, Abkürzungen und Wörter, die Apostrophe enthielten. Bei Wörtern, die Apostrophe enthielten und als Wortakzentminimalpaar auftraten, wurde der Apostroph entfernt. So wurde z.B. „geht's“ zu „gehts“ geändert.

6 Ergebnisse und Evaluation

In diesem Abschnitt werden die Ergebnisse der Arbeit aus Abschnitt 5 vorgestellt und diskutiert. Im Allgemeinen gilt die funktionale Belastung des Wortakzents im Deutschen als gering. Genauere Angaben konnten bisher aber nicht gemacht werden. Das Modell von Niyogi und Surendran ermöglichte es nun, Aussagen über die funktionale Belastung und somit über die Wichtigkeit eines Kontrastes für die kommunikativen Bedürfnisse auf ein empirisch sicheres Fundament zu stellen, nämlich einen genauen statistischen Wert. Die Werte der funktionalen Belastung sind allerdings nur im Verhältnis zu anderen Kontrasten aussagekräftig.

6.1 Ermittelte Werte

Um die funktionalen Belastungswerte des Wortakzents nicht zu abstrakt aussehen zu lassen, wurden noch funktionale Belastungen für weitere Kontraste berechnet. Diese Kontraste waren im Speziellen der Unterschied in der Vokallänge, wie er z.B. Wörter wie „Stadt“ und „Staat“ unterscheidet, und der Stimmhaftigkeitskontrast. Um die Vokallänge, entsprechend Niyogis und Surendrans Modells, zu entfernen, wurden alle gespannten, langen Vokale (/i:/, /e:/, /y:/, /o:/ /u:/) durch die ungespannten, kurzen Vokale (/ɪ/, /ɛ/, /ʏ/, /ɔ/, /ʊ/) ersetzt. Danach wurde einfach das Längungszeichen (/:/) entfernt. Für den Stimmhaftigkeitskontrast wurden die stimmhaften Konsonanten (/g/, /b/, /d/, /z/, /ʒ/, /v/) durch die stimmlosen Konsonanten (/k/, /p/, /t/, /s/, /ʃ/, /f/) ersetzt. Der Stimmhaftigkeitskontrast kann im Deutschen als eher schwach belastet angesehen werden (vgl. dazu Niyogi und Surendran, 2003, Abschnitt 8).

Die funktionale Belastung des Wortakzents im Deutschen gilt im Allgemeinen als niedrig (vgl. Lehiste, 1970). Die Werte, die in dieser Arbeit dazu ermittelt wurden, bestätigen das. In Tabelle 1 sind die funktionalen Belastungen der oben genannten Kontraste aufgelistet. In der Horizontalen sind die funktionalen Belastungen („FL“) der verschiedenen Kontraste, in der Tabelle eingetragen. In der Vertikalen stehen die verschiedenen Korpora, die für die Berechnungen herangezogen worden sind. Die jeweiligen Werte wurden mit 10.000 multipliziert, um sie lesbarer und miteinander vergleichbarer zu machen. Als Vergleich ist noch Tabelle 2 mit einer Auflistung der Anzahl der Minimalpaare, die die Kontraste ausmachen, angefügt.

Es fallen in der Tabelle sofort die Unterschiede in der funktionalen Belastung der Vergleichskontraste (Stimmhaftigkeit und Vokallänge), zwischen CELEX und dem Unitselection-Korpus einerseits und den CQP-Korpora andererseits, auf. Der FL-Wert für die Vokallänge liegt in dem Unitselection-Korpus und in dem CELEX-Korpus bei etwa 19,5. In den CQP-Korpora ist er dagegen deutlich niedriger (13,6 und etwa 11). Die Unterschiede beim Stimmhaftigkeitskontrast sind dagegen geringer. Die Unterschiede zwischen CELEX und dem Unitselection-Korpus einerseits

Korpus	FL-Wortakzent	FL-Vokallänge	Stimmhaftigkeit
CELEX	1,55	19,33	10,90
Unitselection-Korpus	0,48	19,83	12,75
CQP Stuttgarter Zeitung	0,10	13,60	8,93
CQP Bundestagsdebatten	0,05	11,07	5,51

Tabelle 1: Die ermittelten Werte für die funktionalen Belastungen (FL), für $n = 1$ und $T = Wort$. Die errechneten Werte sind hier mit 10.000 multipliziert angegeben.

und den CQP-Korpora andererseits, können dabei auf die Beschaffenheit der Korpora zurückgeführt werden: CELEX und der Unitselection-Korpus sind vollständig mit phonetischen bzw. phonologischen Informationen versehen. Das heißt, dass für jedes Wort, das in diesen Korpora erscheint, eine SAMPA-Transkription vorhanden ist, auf der letztendlich die Berechnungen beruhen. In den CQP-Korpora beschränkt sich bei einem Teil der Wörter die phonologische Information nur auf die Orthographie²³, die die genaue Aussprache nur ungenügend wiedergibt. Gerade bei der Vokallänge beschränken sich die berücksichtigten Wörter auf die, für die in CELEX eine SAMPA-Transkription gefunden werden konnte. Für die Stimmhaftigkeit ist dieser Unterschied etwas geringer, da es in Orthographie unterschiedliche Zeichen für die meisten jeweils stimmhaften und stimmlosen Konsonanten gibt. Ausnahmen davon sind /s/ und /z/, die beide mit dem Schriftzeichen <s> in der Orthographie wiedergegeben werden, und die Auslautverhärtung, die in der Orthographie in der Regel nicht berücksichtigt wird. Trotzdem liegen dadurch die Werte für die funktionale Belastung näher beieinander. Die Werte für die Vergleichskontraste aus CELEX und dem Unitselection-Korpus, können daher als realistischer betrachtet werden als die Werte aus den CQP-Korpora. Wenn man die Unterschiede der verschiedenen Korpora bedenkt, sind die Werte für die funktionalen Belastung erstaunlich gut miteinander vergleichbar.

Was nun die funktionale Belastung des Wortakzents im Vergleich zu den anderen Kontrasten angeht, so ist sie durch alle Korpora hinweg sehr niedrig. Am höchsten ist sie im CELEX-Korpus, wo sie trotzdem von Niyogi und Surendran (2003, Abschnitt 8) als fast gleich null beschrieben wird. Die funktionale Belastung des Wortakzents ist schon in CELEX mehr als zwölf mal niedriger als die der Vo-

²³Besonders für Abkürzungen, lange Komposita und Zahlen ist keine phonologische Information vorhanden.

kallänge und mehr als sieben mal niedriger als die des Stimmhaftigkeitskontrasts. Wenn man sie mit den Belastungswerten aus den anderen Korpora vergleicht, wird die funktionale Belastung des Wortakzents noch niedriger. Im Unitselection-Korpus ist sie drei mal niedriger als in CELEX, wohingegen die anderen Kontraste etwa ähnliche Werte haben.

In den CQP-Korpora ist die Belastung sogar noch vier- bis zehnmal niedriger als im Unitselection-Korpus. Wobei auch zu berücksichtigen ist, dass die anderen Kontraste ebenfalls niedriger liegen, das aber nur leicht und nicht in dem Ausmaß wie beim Wortakzent. Wenn für jedes Wort, das in den CQP-Korpora nicht mit phonologischen Informationen versehen wurde, phonologische Information vorhanden wäre, so müsste davon ausgegangen werden, dass die Werte für die funktionalen Belastungen der Vergleichskontraste in den CQP-Korpora, sich denen im Unitselection-Korpus und in CELEX annähern. Trotzdem ergibt sich jetzt schon in der Stuttgarter Zeitung ein Verhältnis von „1 zu 136“ für die funktionale Belastung des Wortakzents im Vergleich zur funktionalen Belastung der Vokallänge und eine Verhältnis von „1 zu 89“ für die funktionale Belastung des Wortakzents im Vergleich zum Stimmhaftigkeitskontrast. In den Bundestagsdebatten sind Verhältnisse von funktionaler Belastung des Wortakzents zu den funktionalen Belastungen der Vokallänge und der Stimmhaftigkeit jeweils: „1 zu 221“ und „1 zu 110“. Die FL-Werte für den Wortakzent können als akkurat angesehen werden, da, wie in Abschnitt 5.3.3 erklärt, die Häufigkeiten der Minimalpaare von Hand zugeteilt wurden. Die funktionale Belastung des Wortakzents kann also als fast gleich null angesehen werden.

Dass nun die funktionale Belastung in CELEX so viel höher liegt, eben zehnmal, liegt vor allem aber an der Menge von Minimalpaaren. In Tabelle 2 ist, zum Vergleich mit den funktionalen Belastungswerten, die Anzahl von Minimalpaaren für die oben genannten Kontraste angegeben. In CELEX wurde einfach für jedes Betonungsminimalpaar, das sich orthographisch nicht unterscheidet, Vorkommensverteilung von fünfzig zu fünfzig angenommen, was die funktionale Belastung erhöht, da dadurch sehr viel mehr Minimalpaare im Korpus erscheinen als eigentlich vorhanden sind.²⁴

Tabelle 2 zeigt vor allem auch, dass so etwas wie die Anzahl an Minimalpaaren, die in einem Korpus gefunden werden können, nur wenig über tatsächliche funktionale Belastung von Kontrasten aussagt. Die Zahlen, die so gewonnen werden, sind nur in etwa gleich großen Korpora miteinander vergleichbar, im Gegensatz zu den Werten der funktionalen Belastungen, die auch bei unterschiedlich großen Korpora miteinander vergleichbar bleiben. Auch hier sollten wieder nur aus den oben schon genannten Gründen die CQP-Korpora miteinander verglichen werden. Dass

²⁴Aufgrund des logarithmischen Maßes der Entropie erhöht eine größere Anzahl von verschiedenen Wörtern die Entropie mehr als eine Steigerung der Vorkommenswahrscheinlichkeit eines schon vorhandenen Wortes, da die Logarithmusfunktion mit wachsenden Funktionswerten immer schwächer wächst.

Korpus	Wortakzent	Vokallänge	Stimmhaftigkeit
CELEX	428	270	454
CQP-Stuttgarter Zeitung	24	201	512
CQP-Bundestagsdebatten	14	76	152
Unitselection-Korpus	2	52	46

Tabelle 2: Anzahl der Minimalpaare für die untersuchten Kontraste

der Bundestagsdebattenkorpus im Vergleich zum Stuttgarter-Zeitung-Korpus weniger Minimalpaare aufweist, zeigt, dass die Themenbandbreite, bei Berechnungen eine Rolle spielen kann. Die Stuttgarter Zeitung deckt ein viel größeres Themenspektrum ab (z.B. Sport, Kultur, Unterhaltung und Politik) als die Bundestagsdebatten, die nur ein sehr begrenztes Themenspektrum aufweisen (nur Politik). Der Unitselection-Korpus ist viel zu klein um einen Vergleich mit irgendeinem der anderen Korpora zu ermöglichen.

6.2 Theoretische Schlüsse

Die Zahlen zeigen eindeutig, dass die funktionale Belastung des Wortakzents im Deutschen nicht nur niedrig, sondern fast gleich null ist. Es gibt zwar, wie im Anhang A.2 zu sehen ist, eine Menge von möglichen Minimalpaaren, die sich nur durch die Akzentposition unterscheiden, doch ist die Frage dabei, welche dieser Paare überhaupt in der gesprochenen Sprache genutzt werden und wenn, wie oft. Die Anzahl der möglichen Minimalpaare aus einem Wörterbuch herauszusuchen, ist wie oben gezeigt, kein zuverlässiger Anhaltspunkt für die Feststellung der funktionalen Belastung, da die Häufigkeit der Wörter dabei nicht genügend in Betracht gezogen wird. Wenn man die Anzahl der Minimalpaare in den CQP-Korpora betrachtet, könnte man zu dem Schluss kommen, dass die funktionale Belastung des Wortakzents nur etwa fünf- bis zehnmals niedriger ist als die der Vokallänge.

Dass die funktionale Belastung des Wortakzents so niedrig ausfällt, obwohl er eine wichtige Funktion hat, kann teilweise auch durch kleinere Unzulänglichkeiten des Modells für die Beurteilung der Wichtigkeit einer Phonemopposition für eine Sprache erklärt werden. Das Modell berücksichtigt nicht mögliche Bedeutungen von Wörtern. Es berücksichtigt nur Wortpaare, die zusammenfallen. Im Falle eines Wortes, das seine Bedeutung verändert und trotzdem nicht mit einem vorher bekannten Wort zusammenfällt, erkennt das Modell keinen Unterschied. Gerade bei Komposita kann die Akzentposition die Bedeutung verändern. Bei Wie-

se (Wiese, 1996) findet sich dafür ein schönes Beispiel: „**Stadt**planungsbüro“ vs. „Stadt**pl**anungsbüro“. Das zweite der Wörter wird so kaum benutzt und würde wahrscheinlich so in keinem Korpus auftreten und würde folglich nicht in der Berechnung berücksichtigt. Ein psychologisches wäre sinnvoller, wie z.B. die Bestimmung der Vergrößerung der „internen“ Bearbeitungszeit eines Hörers einer Äußerung ohne den zu untersuchten Kontrast. (King, 1967, S. 849) verweist ebenfalls auf die Möglichkeit einer psychologischen Redefinition des Maßes. Allerdings wäre etwas derartiges wohl nicht durchführbar oder bezahlbar für eine vollständige Untersuchung der Struktur einer Sprache. In einem solchen Maß wäre es möglich, dass der Wortakzent eine höhere Wichtigkeit zugemessen bekommt.

Als Zweites können, als Grund für die Niedrigkeit des Belastungswertes, Unzulänglichkeiten in den Korpora gezählt werden. Wenn die Funktion des Akzents im Deutschen die Signalisierung von Grenzen in Einheiten verschiedener Ebenen ist, macht die Wahl von $T = Wort$ Grenzsinalen für Wörter überflüssig, da die Wörter schon voneinander abgegrenzt sind, wenn die funktionale Belastung berechnet wird. Die einzigen Wörter, die sich dann nur durch die Akzentposition unterscheiden, sind die, bei denen es sich zwar orthographisch um ein Wort handelt, phonologisch aber um eine Phrase (vgl. Wiese, 1996, S. 296). Geschickter wäre eine Untersuchung von phonologischen Phrasen oder phonologischen Wörtern mit Kontext, also $n > 1$, so dass der Grenzmarkierungscharakter des Akzents im Deutschen zum Tragen kommen könnte.²⁵ Es wäre also ideal, phonologische Einheiten als T-Objekte untersuchen zu können. CELEX ist allerdings an der Orthographie ausgerichtet und die Struktur als Wortliste erlaubt auch nur $T = Wort$ und $n = 1$. Bei den CQP-Korpora verhält es sich ähnlich, da die phonologische Information für die vorliegenden Berechnungen aus CELEX gewonnen wurden.

Als idealer Korpus, würde sich daher ein Korpus wie der Unitselection-Korpus erweisen, wenn er einen größeren Umfang von etwa 1-2 Millionen Wörtern hätte. In der Regel aber haben phonologische bzw. phonetische Korpora einen sehr geringen Umfang und sind nur wenig umfangreicher als der hier verwendete Unitselection-Korpus. Es bleibt also nur zu warten, bis ein Korpus, der umfangreich genug ist, zugänglich ist um die Untersuchungen, die in dieser Arbeit gemacht wurden, zu wiederholen und adäquatere Aussagen über die Wichtigkeit des Wortakzents zu machen. Die Frage ist, um wieviel die funktionale Belastung des Wortakzents höher ausfallen würde. Es besteht die Möglichkeit, dass der Wert trotz adäquaterer Korpora, und damit adäquaterer Parameterwahl, nicht viel höher ausfiele. Der Grund dafür könnte sein, dass es sich beim Akzent um eine suprasegmentale Eigenschaft handelt und daher nicht sehr gut mit segmentalen Eigenschaften vergleichbar ist.

²⁵„Könnte“ nur deswegen, weil durch die Möglichkeit phonologische Phrasen zu haben, die sich nur durch Akzentposition unterscheiden, noch nicht gesagt ist, dass diese auch so in der Sprache benutzt werden. Das erste Problem könnte also durchaus wieder auftreten und die funktionale Belastung sehr niedrig halten.

A Appendix

A.1 Skript zur Ermittlung der funktionalen Belastung

```
#!/usr/local/bin/perl
# Skript zur Berechnung der funktionalen Belastung.
# Das Skript braucht dafür nur eine Worthäufigkeitsliste
# in Form einer Datei, die auf der Kommandozeile
# mit übergeben wird.
# Die Daten sollten folgende Form haben:
#
#                               [Wort1]\[ Häufigkeit]
#                               [Wort2]\[ Häufigkeit]
#                               [Wort3]\[ Häufigkeit]
#                               usw ...

use strict;
my $in;
my @fldr;
my %wrdlst1; #Wortliste mit Kontrast
my %wrdlst2; #Wortliste ohne Kontrast
my $ref1 = \%wrdlst1;
my $ref2 = \%wrdlst2;
my $check1; # Anzahl aller Wörter in wrdlst1
my $check2; # Anzahl aller Wörter in wrdlst2
my $lang1; #Entropie von wrdlst1
my $lang2; #Entropie von wrdlst2
my $fl;

open (FRQLST, "$ARGV[0]") || die "\nKann $ARGV[0] nicht öffnen!!\n";

while ($in = <FRQLST>) {
    chomp $in;
    @fldr = split (/\\/, $in);
    if ($wrdlst1{"$fldr[0]"}){
        $wrdlst1{"$fldr[0]"}+= $fldr[1];}# unveränderter Hash
    else {$wrdlst1{"$fldr[0]"}= $fldr[1]};

    $fldr[0] =~ s/'//; # Wortveränderung hier!

    if ($wrdlst2{"$fldr[0]"}){
        $wrdlst2{"$fldr[0]"} += $fldr[1];}# veränderter Hash
    else {$wrdlst2{"$fldr[0]"} = $fldr[1]};
};

close (FRQLST) || die "\nKann $ARGV[0] nicht schreiben";
```

```

$check1 = &count_all_wrd($ref1);
$check2 = &count_all_wrd($ref2);
if ($check1 != $check2) {
    print "Fehler! Listen ungleich lang!\n";}
else {
    $lang1 = &compute_ent($ref1,$check1);
    $lang2 = &compute_ent($ref2,$check2);
}
$f1 = ($lang1-$lang2)/$lang1; #FL-Gleichung
print "FL = $f1\n\Bei $check1 Wörtern.\n\n";

# Zählt die Gesamtzahl aller Wörter des Korpus aus der
#Wortfrequenzliste
sub count_all_wrd {
    my $readhsh = $_[0]; # $_[0] = Referenz auf Hash
    my $count;
    my $result = 0;

    foreach $count (values %{$readhsh}){
        $result = $result + $count;
    };
    return $result;
};

#errechnet die Entropie einer Liste
sub compute_ent{
    my $shshlst = $_[0]; # Referenz auf Hash
    my $allwrd = $_[1]; # Gesamtzahl aller wörter
    my $sum = 0;
    my $zwschn; #zum zwischenspeichern
    my $wert;

    foreach $wert(values %{$shshlst}){
        if ($wert != 0){
            $zwschn = ($wert/$allwrd)*(log($wert/$allwrd)/log(2));
            $sum = $sum + $zwschn;}
    };
    return $sum;
};

```

A.2 Liste der Wortakzentminimalpaare aus CELEX

A.2.1 Minimalpaare aus dem Lemmalexikon

Hier ist die Minimalpaarliste aus dem Lemmalexikon aus CELEX wiedergegeben. In jeder Zeile steht dabei ein Wortpaar, welches durch mehrere Leerzeichen getrennt ist. Die erste Spalte enthält die Orthographie des Wortes. <ß> wird durch <ss> wiedergegeben und die Umlaute <ä>, <ö>, <ü> werden als <ae>, <oe>, <ue> wiedergegeben.

CELEX weist in seiner phonetischen Umschrift noch einige Abweichungen vom Standard-SAMPA auf: Vokalisiertes <r> wird im CELEX auch mit /r/ wiedergegeben und nicht wie sonst mit /6/. SAMPA /9/ und /2/ werden als /|/ und /|/ wiedergegeben.

Worte, die so im Deutschen nicht gebraucht werden oder die es so nicht gibt²⁶, sind auf ihrer linken Seite mit einem Stern (*) gekennzeichnet. Bei manchen „gesternteten“ Wörtern kann es sich dabei auch lediglich um Aussprachevarianten handeln, bei denen es keinen Bedeutungsunterschied gibt.

*imperativ\Im-pe:-ra-'ti:f	Imperativ\ 'Im-pe:-ra-ti:f
umschiffen\Um-'SI-f@n	umschiffen\ 'Um-SI-f@n
*umschlingen\Um-'SII-N@n	umschlingen\ 'Um-SII-N@n
*umschlagen\Um-'Sla:-g@n	umschlagen\ 'Um-Sla:-g@n
umspannen\Um-'Spa-n@n	umspannen\ 'Um-Spa-n@n
umspringen\Um-'SprI-N@n	umspringen\ 'Um-SprI-N@n
*umspulen\Um-'Spu:-l@n	umspulen\ 'Um-Spu:-l@n
umschreiben\Um-'Srai-b@n	umschreiben\ 'Um-Srai-b@n
umstecken\Um-'StE-k@n	umstecken\ 'Um-StE-k@n
umstellen\Um-'StE-l@n	umstellen\ 'Um-StE-l@n
*umstechen\Um-'StE-x@n	umstechen\ 'Um-StE-x@n
*umstempeln\Um-'StEm-p@ln	umstempeln\ 'Um-StEm-p@ln
*umstuelpen\Um-'StYl-p@n	umstuelpen\ 'Um-StYl-p@n
umstehen\Um-'Ste:-@n	*umstehen\ 'Um-Ste:-@n
*umstricken\Um-'StrI-k@n	umstricken\ 'Um-StrI-k@n
umbinden\Um-'bIn-d@n	umbinden\ 'Um-bIn-d@n
umbauen\Um-'bau-@n	umbauen\ 'Um-bau-@n
*umblasen\Um-'bla:-z@n	umblasen\ 'Um-bla:-z@n
umfassen\Um-'fa-s@n	umfassen\ 'Um-fa-s@n
umguerten\Um-'gUr-t@n	umguerten\ 'Um-gUr-t@n
umgehen\Um-'ge:-@n	umgehen\ 'Um-ge:-@n
umgeben\Um-'ge:-b@n	*umgeben\ 'Um-ge:-b@n
umgreifen\Um-'grai-f@n	umgreifen\ 'Um-grai-f@n
umhaengen\Um-'hE-N@n	umhaengen\ 'Um-hE-N@n
umkleiden\Um-'klai-d@n	umkleiden\ 'Um-klai-d@n
umlagern\Um-'la:-g@rn	umlagern\ 'Um-la:-g@rn
umlaufen\Um-'lau-f@n	umlaufen\ 'Um-lau-f@n

²⁶Es sind zu jedem Wortpaar mehrere Muttersprachler, in der Regel zwei, befragt worden.

umlegen\Um-'le:-g@n	umlegen\'Um-le:-g@n
umnaehen\Um-'nE:-@n	umnaehen\'Um-nE:-@n
umpflanzen\Um-'pfla-n-ts@n	umpflanzen\'Um-pfla-n-ts@n
umpfluegen\Um-'pfly:-g@n	umpfluegen\'Um-pfly:-g@n
umrahmen\Um-'ra:-m@n	umrahmen\'Um-ra:-m@n
umreißen\Um-'rai-s@n	umreißen\'Um-rai-s@n
umreiten\Um-'rai-t@n	umreiten\'Um-rai-t@n
umziehen\Um-'tsi:-@n	umziehen\'Um-tsi:-@n
umwickeln\Um-'vI-k@ln	*umwickeln\'Um-vI-k@ln
umwinden\Um-'vIn-d@n	*umwinden\'Um-vIn-d@n
umwandeln\Um-'van-d@ln	umwandeln\'Um-van-d@ln
umwehen\Um-'ve:-@n	umwehen\'Um-ve:-@n
umsaeumen\Um-'zOy-m@n	umsaeumen\'Um-zOy-m@n
umsegeln\Um-'ze:-g@ln	umsegeln\'Um-ze:-g@ln
*ungeachtet\Un-g@-'ax-t@t	ungeachtet\'Un-g@-ax-t@t
unterschlagen\Un-t@r-'Sla:-g@n	unterschlagen\'Un-t@r-Sla:-g@n
unterstellen\Un-t@r-'StE-l@n	unterstellen\'Un-t@r-StE-l@n
unterstuetzen\Un-t@r-'StY-ts@n	*unterstuetzen\'Un-t@r-StY-ts@n
unterstehen\Un-t@r-'Ste:-@n	unterstehen\'Un-t@r-Ste:-@n
unterbinden\Un-t@r-'bIn-d@n	unterbinden\'Un-t@r-bIn-d@n
unterbreiten\Un-t@r-'brai-t@n	*unterbreiten\'Un-t@r-brai-t@n
untergraben\Un-t@r-'gra:-b@n	untergraben\'Un-t@r-gra:-b@n
unterhalten\Un-t@r-'hal-t@n	unterhalten\'Un-t@r-hal-t@n
unterlegen\Un-t@r-'le:-g@n	unterlegen\'Un-t@r-le:-g@n
untermengen\Un-t@r-'mE-N@n	untermengen\'Un-t@r-mE-N@n
untermischen\Un-t@r-'mI-S@n	untermischen\'Un-t@r-mI-S@n
Unternehmen\Un-t@r-'ne:-m@n	*unternehmen\'Un-t@r-ne:-m@n
untertauchen\Un-t@r-'tau-x@n	untertauchen\'Un-t@r-tau-x@n
unterziehen\Un-t@r-'tsi:-@n	unterziehen\'Un-t@r-tsi:-@n
untersetzen\Un-t@r-'zE-ts@n	untersetzen\'Un-t@r-zE-ts@n
*Anerbieten\an-Er-'bi:-t@n	anerbieten\'an-Er-bi:-t@n
Argot\ar-'go:\0	Argo\'ar-go:
blutarm\'blu:t-arm	blutarm\'blu:t-'arm
durchschimmern\dUrx-'SI-m@rn	durchschimmern\'dUrx-SI-m@rn
durchscheinen\dUrx-'Sai-n@n	durchscheinen\'dUrx-Sai-n@n
durchschauen\dUrx-'Sau-@n	durchschauen\'dUrx-Sau-@n
durchschiessen\dUrx-'Si:-s@n	durchschiessen\'dUrx-Si:-s@n
durchschlafen\dUrx-'Sla:-f@n	durchschlafen\'dUrx-Sla:-f@n
durchschlagen\dUrx-'Sla:-g@n	durchschlagen\'dUrx-Sla:-g@n
durchschleichen\dUrx-'Slai-x@n	durchschleichen\'dUrx-Slai-x@n
durchschneiden\dUrx-'Snai-d@n	durchschneiden\'dUrx-Snai-d@n
durchschreiten\dUrx-'Srai-t@n	durchschreiten\'dUrx-Srai-t@n
durchstechen\dUrx-'StE-x@n	durchstechen\'dUrx-StE-x@n
*durchstehen\dUrx-'Ste:-@n	durchstehen\'dUrx-Ste:-@n
durchstossen\dUrx-'Sto:-s@n	durchstossen\'dUrx-Sto:-s@n
durchstreichen\dUrx-'Strai-x@n	durchstreichen\'dUrx-Strai-x@n
durchstroemen\dUrx-'Str :-m@n	durchstroemen\'dUrx-Str :-m@n
durchschwimmen\dUrx-'SvI-m@n	durchschwimmen\'dUrx-SvI-m@n
durchbacken\dUrx-'ba-k@n	durchbacken\'dUrx-ba-k@n

durchbeissen\dUrx-'bai-s@n	durchbeissen\'dUrx-bai-s@n
durchblasen\dUrx-'bla:-z@n	durchblasen\'dUrx-bla:-z@n
durchbluten\dUrx-'blu:-t@n	durchbluten\'dUrx-blu:-t@n
durchbohren\dUrx-'bo:-r@n	durchbohren\'dUrx-bo:-r@n
durchdenken\dUrx-'dEN-k@n	durchdenken\'dUrx-dEN-k@n
durchdringen\dUrx-'drI-N@n	durchdringen\'dUrx-drI-N@n
durchfallen\dUrx-'fa-l@n	durchfallen\'dUrx-fa-l@n
durchfahren\dUrx-'fa:-r@n	durchfahren\'dUrx-fa:-r@n
durchfeiern\dUrx-'fai-@rn	durchfeiern\'dUrx-fai-@rn
durchfegen\dUrx-'fe:-g@n	durchfegen\'dUrx-fe:-g@n
durchfliegen\dUrx-'fli:-g@n	durchfliegen\'dUrx-fli:-g@n
durchfliessen\dUrx-'fli:-s@n	durchfliessen\'dUrx-fli:-s@n
durchfluten\dUrx-'flu:-t@n	durchfluten\'dUrx-flu:-t@n
durchfrieren\dUrx-'fri:-r@n	durchfrieren\'dUrx-fri:-r@n
durchgehen\dUrx-'ge:-@n	durchgehen\'dUrx-ge:-@n
durchgliedern\dUrx-'gli:-d@rn	*durchgliedern\'dUrx-gli:-d@rn
durchgluehen\dUrx-'gly:-@n	durchgluehen\'dUrx-gly:-@n
*durchhauen\dUrx-'hau-@n	durchhauen\'dUrx-hau-@n
durchkreuzen\dUrx-'krOy-ts@n	durchkreuzen\'dUrx-krOy-ts@n
durchkriechen\dUrx-'kri:-x@n	durchkriechen\'dUrx-kri:-x@n
durchleuchten\dUrx-'lOyx-t@n	durchleuchten\'dUrx-lOyx-t@n
durchlueften\dUrx-'lYf-t@n	durchlueften\'dUrx-lYf-t@n
durchlaufen\dUrx-'lau-f@n	durchlaufen\'dUrx-lau-f@n
durchmessen\dUrx-'mE-s@n	durchmessen\'dUrx-mE-s@n
durchreisen\dUrx-'rai-z@n	durchreisen\'dUrx-rai-z@n
durchzechen\dUrx-'tsE-x@n	durchzechen\'dUrx-tsE-x@n
durchziehen\dUrx-'tsi:-@n	durchziehen\'dUrx-tsi:-@n
durchwaermen\dUrx-'vEr-m@n	*durchwaermen\'dUrx-vEr-m@n
durchwirken\dUrx-'vIr-k@n	durchwirken\'dUrx-vIr-k@n
durchwachen\dUrx-'va-x@n	*durchwachen\'dUrx-va-x@n
durchwaten\dUrx-'va:-t@n	*durchwaten\'dUrx-va:-t@n
*durchweichen\dUrx-'vai-x@n	durchweichen\'dUrx-vai-x@n
durchwandern\dUrx-'van-d@rn	durchwandern\'dUrx-van-d@rn
durchweben\dUrx-'ve:-b@n	*durchweben\'dUrx-ve:-b@n
durchwuehlen\dUrx-'vy:-l@n	durchwuehlen\'dUrx-vy:-l@n
durchsetzen\dUrx-'zE-ts@n	durchsetzen\'dUrx-zE-ts@n
durchsegeln\dUrx-'ze:-g@ln	durchsegeln\'dUrx-ze:-g@ln
durchsieben\dUrx-'zi:-b@n	durchsieben\'dUrx-zi:-b@n
durchsuchen\dUrx-'zu:-x@n	durchsuchen\'dUrx-zu:-x@n
gegenueber\ge:-g@n-'y:-b@r	*Gegenueber\'ge:-g@n-y:-b@r
hinterbringen\hIn-t@r-'brI-N@n	*hinterbringen\'hIn-t@r-brI-N@n
hintergehen\hIn-t@r-'ge:-@n	*hintergehen\'hIn-t@r-ge:-@n
*Kopfstehen\kOpf-'Ste:-@n	kopfstehen\'kOpf-Ste:-@n
Cafe\ka-'fe:	Kaffee\'ka-fe:
misstrauen\mIs-'trau-@n	Misstrauen\'mIs-trau-@n
Magie\ma-'gi:	Maggi\'ma-gi:
perfekt\pEr-'fEkt	Perfekt\'pEr-fEkt
*plural\plu:-'ra:l	Plural\'plu:-ra:l
tuerkis\tYr-'ki:s	*Tuerkis\'tYr-ki:s

Tailleur\ta-'j :r	*Tailleur\'ta-j :r
Tokaier\to:-'kai-@r	*Tokajer\'to:-kai-@r
widerstreben\vi:-d@r-'Stre:-b@n	*Widerstreben\'vi:-d@r-Stre:-b@n
ueberessen\y:-b@r-'E-s@n	*ueberessen\'y:-b@r-E-s@n
*ueberschlagen\y:-b@r-'Sla:-g@n	ueberschlagen\'y:-b@r-Sla:-g@n
ueberspringen\y:-b@r-'SprI-N@n	ueberspringen\'y:-b@r-SprI-N@n
ueberstechen\y:-b@r-'StE-x@n	*ueberstechen\'y:-b@r-StE-x@n
uebersteigen\y:-b@r-'Stai-g@n	*uebersteigen\'y:-b@r-Stai-g@n
ueberstehen\y:-b@r-'Ste:-@n	ueberstehen\'y:-b@r-Ste:-@n
ueberstreichen\y:-b@r-'Strai-x@n	*ueberstrichen\'y:-b@r-Strai-x@n
ueberstroemen\y:-b@r-'Str :-m@n	*ueberstroemen\'y:-b@r-Str :-m@n
ueberarbeiten\y:-b@r-'ar-bai-t@n	*ueberarbeiten\'y:-b@r-ar-bai-t@n
ueberbinden\y:-b@r-'bIn-d@n	*ueberbinden\'y:-b@r-bIn-d@n
ueberbauen\y:-b@r-'bau-@n	*ueberbauen\'y:-b@r-bau-@n
ueberdecken\y:-b@r-'dE-k@n	*ueberdecken\'y:-b@r-dE-k@n
ueberfaerben\y:-b@r-'fEr-b@n	*ueberfaerben\'y:-b@r-fEr-b@n
ueberfallen\y:-b@r-'fa-l@n	ueberfallen\'y:-b@r-fa-l@n
ueberfahren\y:-b@r-'fa:-r@n	*ueberfahren\'y:-b@r-fa:-r@n
ueberfliessen\y:-b@r-'fli:-s@n	ueberfliessen\'y:-b@r-fli:-s@n
ueberfluten\y:-b@r-'flu:-t@n	*ueberfluten\'y:-b@r-flu:-t@n
ueberfuehren\y:-b@r-'fy:-r@	*ueberfuehren\'y:-b@r-fy:-r@
uebergehen\y:-b@r-'ge:-@n	uebergehen\'y:-b@r-ge:-@n
uebergeben\y:-b@r-'ge:-b@n	*uebergeben\'y:-b@r-ge:-b@n
uebergiessen\y:-b@r-'gi:-s@n	uebergiessen\'y:-b@r-gi:-s@n
ueberhaengen\y:-b@r-'hE-N@n	ueberhaengen\'y:-b@r-hE-N@n
*ueberhalten\y:-b@r-'hal-t@n	ueberhalten\'y:-b@r-hal-t@n
*ueberheben\y:-b@r-'he:-b@n	ueberheben\'y:-b@r-he:-b@n
ueberholen\y:-b@r-'ho:-l@n	*ueberholen\'y:-b@r-ho:-l@n
ueberkommen\y:-b@r-'kO-m@n	*ueberkommen\'y:-b@r-kO-m@n
*ueberkochen\y:-b@r-'kO-x@n	ueberkochen\'y:-b@r-kO-x@n
ueberlassen\y:-b@r-'la-s@n	ueberlassen\'y:-b@r-la-s@n
ueberladen\y:-b@r-'la:-d@n	*ueberladen\'y:-b@r-la:-d@n
ueberlaufen\y:-b@r-'lau-f@n	ueberlaufen\'y:-b@r-lau-f@n
ueberlegen\y:-b@r-'le:-g@n	ueberlegen\'y:-b@r-le:-g@n
uebermalen\y:-b@r-'ma:-l@n	uebermalen\'y:-b@r-ma:-l@n
uebernehmen\y:-b@r-'ne:-m@n	*uebernehmen\'y:-b@r-ne:-m@n
ueberragen\y:-b@r-'ra:-g@n	ueberragen\'y:-b@r-ra:-g@n
*ueberrieseln\y:-b@r-'ri:-z@ln	ueberrieseln\'y:-b@r-ri:-z@ln
ueberzeichnen\y:-b@r-'tsaix-n@n	ueberzeichnen\'y:-b@r-tsaix-n@n
ueberziehen\y:-b@r-'tsi:-@n	ueberziehen\'y:-b@r-tsi:-@n
uebertun\y:-b@r-'tu:n	uebertun\'y:-b@r-tu:n
ueberwerfen\y:-b@r-'vEr-f@n	ueberwerfen\'y:-b@r-vEr-f@n
*ueberwallen\y:-b@r-'va-l@n	ueberwallen\'y:-b@r-va-l@n
ueberwiegen\y:-b@r-'vi:-g@n	*ueberwiegen\'y:-b@r-vi:-g@n
uebersetzen\y:-b@r-'zE-ts@n	uebersetzen\'y:-b@r-zE-ts@n
uebersehen\y:-b@r-'ze:-@n	*uebersehen\'y:-b@r-ze:-@n

A.2.2 Ergänzende Minimalpaare aus dem Vollwortlexikon

Ergänzend folgt hier noch eine kurze Liste aus Minimalpaaren aus dem Vollwortteil von CELEX. Es wurden nur Verb–Nomen-Paare ausgewählt, da die Partikelverbpaare schon voll aus dem Lemmateil aus CELEX oben aufgelistet sind. Paare, die, wie oben, nur Aussprachevarianten darstellen oder die es im Deutschen so nicht gibt, sind wieder mit einem Stern <*> vor dem entsprechenden Wort gekennzeichnet.

entleihn\Ent-'lain	Entlein\'Ent-lain
*irregulaere\I-re:-gU-'lE:-r@	Irregulaere\'I-re:-gU-lE:-r@
*irregulaerem\I-re:-gU-'lE:-r@m	Irregulaerem\'I-re:-gU-lE:-r@m
*irregulaeren\I-re:-gU-'lE:-r@n	Irregulaeren\'I-re:-gU-lE:-r@n
*irregulaerer\I-re:-gU-'lE:-r@r	Irregulaerer\'I-re:-gU-lE:-r@r
*irregulaeres\I-re:-gU-'lE:-r@s	Irregulaeres\'I-re:-gU-lE:-r@s
*imperative\Im-pe:-ra-'ti:-v@	Imperative\'Im-pe:-ra-ti:-v@
*imperativen\Im-pe:-ra-'ti:-v@n	Imperativen\'Im-pe:-ra-ti:-v@n
*imperatives\Im-pe:-ra-'ti:-v@s	Imperatives\'Im-pe:-ra-ti:-v@s
*imperativ\Im-pe:-ra-'ti:f	Imperativ\'Im-pe:-ra-ti:f
umschweife\Um-'Svai-f@	Umschweife\'Um-Svai-f@
umfange\Um-'fa-N@	Umfange\'Um-fa-N@
umkreise\Um-'krai-z@	Umkreise\'Um-krai-z@
umkreisen\Um-'krai-z@n	Umkreisen\'Um-krai-z@n
unterschiede\Un-t@r-'Si:-d@	Unterschiede\'Un-t@r-Si:-d@
unterschieden\Un-t@r-'Si:-d@n	Unterschieden\'Un-t@r-Si:-d@n
unterschied\Un-t@r-'Si:t	Unterschied\'Un-t@r-Si:t
unterbaue\Un-t@r-'bau-@	Unterbaue\'Un-t@r-bau-@
unterbauten\Un-t@r-'bau-t@n	Unterbauten\'Un-t@r-bau-t@n
unterdruecke\Un-t@r-'drY-k@	Unterdruecke\'Un-t@r-drY-k@
unterdruecken\Un-t@r-'drY-k@n	Unterdruecken\'Un-t@r-drY-k@n
unterlasse\Un-t@r-'la-s@	Unterlasse\'Un-t@r-la-s@
unterlagen\Un-t@r-'la:-g@n	Unterlagen\'Un-t@r-la:-g@n
unterlaufe\Un-t@r-'lau-f@	Unterlaufe\'Un-t@r-lau-f@
*unternehmen\'Un-t@r-ne:-m@n	Unternehmen\'Un-t@r-'ne:-m@n
unterrichte\Un-t@r-'rIx-t@	Unterrichte\'Un-t@r-rIx-t@
unterrichten\Un-t@r-'rIx-t@n	Unterrichten\'Un-t@r-rIx-t@n
unterteile\Un-t@r-'tai-l@	Unterteile\'Un-t@r-tai-l@
unterteilen\Un-t@r-'tai-l@n	Unterteilen\'Un-t@r-tai-l@n
untertiteln\Un-t@r-'ti:-t@ln	Untertiteln\'Un-t@r-ti:-t@ln
andrucke\'an-drU-k@	Andrucke\'an-'drU-k@
andrukken\'an-drU-k@n	Andrukken\'an-'drU-k@n
abteile\'ap-tai-l@	Abteile\'ap-'tai-l@
abteilen\'ap-tai-l@n	Abteilen\'ap-'tai-l@n
*abtausche\'ap-tau-S@	*Abtausche\'ap-'tau-S@
abtauschen\'ap-tau-S@n	*Abtauschen\'ap-'tau-S@n
*Butterbrote\bU-t@r-'bro:-t@	Butterbrote\'bU-t@r-bro:-t@
*vorzeiten\'fo:r-'tsai-t@n	Vorzeiten\'fo:r-tsai-t@n
missbrauche\mIs-'brau-x@	Missbrauche\'mIs-brau-x@

misstrauen\mIs-'trau-@n	Misstrauen\'mIs-trau-@n
mobiles\mo:-'bi:-l@s	Mobiles\'mo:-bi:-l@s
*naive\'na-i:-v@	Naive\na-'i:-v@
*naiven\'na-i:-v@n	Naiven\na-'i:-v@n
*naiver\'na-i:-v@r	Naiver\na-'i:-v@r
Pastinake\pas-ti:-'na:-k@	*Pastinake\'pas-ti:-na:-k@
Pastinaken\pas-ti:-'na:-k@n	*Pastinaken\'pas-ti:-na:-k@n
plurale\plu:-'ra:-l@	Plurale\'plu:-ra:-l@
pluralen\plu:-'ra:-l@n	Pluralen\'plu:-ra:-l@n
plurales\plu:-'ra:-l@s	Plurales\'plu:-ra:-l@s
plural\plu:-'ra:l	Plural\'plu:-ra:l
rentiere\rEn-'ti:-r@	Rentiere\'rEn-ti:-r@
rentieren\rEn-'ti:-r@n	Rentieren\'rEn-ti:-r@n
zustande\tsu:-'Stan-d@	Zustande\'tsu:-Stan-d@
zufiel\'tsu:-fi:l	Zufiel\'tsu:-fi:l
zuviel\tsu:-'fi:l	zufiel\'tsu:-fi:l
zumute\tsu:-'mu:-t@	zumute\'tsu:-mu:-t@
*Wallache\'va-la-x@	Walache\va-'la-x@
*Wallachen\'va-la-x@n	Walachen\va-'la-x@n
widerstaende\vi:-d@r-'StEn-d@	Widerstaende\'vi:-d@r-StEn-d@
widerstaenden\vi:-d@r-'StEn-d@n	Widerstaenden\'vi:-d@r-StEn-d@n
widerstand\vi:-d@r-'Stant	Widerstand\'vi:-d@r-Stant
*widerstreite\vi:-d@r-'Strai-t@	Widerstreite\'vi:-d@r-Strai-t@
*widerstreiten\vi:-d@r-'Strai-t@n	Widerstreiten\'vi:-d@r-Strai-t@n
widerstreben\vi:-d@r-'Stre:-b@n	*Widerstreben\'vi:-d@r-Stre:-b@n
wiederaufbaue\vi:-d@r-'auf-bau-@	*Wiederaufbaue\'vi:-d@r-auf-bau-@
widerrede\vi:-d@r-'re:-d@	Widerrede\'vi:-d@r-re:-d@
widerreden\vi:-d@r-'re:-d@n	Widerreden\'vi:-d@r-re:-d@n
widerrufe\vi:-d@r-'ru:-f@	Widerrufe\'vi:-d@r-ru:-f@
widerrufen\vi:-d@r-'ru:-f@n	Widerrufen\'vi:-d@r-ru:-f@n
ueberschauen\y:-b@r-'Sau-@n	*Ueberschauen\'y:-b@r-Sau-@n
uebereile\y:-b@r-'ai-l@	Uebereile\'y:-b@r-ai-l@
ueberblicke\y:-b@r-'blI-k@	*Ueberblicke\'y:-b@r-blI-k@
ueberblicken\y:-b@r-'blI-k@n	*Ueberblicken\'y:-b@r-blI-k@n
ueberdache\y:-b@r-'da-x@	Ueberdache\'y:-b@r-da-x@
ueberdrucke\y:-b@r-'drU-k@	Ueberdrucke\'y:-b@r-drU-k@
ueberfuelle\y:-b@r-'fY-l@	Ueberfuelle\'y:-b@r-fY-l@
ueberfrachten\y:-b@r-'frax-t@n	*Ueberfrachten\'y:-b@r-frax-t@n
ueberkleide\y:-b@r-'klai-d@	Ueberkleide\'y:-b@r-klai-d@
ueberpflanze\y:-b@r-'pflan-ts@	Ueberpflanze\'y:-b@r-pflan-ts@
ueberpflanzen\y:-b@r-'pflan-ts@n	Ueberpflanzen\'y:-b@r-pflan-ts@n
uebertrage\y:-b@r-'tra:-g@	Uebertrage\'y:-b@r-tra:-g@
ueberwerte\y:-b@r-'ve:r-t@	Ueberwerte\'y:-b@r-ve:r-t@
ueberwerten\y:-b@r-'ve:r-t@n	Ueberwerten\'y:-b@r-ve:r-t@n

Literatur

- Baayen, R. H., Piepenbrock, R., Gulikers, L., 1995. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
- Carter, D. M., 1987. An information-theoretical analysis of phonetic dictionary access. *Computer, Speech and Language* 2, 1–11.
- Clark, J., Yallop, C., Fletcher, J., 2007. *An Introduction to Phonetics and Phonology*. Blackwell Publishing.
- Hockett, C., 1955. *A manual of Phonology*. Waverly Press, Baltimore.
- Hockett, C., 1966. The quantification of functional load. *Word* 23, 300–320.
- King, R. D., December 1967. Functional load and sound change. *Language* 43, 831–852.
- Kohler, K., 1977. *Einführung in die Phonetik des Deutschen*. Erich Schmitt Verlag.
- Ladefoged, P., 2001. *A Course in Phonetics*. Harcourt College Publishers, Fort Worth.
- Lehiste, I., 1970. *Suprasegmentals*. MIT-Press, Cambridge, Mass.
- Martinet, A., 1981. Sprachökonomie und Lautwandel. Eine Abhandlung über die diachronische Phonologie. Aus dem Französischen von Claudia Fuchs. Klett-Cotta, Stuttgart.
- Mathesius, V., 1931. Zum Problem der Belastungs- und Kombinationsfähigkeit der Phoneme. *Travaux du cercle linguistique de Prague* 4, 148–152.
- Mengel, A., 2001. *Deutscher Wortakzent. Symbole Signale*.
URL <http://www.andreasmengel.de/pubs/deutscher-wortakzent.pdf>
- Meyerstein, R. S., 1970. *Functional Load*. Mouton, The Hague.
- Niyogi, P., Surendran, D., 2003. Measuring the usefulness (functional load) of phonological contrasts.
URL http://www.cs.uchicago.edu/files/tr_authentic/TR-2003-12.ps
- Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., Säuberlich, B., 2003. Restricted unlimited domain synthesis. In: *Proceedings of Eurospeech-2003 (Geneva)*. pp. 1321–1324.
- Shannon, C. E., 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 379–423, 623–656.
- Shannon, C. E., 1951. Prediction and entropy of printed english. *The Bell System Technical Journal* 30, 50–64.
- Trubetzkoy, N., 1939. *Grundzüge der Phonologie*. *Travaux du cercle linguistique de Prague* 7.
- Wang, W., 1967. The measurement of functional load. *Phonetica* 16, 36–54.
- Wängler, H.-H., 1974. *Grundriss der Phonetik des Deutschen*. N.G. Elwert Verlag.
- Wiese, R., 1996. *The Phonology of German*. Clarendon Press.