This is a preprint version of:

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A. and Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* 62(1), 10-20.

Link to publisher: http://dx.doi.org/10.1026/0033-3042/a000029

dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung

Julian Heister[1], Kay-Michael Würzner[1,2], Johannes Bubenzer[1], Edmund Pohl[1], Thomas Hanneforth[1], Alexander Geyken[2], Reinhold Kliegl[1, 2]

Universität Potsdam, Potsdam[1] &

Berlin-Brandenburgische Akademie der Wissenschaften, Berlin[2]

Kurztitel: dlexDB - eine lexikalische Datenbank

Korrespondenz:

Reinhold Kliegl (email: kliegl@uni-potsdam.de)

Department Psychologie, Universität Potsdam

Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

Telefon: 0331 977 2868 Fax: 0331 977 2793

Zusammenfassung

Mit der lexikalischen Datenbank dlexDB stellen wir der psychologischen und linguistischen Forschung im World Wide Web online statistische Kennwerte für eine Vielzahl von verarbeitungsrelevanten Merkmalen von Wörtern zur Verfügung. Diese Kennwerte umfassen die durch CELEX (Baayen, Piepenbrock und Gulikers, 1995) bekannten Variablen der Häufigkeiten von Wortformen und Lemmata in Texten geschriebener Sprache. Darüber hinaus berechnen wir eine Reihe neuer Kennwerte wie die Häufigkeiten von Silben, Morphemen, Zeichenfolgen und Mehrwortverbindungen sowie Wortähnlichkeitsmaße. Die Datengrundlage bildet das Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS) mit über 100 Millionen laufenden Wörtern. Wir illustrieren die Validität dieser Kennwerte mit neuen Ergebnissen zu ihrem Einfluss auf Fixationsdauern beim Lesen von Sätzen.

Schlüsselwörter: Korpuslinguistik, Lexikalische Datenbank, dlex, dlexDB, CELEX, Blickbewegungen, Lesen, Parafovea

dlexDB - A lexical database for the psychological and linguistic research

Keywords: corpus linguistics, lexical database, dlex, dlexDB, CELEX, eye movement, reading, parafovea

Abstract

The lexical database dlexDB supplies in form of an online database frequency-based norms of numerous process-related word properties for psychological and linguistic research. These values include well known variables such as printed frequency of word form and lemma as documented also in CELEX (Baayen, Piepenbrock and Gulikers, 1995). In addition, we compute new values like frequencies based on syllables, and morphemes as well as frequencies of character chains, and multiple word combinations. The statistics are based on the Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS) with over 100 million running words. We illustrate the validity of these norms with new results about fixation durations in sentence reading.

dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung

Die zunehmende Verfügbarkeit von Texten in elektronischer Form erleichtert die Verwendung und erweitert die Einsatzgebiete von Sprachdaten für die psychologische und linguistische Forschung. Als Datengrundlage für diese Statistiken dienen Korpora, große digitalisierte Textsammlungen, aus denen mithilfe informationstechnischer Methoden Wortstatistiken gewonnen werden. Der Großteil der linguistischen und psycholinguistischen Forschung verwendet derartige korpusbezogene Statistiken über die Eigenschaften bestimmter Wörter. Die steigende Rechenkapazität moderner Computer ermöglicht es mittlerweile große Datenmengen so effizient zu verarbeiten, dass neben dem Auszählen von Häufigkeiten elaborierte Methoden für die Erhebung von Maßen über Wortkombinationen oder die sublexikalische Strukturierung verwendet werden können.

Wir beobachten auf der einen Seite eine Entwicklung in die Breite mit vielen Studien im Bereich der experimentellen Psychologie und Psycholinguistik, die auf solches Korpusmaterial zurückgreifen. Auf der anderen Seite zeigt sich eine Entwicklung in die Tiefe: Speziellere Variablen werden untersucht, deren Erhebung für große Datensätze für den Experimentalpsychologen oder -linguisten nicht einfach durchzuführen ist und korpus- bzw. computerlinguistisches Fachwissen voraussetzt. Mit der lexikalischen Datenbank dlexDB wollen wir dieser Entwicklung gerecht werden und stellen der deutschsprachigen Forschung eine breite Auswahl an zum Teil hochspeziellen Daten zur Verfügung, die über ein Webinterface auf vielseitige Art und Weise abfragbar sind. Im Folgenden stellen wir (1) die Datengrundlage, (2) die Methoden zur Datenerhebung und (3) die in der Datenbank enthaltenen Wortinformationen vor. Wir belegen außerdem die Validität der Wortstatistiken

(4) durch einen Vergleich mit dem Abdeckungsgrad von CELEX und (5) am Beispiel eines neuen Befundes aus der Blickbewegungsforschung beim Lesen.

Datengrundlage

Der De-facto-Standard für deutschsprachige Wortnormen ist CELEX (Baayen et al., 1995). Das von CELEX verwendete Korpus hat eine Größe von ca. 6 Millionen laufenden Wörtern (*Tokens*), die sich aus 290.000 verschiedenen Wörtern (*Types*) zusammensetzen. Es besteht größtenteils aus Zeitungsartikeln, die zwischen 1949 und 1975 erschienen sind, und einigen belletristischen Werken in voller Länge (z.B. Böll und Grass). Neben der aus heutiger Sicht geringen Korpusgröße besteht die wichtigste Unzulänglichkeit von CELEX in der unausgewogenen und für die synchrone Analyse veralteten Textauswahl. Als Datengrundlage für dlexDB verwenden wir das Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS; siehe Geyken, 2007). Das DWDS-Kernkorpus umfasst 122.816.010 Tokens und 2.224.542 Types[1]. Es ist zeitlich gleichmäßig über das gesamte 20. Jahrhundert gestreut und nach Textsorten (Belletristik, Zeitungsartikel, Prosa, Gebrauchstexte und zu einem geringen Teil Transkriptionen gesprochener Sprache) ausgewogen. Das DWDS-Kernkorpus ist in dieser Form seit 2007 online unter http://www.dwds.de verfügbar. Es ermöglicht u. a. auch diachrone Fragestellungen, da die Ergebnisausgabe einen Einblick in den Textkontext erlaubt sowie das Erscheinungsjahr des Dokuments angibt (siehe Beispiel in Lüdeling, Evert & Baroni, 2007).

Datenerhebung

Die Annotierung eines Korpus der Größe des DWDS-Kernkorpus ist von Hand nicht zu bewältigen. Wir greifen deshalb auf automatische Werkzeuge zurück. Die Bestimmung der Grundform der Wörter (Lemmatisierung) und die morphologische Zerlegung werden von der

1 Stand: 01.06.2010, inklusive Satzzeichen.

TAGH-Morphologie (Geyken & Hanneforth, 2006) vorgenommen. Tabelle 1 enthält ein Glossar für die Korpusanalyse wichtiger Begriffe. Die morphologische Analyse erfolgt zunächst kontextunabhängig, d.h. das Wort *beachtete,* beispielsweise, wird sowohl als Verb (1. bzw. 3. Person Singular Präteritum) als auch als Adjektiv analysiert. Der Part-of-Speech-Tagger *moot* (Jurish, 2003) wählt dann anhand des Kontexts die wahrscheinlichste Kategorie aus. Die Entscheidungen von *moot* basieren auf einem *Hidden-Markov-Modell* (Rabiner, 1989), das die Wahrscheinlichkeiten von drei aufeinander folgenden Wörtern (Wort-Trigramme) repräsentiert.

Tabelle 1 hier einfügen

Die Korrektheit des Taggers liegt bei 97,5% und damit im Bereich der besten publizierten Ansätze für das Deutsche. Die verschiedenen Wortkategorien, die Morphologie und Tagger verwenden, beruhen auf dem Stuttgart-Tübingen Tagset (STTS; Schiller, Teufel, Thielen & Stöckert, 1999). Das STTS hat sich als Standard für die deutsche Sprache durchgesetzt und wird in allen Bereichen der automatischen Sprachverarbeitung eingesetzt (vgl. z.B. Skut, Krenn, Brants & Uszkoreit, 1997; Hinrichs, Bartels, Kawata, Kordoni & Telljohann, 2000; Brants, Dipper, Eisenberg, Hansen-Schirra, König, Lezius et al., 2004). Es enthält 54 verschiedene Wortkategorien und kann z.B. unter http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html eingesehen werden. Das oben angegebene Beispiel *beachtete* fällt je nach Kontext in die Klassen finites Vollverb oder attributives Adjektiv.

Tabelle 2 hier einfügen

Die Silbentrennung wird auf Basis eines Ansatzes von Bouma (2003) durchgeführt. Gegeben die allgemeine Silbenstruktur *Ansatz Nukleus Koda,* wird zunächst der längste mögliche Nukleus einer Silbe markiert. Dann werden nach dem *maximal onset principle* (vgl. Fallows, 1981) Folgen von *Ansatz Nukleus Konsonant** als Silben markiert[2]. Die verwendete Liste der Nuklei und Ansätze ist in der Dokumentation der Website aufgeführt. Die Evaluation des Ansatzes erfolgte mit Hilfe der in CELEX enthaltenen Silbentrennungen und ergab eine Korrektheit von 88%, was der in Bouma (2003) angegebenen Korrektheit für das Holländische entspricht. Die Mehrzahl der Fehler entsteht bei Komposita (z.B. **Gärung-schemie* vs. *Gärungs-chemie*) und Präfigierung von Verben, die mit Vokalen beginnen (z.B. **ve-rarmen* vs. *ver-armen*). Deshalb wird die morphologische Zerlegung benutzt, um die Silbentrennung zu verbessern: Jede starke Morphemgrenze ist auch eine Silbengrenze, jedes Präfix ist auch eine Silbe. Dadurch erhöht sich die Korrektheit auf 96%.

Website

Die lexikalische Datenbank dlexDB ist über die Website http://dlexdb.de öffentlich zugänglich. Die Weboberfläche von dlexDB ist an eine relationale Datenbank (MySQL) angeschlossen. Eine in Python programmierte Datenbank-API dient als Bindeglied zwischen der Weboberfläche und der Datenbank. Der psychologischen und linguistischen Forschung steht damit ein mächtiges Suchwerkzeug für die Abfrage einer Vielzahl an Variablen zur Verfügung, mit denen Antworten auf die zwei wichtigsten Fragen effizient gegeben werden: (1) Suche nach Wörtern, die bestimmte Kriterien erfüllen (*Filterabfrage*) und (2) Abruf von Häufigkeitsstatistiken für eine Liste von Wörtern (*Listenabfrage*).

Abb. 1 hier einfügen

---

2 Der Asterisk steht für 0-n-malige Wiederholung.

Filterabfrage

Das erste große Anwendungsfeld von dlexDB ist die Hilfestellung bei der Suche nach Wort- und Satzmaterial für experimentalpsychologische und psycholinguistische Experimente. Über die Suchmaske kann der Nutzer Einschränkungen für jede in dlexDB vorhandene Variable vornehmen. Dazu zählen zunächst neben der Wortlänge und der Wortkategorie die Häufigkeit eines Wortes, die allerdings nicht nur auf Token-, sondern auch auf Satz- und Dokumentebene (*contextual diversity*; siehe Tabelle 1 und 2) bereit gestellt wird. Welche dieser Variablen die Zugriffszeiten bei der Worterkennung in lexikalischen Entscheidungs- und Benennungsaufgaben, aber auch Fixationsdauern beim Lesen am besten widerspiegelt und damit indirekt auch wichtige Hinweise auf die Organisation des menschlichen Lexikons liefert, ist eine aktuelle Fragestellung. Für eine Einführung in diese Fragestellungen verweisen wir auf die in der Tabelle 2 zitierten Arbeiten.

Informationen können als *Type* (Wortformen) oder *Lemma* (Grundformen) abgefragt werden. Darüber hinaus kann auch nach *annotierten Types* gesucht werden, d.h. Wortformen können zusätzlich nach ihrer Wortkategorie (Part-of-Speech: *PoS*) im Korpus unterschieden werden. Beispielsweise kommt der Type *Sinn* sowohl als Eigenname (NE) als auch als gewöhnliches Nomen (NN) vor. In der *Annotierte-Types-Abfrage* lassen sich die Häufigkeiten für diese beiden Verwendungen von *Sinn* getrennt abfragen (siehe Tabelle 3). In der *Type-Abfrage* erhält man die Summe der beiden Einzelhäufigkeiten und in der *Lemma-Abfrage* die Summe der Einzelhäufigkeiten aller Wortformen von *Sinn* (*Sinne, Sinnen*, *Sinnes* usw.).

Tabelle 3 hier einfügen

Neben der orthografischen Repräsentation und der Häufigkeit kann der Benutzer Suchbedingungen auf die phonologische Repräsentation, auf die Silbenstruktur und auf die morphologische Zerlegung eines Wortes setzen. Auf diesen Repräsentationen ist auch eine Abfrage mit regulären Ausdrücken möglich, um beispielsweise Wörter mit bestimmten Präfixen, Infixen oder Suffixen zu finden. Abbildung 1 zeigt einen Ausschnitt der *Filtersuche* mit einer Beispielanfrage für alle Wörter, die mit *Ver* beginnen und mit *ungen* enden. dlexDB bietet neben Darstellungen als Type und Lemma und der Angabe des PoS die morphologische und Silbenzerlegung an. Tabelle 4 zeigt die verschiedenen Repräsentationsebenen für das Beispiel *Versicherungsvertretern*[3].

Tabelle 4 hier einfügen

---

dlexDB liefert auch Statistiken für viele andere linguistische Variablen, die sich in der Leseforschung in spezifischer Form etabliert haben. Neben der initialen Buchstaben-Bigrammhäufigkeit (kumulierte Häufigkeit der Wörter, die mit den gleichen zwei Buchstaben beginnen, unabhängig von der Wortlänge)werden die Variablen *informativeness* und *familiarity* in der Leseforschung verwendet (Kennedy, Pynte & Ducrot, 2002; siehe auch Lima & Inhoff, 1985). Informativität (*informativeness*) bezeichnet die Anzahl aller Wörter (*Types*), die die ersten drei Buchstaben teilen und dieselbe Wortlänge haben (z.B. *Riese, Riech, Riege*, usw.); Familiarität (*familiarity*) die kumulierte Häufigkeit dieser Wörter.

Die Definition solcher Variablen erfolgt immer vor dem Hintergrund theoretischer Überlegungen. Eine Voraussetzung für einen produktiven interdisziplinären Dialog zwischen Linguistik, Psycholinguistik und experimenteller Psychologie ist es, dass diese Variablen mit geringem Aufwand auf der gleichen Grundgesamtheit von Texten berechnet und in

3 Eine Erläuterung der einzelnen Trenner in der morphologischen Zerlegung findet sich in der Dokumentation der Website.

experimentellen Kontexten evaluiert werden können. Wir denken, dass dlexDB ein wichtiges Werkzeug hierfür sein wird.

Listenabfrage

Über die *Listenabfrage* können sich die Nutzer für selbst generierte Wortlisten Wortstatistiken ausgeben lassen. Dies ermöglicht es für große Datensätze einfach und schnell an die gewünschten Variablen aus dlexDB zu gelangen. Analog zur Suche nach einzelnen Wörtern stehen auch in der *Listenabfrage* alle in dlexDB enthaltenen Variablen zur Auswahl. Auf der Website wird in der Ausgabe der *Listenabfrage* jedem Type entsprechend der Eingabereihenfolge eine ID zugeordnet, um eine spätere Zuordnung zu vereinfachen. Dies ist vor allem notwendig, wenn in der *Annotierte-Types-Abfrage* mehrere PoS-Klassifizierungen möglich sind.

Für alle Suchen besteht die Option, die Groß-/Kleinschreibung zu ignorieren. Die Ausgabevariablen können vom Benutzer selbst zusammengestellt werden und einer Ausgabeliste hinzugefügt werden, wobei aus mehreren Normierungsvarianten (logarithmierte oder unlogarithmierte, absolute und normalisierte Normen) ausgewählt werden kann. Die Ergebnisse können online betrachtet, nach Ausgabevariablen sortiert und als csv-Datei ausgegeben werden. Jede Suchabfrage kann in Gestalt einer kurzen XML-Datei abgespeichert und in einer Publikation zitiert werden, um so anderen Benutzern zur Verfügung zu stehen.

Die Website dlexdb.de enthält neben den Suchfunktionen weitere Informations- und Dokumentationsseiten, die Hilfestellungen zu Such- und Ausgabemöglichkeiten enthalten und die verschiedenen Variablen und das der Datenbank zugrunde liegende Korpus beschreiben sowie einen Überblick über das Projekt geben. In der aktuellen Ausbaustufe von dlexDB sind Abfragen von Wort-Unigrammen (Type, Lemma, annotiertes Type) bereits um Zeichen- und Wortbigramme- und -trigramme erweitert. Die gegenwärtige Korpusgröße von

122 Millionen laufenden Wörtern umfasst 2,3 Millionen Types, wohingegen die Anzahl der verschiedenen Wort-Bigramme bereits 24 Millionen beträgt (Trigramme 64, Pentagramme 111 Millionen).

Korpusgröße und Wortabdeckungsrate

Das dlexDB zugrunde liegende Korpus DWDS basiert auf einer wesentlich größeren Datengrundlage als CELEX. Eine wichtige Frage ist natürlich, wie groß ein Korpus sein muss, um Worthäufigkeiten möglichst genau zu schätzen (Burgess and Livesay, 1998). Dass die direkte Verwendung des Web als Korpus nur eingeschränkt sinnvoll ist, zeigen Lüdeling et al. (2007) am Beispiel der Wortendung -*itis* (siehe aber auch Blair, Urland und Ma, 2002; Griffiths, Steyvers und Firl, 2007). Brysbaert und New (2009) kommen zu dem Schluss, dass eine Korpusgröße von 16 bis 30 Millionen ausreicht, um Häufigkeitsnormen adäquat zu schätzen, und dass eine weitere Vergrößerung des Korpus wenig Zugewinn bringt. Um die beiden Korpora dlexDB und CELEX zu vergleichen, führen wir im Folgenden die Begriffe *Abdeckungsgrad* und *Type/Token-Verhältnis* ein.

Wir vergleichen dlexDB und CELEX mit zwei größeren Korpora und quantifizieren die Unterschiede anhand der *coverage probability* (Abdeckungsrate). Die Abdeckungsrate umfasst die relative Häufigkeit von Types, die in einem anderen sogenannten Evaluationskorpus vorkommen. Die *type coverage probability* (Type-Abdeckungsrate) berechnet sich aus der Schnittmenge aller Types aus dem Ausgangskorpus (dlexDB oder CELEX) in Relation zur Menge aller Types aus dem Evaluationskorpus ($types_{eval}$). Für die *token coverage probability* (*Token*-Abdeckungsrate) werden die Frequenzen aller Types des Ausgangskorpus im Evaluationskorpus ($freq_{eval}$) aufsummiert und an der Anzahl aller Tokens ($token_{eval}$) des Evaluationskorpus relativiert. Wir berechnen Abdeckungsraten für zwei

Evaluationskorpora: einerseits die deutschsprachige Wikipedia und andererseits ein Zeitungskorpus aus dem Bestand der DWDS-Korpora.

$$coverage_{type} = \frac{|types_{corpus} \cap types_{eval}|}{|types_{eval}|} \qquad coverage_{token} = \frac{\sum_{w \in types_{corpus}} freq_{eval}(w)}{|token_{eval}|}$$

Die Wikipedia besteht aus ca. 250 Millionen Tokens und hat mit ca. 5 Millionen Types ein sehr hohes Type/Token-Verhältnis[4]. Das erklärt sich daraus, dass eine Enzyklopädie viele Themen mit speziellem Vokabular behandelt und übermäßig viele Personen und Namen verzeichnet sind. Das hier verwendete Zeitungskorpus hat mit ca. einer halben Millionen Types auf 14.5 Millionen Tokens ein geringeres Type/Token-Verhältnis, was durch die Orientierung von Zeitungstexten an der Umgangssprache zu erklären ist.

Da CELEX auf einer kleineren Datengrundlage basiert, ist es nicht überraschend, dass sich für dlexDB deutlich höhere Abdeckungsgrade ergeben (Tabelle 5). Die geringen Abdeckungsrate für die Wikipedia im Gegensatz zum Zeitungskorpus sind dem hohen Type/Token-Verhältnis der Wikipedia zuzuschreiben. Dies gilt insbesondere für die niedrigen Type-Abdeckungsrate sowohl für CELEX (.05 und .22) als auch für dlexDB (.23 und .51), da häufige Types sehr vielen, seltenen Types gegenüber stehen.

Ordnet man alle Types mit gleicher Häufigkeit einer Häufigkeitsklasse zu und sortiert sie nach ihrer konstituierenden Häufigkeitsklasse, ergibt sich der Frequenzrang eines Types (*„der“* erhält z.B. den Rang 1). In Abbildung 2 ist die Type-Abdeckungsrate (in Prozent) gegen den normalisierten Frequenzrang abgetragen. Die Grafiken geben an, wie viel Prozent eines bestimmten Frequenzbereichs in dlexDB (durchgezogene Linie) beziehungsweise CELEX (gestrichelte Linie) enthalten sind. dlexDB hat in allen Frequenzbereichen

4 Stand: 01.10.2009, inklusive Satzzeichen.

gegenüber CELEX sowohl in der Wikipedia (Abb. 2, links) als auch in dem Zeitungskorpus (Abb. 2, rechts) eine höhere Abdeckungsrate. Ein Abfall der Type-Abdeckungsrate zeigt sich für CELEX und dlexDB erst für seltene Types, d.h. seltene Types haben bezogen auf die verschiedenen Korpora sehr unterschiedliche Frequenzränge. Für die Praxis ist dies von geringer Bedeutung, da Unterschiede zwischen extrem seltenen Types kaum eine Rolle spielen.

Abb. 2 hier einfügen

Illustration von dlexDB am Beispiel der Analyse von Fixationsdauern beim Lesen

Einfluss lexikalischer Variablen beim Lesen

Durch die Vergrößerung eines Korpus in Type- und Tokenanzahl lassen sich die Worthäufigkeiten feiner auflösen. Häufigkeitsunterschiede, die bei einem kleinen Korpus nicht sichtbar sind, spielen allerdings für viele experimentelle Studien, die Häufigkeitsklassen als feste Faktoren in ihrem Design verwenden, eher eine untergeordnete Rolle (siehe aber Balota, Cortese, Sergent-Marschall, Spieler & Yap, 2004). Für die statistische Evaluation von computationalen Blickbewegungsmodellen werden jedoch fein abgestufte Häufigkeitsnormen von bisher drei Korpora hinzugezogen (Britisches Englisch: British National Corpus (2007); US Englisch: Brown Corpus (Kucera & Francis, 1967); Deutsch: CELEX (Baayen et al., 1995). Solche fein aufgelösten Häufigkeitsnormen liefern auch die Grundlage für multivariate Analysen von Blickbewegungen beim Lesen (z.B. Kliegl, Grabner, Rolfs & Engbert, 2004).

Bezogen auf die Bewegungen der Augen besteht das Lesen eines einfachen Textes aus einer Abfolge von Fixationen, die in der Regel 150 bis 300 ms dauern, und kurzen

ballistischen Sprüngen von ca. 30 ms Dauer mit einer Länge von ca. 3 bis 9 Buchstaben (*Sakkaden*). Neben den vorwärts gerichteten Sakkaden vom fixierten Wort *n* zum nächsten Wort *n+1* (~50%) werden viele Wörter auch übersprungen (*Skipping*, ~20%), mehrfach fixiert (*Refixation*, ~15%) oder erst im zweiten Durchlauf angesteuert (~15% *Regressionen*) (Kliegl, Nuthmann & Engbert, 2006; Rayner, 1998). Grundsätzlich gilt: Je schwieriger ein Wort (z.B. geringe Häufigkeit, niedrige Vorhersagbarkeit), desto (a) länger die Fixationsdauer, (b) geringer die Übersprungswahrscheinlichkeit und (c) höher die Refixations- und Regressionswahrscheinlichkeit. Neben der Vielzahl an experimentellen Studien (Überblick in Rayner, 1998) bestätigen auch korpusbasierte Statistiken (Regressionsanalysen) diese Befunde (z.B. Kliegl et al. 2006; Kennedy & Pynte, 2005).

Spezifischer Einfluss von Wortanfangsbigrammen auf Lesefixationen

Es wird allgemein angenommen, dass globale visuelle Eigenschaften wie Wortumriss und Wortlänge in Leserichtung in einer Entfernung bis zu 10 bis 12 Buchstaben entdeckt werden. Die Identität von Buchstaben wird allerdings wahrscheinlich nur bis zu 7 Buchstaben in Leserichtung verhaltenswirksam (McConkie & Rayner, 1975; Rayner, 1975; Henderson & Ferreira, 1990). Schon die Präsentation der ersten drei Buchstaben des Wortes *n+1* ermöglicht eine fast normale Leserate (Inhoff, 1989; Rayner, McConkie & Zola, 1980; Rayner, Well, Pollatsek & Bertera, 1982). Neben den sehr gut etablierten Befunden zum Einfluss der Eigenschaften des fixierten Wortes, dass häufige Wörter schneller verarbeitet und also kürzer fixiert werden, sind in den letzten Jahren die Effekte der lexikalischen oder sublexikalischen Häufigkeit des parafovealen Wortes *n+1* und des zurückliegenden Wortes *n-1* auf die Fixationsdauer des aktuell fixierten Wortes *n* ein Schwerpunkt der Forschung zur Blicksteuerung beim Lesen (Abb. 3; Kliegl, et al., 2006; Starr & Rayner, 2001) und ihrer computationalen Modellierung (Engbert, Nuthmann, Richter & Kliegl, 2005). Inzwischen

liegen umstrittene Befunde zu orthographischen Einflüssen von Wort *n+2* auf die Fixationsdauer auf Wort *n* (Pynte, Kennedy, & Ducrot, 2004) und lexikalische Einflüsse von Wort *n+2* auf Wort *n+1* vor (Kliegl, Risse & Laubrock, 2007).

Abb. 3 hier einfügen

Bezogen auf die parafovealen Einflüsse des nächsten Wortes *n+1* ist die Hauptfrage, ob sich der Einfluss dieser Wörter auf ihre visuellen und sublexikalischen Merkmale beschränkt (z.B. Wortform und Anfangsbuchstaben) oder ob er sich auf lexikalische und semantische Eigenschaften ausweitet (Inhoff, Radach, Starr & Greenberg, 2000; Kennedy et al., 2002; Rayner, White, Kambe, Miller & Liversedge, 2003). Dieser sogenannte parafoveale Einfluss von Wort *n+1* (*parafoveal on foveal effect*) wird kontrovers diskutiert (z.B. Kennedy, 2000; Kliegl et al., 2006; Kiegl, 2007; Rayner, Pollatsek, Drieghe, Slattery & Reichle, 2007).

In einer Studie von Kennedy et al. (2002) verkürzten sich Fixationsdauern auf Wort *n*, wenn Wort *n+1* eine niedrige Familiarität oder Regularität besitzt, d.h. wenn nur wenige oder seltene Wörter mit denselben Buchstaben anfangen. Dieser Befund entspricht dem *magnetischen Anziehungseffekt* (Hyönä, 1995; Hyönä & Bertram, 2004), dass irreguläre Buchstabenfolgen in der Parafovea Sakkaden „anziehen“. In White und Liversedge (2004) deutet sich an, dass dieses auch über Wörter hinweg möglich ist. Allerdings finden sie nur erhöhte Skippings vor seltenen Wortanfängen, aber keine kürzeren Fixationsdauern. Im Allgemeinen sind Fixationsdauern im Vergleich zu Fällen, in denen Wort *n+1* fixiert wird, vor übersprungenen kurzen oder häufigen Wörtern kürzer und vor übersprungenen langen oder seltenen Wörtern länger (siehe Abb. 2, Kliegl & Engbert, 2005).

Der Einfluss des vorangegangenen Wortes *n-1* auf die aktuelle Fixation ist weniger umstritten. Es wird z.B. angenommen, dass die verlängerten Fixationsdauern nach seltenen

Wörtern eine Nachverarbeitung des Wortes widerspiegeln, d.h. dass das letzte Wort vor dem Verlassen nicht vollständig verarbeitet wurde (*spillover, lag-effect*). In den Analysen von Kliegl et al. (2006) ist der Einfluss der Häufigkeit von Wort *n-1* in etwa so stark wie der des Wortes *n*: In früheren Studien erreichte dieser Einfluss nur knapp 40% des Einflusses von Wort *n* (Balota, Pollatsek & Rayner, 1985; Schroyens, Vitu, Brysbaert & d'Ydewalle, 1999; Henderson & Ferreira, 1990). Der simultane Einfluss von Wort *n*-2 ist bisher nicht betrachtet worden.

Als einen ersten neuen konkreten Anwendungsfall von dlexDB berichten wir hier Ergebnisse zu den unterschiedlichen Einflüssen von Lemma-, Dokument-, Wortanfangsbigramm- und Tokenhäufigkeiten des fixierten Wortes *n* und der benachbarten Wörter *n-1, n-2, n*+1, *n+2* auf die Fixationsdauer. Unter der Annahme, dass Wörter über mehrere Fixationen verteilt verarbeitet werden, sollten lemmatisierte Worthäufigkeiten und Wortanfangsbigramme einen stärkeren Einfluss auf die aktuelle Fixationsdauer haben, wenn ein Wort zum ersten Mal in der Wahrnehmungsspanne erscheint, d.h. wenn das Wort rechts vom fixierten Wort steht. Erst wenn ein Wort tatsächlich fixiert wird oder gerade mit einer Sakkade verlassen wurde, sollte die Token- und Dokumenthäufigkeit aufgrund der konkreten grammatischen Einschränkungen zum Tragen kommen. Die Erwartung ist also, dass die parafoveale Lemma- oder Bigrammhäufigkeit des Wortes *n*+1 einen stärkeren Einfluss auf eine Fixation auf dem Wort *n* hat als die Tokenhäufigkeit, weil sich die Unterschiede zwischen beiden meistens am Wortende finden und deshalb für die parafoveale Vorverarbeitung von geringerer Bedeutung sein sollten. Umgekehrt sollten bei der Verarbeitung des Wortes *n*-1 die syntaktischen Differenzierungen, die sich in der Wortform niederschlagen, eine größere Rolle spielen. Die Token- oder Dokumenthäufigkeit des Wortes *n*-1 sollte daher einen stärkeren Einfluss haben als seine Lemma- oder Wortanfangshäufigkeit, weil sich zu diesem Zeitpunkt vor allem die Nachwirkung aktueller

Verarbeitungsschwierigkeit, die sich nicht zuletzt aus syntaktischen Prozessen ableitet, in der Fixationsdauer zeigen sollte. Wenn wir Einflüsse von Wort *n+2* finden, dann sollten diese vor allem von den initialen Bigrammhäufigkeiten stammen.

Zur Prüfung dieser Hypothesen verwenden wir 91 767 Lesefixationen, die die einzige Fixation auf einem Wort waren und zwischen zwei in Leserichtung ausgeführten Sakkaden lagen (vgl. Abbildung 3). Die Fixationen stammen von 267 Lesern unterschiedlichen Alters des Potsdam Satz Corpus (PSC), der sich aus 144 einzelnen Sätzen und 550 verschiedenen Wörtern zusammensetzt (vgl. Kliegl et al. 2004). Kliegl et al. (2006) belegen den Einfluss der Tokenhäufigkeit von Wort *n-1*, *n* und *n+1* auf die Fixationsdauer auf Wort *n* bei gleichzeitiger statistischer Kontrolle einer großen Zahl anderer Einflussgrößen wie die Länge und die Vorhersagbarkeit der Wörter aus dem vorhergehenden Satzkontext (siehe auch Kliegl, 2007). Auch die Amplituden der vorhergehenden und nachfolgenden Sakkaden und die Position der Fixation im Wort beeinflussen die Fixationsdauer und wurden im Regressionsmodell berücksichtigt.

Mit der *Listenabfrage* in dlexDB haben wir Token-, Dokument-, Lemma- und initiale Bigrammhäufigkeit für alle Wörter des PSC erhoben. Die dlexDB-Tokenhäufigkeit führte zu keinen anderen Befunden als denen, die in früheren Veröffentlichungen auf der Grundlage von CELEX berichtet wurden. Das methodische Problem für den Nachweis der wortpositionsabhängigen Einflusses von Lemma- und Token- und Dokumenthäufigkeit auf Fixationsdauern ist die hohe Korrelation der drei Häufigkeiten in natürlichen Texten. Sie liegt für unsere Daten über 0.94. In Heister, Würzner und Kliegl (2010) zeigen wir anhand der Analyse von Regressionsresiduen jedoch einen möglichen Umgang mit derartigen Kollinearitäten.

In einem statistischen Modell untersuchen wir den Einfluss von 162 Prädiktoren auf die logarithmierten Fixationsdauern. Das Modell enthält Polynome dritter Ordnung der *Token-*

und initialen Bigrammhäufigkeit, Vorhersagbarkeit und Länge der fünf Wörter, Polynome dritter Ordnung für Input- und Output-Sakkadenamplitude, ein Polynom zweiter Ordnung für die relative Fixationsposition der Fixation im Wort und fünf theoretisch motivierte Interaktionen zwischen diesen Variablen (vgl. Kliegl et al., 2006). Dazu nehmen wir zwei Moderatorvariablen auf, die indizieren, ob das letzte Wort oder nächste Wort übersprungen wird, und deren Interaktionen mit allen Prädiktoren der jeweiligen Start-, Ziel- und übersprungenen Wörter (vgl. Kliegl, 2007).

Diese festen Effekte haben wir gleichzeitig mit den Varianzen der mittleren Fixationsdauern über die zufälligen Faktoren Leser, Sätze und Wörter in einem partiell gekreuztem linear gemischten Modell (LMM) geschätzt. Mit diesem Verfahren werden die über die drei zufälligen Faktoren induzierten Korrelationen in den Daten berücksichtigt.

Das LMM liefert Koeffizienten für den Einfluss von Token- und initialer Bigrammhäufigkeit. Alle Häufigkeitsnormen wurden um 1 erhöht und dann log-transformiert, Token- und initiale Bigrammhäufigkeiten sind normalisiert (pro Million). Die LMM wurden mit dem *lmer*-Programm (*lme4*-package, Bates & Maechler, 2009) geschätzt; die Abbildung mit dem *ggplot2*-Progamm erstellt (Wickham, 2009). Beide Programme sind Teil der *Open-Access* R-Umgebung für Statistik und Grafik (R Development Core Team, 2009). Wir beschränken uns hier auf die Darstellung eines LMM für Token- und initiale Bigrammhäufigkeit. Die Details der Analyse sind in Heister et al. (2010) beschrieben.

Die quadratischen Einflüsse der Token- und der initialen Bigrammhäufigkeiten von Wort *n*, *n-1,n-2,n+1* und *n+2* auf die Fixationsdauern auf Wort *n* sind in Abbildung 4 zu sehen. In der oberen Reihe sind die Einflüsse für alle Fixationen dargestellt, wenn das nachfolgende Wort *n+1* fixiert wird (66 % der Daten), während in der unteren Reihe die entsprechenden Einflüsse abgebildet sind, wenn Wort *n+1* übersprungen wird (34% der Daten). Die Ergebnisse entsprechen den oben skizzierten Erwartungen: Neben Einflüssen

des aktuellen Wortes auf die Fixationsdauer zeigen sich Einflüsse der Häufigkeiten, sowohl von Wörtern links der Fixation (Wörter *n-1* und *n-2*) als auch von den noch nicht fixierten Wörtern *n+1* und *n+2*.

Abb. 4 hier einfügen

Betrachtet man Fixationen, auf die ein Skipping folgt, ändert sich das Bild für den Einfluss von Wort *n+2* (Abb. 4, untere Reihe, rechts). Für Wort *n+2* ist der Einfluss der initalen Bigrammhäufigkeiten entgegen der orthodoxen Richtung von Häufigkeitsmaßen positiv. Dieser positive Koeffizient zeigt sich deutlich in der grafischen Darstellung in Abbildung 4 für Wort *n+2*.

In den Analysen deutet der Einfluss der Tokenhäufigkeit von Wort *n+2* auf eine zumindest teilweise lexikalische Integration von Wort *n+2* hin. Dieser Einfluss wird jedoch nur bedeutsam, wenn Wort *n+1* übersprungen wird (siehe Heister et al., 2010; vgl. Kliegl et al., 2007). Die Verkürzung von Fixationsdauern vor Skippings durch seltene initiale Bigrammhäufigkeiten von Wort *n+2* interpretieren wir ähnlich wie *Hyönä* and Bertram (2004) als *magnetische Anziehung:* Seltene unregelmäßige Wortanfänge ziehen das Auge an, da sie entweder einen hohen Informationsgehalt besitzen oder schwer zu identifizieren sind (siehe auch White & Liversedge, 2004). Erstaunlicherweise wirkt diese Anziehung in unseren Daten über Wörter hinweg und führt zum Überspringen eines Wortes. Das Aufzeigen dieses differenziellen Einflusses von initialer Bigrammhäufigkeit unter Berücksichtigung einer Vielzahl anderer linguistischer Variablen ist möglich, weil zwei Voraussetzungen erfüllt sind: (1) eine sehr große Anzahl von Blickbewegunsdaten und (2) eine hohe Reliabilität der linguistischen Kennwerte, die aus großen und gepflegten Korpora wie dem DWDS gewonnen werden können.

Ausblick

Neuere Ergebnisse sowohl der linguistischen Forschung als auch der Leseforschung legen es nahe, dass in Zukunft noch speziellere als bisher verfügbare Variablen gefordert sind. Das dlex-Projekt will mit der lexikalischen Datenbank dlexDB diese Lücke schließen und der deutschsprachigen psycholinguistischen Forschung mit der Hilfe bewährter Methoden aus der Computerlinguistik ein großes Korpus hoher Datenqualität online zur Verfügung stellen. Für das Englische bietet das *English Lexicon Project* (ELP; Balota, Yap, Cortese, Hutchison, Kessler, Loftis et al., 2007) auf ihrer Website vergleichbare Wortinformationen wie das dlex-Projekt für die deutsche Sprache. Ähnlich dem ELP will dlexDB neben den bekannten lexikalischen Häufigkeitsnormen in Zukunft auch weitere neue Variablen bereit stellen und dazu experimentell erhobene Daten online zur Verfügung stellen.

Eine wichtige Frage ist die Angemessenheit des zugrunde liegenden Korpus. Zum jetzigen Zeitpunkt erscheint uns das DWDS-Kernkorpus aufgrund seiner sorgfältigen Zusammenstellung als das am besten geeignete Korpus. Wir haben gezeigt, dass das DWDS einen höheren Abdeckungsgrad gegenüber CELEX sowohl mit Bezug auf ein hochspezifisches Korpus (Wikipedia) als auch mit Bezug auf ein Zeitungskorpus besitzt. Entgegen unserer Erwartungen und Voranalysen zeigen sich in den Analysen der Blickbewegungsdaten des Potsdam Satz Corpus (PSC) trotz des Größenvorteils bisher nur geringe Unterschiede zu CELEX. Dies liegt vermutlich zum einen am Satzmaterial des PSC, das eine Sammlung möglichst repräsentativer Sätze ist und nicht in erster Linie für die experimentelle Überprüfung spezifischer linguistischer Variablen konzipiert ist.

Eine der Stärken von dlexDB ist neben seiner vollständigen linguistischen Annotation die simultane Extrahierung verschiedener linguistischer Variablen. Auf Blickbewegungsdaten beim Lesen haben parafoveale initiale Bigrammhäufigkeiten vor

Skippings einen qualitativ andersartigen Einfluss als wortbasierte Häufigkeitswerte wie Token-, Lemma- und Dokumenthäufigkeit. Durch detaillierte Analysen (z.B. durch die Untersuchung von Interaktionseffekten der Häufigkeiten benachbarter Wörter) und sorgfältig geplante experimentelle Variationen werden wir weitere Einsatzmöglichkeiten und Vorteile von dlexDB sichtbar machen.

Die offensichtlichen Vorteile von dlexDB gegenüber CELEX liegen in der sorgfältigen Annotation des zugrundeliegenden Korpus DWDS, in der besseren Handhabung über das Webfrontend sowie in der Vielzahl an flexiblen Such- und Ausgabemöglichkeiten. Die Erweiterung auf Wort-N-gramme (N=4) stellt sowohl in theoretischer als auch technischer Hinsicht eine Herausforderung dar. Wir denken, dass dlexDB in seiner Vielfalt an Abfrage- und Erweiterungsmöglichkeiten sowie Anpassungsfähigkeit an aktuelle Fragestellungen der deutschsprachigen Forschung für künftige experimentalpsychologische, psycholinguistische und linguistische Fragestellungen ein nützliches Werkzeug zur Verfügung stellt.

Literaturverzeichnis

Adelman, J. S., Brown, G. D. & Quesada, J. F. (2006). Contextual diversity not word frequency determines word naming and lexical decision times. *Psychological Science, 17* (9), 814–823.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H. & Yap, M. J. (2004). Visual word recognition of single-syllable words, *Journal of Experimental Psychology: General*, *133* (2), 283-316.

Balota, D. A., Pollatsek, A. & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17,* 364 –390.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B. & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39, 445-459.*

Bates, D. & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-31. URL http://cran.r-project.org/package=lme4

Beauvillain, C. (1996). The integration of morphological and whole-word form information during eye fixations on prefixed and suffixed words. *Journal of Memory and Language, 35* (6)*,* 801-820.

Blair I. V., Urland, G. R. & Ma, J. E. (2002). Using internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, and Computers, 34* (2), 286-9.

Bouma, G. (2003). Finite state methods for hyphenation. *Natural Language Engineering, 9* (1), 5–20.

Brants, S., Dipper, S. Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G. & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Interpretation of German Language and Computation, 2* (4), 597-620.

British National Corpus (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Version 3 (*BNC XML Edition*). http://www.natcorp.ox.ac.uk

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments and Computers, 41*, 977-990.

Burgess, C. & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments and Computers*, *30*, 272-277.

Bussmann, H. (1983). Lexikon der Sprachwissenschaft. Stuttgart: Kröner.

Coltheart, M., Davelaar, E. J., Jonasson, J. T. & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Hrsg.), *Attention and Performance*, (Bd. 6, S. 535–555). Hillsdale: Lawrence Erlbaum Associates.

Engbert, R., Nuthmann, A., Richter, E. & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777-813.

Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics, 17*, 309–317.

Geyken, A. (2007). The DWDS Corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum, (Hrsg.), *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. London: Continuum Press.

Geyken, A. und Hanneforth, T. (2006). TAGH: A complete morphology for German based on weighted finite state automata. In *Finite State Methods and Natural Language Processing* (Lecture Notes in Computer Science, Bd. 4002, S. 55-66). Berlin: Springer.

Griffiths, T. L., Steyvers, M. & Firl, A. (2007). Google and the mind. *Psychological Science, 18,* 1069-1076.

Henderson, J. M. & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 417– 429.

Heister, J., Würzner, K. M. & Kliegl, R. (2010). Differential effects of sublexical frequency measures and token frequency in reading. (Manuskript in Vorbereitung)

Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V. & Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster (Hrsg.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.

Hyönä, J. (1995). Do irregular letter combinations attract readers' attention? Evidence from fixation locations in words. *Journal of Experimental Psychology: Human Perception and Performance,* 21, 68-81.

Hyönä, J. & Bertram, R. (2004). Do frequency characteristics of non-fixated words influence the processing of non-fixated words during reading? *European Journal of Cognitive Psychology, 16*, 104-127.

Inhoff, A. W. (1989). Parafoveal processing of words and saccade computation during eye fixations in reading. *Journal of Experimental Psychology: Human Perception & Performance, 15, 544-555.*

Inhoff, A. W., Radach, R., Starr, M. & Greenberg, S. (2000). Allocation of visuo-spatial attention and saccade programming during reading. In A. Kennedy, R. Radach, D.

Heller & J. Pynte (Hrsg.), *Reading as a perceptual process* (S. 221–246). Amsterdam: Elsevier.

Inhoff, A. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40* (6), 431-439.

Jurish, B. (2003). *Part-of-Speech tagging with finite state morphology*. Poster präsentiert auf der Konferenz Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives, Berlin.

Kennedy, A. (2000). Parafoveal processing in word recognition. *The Quarterly Journal of Experimental Psychology, 53* (A)*,* 429–455.

Kennedy, A. & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research, 45,* 153–168.

Kennedy, A., Pynte, J. & Ducrot, S. (2002). Parafoveal-on-foveal interactions in word recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55* (A)*,* 1307–1337.

Kliegl, R. (2007). Towards a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, und Reichle (2007). *Journal of Experimental Psychology: General, 138*, 530-537.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16,* 262-284.

Kliegl, R. & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review, 12,* 132-138.

Kliegl, R., Nuthmann, A. & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135,* 12-35.

Kliegl, R., Risse, S. & Laubrock, J. (2007). Preview benefit and parafoveal-on-foveal effects from word n+2. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1250-1255.

Kucera, H. & Francis, W. N. (1967). *Computational analysis of presentday American English*. Providence, RI: Brown University Press.

Lima, S. D. & Inhoff, A. W. (1985). Lexical access during eye fixations in reading: Effects of word initial letter sequence. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 272–285.

Lüdeling, A., Evert, S. & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf & C. Biewer, (Hrsg.), *Corpus Linguistics and the Web* (S. 7–24). Amsterdam: Rodopi.

McConkie, G. W. & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17,* 578 – 586.

Murray, W. S. & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*, 721-756.

Novick, L. R. & Sherman, S. J. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods, Instruments, and Computers*, *36* (3), 397-401.

Pynte, J., Kennedy, A. & Ducrot, S. (2004). The influence of parafoveal typographical errors on eye movements in reading. *European Journal of Cognitive Psychology*, *16*, 178-202.

R Development Core Team (2009). *R: A language and environment for statistical computing. (version 2.9.0).* [Software]. Vienna, Austria: R Foundation for Statistical Computing. http://www. r-project.org

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE*, 77 (2), 257-286.

Rayner, K. (1975). Eye movements and the perceptual span in reading. *Cognitive Psychology, 7*, 65–81.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rayner, K. & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14* (3), 191-201.

Rayner, K., McConkie, G. W. & Zola, D. (1980). Integrating information across eye movements. *Cognitive Psychology, 12* (2), 206-26.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. & Reichle, E. D. (2007) Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, und Engbert (2006). *Journal of Experimental Psychology, General, 136* (3), 520-529.

Rayner, K., White, S. J., Kambe, G., Miller, B. & Liversedge, S. P. (2003). On the processing of meaning from parafoveal vision during eye fixations in reading. In J. Hyönä, R. Radach, & H. Deubel (Hrsg.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (S. 213-234). Amsterdam: Elsevier.

Rayner, K., Well A. D, Pollatsek A. & Bertera, J. H. (1982). The availability of useful information to the right of fixation in reading. *Perception & Psychophysics, 31* (6), 537-50.

Schiller A., Teufel, S., Thielen, C. & Stöckert, C. (1995). *Guidelines für das Taggen deutscher Textkorpora mit STTS.* Technical Report, IMS-CL, University Stuttgart. http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Schroyens, W., Vitu, F., Brysbaert, M. & d'Ydewalle, G. (1999). Eye movement control during reading: foveal load and parafoveal processing. *Quarterly Journal Experimental Psychology A., 52 (*4), 1021-46.

Skut, W., Krenn, B., Brants, T. & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of ANLP-97*. Washington, DC.

Starr, M. & Rayner, K. (*2001*). Eye movements during reading: some current controversies. *Trends in Cognitive Science, 30*, 147-157.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. [Software] R package version 0.8.1. http://had.co.nz/ggplot2/

White, S.J., & Liversedge, S.P. (2004). Orthographic familiarity influences initial eye fixation positions in reading. *European Journal of Cognitive Psychology, 16,* 52-78.

Yarkoni, T., Balota, D. A. & Yap, M. J. (*2008*). Beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971-97.

ANMERKUNG

Gefördert durch Deutsche Forschungsgemeinschaft (KL 955/12-1). Korrespondenz: Reinhold Kliegl, Department Psychologie, Universität Potsdam, Karl-Liebknecht-Str. 24-25, 14465 Potsdam-Golm, email: kliegl@uni-potsdam.de.

Tabelle 1

Erläuterung wichtiger linguistischer Begriffe

Wortform	Grammatikalisch bestimmte Form eines Wortes (*fahren, fährt, gefahren, fuhr, etc.*)
Type	Jede Wortform, die im Textkorpus mindestens einmal vorkommt
Token	Konkretes Vorkommen einer Wortform im Textkorpus
Lemma	Grundform eines Wortes (*fahren* für *fährt, gefahren, fuhr, etc.*)
Morphem	Kleinste bedeutungstragende Einheit in der Sprache
Phonem	Kleinste bedeutungsunterscheidende Lauteinheit in der Sprache
N-gramm	Folge von N Wörtern oder Zeichen (Wortbigramme: *runter gehen*, *Licht anschalten*; Zeichenbigramme: *sp*, *ck*, *tz*)
Dokumenthäufigkeit	Anzahl verschiedener Dokumente, in denen ein Wort vorkommt
Wortnachbarn	Anzahl der Wörter, die sich in einem Buchstaben von einem Wort unterscheiden (*Reise – Meise*, *liegen - lieben*)
häufigere Wortnachbarn	Anzahl der Wörter, die sich in einem Buchstaben von einem Wort unterscheiden und häufiger vorkommen (*halten* ist ein häufigerer Nachbar von *halsen*)

Tabelle 2

Überblick über die in dlexDB enthaltenen linguistischen Variablen

Variable	Literatur
Worthäufigkeit	Rayner & Duffy (1986); Inhoff & Rayner (1986)
Dokumenthäufigkeit (contextual diversity)	Adelman, Brown & Quesada (2006)
Lemmahäufigkeit	Brysbaert & New (2009); Beauvillain (1996)
Orthografische Nachbarschaft	Coltheart et al. (1977); Yarkoni, Balota & Yap (2008)
Initiale Buchstabenbigrammhäufigkeit	Lima & Inhoff (1985)
Buchstabenbigrammhäufigkeit	Novick & Sherman (2004)
Familiarität/Informativität	Kennedy, Pynte & Ducrot (2002)
Frequenzrang	Murray & Forster (2004)

Tabelle 3

Vergleich der Häufigkeiten von *Sinn* je nach Suchanfrage: Type, Lemma und annotiertes Type

Suche	Eingabe	PoS	Lemma	Häufigkeit
Type	*Sinn*	NN/NE	*Sinn*	16028
Annotiertes Type	*Sinn*	NN	*Sinn*	16000
	Sinn	NE	*Sinn*	28
Lemma	*Sinn*		*Sinn (Sinn, Sinne, Sinnes, Sinns, ...)*	36948

Anmerkungen: PoS: Part-of-Speech (Wortkategorie); NN: normales Nomen; NE: Eigennamen. Abkürzungen aus der STTS Tag Tabelle (Schiller et al., 1999).

Tabelle 4

Annotationsbeispiel für *Versicherungsvertretern*

Ebene	Beispiel
Lemmatisierung	Versicherungsvertreter
Wortart (*Part-of-Speech*)	normales Nomen (NN)
morphologische Zerlegung	ver|sicher/V~ung/n\s#ver|tret/V~er=n
Silbenstruktur	ver-si-che-rungs-ver-tre-tern

Tabelle 5

Token und *type coverage probability* von dlexDB und CELEX im Vergleich zu Wikipedia und einem Zeitungskorpus

	Wikipedia		Zeitungskorpus	
	Token	Type	Token	Type
CELEX	0.85	0.05	0.84	0.22
dlexDB	0.96	0.23	0.94	0.51

Abbildung 1. Ausschnitt der Website http://dlexdb.de für die Suche mit einem regulären Ausdruck

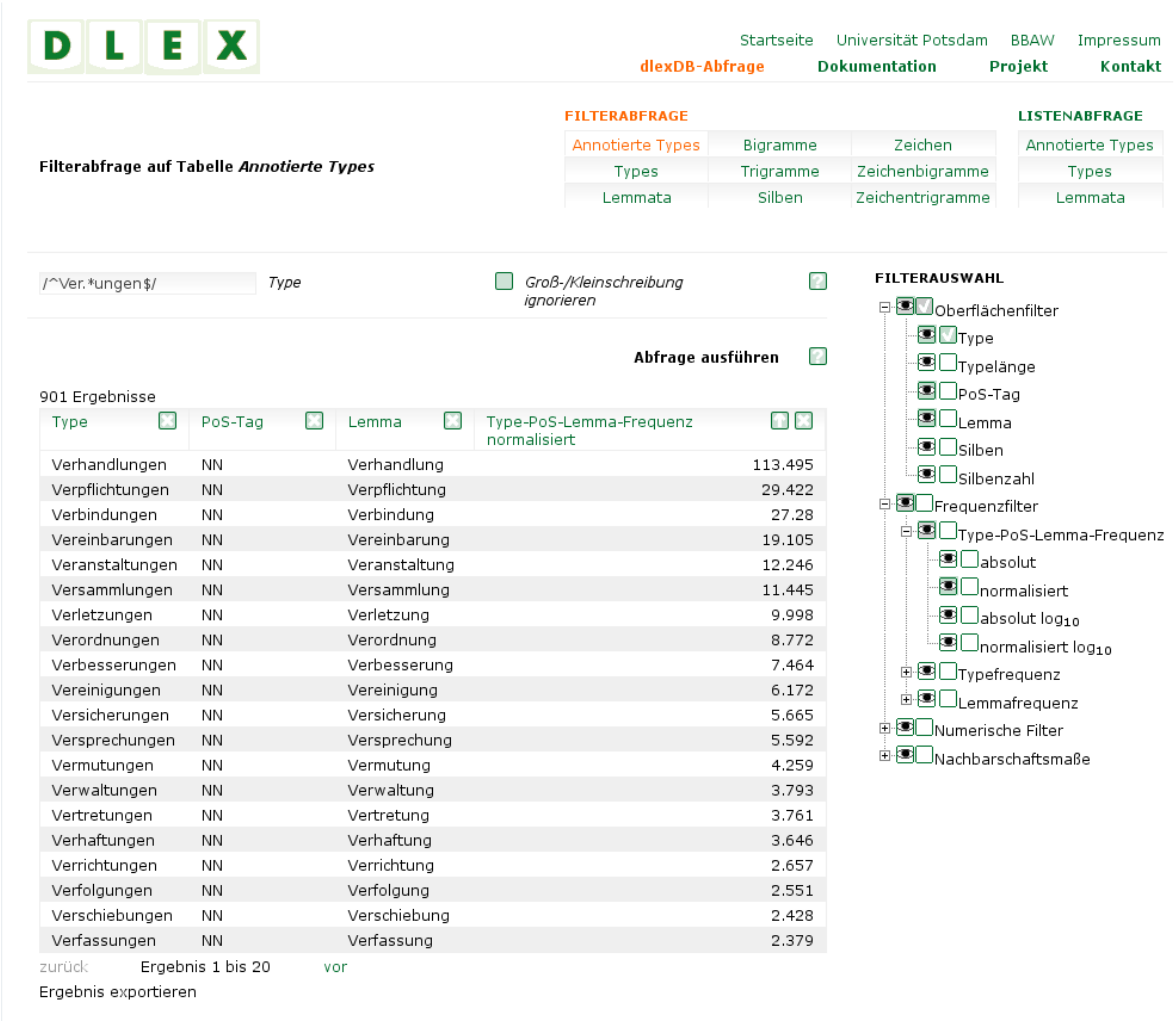


Abbildung 2. *Type*-Abdeckungsrate (*type coverage probability*) in Abhängigkeit des Frequenzranges für dlexDB (durchgezogene Linie) und CELEX (gestrichelte Linie). Links für Wikipedia-Frequenzränge, rechts für Frequenzränge eines Zeitungskorpus. Glättung erfolgte mit der loess-Funktion in ggplot2.

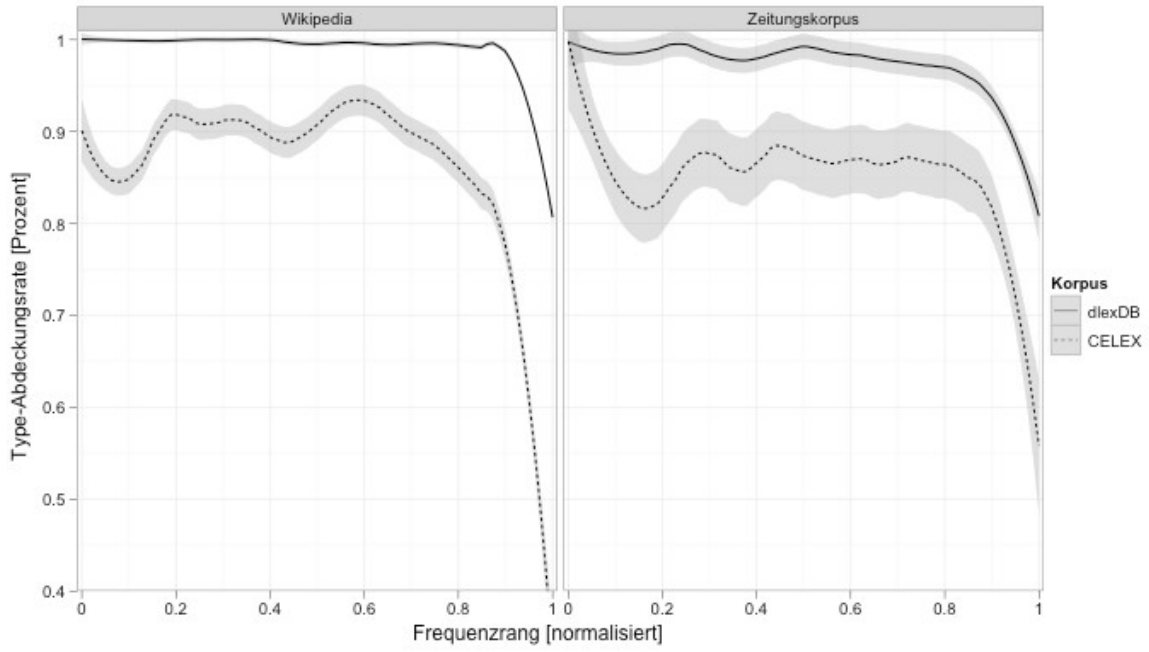


Abbildung 3. Fixationsmuster und Worteinflüsse der verteilten Verarbeitung auf eine Fixation (•) auf Wort n: Pfeile zeigen die möglichen eingehenden und ausgehenden Sakkaden an (Wort *n-1*, bzw. Wort *n+1* wird fixiert oder übersprungen). Einflüsse wie Wortlänge, Worthäufigkeit von Wort *n-2*, *n-1*, *n*, *n+1* und *n+2* auf die Fixationsdauer auf Wort *n* (vgl. Abb. 3 in Kliegl et al., 2006).

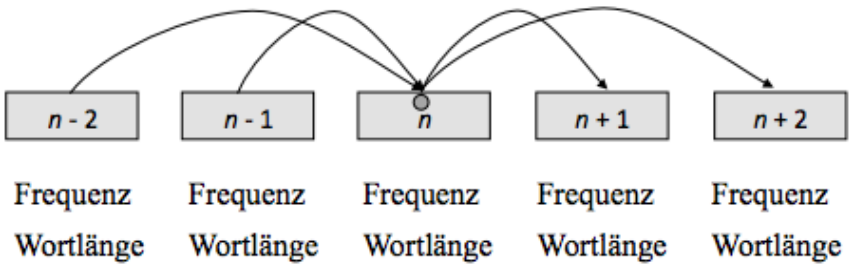
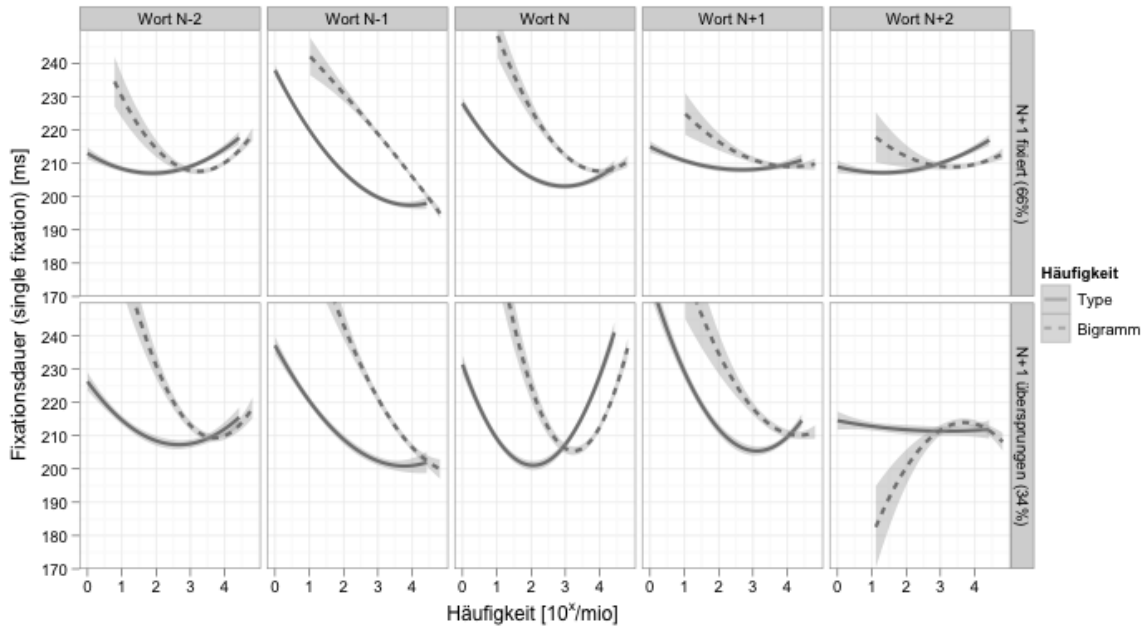


Abbildung 4. Fixationsdauern auf Wort *n* in Abhängigkeit von der Token-, und initialen Bigrammhäufigkeit der Worte *n, n-1, n+1, n-2, n+2*. Oben: Fixationsdauern als quadratische Funktion der Häufigkeiten, wenn Wort *n*+1 fixiert wird. Unten: Fixationsdauern als quadratische Funktion der Häufigkeiten, wenn Wort *n*+1 übersprungen wird. Fixationsdauern und Häufigkeiten wurden log-transformiert.



Originalitätserklärung

Ich erkläre hiermit durch meine Unterschrift – auch im Namen meiner MitautorInnen –, dass das der Psychologischen Rundschau eingereichte Manuskript:

„dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung“

unser geistiges Eigentum ist, dass wir das uneingeschränkte Copyright besitzen, dass es bisher weder als Ganzes noch in Teilen in deutscher Sprache publiziert worden ist und dass es keine Rechte Dritter verletzt. Die Abbildungen und Tabellen stammen – soweit nicht anders vermerkt – von uns.

Wenn die Arbeit bereits in einer anderen Sprache eingereicht oder publiziert wurde oder wenn in andere Publikationen bzw. eingereichte Manuskripte Teilbefunde oder Teilaspekte des oben aufgeführten Manuskriptes eingegangen sind, so habe ich Kopien dieser Arbeiten beigefügt. Ferner verpflichte ich mich, bis zu einer Entscheidung über die Annahme des eingereichten Manuskriptes dieses oder eine modifizierte Form davon keiner anderen Zeitschrift oder keinem anderen Verlag anzubieten.

Ich erkläre mich damit einverstanden, dass mit der Annahme des Manuskripts und dessen Veröffentlichung durch den Hogrefe-Verlag das Verlagsrecht für alle Sprachen und Länder an den Verlag übergeht. Dies betrifft unter anderem auch das ausschließliche Recht der photomechanischen Wiedergabe oder einer anderweitigen Vervielfältigung bzw. Verbreitung.