Universität Paderborn

Methoden des Automatic Abstracting: Untersuchung der Verwendbarkeit Assoziativer Wortnetze im Information Retrieval

Diplomarbeit

vorgelegt von
Gero Zahn
geboren am: 11.10.1970
Mat.-Nr.: 3921270

Universität Paderborn
Fakultät für Elektrotechnik, Informatik und Mathematik
Sommersemester 2004

1. Gutachter: PD Dr. Benno Maria Stein,
Institut für Informatik

2. Gutachter: Prof. Dr. Manfred Wettler,
Fakultät für Kulturwissenschaften / Institut für Psychologie

Inhaltsverzeichnis

Vorwort und Danksagung

„Was lange währt, wird endlich gut“, sagt man. Dies bezieht sich allerdings nicht auf die Bearbeitungszeit dieser Diplomarbeit, die sich im normalen Rahmen bewegte, sondern vielmehr auf die Zeitdauer bis zum Beginn ihrer Bearbeitung. Allzu viel stand ihr für einige Jahre im Wege. Um so erfreuter war ich, dass ich in Herrn Prof. Dr. Manfred Wettler vom Institut für Psychologie der Fakultät für Kulturwissenschaften einen Mentor gefunden habe, der mir dieses Thema als interdisziplinäre Diplomarbeit auf der Grenze zwischen meinem Nebenfach Psychologie und meinem Hauptfach Informatik anbot. Ebensolcher Dank gebührt Herrn PD Dr. Benno Maria Stein aus „meiner“ Fakultät für Elektrotechnik, Informatik und Mathematik aus der Arbeitsgruppe „Wissensbasierte Systeme“, der die Betreuung auf Informatikseite übernahm.

Für die technische Umsetzung habe ich mich mit meinen eigenen Wurzeln überworfen und den einer Informatik-Diplomarbeit geziemenden L_AT_EX-Satz zugunsten der Erstellung in *OpenOffice.org 1.1.0* bzw. *StarOffice 7* links liegen gelassen. So sehr mich die Abkehr von L_AT_EX einerseits schmerzt, so sehr freut es mich im gleichen Schritt, mich gegen die Verwendung von *Microsoft Word* entschieden zu haben. Mit *OpenOffice.org* liegt eine Betriebsystem übergreifende und zudem kostenlose Alternative vor, die mindestens ebenso leistungsfähig und zugleich laufstabiler ist als der „Industriestandard“. *Plus the good feeling while using it.* Obwohl: Natürlich ist man hüben wie drüben angewiesen auf Artikel in der Fachpresse sowie auf anderweitige Sekundärliteratur und Internet-Seiten, um die üblichen Probleme zu erkennen, zu umschiffen, und nicht zuletzt um gangbare Workarounds zu finden. Dennoch denke ich, den richtigen Weg beschritten zu haben – auch wenn mit L_AT_EX sicher einiges ganz anders gelaufen wäre.

Keinesfalls möchte ich dieses Vorwort beschließen, ohne meiner lieben Ehefrau für ihre ständige Unterstützung zu danken: Egal, ob es einerseits darum ging, mir anderweitige Arbeit vom Halse zu halten, oder andererseits darum, mich hinsichtlich dieser Diplomarbeit auf den rechten Weg zurück zu führen ...

Gero Zahn, Juli 2004

Einleitung – Kernwörter und Autoabstracting

In heutiger Zeit liegen Texte mehr und mehr in digitaler Form vor. Spätestens durch die nachhaltige Verbreitung des Internets, des darauf basierenden World Wide Webs (WWW) und die Verwendung von Internet-Suchmaschinen lassen sich zu jedem gewünschten Themengebiet nahezu beliebige Mengen an Textmaterial auffinden. Jedoch: Die Sichtung des gefundenen Materials ist angesichts der Informationsflut kaum mehr zu bewältigen. Das heute scheinbar simple Auffinden von Texten entpuppt sich allzu leicht als sprichwörtliche Suche nach der Nadel im Heuhaufen.

Üblicherweise findet eine Suchmaschine wie Google bei der Eingabe von Suchwörtern hunderte oder gar tausende von WWW-Seiten, die diese Suchwörter enthalten. Das dabei verwendete so genannte „Boolesche Retrieval", das Dokumente auffindet, für das bestimmte Eigenschaften gelten (in diesem Fall das Enthaltensein der eingegebenen Suchwörter), liefert jedoch keinerlei Auskunft darüber, ob die Suchwörter auf den gefundenen WWW-Seiten nur beiläufig verwendet wurden oder ob es sich dabei um die Kernaspekte der gefundenen Seiten handelt.

Auch der Google-immanente Mechanismus des so genannten „PageRank" (vgl. [PBMW1998]) ist mitnichten eine Bewertung, wie gut die gefundenen Seiten zu den eingegebenen Suchwörtern passen: Hierbei wird eine WWW-Seite völlig unabhängig von ihrem textuellen Inhalt rein anhand der bloßen Anzahl anderer auf sie verweisender WWW-Seiten bewertet. Wie bei wissenschaftlichen Arbeiten üblich, bei denen ein häufig zitiertes Schriftstück offensichtlich von großer fachlicher Relevanz ist (weil sich viele andere Texte „darauf berufen"), bildet der Google-PageRank exakt diesen Zusammenhang automatisiert für WWW-Seiten nach: Eine WWW-Seite, auf die viele andere WWW-Seiten mit Links verweisen (d.h. „sich auf sie berufen"), wird höchstwahrscheinlich von großer Bedeutsamkeit sein. — Es war über kurz oder lang jedoch zu erwarten, dass findige Geschäftemacher den ursprünglich durchaus praxistauglichen Google-PageRank für ihre Zwecke missbrauchen würden[1]. Google ist dadurch in heutiger Zeit aufgrund dieser PageRank-Manipulationen je nach Thema der Suche kaum noch sinnvoll verwendbar.

Somit bleibt weiterhin die Frage im Raum, in wieweit eine gefundene WWW-Seite tatsächlich relevant für den Suchenden ist. Faktisch müsste er daher jeden der gefundenen Texte öffnen und lesen, um die behandelte Thematik einschätzen zu können.

1 Getreu dem Motto „Google will vielfach untereinander verlinkte Seiten, also geben wir Google vielfach untereinander verlinkte Seiten!" wurden Meta-Portale mit entsprechend großen Link-Zahlen automatisch generiert und der PageRank somit manipuliert. Daraus resultierte, dass ein Suchender, der beispielsweise die genaue Typ-Bezeichnung eines Elektrogeräts eingibt, nicht mehr zur eigentlich gesuchten Produktseite des Herstellers gelangt, sondern vielmehr zu einem Kaufportal, das seinerseits zu Amazon oder zu Ebay führt.

Um diesen langwierigen Prozess abzukürzen, enthalten wissenschaftliche Texte seit jeher häufig einen *Abstract*, eine den Inhalt des ggf. weitaus längeren Schriftstücks in wenigen Sätzen abstrahierende Kurz-Zusammenfassung. Der Suchende kann sich also auf das Lesen des Abstracts beschränken, um den Kern des Textes zu erfassen und dessen Relevanz für die aktuelle (Literatur-) Suche einzuschätzen. Zusätzlich kann sich das Boolesche Retrieval, d.h. in diesem Falle die Suche nach Texten, in denen bestimmte Schlüsselwörter enthalten sind, auf die Abstracts statt auf die Volltexte beschränken, sofern die Abstracts sämtliche potenziell gesuchten Schlüsselwörter enthalten. Derartige potenzielle Such- bzw. Schlüsselwörter sollen im weiteren Verlauf dieser Arbeit als *Kernwörter eines Textes* bezeichnet werden.

Idealerweise wird der Abstract bereits beim Verfassen eines Textes vom Autor unter der Berücksichtigung der ihm durchaus bekannten Kernwörter seines Textes verfasst. Der Abstract kann dann direkt in Bibliographien oder entsprechende Abstract-Suchmaschinen eingepflegt werden. Liegt jedoch kein Abstract zu einem Text vor, besteht der übliche Weg darin, dass ein Editor bzw. Redakteur den Text liest und daraufhin selbstständig einen Abstract verfasst. In wieweit dieser Abstract den Vorstellungen des Autors gerecht wird, mag von Fall zu Fall variieren. Dies ist insbesondere dadurch gegeben, dass das Verfassen des Abstracts hohe Ansprüche an die Fachkenntnis, Aufmerksamkeit und Konzentration des bearbeitenden Redakteurs stellt.

Um der ständig wachsenden Flut an wissenschaftlichen Texten bei gleichzeitiger Entlastung der Redakteure Herr zu werden, entstanden bereits in den späten 1950er Jahren (beispielsweise von H.P. Luhn, vgl. [LUH1958]) erste Ansätze, fehlende Abstracts vollautomatisch, d.h. ganz ohne menschliche Redakteure zu erstellen. Dieser Vorgang wird als *Automatic Abstracting* bzw. kurz *Autoabstracting* bezeichnet, wörtlich übersetzt also „automatisches Abstrahieren“ bzw. „automatisches Zusammenfassen“.

Derartige Verfahren haben durch die Verbreitung des WWW ungemein an Bedeutung gewonnen: WWW-Seiten enthalten bekanntlich in aller Regel keinen Abstract. Genau deswegen ist es so schwierig zu entscheiden, ob eine von einer Suchmaschine gelieferte WWW-Seite die verwendeten Suchwörter als Kernwörter enthält (womit die Seite höchstwahrscheinlich eine sehr relevante Fundstelle bildet) oder ob die Suchwörter auf der gefundenen WWW-Seite nur beiläufig verwendet werden.

Natürlich kann der Betreiber einer Website[2] beim Erstellen der HTML-Dokumente unsichtbar „Keywords“ d.h. Schlüsselwörter bzw. Kernwörter definieren, jedoch ist er nicht dazu gezwungen. Und selbst wenn er Keywords in die HTML-Dokumente seiner

2 Die Begriffe „Homepage“, „Website“ und „Internet-Portal“ können problemlos synonym verwandt werden. Gemeint ist in jedem Falle eine Gruppe zusammengehöriger WWW-Seiten: Der Begriff „Homepage“ suggeriert im englischen Wortsinn zwar nur eine einzelne Seite, praktisch handelt es sich aber bereits um einige (wenige) Seiten, meist privater Natur. Eine „Website“ stellt eine „große Homepage“ dar, meist mit kommerziellem Hintergrund. Ein „Internet-Portal“ schließlich bildet eine große, meist in Themenbereiche untergliederte „große Website“.

Website einbettet, beziehen sie sich üblicherweise eher auf die komplette Website und weniger auf die Inhalte der einzelnen WWW-Seiten. Die Auswertung der Keywords hinsichtlich der eingegebenen Suchwörter macht für eine Suchmaschine daher häufig nur dann Sinn, wenn nach eher allgemeinen Suchwörtern gesucht wird und so die Startseiten von Websites geliefert werden sollen.

Wünschenswert wäre ein gleichermaßen von Autoren und menschlichen Redakteuren unabhängiges, vollautomatisches Verfahren, das mit einer hinreichenden Genauigkeit die Kernwörter aus einem beliebigen Text extrahiert. Die Ergebnisse des Booleschen Retrievals, d.h. die WWW-Seiten, in denen die eingegebenen Suchwörter vorkommen, ließen sich durch Betrachtung der Übereinstimmungen zwischen den verwendeten Suchwörtern und den berechneten Kernwörtern der WWW-Seiten hinsichtlich ihrer Relevanz einstufen. Dies ginge über den PageRank hinaus, der ja lediglich die „Prominenz" einer gefundenen WWW-Seite bewertet, nicht ihren eigentlichen Inhalt. — Andererseits lassen sich mit Hilfe der Kernwörter eines Textes, sobald sie als solche ausgemacht sind, auch vollständige Abstracts generieren, was dem eigentlichen Vorgang des Autoabstractings entspricht. Dies gelingt bereits H.P. Luhn im Jahre 1958.

Diese Diplomarbeit beschränkt sich auf die Berechnung der Kernwörter als Basistechnik des Autoabstracting einerseits und weiter gedacht des Information Retrieval andererseits. Hier soll also erwogen werden, ob Kernwörter hinreichend genau computergestützt berechnet werden können. Die weitere Verwendung der berechneten Wörter[3] übersteigt den Umfang dieser Arbeit.

Im ersten Teil werden verschiedene klassische Methoden der Kernwortsuche in Texten exemplarisch dargestellt.

Anschließend entwickele ich im zweiten Teil einen ungewöhnlichen Ansatz: Die Verwendung von Assoziativen Wortnetzen und der Simulation menschlichen Assoziationsvermögens zur Berechnung der Kernwörter von Testtexten mit Hilfe des Softwarepakets ATA2 der Arbeitsgruppe Kognitive Psychologie um Prof. Dr. M. Wettler.

Um die Qualität der mit ATA2 berechneten Kernwörter zu überprüfen, wurde eine Studie mit Versuchsteilnehmern durchgeführt, die die Kernwörter einiger Testtexte markieren sollten. Der dritte Teil beschreibt die Versuchsdurchführung im Rahmen einer datenbankgestützten Internet-Anwendung.

Der vierte Teil vergleicht die ATA2-Berechnungen mit den Resultaten der Wortmarkierungsstudie. Im fünften Teil wird eine abschließendes Sicht und ein Ausblick geschildert.

3 Teilweise erscheint in dieser Arbeit der alternative Plural „Worte", vor allem in den Quelltexten der Wortmarkierungsstudie und im Anhang. Trotz leicht variierender Bedeutungen werden beide Formen synonym verwendet.

1 Grundlagen des Autoabstracting

Die Versuche, Texte automatisch für eine effiziente Suche nach zu einer Aufgabenstellung thematisch passenden Texten vorzuverarbeiten, reichen zurück bis in die späten 1950er Jahre. Die zugrunde liegende Disziplin wird mit dem Anglizismus *Information Retrieval* bezeichnet, ins Deutsche übersetzt also das „Auffinden von Informationen" – genauer gesagt, das *Wieder*auffinden von existierenden, aber zurzeit nicht verfügbaren Informationen.

In diesem Kapitel wird anfänglich umrissen, was Information Retrieval ausmacht und inwieweit das Autoabstracting bzw. die Suche nach Schlüsselwörtern Bedeutung für das so genannte *Boolesche Retrieval* hat. Im Anschluss daran werden exemplarisch einige klassische Verfahren zur automatischen Schlüssel- bzw. Kernwortsuche in Ansätzen vorgestellt.

1.1 Information Retrieval

Im Information Retrieval wird Wert auf die Unterscheidung *Daten* und *Informationen* gelegt (vgl. [FER2003] S. 27): Jede beliebige Folge von Buchstaben, Zeichen oder auch nur aus den Binärziffern 1 und 0 stellt *uninterpretierte Daten* dar. Sie kann je nach Kontext unterschiedliche *Informationen* darstellen.

Beispielsweise ist die Zeichenkette „701011" eine mehrdeutige Datenmenge, die einerseits im Kontext eines Telefonverzeichnisses eine sechsstellige Rufnummer darstellen könnte, andererseits im Kontext einer Geburtstagssammlung aber auch das Geburtsdatum des Autors – nämlich den 11. Oktober 1970 in amerikanischer YYMMDD-Schreibweise[4]. Es gibt daneben noch weitere Deutungen, z. B. Seriennummern, Bestellnummern, und vieles andere mehr, die sich mit derselben Zeichenkette beschreiben lassen. *Daten* sind also immer uninterpretiert und implizit mehrdeutig, *Informationen* entsprechen *interpretierten Daten*[5] im jeweiligen Kontext.

4 YY – zweistellige Jahreszahl ohne „19" oder „20", „MM" - zweistellige Monat, „DD" - zweistelliger Tag

5 Mitunter wird zu „Daten" und „Informationen" als drittes Moment das „Wissen" hinzugefügt: *Wissen* bezeichnet Informationen im Moment ihrer Wahrnehmung, Verarbeitung oder Nutzung durch einen Menschen. An dieser Stelle soll aber die Dualität „Daten" vs. „Informationen" d.h. „uninterpretiert" vs. „interpretiert" genügen. — In anderen Definitionen werden die Begriffe „Informationen" und „Wissen" umgekehrt verwendet, d.h. *Wissen* entspricht interpretierten *Daten*, während *Informationen* die Summe des aktuell benötigten *Wissens* darstellen.

Beim *Information Retrieval* geht es also um das Auffinden von inhaltlich möglichst relevanten Informationen, basierend auch auf teilweisen Übereinstimmungen mit der Suchanfrage. Im Gegensatz dazu beschäftigt sich das *Data Retrieval* mit der exakten Suche nach rein datenbasierten exakten Übereinstimmungen mit der Suchanfrage (vgl. [FER2003] S. 31).

1.1.1 Dokumentrepräsentation

Ein *Information Retrieval System* (IRS) enthält eine Sammlung von *Dokumenten*. Ein einzelnes Dokument könnte je nach Anwendungsfall des IRS ein komplettes Buch repräsentieren oder auch nur einen kurzen Text. Jedes Dokument wird durch einen genormten Datensatz repräsentiert, der aus einer Ansammlung von *Feldern* bzw. allgemeiner *Termen* besteht. Beim „klassischen" Information Retrieval in einer wissenschaftlichen Literaturdatenbank sind dies der Dokumenttitel, der Dokumenttyp, das Erscheinungsjahr, die Namen der Autoren, die Sprache des Dokuments, die Anzahl der Referenzen und vieles andere mehr. – Das IRS enthält also i.d.R. nicht die Dokumente selbst, sondern vielmehr jeweils eine Art „standardisiertes Surrogat", eine Reduktion auf die charakteristischen Eigenschaften, die maschinell effizient durchsuchbar sind.

Das Information Retrieval ist nicht auf Textdokumente beschränkt. Denkbar ist z.B. auch ein IRS für Klang-Dokumente, bei denen als Terme etwa der (Datei-)Name des Klangs und eine Beschreibung des zu Gehör gebrachten Geräuschs in Frage kommen. Lediglich die Arten der Terme variieren je nach Dokumenttyp des IRS, d.h. die Struktur des genormten Datensatzes pro Dokument ist jeweils vollkommen unterschiedlich.

Ein praktisches Beispiel für ein derart ungewöhnliches IRS ist eine relativ junge Entwicklung des Fraunhofer Instituts für Integrierte Schaltungen (FIIS) namens „Query by Humming" (QbH, auf deutsch in etwa „Suchanfrage durch Vorsummen", vgl. [FIIS2003]). Das zu Grunde liegende IRS enthält eine Sammlung von Musikstücken, die das System nach dem Vorsummen eines charakteristischen Melodie-Ausschnitts nach passenden Titeln durchsucht. Dies soll einerseits Verkäufer im Musikfachhandel entlasten, andererseits auch eine „intelligente Stereoanlage" ermöglichen, die serverbasiert ohne das Einlegen von Tonträgern und ohne umständliche, konventionelle Suche nach dem gewünschten Liedtitel auskommt.

Das „Query by Humming"-IRS verdeutlicht, dass das jeweilige IRS eine zu den Termen der gespeicherten Dokumente passende Benutzerschnittstelle bereitstellen muss: Das jeweilige System akzeptiert eine Suchanfrage als Eingabe und liefert daraufhin diejenigen im IRS enthaltenen Dokumente bzw. Dokumentverweise, die „zu der Suchanfrage passen". Im am häufigsten anzutreffenden Fall eines IRS, das Texte speichert,

muss das System eine Schnittstelle zur Eingabe von Suchwörtern bereitstellen, um dann die textuellen Terme der im System gespeicherten Dokumente mit den Suchwörtern zu vergleichen.

Zu den Termen von Literaturdatenbanken zählen in der Regel die *Schlagwörter*, d.h. eine kleine Anzahl von z.B. zehn Wörtern, durch die das Textdokument thematisch gut repräsentiert wird. Eine Suche nach Schlagwörtern liefert auf einfache Weise thematisch passende Dokumente.

Soll das IRS kurze Texte enthalten, kann zusätzlich zu den Schlagwörtern oder auch stattdessen der Volltext des Textes bzw. dessen komplettes Vokabular in den Termen des IRS gespeichert werden. Eine Suchanfrage durch Eingabe eines Suchwortes liefert dann alle Texte, in denen das eingegebene Wort an beliebiger Stelle im Text vorkommt. Dies repräsentiert die prinzipielle Funktionsweise der meisten Internet-Suchmaschinen wie z. B. Google: Im Index der jeweiligen Suchmaschine sind als Terme der WWW-Dokumente neben dem Volltext-Vokabular der Seiten auch die vom Autor unsichtbar in den HTML-Text eingebetteten Schlüsselwörter gespeichert, sowie insbesondere ein Verweis auf das eigentliche Dokument mitsamt seiner Adresse im WWW.

1.1.2 Boolesches Retrieval

Das gebräuchlichste Retrieval-Verfahren ist das so genannte *Boolesche Retrieval*. Der Name geht zurück auf den englischen Mathematiker und Philosophen George Boole (1815-1864). Die nach ihm benannte *Boolesche Algebra* stellt einen mathematischen Formalismus der Aussagenlogik dar: Den Kern bilden die Wahrheitswerte „WAHR“ und „FALSCH“ sowie deren Verknüpfung mit den elementaren Operationen „NICHT“, „UND“ und „ODER“, aus denen sich komplexere Operationen ableiten lassen.

Ein IRS für Textdokumente, dessen Benutzerschnittstelle ein Boolesches Retrieval realisiert, interpretiert die eingegebenen Suchtwörter als so genannte *Attribute*, die für einen Text wahr oder falsch sein können. Das heißt: Ein Suchwort ist in einem Textdokument (bzw. in den zu ihm gespeicherten Termen) entweder enthalten oder nicht enthalten. Das dem Suchwort entsprechende Attribut teilt die Menge der im IRS gespeicherten Textdokumente daher disjunkt in diejenigen Textdokumente, in deren Termen das Suchwort enthalten ist, sowie diejenigen, in deren Termen das Suchwort *nicht* enthalten ist. Die beiden resultierenden Teilmengen sind prinzipiell ungeordnet.

Werden mehrere Suchwörter eingegeben, ergeben sich somit entsprechend viele Attribute und ebenso viele, sich ggf. überlagernde Dokument-Teilmengen. Werden die Suchattribute nun mit booleschen Operatoren verknüpft, entspricht dies Mengenoperationen auf den Attribut-Teilmengen. Abbildung 1 auf S. 8 visualisiert dies.

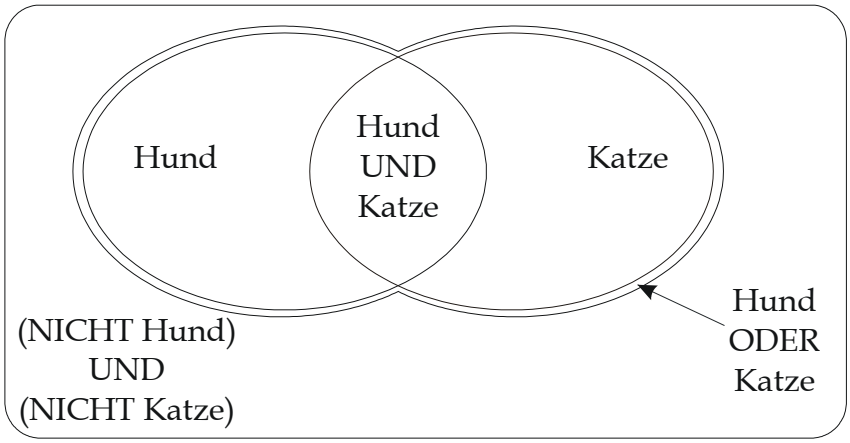

Abbildung 1: Boolesche Operationen auf Attribut-Teilmengen

Am Beispiel: Das Suchwort „Hund“ definiert das Attribut für diejenige Teilmenge aller Dokumente, die das Wort „Hund“ enthalten. Alle anderen Dokumente enthalten das Wort „Hund“ nicht. Ähnliches gilt für das Suchwort „Katze“. Die Verknüpfung „Hund UND Katze“ entspricht der Schnittmenge der beiden Mengen, während „Hund ODER Katze“ der Vereinigungsmenge der beiden Mengen entspricht. Alle anderen Dokumente, die im IRS gespeichert aber in keiner der beiden Mengen enthalten sind, enthalten weder „Hund“ noch „Katze“, d.h. „(NICHT Hund) UND (NICHT Katze)“. Anders betrachtet ist dies (boolesch-)algebraisch äquivalent zu „NICHT (Hund ODER Katze)“, was auch in der Mengendarstellung anschaulich visualisiert wird.

Wie die Verwendung von Information Retrieval Systemen allgemein ist auch das Boolesche Retrieval nicht auf Textdokumente beschränkt: Die Bildung von disjunkten Teilmengen der im IRS enthaltenen Dokumente anhand der aus Suchtermen resultierenden Attribute, die Verknüpfung der Attribute mit Hilfe von booleschen Operatoren, entsprechend Mengenoperationen auf den resultierenden Dokumentmengen – all das ist nicht zwingend an Textdokumente und Suchwörter gebunden.

1.1.2.1 Schwächen des Booleschen Retrieval

Boolesches Retrieval scheint sich speziell im Falle des Volltext-IRS, d.h. mit komplettem Textvokabular, als *das* Verfahren schlechthin etabliert zu haben: Die Suche von Wörtern in den Volltexten einer großen Menge von Textdokumenten ist im Zeitalter von WWW-Seiten und darauf operierenden Internet-Suchmaschinen wie Google eine alltägliche Tätigkeit nahezu jedes Internetbenutzers geworden.

Nichtsdestotrotz hat es seine Schwächen: Die gesamte Mächtigkeit der booleschen Operatoren ist nicht für jeden Suchenden vollständig überschaubar. Eine klare Eingrenzung der Suchergebnisse erfordert, dass der Benutzer die Operatoren und Umformungen der formalen Aussagenlogik beherrscht und korrekt anwenden kann.

Allerdings ist beispielsweise bereits die Unterscheidung der Operatoren „Oder“ („Hund ODER Katze“ = Vereinigungsmenge) und „Exklusiv-Oder“ („ENTWEDER Hund ODER Katze“ = „(Hund ODER Katze) ABER NICHT (Hund UND Katze)“ = Vereinigungsmenge abzüglich Schnittmenge) für einige Menschen kaum mehr intuitiv verständlich. Spätestens Umformungen wie das weiter oben am Beispiel angewandte erste DeMorgansche Gesetz („(NICHT Hund) UND (NICHT Katze)“ = „NICHT (Hund ODER Katze)“) und deren Anwendung zur Formulierung komplexer Suchanfragen sind, obgleich ebenfalls elementar, für viele Menschen schwer im Kopf nachvollziehbar und erfordern intensive Konzentration, am besten unter Zuhilfenahme von Stift und Papier.

Aus diesem Grunde implementieren viele Internet-Suchmaschinen gar nicht den kompletten Umfang der Booleschen Algebra, sondern nur einen geringen Teil, der dadurch mit effizienterer Laufzeit realisiert werden kann. Häufig kommt eine mehr oder minder simple Suchheuristik zum Einsatz, bei der zuerst die Dokumente gelistet werden, die alle eingegebenen Suchwörter enthalten („UND“-Operator), und danach die Dokumente, die nur einige Suchwörter enthalten („ODER“-Operator).

Eine genauere Steuerung der Suche ist erst in der so genannten „erweiterten Suche“ möglich, die jedoch eher selten Verwendung findet und in der Regel ebenfalls nicht die volle Mächtigkeit der Booleschen Algebra realisiert.

Trotz der unerwarteten Komplexität ist das Boolesche Retrieval für einen unkundigen Benutzer mit einfachen Suchanfragen leicht überschaubar: Offensichtlich gilt, dass das IRS die Dokumente meldet, in denen das Suchwort enthalten ist, und die ausblendet, in denen es nicht enthalten ist. Wenn das IRS eine Verfeinerung der Suche gestattet, d.h. eine anschließende Suche, die auf die vorherigen Suchergebnisse beschränkt ist, lässt sich so auch von Benutzern mit wenigen Vorkenntnissen effizient suchen.

1.1.3 Schlüsselwörter vs. Kernwörter

Werden die Volltexte der Dokumente im IRS gehalten, gibt es pro Dokument keine gesonderten Schlagwörter bzw. *Schlüsselwörter*. Jedes im Text enthaltene Wort kann als Suchwort verwendet werden, um ein Suchattribut des Booleschen Retrieval zu bilden. Obgleich diese Möglichkeit für große Dokumentmengen noch zu Anfang der 1980er Jahre für technisch undenkbar, weil zu aufwändig und somit letztlich zu teuer gehalten wurde (vgl. [SMG1983], S. 53), ist dies im Falle heutiger Internet-Suchmaschinen Standard.

Jedoch liefert ein aus einem Suchwort resultierendes Suchattribut beim Booleschen Retrieval prinzipiell eine ungeordnete Ergebnismenge, deren Dokumente willkürlich sortiert sind: Die Relevanz der Suchwörter innerhalb der gefundenen Dokumente ist nicht genauer spezifiziert.

Es ist daher unklar, ob beispielsweise das Suchwort „Hund" im jeweiligen Dokument nur beiläufig verwendet wird, z. B. in *„[...] In der Fernsehsendung »Die Montagsmaler« wurde häufig sofort »Hund, Katze, Maus!« gerufen, um diese möglichen Lösungen sofort auszuschließen. [...]"*, oder ob das Suchwort den Kern des Textes darstellte, z. B. in *„[...] Ein Hund, der hinsichtlich bestimmter Rassemerkmale gezüchtet wurde, kann im Laufe seines Lebens körperliche Probleme erfahren. [...]"*. Während der erste Text offensichtlich von Ratesendungen im Fernsehen handelt und kaum als sinnvolles Suchresultat auf das Suchwort „Hund" eingestuft werden kann, handelt der zweite Text offensichtlich tatsächlich von Hunden.

Es kann also sinnvoll sein, die Suche nicht auf den Volltexten der Dokumente durchzuführen, sondern auf geeignete Weise den „Kern der Dokumente" zu ermitteln. Bei der Aufnahme in das IRS, der so genannten *Indexierung*[6] werden dazu die Schlagwörter der Dokumente bestimmt und als Schlüsselwörter bzw. *Indexwörter* in speziellen Termen gespeichert. Dies kann manuell durch geschulte Lektoren oder durch automatische Verarbeitungsprozesse erfolgen.

Insbesondere im Falle der manuellen Indexierung gibt es zwei unterschiedliche Arten der Auswahl von Wörtern für die Aufnahme in den IRS-Index (vgl. [SMG1983], S. 54): *Kontrollierte* bzw. *unkontrollierte Indexwörter*.

Im Falle von *unkontrollierten Indexwörtern* kann grundsätzlich jedes Wort des Textes in den Index aufgenommen werden, was zu einer exakten Indexierung einerseits, aber zu einer großen thematischen Vielfalt andererseits führt. Dadurch wird es mitunter vorkommen, das zwei Dokumente mit zwei unterschiedlichen Synonymen desselben Wortes indexiert werden, z. B. „Welle" und „Woge". Das Retrieval mit dem Suchwort „Welle" würde nicht das unter dem Wort „Woge" indexierte Dokument liefern.

Im Gegensatz dazu werden bei der Verwendung von *kontrollierten Indexwörtern* mehrere Synonyme auf ein identisches Indexwort zurückgeführt, so dass im o.a. Beispiel beide Texte mit demselben Wort „Welle" indexiert würden. Dies kann automatisiert oder manuell unter Zuhilfenahme eines *Thesaurus*[7] erfolgen. Obwohl dies ggf. zu einer inhaltlichen Vereinfachung und Egalisierung führt, würde in diesem Falle das Beispiel-Suchwort „Welle" korrekterweise beide Dokumente liefern.

6 Obgleich sich das Wort „Indexierung" merkwürdig liest, ist dies die richtige Schreibweise. „Indizieren" steht für „auf etwas hinweisen; ratsam erscheinen lassen; auf den Index setzen" (Quelle: Wahrig Deutsches Wörterbuch).

7 Ein Thesaurus bildet im einfachsten Fall ein kontrolliertes Vokabular, in dem Begriffe durch Relationen miteinander verbunden sind. Vorstellbar ist dies als Ersetzungstabelle, mit der im o.a. Beispiel der Begriff „Woge" mit dem Begriff „Welle" in Bezug gesetzt wird, sodass „Welle" statt „Woge" für die Indexierung verwendet wird.

Zusätzlich können bei der Verwendung kontrollierter Indexwörter auch Wörter in den Index aufgenommen werden, die als solche nicht im Text enthalten sind. Derartige Abstraktionen dienen der Qualität des Indexes, indem statt ggf. vieler Einzelwörter wenige und dafür inhaltlich abstrahierende künstliche Schlüsselwörter in den Index aufgenommen werden.

In beiden genannten Fällen ist es unumgänglich, die Suchwörter der Suchanfragen in einer Vorverarbeitung auf passende Indexwörter zurückzuführen. Denn in o.a. Beispiel würde eine Suche nach dem Wort „Woge“ sonst ggf. kein einziges Dokument liefern, wenn es bei der Indexierung in allen Dokumenten durch „Welle“ ersetzt worden wäre: Erst die Suche nach dem Synonym „Welle“ liefert die gewünschten Ergebnisse.

Die notwendige Optimierung der Suchwörter kann durch entsprechend geschultes Personal erfolgen, das dieselben Ersetzungen wie bei der Indexierung verwendet (etwa denselben Thesaurus), um ungebräuchliche auf gebräuchliche Synonyme oder künstliche Indexwörter zurückzuführen. – Ferber, Wettler und Rapp haben in [FWR1995] eine automatische, auf assoziativen Wortnetzen basierende Methode erforscht und vorgestellt, um diesen Vorgang mit Hilfe von simulierten Wortassoziationen auf die eingegebenen Suchwörter zu unterstützen.

Im weiteren Verlauf soll angenommen werden, dass ein *Kernwort* ein Wort darstellt, das im ursprünglichen Text enthalten war, während ein *Schlüsselwort* auch ein zusätzlich eingefügtes Indexwort darstellen kann. Alle weiteren Betrachtungen beziehen sich, sofern nicht anders angegeben, auf im Text enthaltene Kernwörter.

Im Falle von automatischer Indexierung, beim *Autoindexing,* kommen häufig unkontrollierte Indexwörter zum Einsatz. Die intellektuelle Leistung geschulter Lektoren, die den Kern eines Textes erfassen und Synonyme ohne gravierende thematische Egalisierung zusammenfassen bzw. sogar thematisch korrekte abstrahierende Schlüsselwörter hinzufügen, ist für automatische Indexierungssysteme häufig nicht realisierbar. Daher wird in diesem Fall größerer Wert auf die unveränderte Übernahme der Kernwörter in den Index gelegt, und die Diversifikation des IRS-Index wird gebilligt.

1.1.3.1 Abstracts und Autoabstracting

Ein *Abstract* bildet eine Kurzfassung eines Textes, die dazu dient, in wenigen Sätzen den Kern des zugehörigen Textes wiederzugeben. Wann immer ein Suchender mit einer beliebigen Suchmethode einen Text findet, kann er so innerhalb kürzester Zeit entscheiden, ob der gefundene Text zum gewünschten Themengebiet gehört oder nicht.

Der Abstract wird bei wissenschaftlichen Texten in der Regel vom Autor des Dokumentes verfasst. Liegt kein Abstract vor, soll das IRS aber Abstracts der gespeicherten Dokumente enthalten, muss beim Indexieren auch ein Abstract verfasst werden.

Es ist offensichtlich, dass der Abstract die Kernwörter und ggf. auch abstrahierende Schlüsselwörter des Textes enthalten muss, um dessen thematische Einordnung zu ermöglichen. Das heißt, ein entsprechend geschulter und qualifizierter Lektor, der die Kernwörter auswählt und ggf. um zusätzliche Schlüsselwörter ergänzt, muss eben diese Wörter in prägnanter Form in einem Kurztext verarbeiten, um den Themenbezug des Textes auf einfache Weise darzustellen. Diese Tätigkeit lässt sich treffend mit der englischsprachigen Verlaufsform *Abstracting* bezeichnen.

Auch beim Autoindexing, also letztlich bei einer maschinellen Suche nach Kernwörtern, kann es je nach IRS erforderlich sein, einen Abstract automatisch zu generieren. Der Vorgang des automatischen Abstracting wird als *Autoabstracting* bezeichnet.

Da das freie Schreiben von Texten, insbesondere von kurzen, prägnanten Abstracts, kaum maschinell durchführbar ist, liegt es bei der Verwendung von unkontrollierten Indexwörtern nahe, im Text enthaltene Sätze unverändert in den Abstract aufzunehmen, statt neue zu verfassen. Als einfache Heuristik kann angenommen werden, dass diejenigen Sätze, in denen die zuvor als solche herausgestellten Kernwörter gehäuft auftreten, diese miteinander in Bezug setzen – und damit Sinn bildend für den gesamten Text sind. Diese Sätze sind somit geeignet, in den Abstract aufgenommen zu werden. Das Resultat dieses Vorgangs wird demnach als *Autoabstract* bezeichnet.

Dieser simple Ansatz ist verständlicherweise leichter beschrieben als maschinell implementiert. Die verschiedenen Ansätze zur Auswahl der für den Autoabstract relevanten Passagen des Originaltextes, ebenfalls beginnend in den späten 1950er Jahren, gehen weit über den in dieser Arbeit behandelten Themenbereich hinaus. An dieser Stelle soll daher genügen, dass dem Verfassen des Autoabstracts notwendigerweise die automatische Suche nach den Kernwörtern des Textes vorausgeht.

1.2 Kernwortsuche

Nachfolgend werden exemplarisch einige Methoden der automatischen Kernwortsuche kurz umrissen. Eine genauere Abschätzung der Leistungsfähigkeit würde implizieren, sämtliche Algorithmen zu implementieren und auf geeignete Testtexte anzuwenden. Dies würde den Umfang dieses Kapitels und letztlich dieser Arbeit jedoch sprengen, zumal die im weiteren Verlauf verwendeten Testtexte (vor allem ohne umgebende Textsammlung ähnlicher Thematik und aufgrund ihrer Kürze) nicht als Eingabe für die beschriebenen Methoden dienen können. Ein konkreter Vergleich wäre daher *a priori* schwerlich möglich.

1.2.1 Worthäufigkeit (H.P. Luhn)

H.P. Luhn muss zu den Pionieren des Autoabstractings gezählt werden. Er beschäftigte sich bereits gegen Ende der 1950er Jahre mit der automatischen Verarbeitung von Texten. Sein Hauptaugenmerk liegt dabei auf der Häufigkeit der Wörter innerhalb des zu verarbeitenden Textes. Er beschreibt dies wie folgt:

> *»Die Rechtfertigung dafür, die Wichtigkeit von Wörtern an ihrer Verwendungshäufigkeit zu messen, basiert auf der Tatsache, dass ein Autor normalerweise bestimmte Wörter wiederholt, während er seine Argumente entwickelt oder variiert, und während er einen Aspekt des Themas ausarbeitet. Dieses Mittel der Betonung kann als Indikator der Wichtigkeit verwendet werden.«*
>
> [LUH1958] *(Übers. d. A.)*[8]

Die Häufigkeitsverteilung von Wörtern in der Sprache wird grob durch das empirische *Zipfsche Gesetz* nach dem Mathematiker George K. Zipf (1902-1950) beschrieben. Sortiert man alle Wörter eines *Textkorpus*[9] absteigend nach Ihrer Häufigkeit und nummeriert die Wörter in aufsteigender Folge, sodass das häufigste Wort den Rang 1 erhält, das nächst weniger häufige Wort den Rang 2 usw., und bezeichnet man die Häufigkeit eines Wortes w mit $h(w)$ sowie dessen Rang mit $r(w)$, so gilt:

$h(w) \cdot r(w) \approx const$ (z.B. in [FER2003], S. 67)

Anders ausgedrückt: Natürlich ist die Rangplatzierung eines Wortes um so niedriger, je größer seine Häufigkeit ist. Wichtig ist jedoch, dass sich Rangplatzierungen und Häufigkeiten der jeweiligen Wörter in etwa linear gegensätzlich zueinander entwickeln. Umgekehrt gilt, dass seltene Wörter nur einen kleinen Anteil des Textkorpus ausmachen, während häufige Wörter den größten Anteil des Korpusvolumens bilden.

8 Originaltext: *"The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance."* [LUH1958], S. 160

9 Ein Textkorpus (bzw. kurz „Korpus") stellt allgemein eine Textsammlung von mitunter beträchtlichem Umfang dar. Dabei könnte es sich beispielsweise um einen oder mehrere komplette Jahrgänge einer Zeitschrift handeln oder um den Gesamttext eines mehrbändigen Lexikons. Bei einem korpusbasierten Experiment geht üblicherweise dessen Gesamtumfang in den Versuch ein, ohne die Einzeltexte im Detail zu betrachten. Wird für ein Experiment ein entsprechend diversifizierter Korpus verwendet, können die daraus gewonnenen Beobachtungen die Basis für empirische Schlussfolgerungen über Eigenschaften der (Schrift-)Sprache im Allgemeinen sein.

Ferber schreibt dazu:

> *»Nimmt man nun (unrealistischerweise) an, dass jedes einzelne Wort im Korpus in etwa gleich verteilt ist, zeigt sich, dass wegen der starken Häufigkeitsunterschiede die häufigen Wörter in fast jedem Textteil erwartet werden können. Andererseits treten seltene Wörter nur in sehr wenigen Texten auf. Häufige Terme sind also keine guten Suchterme, weil sie nicht spezifisch für einen Text sind. Bei seltenen Termen kann man nicht erwarten, dass sie in allen relevanten Texten vorkommen. Übrig bleiben bei der Suche nach geeigneten Suchtermen die Terme mittlerer Häufigkeit [...].«*
>
> [FER2003], S. 67

Dies ist zugleich Luhns empirische Feststellung, wonach Wörter mit großer Häufigkeit kaum Relevanz hinsichtlich des Textsinns enthalten, ebenso wenig Wörter mit sehr geringer Häufigkeit.

Die Entfernung sehr häufiger Wörter kann, wie Luhn vorschlägt, durch den Vergleich mit einer Liste der häufigsten Wörter der verwendeten Sprache d.h. mit einer sog. „Stoppwort-Liste“ geschehen. Alternativ kann die Worthäufigkeit der einzelnen Wörter analysiert werden, wodurch in einem Durchgang sowohl sehr häufige als auch sehr seltene Wörter entfernt werden. Geeignete Ober- und Untergrenzen für die minimale und maximale Worthäufigkeit sollen laut Luhn experimentell festgestellt werden.

Die rein lokale Bestimmung der Ober- und Untergrenzen kann leicht zu unerwünschten Effekten führen, etwa bei Spezialtexten mit besonderer Wortwahl. Das Zipfsche Gesetz beschreibt zuverlässiger die Worthäufigkeitsverteilung in größeren Textkorpora als in kleinen Wort-Vereinigungsmengen von mehr oder minder kurzen Texten. Daher bietet es sich an (vgl. [SMG1983], S. 61), die Worthäufigkeitsanalyse über eine Sammlung von thematisch verwandten Dokumenten durchzuführen – etwa der Texte, die aktuell indexiert werden sollen. Das heißt, die Entscheidung, ob ein Wort zu häufig oder zu selten vorkommt und daher aus der Betrachtung genommen wird, fällt nicht mehr anhand des gerade bearbeiteten Textes allein, sondern aufgrund aller Texte der Sammlung, in der der Text sich befindet. Das Gelingen dieser Methode wird durch die o.a. thematische Verwandtschaft der Texte gewährleistet.

Trotz dieser modifizierten Vorauswahl gilt für die weitere Verarbeitung der verbliebenen Wörter mit Luhns Methode, dass grundsätzlich die Worthäufigkeiten innerhalb des jeweils zu untersuchenden Textes als Merkmal der Wortwichtigkeit herangezogen wird. Die Aussage, ob ein Wort zu den wichtigeren oder unwichtigeren Wörtern des Textes zählt, liegt also im Text selbst – nicht in der umgebenden Textsammlung. Das Maß der Worthäufigkeit stellt also einen *lokalen Gewichtungsfaktor* dar.

Luhn nimmt in der Menge der verbliebenen Wörter mit mittleren Häufigkeiten an, dass wiederum die Wörter mit mittlerer Häufigkeit zu denen mit dem meisten Anteil an der Textbedeutung (*„resolving power of significant words*“, [LUH1958] S. 161) gehören. Auf diese Weise unterscheidet er wichtige von weniger wichtigen Wörtern.

Mithilfe der als wichtig klassifizierten Wörter bewertet Luhn im Anschluss daran zusätzlich die Wichtigkeit der einzelnen Sätze des Textes. Er verwendet dazu eine Methode, wie viele der wichtigen Wörter in geringem Wortabstand innerhalb eines Satzes vorkommen. Die wichtigsten Sätze nimmt er in den Autoabstract des Textes auf.

1.2.1.1 Gewichtete Worthäufigkeit

Die recht simple Worthäufigkeit beinhaltet das Problem, das sie ein absolutes und kein relatives Maß darstellt. Dieses Manko lässt sich beheben, indem man die absolute Häufigkeit jedes Wortes mit der absoluten Häufigkeit des häufigsten Wortes in Beziehung setzt. Sei w also ein beliebiges Wort des Textes, und sei w^{max} dasjenige Wort mit der maximalen Häufigkeit, so kann das Wichtigkeitsmaß z.B. wie folgt berechnet werden:

$$sig(w) = \frac{h(w)}{h(w^{max})}$$ (vgl. [FER2003], S. 70)

Die Wichtigkeit $sig(w)$ der Wörter des Textes („*sig*“ wie „significance“, engl. für „Wichtigkeit“) variiert somit zwischen 0 und 1. Die Worthäufigkeit wird somit von einem absoluten zu einem relativen Maß.

1.2.2 (Inverse) Dokumenthäufigkeit

Im Gegensatz zu der reinen Worthäufigkeit im Text oder im Korpus der umgebenden Textsammlung kann die *Dokumenthäufigkeit* betrachtet werden (vgl. [SMG1983], S. 63). Hierbei wird angenommen, dass die Wichtigkeit eines Wortes innerhalb einer Sammlung von thematisch verwandten Dokumenten von der Anzahl der Dokumente abhängt, in denen das Wort vorkommt. Hierbei wird die häufige Verwendung eines Wortes innerhalb eines Textes nicht betrachtet: Es ist unbedeutend, ob ein Wort einmalig oder mehrfach in einem der Dokumente auftritt.

Grundsätzlich wird die *Inverse Dokumenthäufigkeit*[10] wie folgt berechnet:

$$idf(w) = \frac{1}{d(w)}$$ mit $d(w)$ als Anzahl der Dokumente, in denen w enthalten ist.

Mitunter werden auch andere Berechnungsvorschriften verwendet, etwa

10 Die englische Bezeichnung lautet „*inverted document frequency*“, daher wird häufig auch die Bezeichnun „*invertierte Dokumenthäufigkeit*“ verwendet.

$$idf(w)=\ln\left(\frac{m}{d(w)}\right) \text{ oder } idf(w)=\ln\left(\frac{m-d(w)}{d(w)}\right) \text{ (vgl. [FER2003], S. 68f)}$$

mit m als Anzahl der Texte in der aktuellen Sammlung. Dadurch wird auch die (inverse) Dokumenthäufigkeit von einem absoluten zu einem relativen Maß, nämlich bezogen auf die Größe der aktuellen Textsammlung.

Ferber schreibt:

> *Diesen Formeln ist gemeinsam, dass ihr Wert mit wachsendem [...] [d(w)] monoton fällt. Der Logarithmus dämpft große Werte, schwächt also in diesen Formeln die Gewichte seltener Terme wieder etwas ab.*
>
> [FER2003], S. 69

Die inverse Dokumenthäufigkeit kann verwendet werden, um die Bedeutung von Termen zu gewichten. Offensichtlich gilt, dass ein Wort mit großer Dokumenthäufigkeit, das also in sehr vielen Dokumenten enthalten ist (entsprechend geringer *inverser* Dokumenthäufigkeit), in geringerem Umfang zum Sinn eines spezifischen Einzeldokuments beiträgt – und umgekehrt.

Das Maß der (inversen) Dokumenthäufigkeit stellt also einen *globalen Gewichtungsfaktor* zur Wortwichtigkeit dar.

1.2.3 Kombinierte lokale und globale Gewichtungseinflüsse

Lokale Gewichtungseinflüsse wie die Worthäufigkeit im jeweils betrachteten Text und globale Gewichtungseinflüsse wie die (inverse) Dokumenthäufigkeit lassen sich kombinieren. Eine häufig anzutreffende Berechnung der Wortwichtigkeit, die in vielen Systemen und Untersuchungen eingesetzt wurde, lautet

$$sig_i(w)=h_i(w)\cdot idf(w) \text{, z. B. } sig_i(w)=\frac{h_i(w)}{d(w)} \text{ (vgl. [FER2003], S. 71).}$$

$sig_i(w)$ berechnet die Wichtigkeit des Wortes w in Text i aus der Worthäufigkeit des Wortes w in Text i, multipliziert mit der inversen Dokumenthäufigkeit von w, d.h. bezogen auf alle Texte der aktuellen Textsammlung.

Natürlich können sowohl für die lokale Worthäufigkeit als auch für die globale Dokumenthäufigkeit relative statt absolute Gewichtungsfaktoren verwendet werden. Wird beispielsweise die absolute Häufigkeit des Wortes w in Text i mit der maxima-

len Häufigkeit des häufigsten Wortes w^{max} in Text i in Bezug gesetzt, und wird die inverse Dokumenthäufigkeit mit der Anzahl der Dokumente m in Bezug gesetzt, ergibt sich beispielsweise:

$$sig_i(w) = \frac{h_i(w)}{h_i(w^{max})} \cdot \ln\left(\frac{m - d(w)}{d(w)}\right)$$

Grundsätzlich werden derartige kombinierte Maße als *TF-IDF-Gewichtungen* bezeichnet, weil sowohl die Worthäufigkeit „TF" (für „term frequency") als auch die inverse Dokumenthäufigkeit „IDF" (für „inverted document frequency") eingehen.

1.2.4 Informationstheoretischer Ansatz (Signal vs. Rauschen)

Unter Zuhilfenahme der Informationstheorie nach Claude E. Shannon (1916-2001) lässt sich ein anderer Ansatz zur Bestimmung der Wichtigkeit eines Wortes definieren. Den Grundsatz bildet wiederum die Erkenntnis, dass der Informationsgehalt eines Wortes umgekehrt zur Wahrscheinlichkeit seines Auftretens in einem gegebenen Text variiert (vgl. [SMG1983], S. 63ff). Dies entspricht insofern bis hier lediglich dem lokalen Gewichtungseinfluss der absoluten Häufigkeit eines Wortes in einem Text.

Ein Wort trägt zum Abbau der *Unsicherheit* über den Inhalt des Textes bei („*uncertainty*", vgl. [SHA1951]): Je häufiger es vorkommt, desto schwächer ist sein Anteil am Inhalt; je seltener es vorkommt, desto stärker ist sein Anteil. Übereinstimmend damit wird in der Informationstheorie der Informationsgehalt $inf(w)$ eines Worts w mit der Formel

$$inf(w) = -\log_2 p(w)$$

berechnet, wobei $p(w)$ der Auftrittswahrscheinlichkeit des Wortes entspricht, d.h. seiner absoluten Wahrscheinlichkeit in einer Textsammlung. — Zur Erinnerung: Dies entspricht nicht der Dokumenthäufigkeit $d(w)$, bei der ein mehrfach innerhalb eines Textes enthaltenes Wort nur einmalig gezählt wird.

Liegt eine Menge von t Wörtern vor, die einen Text als Kernwörter repräsentieren sollen, so trägt jedes dieser Wörter anteilig am Abbau der o.a. Unsicherheit bei. Laut Informationstheorie (vgl. [SHA1951]) lässt sich über die Wortmenge wie folgt der durchschnittliche Informationsgehalt $\overline{inf}$ der Wörter w_1 bis w_t berechnen:

$$\overline{inf} = -\sum_{i=1}^{t} p(w_i) \log_2 p(w_i)$$

Der durchschnittliche Informationsgehalt wird offensichtlich genau dann maximal, wenn alle t betrachteten Wörter mit Wahrscheinlichkeit $1/t$ vorkommen.

Analog dazu lässt sich ein *„Rausch-Maß"* (englisch: „noise"*)* für ein Wort definieren. Die Berechnungsvorschrift lautet bei Salton & McGill (vgl. [SMG1983], S. 65):

$$Noise(w)=\sum_{i=1}^{n}\frac{h_i(w)}{h_*(w)}\log_2\frac{h_*(w)}{h_i(w)} \text{ mit } h_*(w)=\sum_{i=1}^{n}h_i(w)$$

Die Häufigkeit $h_i(w)$ des Wortes w in Text i wird dabei ins Verhältnis gesetzt mit der Summe der Häufigkeiten $h_*(w)$ in allen Texten der Textsammlung.

Ist $h_i(w)$ für alle Texte gleich, d.h. kommt das Wort w in allen n Texten gleich häufig mit $h_i(w)=h_*(w)/n$ vor, so wird $Noise(w)=\log_2 h_*(w)$ maximal. Kommt w aber nur in einem Text (o.B.d.A. Text x) mit $h_x(w)=h_*(w)$ vor, wird $Noise(w)=0$.

Invers zum Rauschen eines Wortes ist offensichtlich seine *Signalstärke*, d.h. sein Potenzial zur Verringerung der o.a. Unsicherheit über den Inhalt des Textes. Die Berechnung der Signalstärke $Signal(w)$ erfolgt simpel durch Subtraktion der Rauschstärke $Noise(w)$ von der maximalen Rauschstärke $\log_2 h_*(w)$. Somit gilt:

$$Signal(w)=\log_2 h_*(w)-Noise(w)$$

Je nach Anwendungsfall kann die globale Signalstärke eines Wortes direkt zur Bestimmung seiner Wichtigkeit verwendet werden. Alternativ kann wiederum der lokale Einfluss der Worthäufigkeit innerhalb eines Textes mit einbezogen werden, während die global über die gesamte Textsammlung berechnete Signalstärke die Worthäufigkeit lediglich gewichtet. Somit ergäbe sich

$$sig_i(w)=h_i(w)\cdot signal(w).$$

1.2.5 Textunterscheidung anhand wichtiger Wörter

In den vorangegangenen Abschnitten wurde als Wichtigkeit eines Wortes seine Fähigkeit bewertet, einen Text zu repräsentieren. Aus anderem Blickwinkel betrachtet gilt: Ein Wort, das einen Text thematisch „ausmacht", ist andererseits ebenfalls ein Wort, das diesen Text thematisch von anderen Texten unterscheidet. Anders ausgedrückt: Als Kernwörter eines Textes eignen sich diejenigen Wörter besonders gut, die ihn von anderen Texten unterscheidbar machen.

Das im Folgenden lediglich ansatzweise vorgestellte Verfahren (in enger Anlehnung an [SMG1983], S. 66ff) basiert auf einer Sammlung von (o.B.d.A.) m Dokumenten. Es sei angenommen, dass für jeden Text eine Menge von Kernwörtern bereits ausgewählt worden ist, d.h. für jeden Text i existiert eine Menge von Kernwörtern D_i.

Sei $sim(D_i, D_j)$ ein *Ähnlichkeitsmaß* („sim" wie „similar", engl. für „ähnlich"), das die Ähnlichkeit zweier Texte $i, j; i \neq j$ anhand ihrer beiden D_i, D_j berechnet. Sind die beiden Kernwortmengen identisch, so sind die Texte vollständig identisch, damit sei $sim(D_i, D_j) = 1$. Sind die beiden Kernwortmengen vollständig unterschiedlich, so sind die Texte vollständig unterschiedlich, damit sei $sim(D_i, D_j) = 0$.

Die genaue Spezifikation des Ähnlichkeitsmaßes sim kann je nach Einsatz variiert werden, sofern es der genannten Definition entspricht. Häufig werden dafür vektorbasierte Entfernungsmaße verwendet (vgl. [FER2003], S. 72ff). Denkbar wäre etwa der Einsatz einer ungewichteten Variante des im späteren Verlauf dieser Arbeit beschriebenen Maßes der *Gewichteten Euklidischen Qualität*, das n-dimensionale euklidische Abstände zwischen Wortmengen berechnet und als Qualitätsmaß heranzieht.

Werden alle Texte der Textsammlung paarweise miteinander auf Ähnlichkeit untersucht, so lässt sich die *mittlere Ähnlichkeit* $\overline{sim}$ herleiten. Sie berechnet sich wie folgt:

$$\overline{sim} = c \cdot \sum_{i=1}^{m} \sum_{j=1}^{m, i \neq j} sim(D_i, D_j) \text{ mit } c = \frac{1}{m \cdot (m-1)}$$

Die Konstante c dient der Normierung: Die „doppelte Summe" wird bei maximaler durchschnittlicher Ähnlichkeit höchstens $m \cdot (m-1)$. Durch die Division durch exakt diese Größe variiert die mittlere Ähnlichkeit ebenfalls zwischen 0 und 1.

Das Maß der durchschnittlichen Ähnlichkeit symbolisiert die paarweise thematische Verwandtschaft aller Texte untereinander, sozusagen die „thematische Dichte" der Sammlung um ein angenommenes „zentrales Dokument". Ein solches Zentraldokument bzw. dessen Menge von Kernwörtern lässt sich derart künstlich generieren, dass jedes Wort der Menge in allen m Dokumenten durchschnittlich häufig vorkommt. Die durchschnittliche Häufigkeit $\overline{h}(w)$ eines Wortes w ergibt sich wie folgt:

$$\overline{h}(w) = \frac{1}{m} \sum_{i=1}^{m} h_i(w)$$

Mithilfe derart selektierter Wörter der Textsammlung ergibt sich ein künstliches *Zentraldokument* $\overline{D}$, und die Berechnung der mittleren Ähnlichkeit vereinfacht sich zu

$$\overline{sim} = c \cdot \sum_{i=1}^{m} sim(D_i, \overline{D}) \text{ mit } c = m \text{ (wiederum zur Normierung).}$$

Sei nun $\overline{sim_w}$ diejenige durchschnittliche Ähnlichkeit der Textsammlung, wenn das Wort w aus allen Texten, d.h. aus deren Kernwortmengen entfernt würde. Wäre w ein Wort, das in allen Texten und deren Kernwortmengen gleichermaßen enthalten wäre, so würde seine Entfernung die durchschnittliche Ähnlichkeit der Dokumente offensichtlich verringern, d.h. die Dokumente „unähnlicher machen“: Da w nicht gut als Unterscheidungskriterium zwischen den Texten geeignet ist, erhöht es bei Verwendung als Kernwort die durchschnittliche Dokumentähnlichkeit. Dies ist hinsichtlich der Unterscheidbarkeit der Dokumente anhand ihrer Kernwörter kaum wünschenswert. Also sollte ein derartiges w nicht als Kernwort verwendet werden.

Entgegengesetzt betrachtet ist es offensichtlich sinnvoll, Kernwörter zu verwenden, deren Einfluss auf die durchschnittliche Dokumentähnlichkeit genau umgekehrt ist: Ein gutes Kernwort soll die durchschnittliche Dokumentähnlichkeit möglichst verringern, d.h. die Dokumente „unähnlicher machen“. Klarer ausgedrückt: Die Verwendung eines guten Kernworts verringert die Ähnlichkeit der Dokumente, die eines schlechten erhöht sie.

Für jedes Wort der Textsammlung lässt sich durch sein testweises Entfernen sein Einfluss auf die durchschnittliche Dokumentähnlichkeit bestimmen. Sie wird wortweise wie folgt berechnet:

$$\Delta\,\overline{sim_w} = \overline{sim_w} - \overline{sim}$$

Das Maß $\Delta\,\overline{sim_w}$ eines Wortes w symbolisiert den Einfluss des Wortes auf die durchschnittliche Dokumentähnlichkeit, wenn es als Kernwort verwendet würde: Ist der Wert für das Wort w negativ, so erhöht das Wort die durchschnittliche Dokumentähnlichkeit (es macht die Dokumente „ähnlicher“) – es sollte daher nicht als Kernwort verwendet werden. Ebenso wenig sind Wörter geeignet, deren $\Delta\,\overline{sim_w}$ etwa 0 ist, d.h. Wörter, die kaum Einfluss auf die Dokumentähnlichkeit haben. Nur Wörter, deren $\Delta\,\overline{sim_w}$ positiv ist, d.h. die positiven Einfluss auf die Dokumentähnlichkeit haben, und somit die Dokumente „unähnlicher machen“, kommen als Kernwörter in Betracht.

Wiederum gilt: Dieses globale Maß kann je nach Anwendungsfall direkt zur Bestimmung der Wichtigkeit von Wörtern verwendet werden. Es bietet sich aber ebenfalls an, den lokalen Einfluss der Worthäufigkeit der Wörter mit einzubeziehen, also etwa

$$sig_i(w) = h_i(w) \cdot \Delta\,\overline{sim_w}\,.$$

1.3 Zusammenfassung und Ausblick

Eine grundsätzliche Schwäche des rein textbezogenen Maßes der Worthäufigkeit ist, dass für die Analyse nur Texte geeignet sind, die eine hinreichende Länge haben. Bei zu kurzen Texten ergeben sich oft nur sehr geringe, mitunter willkürlich variierende Worthäufigkeiten. Es hat sich weiterhin gezeigt, dass die Häufigkeit eines Wortes allein nicht ausreicht, um seine Wichtigkeit und letztlich seine Eignung als Kernwort des Textes zu bestimmen, in dem es enthalten ist.

Darüber hinaus ist für alle hier vorgestellten globalen Maße, die zur Gewichtung der Worthäufigkeit geeignet sind (die Dokumenthäufigkeit, die Signalstärke oder der Einfluss auf die Dokumentähnlichkeit), eine Sammlung von thematisch verwandten Texten erforderlich, um sie zu berechnen.

Ziel dieser Arbeit ist es, ein Verfahren zur Kernwortsuche zu diskutieren, das auch relativ kurze Texte (z.B. die Inhalte von WWW-Seiten) ohne eine umgebende Sammlung von thematisch verwandten Texten verarbeiten kann. Globale Gewichtungseinflüsse sollen unabhängig von den gerade zu verarbeitenden Texten sein, d.h. nur auf entsprechend diversifiziert vorliegenden, großen Textkorpora basieren.

2 Wortassoziationen und Assoziative Wortnetze

Eine *Assoziation* stellt den Zusammenhang zwischen einem *Stimulus* und einer *Reaktion* darauf dar: Nimmt ein Mensch zum Beispiel einen Stimulus in Form eines Wortes wahr, *assoziiert* er daraufhin spontan eines oder mehrere andere Wörter. Diesbezügliche Experimente gehen zurück auf den Assoziationsversuch von Sir Francis Galton (1822-1911) aus dem Jahre 1880 (vgl. [GAL1880]), bei dem Versuchsteilnehmer aufgefordert wurden, zu einem vorgegebenen Stimuluswort jeweils dasjenige Wort zu nennen, das ihnen als erstes einfiel – d.h. das Wort, das sie am stärksten mit dem Reizwort assoziierten.

Assoziation ist jedoch nicht auf Wortassoziation beschränkt, vielmehr können beliebige Stimuli und beliebige Reaktionen an einer Assoziation beteiligt sein. Ein prominentes, häufig zitiertes Beispiel ist Iwan P. Pawlows (1849-1936) Versuchsreihe zur *Klassischen Konditionierung* (vgl. z.B. [ZIM1992], S. 231), bei denen er seinen Laborhunden Glockenklang und Futtergabe als Stimuli präsentierte, deren sichtbare Reaktion auf die ausgebildete Assoziation dieser beiden Reize daraufhin in der einsetzenden Speicheltätigkeit bestand.

Der Mechanismus, der zu einer Assoziation führt, ist immer derselbe. William James (1842-1910) fasste bereits 1890 seine Beobachtungen des menschlichen Assoziationsvermögens als *Gesetz der mentalen Assoziation* bzw. einfacher als *Assoziationsgesetz* zusammen:

> *»[Zwei] Objekte, die gemeinsam wahrgenommen werden, neigen dazu, in der Vorstellung miteinander assoziiert zu werden, so dass, sobald an eines der beiden gedacht wird, höchstwahrscheinlich auch an das andere gedacht wird – wie zuvor in derselben Reihenfolge des Auftretens oder gleichzeitig auftretend. Diese Aussage bezeichnen wir als „Gesetz der mentalen Assoziation der Nähe".«*
>
> *[JAM1890], (Übers. d. A.)*[11]

Die gemeinsame Wahrnehmung von zwei Wörtern führt demnach zur Verstärkung einer Assoziation der beiden Wörter. Assoziationen sind also keinesfalls angeboren, sondern werden mit der Zeit durch fortwährendes, gemeinsames Wahrnehmen von Wörtern erlernt.

11 Originaltext: *"Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity."* [JAM1890], S. 561

Menschen, die in einem ähnlichen Umfeld aufwachsen, erlernen zu gebräuchlichen Wörtern (bzw. damit verknüpften Konzepten) auch ähnliche Assoziationen. Ein Beispiel: Die Präsentation des Wortes „Küche“ führt bei vielen Menschen zu der Erstassoziation „Herd“. Es werden aber zusätzlich dazu noch weitere Wörter assoziiert, etwa „Kochen“, „Spüle“, „Kühlschrank“, „Esstisch“. Die Liste der assoziierten Beispielwörter, deren Reihenfolge sowie deren Assoziationsstärken basieren auf ihrem semantischen Zusammenhang im nordeuropäischen Habitat („kollektives Wissen“), manifestiert durch das vorher erworbene Weltwissen der assoziierenden Person.

Andererseits können schon geringfügige Unterschiede in den Umfeldern von Personen zu gravierend unterschiedlichen Assoziationsstärken führen: Passend zum o.a. Beispiel würde ein Kind, dessen Eltern passionierte Teetrinker sind, das Wort „Kaffeemaschine“ nicht oder nur weitaus schwächer mit dem Reizwort „Küche“ assoziieren, als ein Kind, dessen Eltern regelmäßig Kaffee kochen und konsumieren.

2.1 Assoziatives Lernen

Bezogen auf das soeben zitierte James'sche Assoziationsgesetz müssen zwei Objekte „gemeinsam wahrgenommen werden“, damit sie miteinander assoziiert werden können. Bei Assoziationen zwischen beliebigen wahrnehmbaren Stimuli bezeichnet dies beispielsweise die gleichzeitige visuelle Wahrnehmung bzw. zumindest in geringem zeitlichen Abstand. Dies lässt sich auf die Wahrnehmung mit anderen Sinnen übertragen, z.B. das Erklingen einer Glocke und kurze Zeit später die Futtergabe: Glockenklang und der Geruch bzw. der Geschmack des Futters werden also „gemeinsam wahrgenommen“. Bei Wortassoziationen ist dieser Vorgang schwerer zu umschreiben: Wann werden zwei Wörter gemeinsam wahrgenommen?

Es ist unumgänglich, die „gemeinsame Wahrnehmung“ im Rahmen der Textwahrnehmung bzw. des Textverständnisses zu definieren. In der Wahrnehmungspsychologie hat sich hinsichtlich der gesprochenen oder gedruckten Sprache der Begriff der *Fensterlänge* etabliert. Sie hat ihr lernpsychologisches Pendant in der sog. *Psychischen Präsenz* nach William Stern (1871-1938) von 1897, die erstmals Hermann Ebbinghaus (1855-1909) im Jahre 1885 in seinen Versuchen zur unmittelbaren Reproduktion von Folgen erforschte (vgl. [FOP1965], S. 243ff). Menschen können, ohne gesondertes Training, direkt nach der Präsentation eine Folge von etwa sechs bis acht Elementen fehlerfrei reproduzieren – d.h. etwa sieben Elemente sind kurzzeitig gemeinsam psychisch präsent. Dies gilt jedoch nur für unsinnige Folgen (etwa von Nonsens-Silben oder Einzelbuchstaben): Für sinnvolle Wortfolgen steigt die Leistung auf 10-20 Wörter.

Augenscheinlich gilt, dass nur diejenigen Wörter zu Wortassoziationen führen können, die beim Hören oder Lesen *gemeinsam* psychisch präsent sind. Die von James in seinem Assoziationsgesetz geforderte „gemeinsame Wahrnehmung" lässt sich also bei der Wortassoziation auf nahes gemeinsames Auftreten innerhalb eines gesprochenen oder gedruckten Textes zurückführen. Ein solches nahes gemeinsames Auftreten wird in der Kognitiven Psychologie als *Kookurrenz*[12] von zwei Wörtern bezeichnet.

2.1.1 Konnektionistische Sichtweise

Wie könnte der assoziative Lernprozess physiologisch realisiert sein? Wie lässt sich dieser Lernprozess und eine anschließende „Abfrage", d.h. die Präsentation eines Reizwortes und die Abfrage der assoziierten Wörter auf einem Rechner simulieren?

Ein stimuliertes Wort kann andere Wörter assoziieren, oder anders gesagt *aktivieren*. Die Weiterleitung eines Stimulus über unterschiedliche Assoziationsstärken leitet hin zu einer neuronalen Sichtweise: Modellhaft lässt sich jedes Wort durch ein einzelnes *Neuron* repräsentiert verstehen.

Neuronen bilden die kleinsten funktionalen Einheiten des Nervensystems, die für die Reizverarbeitung und -ausbreitung verantwortlich sind. Jedes einzelne Neuron ist mit teilweise bis zu tausenden anderer Neuronen verknüpft, zusammen bilden sie ein *Neuronales Netzwerk*. Stark vereinfacht gesagt: Ein Neuron besitzt eine Vielzahl von „Eingängen", die *Dendriten*, sowie einen „Ausgang", das *Axon*. Vom Axon eines Neurons aus besteht durch *Synapsen* der Kontakt zu den Dendriten anderer Neuronen. Die Reizleitung zwischen den Neuronen ist ein elektrochemischer Prozess: Jedes Neuron besitzt ein Ruhepotenzial, die über seine Dendriten empfangenen Reize wirken entweder hemmend oder erregend darauf. Übersteigt das Potenzial des Neurons bzw. anders gesagt dessen *Aktivierung* eine bestimmte Schwelle (die sog. *Depolarisationsschwelle*), wird das Neuron mit Hilfe seines Axons sein *Aktionspotenzial* über die Synapsen an die Dendriten der verknüpften Neuronen weiterleiten. Dort läuft derselbe Vorgang auf identische Weise zur selben Zeit ab. Ein entsprechend „verschaltetes" Neuronales Netz vollbringt auf diese Weise sozusagen eine biologische Informationsverarbeitung.

Der assoziative Lernprozess muss mit einer neuronalen „Einstellung" der Aktivierungsausbreitung zwischen den einzelnen Nervenzellen einher gehen. 1949 formulierte Donald O. Hebb (1904-1985) das nach ihm benannte *Hebbsche Postulat*:

12 Im Bereich der Computerlinguistik spricht man statt von „Kookurrenzen" häufig von „Kollokationen". Eine Kollokation beschreibt aber eher zwei direkt aufeinander folgende Wörter, während eine Kookurrenz zweier Wörter auch bei definiertem Wortabstand gegeben ist.

> *»[...] Wenn [...] [eine Nerven-]Zelle A [...] wiederholt oder kontinuierlich an der Erregung einer [Nerven-}Zelle B beteiligt ist, findet ein Wachstumsprozess bzw. eine Stoffwechsel-Veränderung in einer oder in beiden Zellen statt, so dass die Effizienz von A gestärkt wird, B [...] [zu erregen].«*
>
> *[HEB1949] (Übers. d. A.)*[13]

Die benannte „Effizienz von A, [...] B zu erregen" entspricht bezogen auf die Anwendung zur Wortassoziation exakt der Stärke, mit der das Reizwort A ein anderes Wort B assoziiert. Die von Hebb beschriebene Stoffwechsel-Veränderung stellt daher den Lernprozess dar, den das *Neuronale Netz* vollzieht, wenn die zwei Wörter A und B häufig gemeinsam wahrgenommen werden, d.h. wenn sie *kookurrieren*. Das Hebbsche Postulat liefert somit das neuronale Pendant zum weiter oben zitierten James'schen Assoziationsgesetz (S. 22).

2.1.2 Inkrementelles Lernen nach Estes

Das soeben zitierte Hebbsche Postulat, das sich auf reale neuronale Prozesse bezieht, inspirierte verschiedene neuronale Lerntheorien. Dem in dieser Arbeit verwendeten Softwarepaket „ATA2" (mehr dazu ab Kapitel 2.2 auf S. 37, vgl. [BÖH1997]) liegt die von William K. Estes in den 1950er Jahren entwickelte „Stimulus Sampling Theorie" (SST)[14] zu Grunde. Deren theoretische Grundlagen sowie ihre Anwendbarkeit werden für die Wortassoziation soll nachfolgend grob umrissen.

2.1.2.1 Stimulus Sampling Theorie

Klaus Foppa stellt Estes' Theorie folgendermaßen dar:

> *»Lernen besteht für Estes in einer Verknüpfung situativer (S-) und reaktiver (R-) Merkmale. Unter Reizen werden [...] Umweltbedingungen, die ohne Bezug auf das Verhalten des Organismus zu beschreiben sind [...], verstanden. Reaktionen sind alle beobachteten Verhaltensweisen des Individuums, die, je nach Bedarf, in sich wechselseitig ausschließenden, erschöpfenden Reaktionsklassen erfasst werden [...].«*
>
> *[FOP1965] (S. 371)*

13 Originaltext: *"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."* [HEB1949], S. 62

14 Die „Stimulus Sampling Theorie" wird häufig auch als „Stimulus-Auswahl-Theorie" oder „Reiz-Auswahl-Theorie" bezeichnet.

Gemäß dem oben zitierten James'schen Assoziationsgesetz bildet sich im Verlauf des Lernprozesses eine Assoziation zwischen den wahrgenommenen Umweltbedingungen bzw. Reizen aus. Ob die Assoziation gebildet wurde oder nicht, manifestiert sich in der Reaktion.

Zur Verdeutlichung eines solchen assoziativen Zusammenhangs soll hier erneut die Pawlowsche *Klassische Konditionierung* (vgl. [ZIM1992], S. 231) genannt werden. Zur Erinnerung: Im bekanntesten Versuch sollten Pawlows Laborhunde eine Assoziation zwischen Glockenklang und Futtergabe lernen. Der Stimulus „Glocke“ bildet den unkonditionierten Reiz, der Stimulus „Futter“ bildet den konditionierten Reiz. Wenn der Hund häufig genug beide Stimuli gleichzeitig wahrgenommen hatte, bildete sich eine Assoziation, und der Hund erwartete den Stimulus „Futter“ bei alleiniger Darbietung des Stimulus„Glocke“. Die Manifestation dieser Assoziation erfolgte über das beobachtbare Einsetzen der Speicheltätigkeit, wenn der Hund diese Assoziation gebildet hatte.

Estes definiert die *Umwelt*, d.h. die Situation, die das Individuum beim Lernen umgibt, als Vereinigungsmenge von (o.B.d.A.) N möglichen situativen Merkmalen. Dabei handelt es sich einerseits um ganz offensichtliche Eigenschaften der Umwelt (z.B. die Farbe des Fußbodens oder die Größe des Versuchslabors), aber auch um die untersuchten Reize (z.B. der Glockenklang oder die Futtergabe). Das untersuchte Ereignis (z.B. Assoziation), das durch eine eintretende oder ausbleibende Reaktion klassifiziert wird (z.B. die Speicheltätigkeit des Versuchstiers), ist also offensichtlich die Folge aus wiederholt gemeinsam wahrgenommenen Umweltmerkmalen.

Laut Estes wird von der N-elementigen Umwelt zu einem bestimmten Zeitpunkt t nur eine s-elementige Teilmenge bzw. *Stichprobe* der situativen *Merkmale* wahrgenommen. Jedes Merkmal geht mit einer festen Wahrscheinlichkeit θ in die Stichprobe ein[15]. Die durchschnittliche Größe der Stichprobe beträgt somit $N \cdot \theta$.

Das SST vertritt die „Alles oder Nichts“-Theorie, d.h. bereits das einmalige Eintreten des Ereignisses E auf ein Merkmal M hin erzeugt eine Verbindung zwischen M und E. Ebenso löst bereits einmaliges Nicht-Eintreten die Verbindung zwischen M und E und erzeugt eine Verbindung zwischen M und dem entgegengesetzten Ereignis $\bar{E}$.

Betrachtet wird die Wahrscheinlichkeit p, mit der ein beliebiges Merkmal mit dem Ereignis E verbunden ist. Die Merkmale als solche werden in dieser theoretischen Betrachtung nicht unterschieden: Jedes *beliebige* Merkmal ist mit einer bestimmten Wahr-

15 Es gibt eine weitere Betrachtungsweise, wonach die Auswahl der Stichprobe als Zufallsexperiment der Ziehung von s Kugeln aus einer Urne mit N Elementen ohne Zurücklegen verstanden wird, die hier aber nicht relevant ist.

scheinlichkeit mit dem untersuchten Ereignis verbunden. Im Gegensatz dazu gilt die Gegenwahrscheinlichkeit $\bar{p}=1-p$, wonach ein beliebiges Merkmal mit dem inversen Gegenereignis $\bar{E}$ verbunden ist.

Die Wahrscheinlichkeiten p und $\bar{p}$ teilen die N-elementige Umwelt disjunkt nach ihrer Verbindung zu E bzw. $\bar{E}$ auf. Ein Beispiel ist in Abbildung 2 dargestellt: Ausgefüllte Merkmale sind mit E, nicht-ausgefüllte Merkmale mit $\bar{E}$ verbunden.

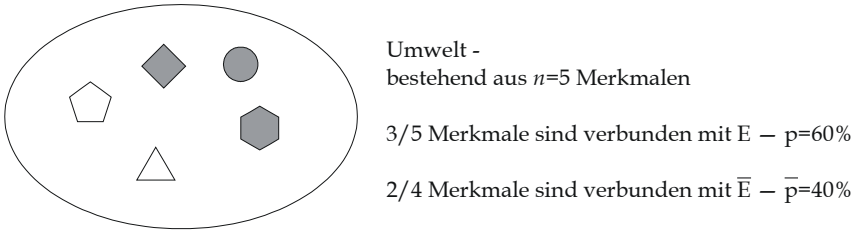

Abbildung 2: Umwelt nach Estes - Basismodell

p und $\bar{p}$ mit $p+\bar{p}=1$ verändern sich über die Zeit, d.h. im Verlauf der Lerndurchgänge. Offensichtlich gilt: Das Verhältnis der Merkmale $p/\bar{p}$, die mit dem Ereignis E oder dem Gegenereignis $\bar{E}$ verbunden sind, verändert sich von Lerndurchgang zu Lerndurchgang, wenn wiederum das Ereignis E oder das Gegenereignis $\bar{E}$ auftritt. Ein einfaches Beispiel ist in Abbildung 3 abgebildet.

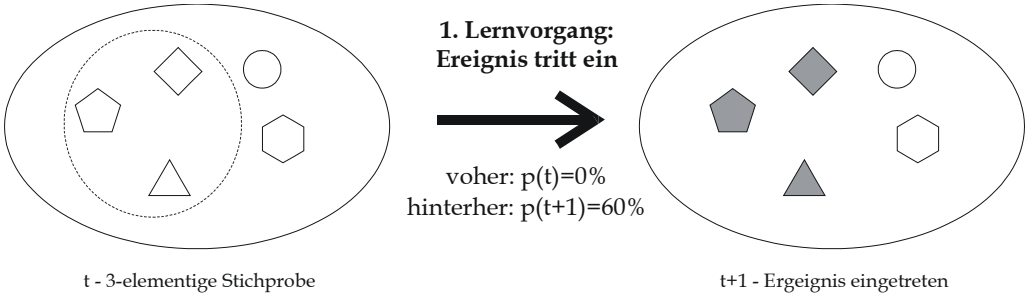

Abbildung 3: Lernen nach Estes - 1. Lernvorgang - Beispiel

Vor dem Lernvorgang war keines der fünf Merkmale mit dem Ereignis E verbunden, bzw. alle fünf waren mit dem Gegenereignis $\bar{E}$ verbunden. Danach sind alle drei Merkmale der Stichprobe mit E verbunden.

Allgemein bezeichnet $p(t)$ die Wahrscheinlichkeit im Zeitpunkt t kurz *vor* einem Lerndurchgang, $p(t+1)$ bezeichnet dagegen die Wahrscheinlichkeit im darauf folgenden Zeitpunkt $t+1$ kurz *nach* einem Lerndurchgang. — Zu jedem Zeitpunkt gilt $\bar{p}(t)=1-p(t)$.

Lernen

Tritt zum Zeitpunkt t das Ereignis E ein, folgt für das Lernen folgende iterative Berechnungsvorschrift:

$$p(t+1)=p(t)+\theta\cdot\bar{p}(t)$$

Das bedeutet: Mit der Wahrscheinlichkeit $p(t)$ sind die Merkmale dieser Stichprobe bereits mit E verbunden, d.h. für diese Merkmale ändert sich bei Eintreten von E nichts. Im Beispiel in Abbildung 4 sind dies die Merkmale „Fünfeck“ und „Raute“.

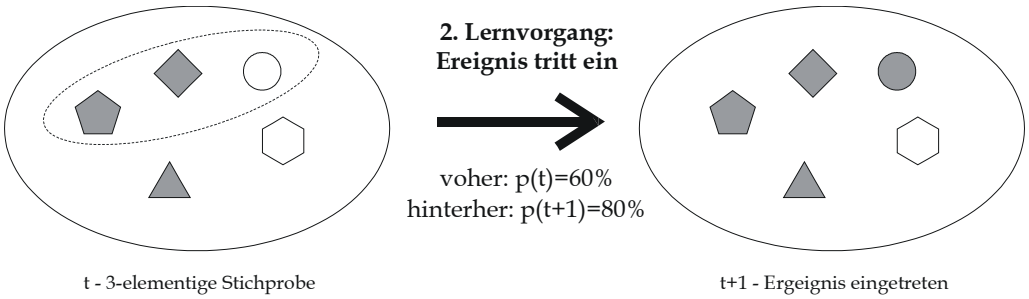

Abbildung 4: Lernen nach Estes - 2. Lernvorgang - Lernen

Mit der Wahrscheinlichkeit θ befinden sich jedoch auch Merkmale in der Stichprobe, die bislang mit der Gegenwahrscheinlichkeit $\bar{p}$ mit $\bar{E}$ verknüpft sind. Diese werden nun gemäß der „Alles oder Nichts“-Theorie ebenfalls mit E verbunden. Im Beispiel in Abbildung 4 ist dies das Merkmal „Kreis“.

Mit $\bar{p}(t)=1-p(t)$ ergibt sich als *Lernregel:* $p(t+1)=p(t)+\theta\big(1-p(t)\big)$

Abschwächen / Hemmen

Tritt zum Zeitpunkt t das Gegenereignis $\bar{E}$ ein, folgt für das Abschwächen bzw. Hemmen folgende iterative Berechnungsvorschrift:

$$p(t+1)=\bar{\theta}\cdot p(t)+\theta\cdot 0$$

Analog zu oben beschreibt die Wahrscheinlichkeit $\bar{\theta}=1-\theta$, dass ein mit Ereignis E verbundenes Merkmal nicht Teil der Stichprobe ist. Für diese Merkmale ändert sich nichts. Im Beispiel in Abbildung 5 ist dies das Merkmal „Sechseck“.

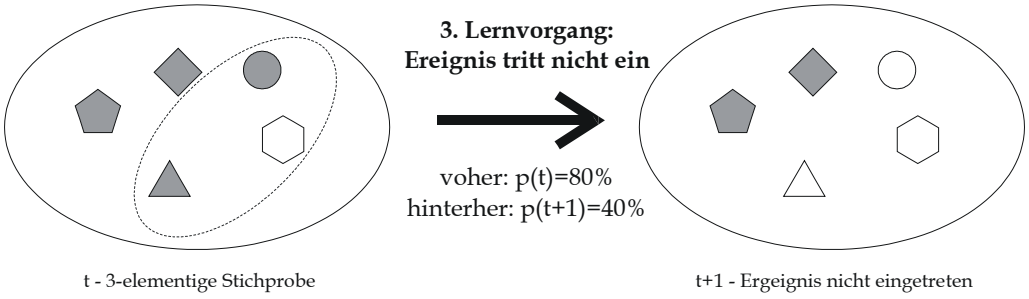

Abbildung 5: Lernen nach Estes - 3. Lernvorgang - Abschwächen

Die anderen Merkmale, die mit Wahrscheinlichkeit θ Teil der Stichprobe sind, müssen nun mit $\bar{E}$ verbunden werden, was innerhalb von p (d.h. der Wahrscheinlichkeit einer Verbindung mit E) 0 entspricht. Im Beispiel in Abbildung 5 sind dies die Merkmale „Dreieck“ und „Kreis“.

Mit $\bar{\theta}=1-\theta$ und $\theta\cdot 0=0$ ergibt sich als *Hemmungsregel:* $p(t+1)=(1-\theta)\cdot p(t)$

2.1.2.2 Anwendung der SST für Wortassoziationen

Durch die SST lässt sich nicht nur die Klassische Konditionierung nach Pawlow erklären, sondern auch auf das Erlernen von Wortassoziationen: Während die Klassische Konditionierung die Folge aus wiederholt gemeinsam wahrgenommenem unkonditionierten und konditionierten Reiz darstellt, stellt die Assoziation eines Wortes die Folge aus wiederholt gemeinsam wahrgenommenen Wörtern dar – konform mit dem weiter oben zitierten James'schen Assoziationsgesetz (siehe S. 22).

Um die soeben hergeleiteten iterativen Lern- und Hemmungsregeln für die Wortassoziation verwenden zu können, bedarf es einer Umdeutung der N-elementigen Umwelt und der $N \cdot \theta$-elementigen Stichproben, die ein Ereignis E oder das Gegenereignis $\bar{E}$ hervorrufen.

Sei V eine definierte Wortmenge, die fortan als *Vokabular* bezeichnet werden soll. Gesucht ist die Assoziationsstärke zwischen je zwei Wörtern des Vokabulars.

Im Sinne der SST handelt es sich um ein Ereignis, wenn ein Wort ein anderes Wort assoziiert. Die Assoziationsstärke von Wort i zu Wort j entspricht dabei für je zwei Wörter $w_i, w_j \in V$ der Wahrscheinlichkeit, mit der das Wort i das Wort j assoziiert.

Jede einzelne dieser Wortassoziationen bildet ein Ereignis, d.h. jedes dieser Ereignisse verfügt über eine gesonderte Wahrscheinlichkeit. Sie soll fortan nicht mehr mit p sondern mit ψ bezeichnet werden. Alle Assoziationsstärken variieren zwischen 0 bzw. 0% (keine Assoziation) und 1 bzw. 100% (stärkstmögliche Assoziation).

Die Assoziationsstärke zwischen je zwei Wörtern i und j mit $i \neq j$ soll mit ψ_{ij} bezeichnet werden. Im weiteren Verlauf sei angenommen, Assoziationsstärken seien *symmetrisch*, d.h. die Stärke, mit der Wort i das Wort j assoziiert, entspricht der Stärke, mit der Wort j das Wort i assoziiert. Formalisiert: $\psi_{ij} = \psi_{ji} \; \forall i, j$. Weiterhin sei angenommen, das kein Wort sich selbst assoziiert, d.h. $\psi_{ii} = 0 \; \forall i$.

Auf diese Weise lässt sich eine quadratische *Assoziationsmatrix* Ψ erstellen, die folgende Form hat:

$$\Psi = \begin{pmatrix} 0 & \psi_{12} & \psi_{13} & \cdots & \cdots & \psi_{1n} \\ \psi_{12} & 0 & \psi_{23} & \cdots & \cdots & \psi_{2n} \\ \psi_{13} & \psi_{23} & 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \psi_{(n-1)n} \\ \psi_{1n} & \psi_{2n} & \cdots & \cdots & \psi_{(n-1)n} & 0 \end{pmatrix}$$

Gut zu erkennen ist die Symmetrie der Assoziationsstärken, wodurch die obere und die untere Dreiecksmatrix spiegelsymmetrisch zueinander sind. Weiterhin zeigt die Nulldiagonale die definitionsgemäß nicht vorhandenen Auto-Assoziation aller Wörter.

Den Ausgangspunkt bildet dabei eine vollständig null-initialisierte Assoziationsmatrix, d.h. es gibt noch keinerlei Assoziationsstärken. Für deren Berechnung wird nun ein möglichst diversifizierter Textkorpus systematisch analysiert: Es erfolgt eine sequenzielle Bearbeitung, bei der jeweils Teilabschnitte betrachtet werden, die eine bestimmte Länge haben. Der aktuell betrachtete Teilabschnitt wird als *Textfenster* bezeichnet, dessen Länge ist folglich die *Fensterlänge*. Einfach gesagt: Das Fenster „wandert wortweise über den Korpus".

Passend zur gemeinsamen Psychischen Präsenz (vgl. Abschnitt 2.1 auf S. 23) werden innerhalb eines Textfensters gemeinsame auftretende Wörter als „gemeinsame Wahrnehmung" im Sinne des James'schen Assoziationsgesetzes interpretiert. Das heißt, die auftretenden Kookurrenzen zwischen den Wörtern werden ausgewertet.

Zur Erinnerung: Beim SST erfolgt ein Lernvorgang durch eine Stichprobe aus den Stimuli der N-elementigen Umwelt, deren Elemente per „Alles oder Nichts"-Prinzip mit dem Ereignis E verbunden werden. Sind letztlich alle Elemente der Umwelt mit dem Ereignis verbunden sind, ergibt sich eine Wahrscheinlichkeit von 1; umgekehrt ergibt sich eine Wahrscheinlichkeit von 0, wenn keine Elemente mit dem Ereignis verbunden sind.

Bezogen auf Wortassoziationen ergibt sich eine Assoziationsstärke ψ_{ij} von 1, wenn das Wort i immer das Wort j assoziiert – was daraus resultiert, wenn jedes im Korpus enthaltene Wort j ausnahmslos immer mit dem Wort i kookurriert. Umgekehrt ergibt sich eine Assoziationsstärke ψ_{ij} von 0, wenn ausnahmslos jedes im Korpus enthaltene Wort j ohne Kookurrenz mit dem Wort i auftritt.

Bei den Wortassoziationen bildet daher jede Fensterbetrachtung einen Lernvorgang, wie zuvor eine Merkmal-Stichprobe der N-elementigen Umwelt. Umgedeutet heißt das: Die „Umwelt" des „Ereignisses" mit der „Wahrscheinlichkeit" ψ_{ij} wird gebildet durch die Vereinigungsmenge aller Fensterbetrachtungen, in denen das Wort j vorkommt. Eine „Stichprobe" enthält exakt eine Fensterbetrachtung mit dem Wort j, mit oder ohne Kookurrenz mit dem Wort i.

Bezüglich der Assoziationsstärke gilt offensichtlich:

1. Werden zwei Wörter gemeinsam wahrgenommen, so steigt dadurch die Assoziationsstärke zwischen diesen beiden Wörtern.

2. Tritt eines dieser beiden Wörter dagegen ohne das andere auf, so sinkt dadurch die Assoziationsstärke vom einzeln auftretenden Wort zu dem anderen Wort. Korrekter gesagt: Die Assoziationsstärken zu *allen* anderen Wörtern, ohne die es auftritt, sinken.

Formalisiert ergeben sich, analog zum „klassischen SST", folgende iterativen Vorschriften[16]:

1. Treten zwei Wörter w_i und w_j (mit $w_i \neq w_j$) als Kookurrenz auf, so ergibt sich als

$$\textit{Lernregel: } \psi_{ij}(t+1) = \psi_{ij}(t) + \theta \cdot (1 - \psi_{ij}(t))$$

2. Tritt ein Wort w_j in einem anderen Kontext, d.h. nicht in Kookurrenz mit dem Wort w_i auf, so ergibt sich als

$$\textit{Hemmungsregel: } \psi_{ij}(t+1) = (1-\theta) \cdot \psi_{ij}(t)$$

Ein geeignetes θ muss je nach Versuch empirisch bestimmt werden. Es entspricht der Lern- bzw. Hemmungsrate, je nach Experiment können auch unterschiedliche Werte verwendet werden.

Der Lernalgorithmus verläuft überaus simpel, wie in Quelltext 1 angegeben.

```
1. Beziehe nächstes Textfenster.
      2. Ermittele alle Kookurrenzen innerhalb des Textfensters.
      3. Für alle Kookurrenzen w_i mit w_j:
            4. Wende Lernregel auf ψ_ij an.
            5. Für alle anderen w_k mit w_k ≠ w_i und w_k ≠ w_j:
                  6. Wende Hemmungsregel auf alle ψ_kj an.
7. Wiederhole ab (1), bis alle Textfenster bearbeitet.
```

Quelltext 1: Lernalgorithmus mit Lern- und Hemmungsregel

In der Praxis lassen sich die Schritte 4-6 in einer Schleife über sämtliche Wörter w_1 bis w_n in der Zeile j des Wortes w_j durchführen: Für kookurrierendes Wort w_i Lernregel anwenden, für identisches Wort w_j untätig bleiben, für alle anderen Wörter Hemmungsregel anwenden. — Zu beachten ist dabei lediglich die Annahme der symmetrischen Assoziationsstärken $\psi_{ij} = \psi_{ji} \forall i, j$, d.h. es muss beim Bearbeiten der kompletten Zeile i der Assoziationsmatrix gleichzeitig auch die komplette Spalte i nachgeführt werden.

In der Arbeitsgruppe Kognitive Psychologie der Universität Paderborn wird derzeit von Petra Seidensticker ein Softwarepaket entwickelt, das diesen Lernalgorithmus in leicht abgewandelter Form direkt implementiert (vgl. [SEI2003]).

16 Die Regeln sind bzgl. der Notation korrigiert, aber ansonsten identisch übernommen aus [WRF1993].

2.1.2.3 Variationen der Fenstertechnik

Die Fensterlänge im soeben vorgestellten Lernalgorithmus kann, gemäß der gemeinsamen Psychischen Präsenz, auf Wunsch je nach Experiment variiert werden: In [RW1991] wird eine Fensterlänge von 18 verwendet, in [WRF1993], [RAP1996] und [BÖH1997] beträgt sie 12.

Das heißt für die letztgenannten Beispiele: Ein Auftreten zweier Wörter mit maximalem Wortabstand von ±12, d.h. im Zentrum eines „Wortfensters" der Länge 23 (12 Wörter links, 12 Wörter rechts, zentrales Wort als Überlappung) wird als Kookurrenz der beiden Wörter gewertet. – Im weiteren Verlauf dieses Abschnitts bezeichne ich die Fensterlänge allgemein mit k.

Indes gibt es zwei grundsätzlich unterschiedliche Vorgehensweisen, die Kookurrenzen innerhalb des aktuell betrachteten Wortfensters zu ermitteln.

Vollständige Fensterauswertung mit Mehrfachbetrachtung

Der naive Ansatz ermittelt durch kreuzweises Verknüpfen aller k Wörter sämtliche Kookurrenzen im aktuellen Wortfenster. Wenn das Wortfenster im Anschluss daran um ein Wort nach rechts verschoben wird, startet der Algorithmus erneut und ermittelt im neuen Fenster auf dieselbe Weise wiederum alle Kookurrenzen.

Die Besonderheit dieser Vorgehensweise liegt auf der Hand: Wörter, die sehr nahe zusammen stehen, verbleiben für mehrere Fensterverschiebungen gemeinsam im Wortfenster und werden somit mehrfach betrachtet.

Eine „nahe Kookurrenz" zweier direkt benachbarter Wörter mit Abstand 1 wird bei Fensterlänge k folglich $k-1$ mal betrachtet. Im Gegensatz dazu wird eine „weite Kookurrenz" mit Abstand $k-1$ (die also gerade noch ins Fenster passt), exakt einmal betrachtet. – Allgemein gilt: Kookurrenzen mit Abstand x (wobei definitionsgemäß $x<k$) werden bei dieser Vorgehensweise $k-x$ mal betrachtet.

Obgleich die Überbetonung naher Kookurrenzen auf den ersten Blick unsinnig erscheint, kann sie aus psychologischer Sicht durchaus sinnvoll sein: Wörter, die häufig nahe zusammen vorkommen, bilden stärkere Assoziationen als Wörter, die häufig weit auseinander vorkommen (sofern sie immer noch so nahe zusammen stehen, um überhaupt als Kookurrenz gewertet zu werden). – Dass manche Untersuchungen diese Auswertungsmethode verwenden, basiert also nicht auf einer „algorithmischen Unachtsamkeit" der jeweiligen Versuchsleiter, sondern ist durchaus so gewollt.

Eine derartige Auswertungsmethode impliziert quadratische Laufzeit bezogen auf die Fenstergröße, jedoch bei linearem Zusammenhang mit der Korpusgröße. Bei Fensterlänge k und Korpusgröße K ergibt sich $O\left(k^2 \cdot (K-k)\right)$ bzw. (wegen $k = const. \ll K$) $O(k^2 \cdot K)$. Wegen $k = const.$ gilt aber auch $k^2 = const. \ll K$, also bleibt es letztlich bei (schlechter) linearer Laufzeit $O(c \cdot K) = O(K)$.

Vermeidung von Mehrfachbetrachtungen

Ein anderer Ansatz vermeidet Mehrfachbetrachtungen von „nahen Kookurrenzen". Hierbei wird nach dem Weiterschieben des Fensters nur das jeweils neu ins Fenster aufgenommene Wort mit den noch verbliebenen Wörtern des Fensters verknüpft.

Dadurch, dass immer nur das neue Wort in Kookurrenz mit den noch im Fenster befindlichen Wörtern betrachtet wird, d.h. insbesondere nur in eine Richtung (i.e. „nach hinten") verknüpft wird, sind Mehrfachbetrachtungen derselben Kookurrenz ausgeschlossen. Trotzdem werden aufgrund der sequenziellen wortweisen Vorgehensweise ausnahmslos alle Kookurrenzen jeweils einmal betrachtet.

Wiederum startet der Algorithmus mit den ersten k Wörtern im Fenster, es gibt danach noch $K-k$ Fensterverschiebungen. Bei jedem dieser $1+(K-k)$ Fensterbetrachtungen werden konstant $k-1$ Kookurrenzen gebildet. Insgesamt ergibt sich also $O\left(k \cdot (K-k)\right)$ bzw. annähernd $O(k \cdot K)$, und da weiterhin $k = const. \ll K$ gilt, bleibt eine (bessere) lineare Laufzeit von $O(c \cdot K) = O(K)$.

Der Laufzeit-Unterschied zu der aufwändigeren ersten Methode liegt in einem konstanten Faktor von etwa $k^2 - k \approx k^2$, d.h. die Laufzeit verschlechtert sich quadratisch mit der verwendeten Fenstergröße. Dieser Unterschied ist bei großen Korpora und hinreichend großen Fensterlängen zwar durchaus „spürbar", aber durch schnellere Hardware kompensierbar: Es bleibt in beiden Fällen ein linearer Aufwand bezogen auf die Korpusgröße.

2.1.2.4 Verallgemeinerung

Die in Abschnitt 2.1.2.2 definierte Assoziationsmatrix und der darauf arbeitenden SST-Lernalgorithmus lassen sich nicht nur für das Lernen von Wortassoziationen verwenden, sondern generell auf das Lernen von Assoziationen. In diesem Fall entsprächen die Zeilen bzw. Spalten der Assoziationsmatrix nicht Wörtern, sondern beliebigen Stimuli, die sich gegenseitig assoziieren können (z.B. Glockenklang, Lichtsignal, Tonsignal, Futtergabe, Elektroschock usw.).

Der Lernalgorithmus in Quelltext 1 auf S. 31 verwendet eine Schleife, um sämtliche Textfenster des verwendeten Textkorpus' der Reihe nach zu bearbeiten und die darin enthaltenen Kookurrenzen zu ermitteln. Werden andere Stimuli verwendet, die nicht

in sequenzieller zeitlicher Abfolge (z.B. in einem „Zeitfenster“) präsentiert werden, muss die umliegende Schleife durch eine fallweise Betrachtung ersetzt werden, um anderweitig sämtliche gemeinsamen Wahrnehmungen von je zwei Stimuli in den Lernprozess eingehen zu lassen.

Der Kern des Verfahrens besteht jedoch in der Methode der Lern- und der Hemmungsregel, die algorithmische Aufbereitung kookurrierender Wörter bzw. gemeinsam wahrgenommener Stimuli ist dagegen zweitrangig.

2.1.3 Assoziationsnormen über Kookurrenzhäufigkeiten

Statt der im SST verwendeten sequenziellen Lernmethode (vgl. Abschnitt 2.1.2.1) soll nun eine Methode vorgestellt werden, die indirekt ohne schrittweises Lernen die Assoziationsstärken zwischen den Wörtern berechnet. Diese ist in ATA2 (vgl. [BÖH1997]) implementiert, das für diese Arbeit Verwendung fand.

Von der Lern- und Hemmungsregel des SST ausgehend ergibt sich bei identischem θ laut [WRF1993] (vgl. [FOP1965]), dass der Erwartungswert der Assoziationsstärke zwischen w_i und w_j für $t \to \infty$ der bedingten Wahrscheinlichkeit von w_j bei gegebenem w_i entspricht. Also:

$$\psi_{ij}(t \to \infty) = p(w_j|w_i) = p(w_i|w_j) \text{ (wegen Symmetrie)}$$

Diese bedingte Wahrscheinlichkeit lässt sich durch Auszählen von Kookurrenzhäufigkeiten innerhalb des Korpus näherungsweise bestimmen. Das heißt, statt der SST-Lernphase mit Lern- und Hemmungsregel *zählt* der Algorithmus vorerst lediglich die Kookurrenzen der Wörter untereinander. Der Zählalgorithmus in Quelltext 2 arbeitet wiederum mit einer der beiden o.a. Fenstertechniken (vgl. Abschnitt 2.1.2.3).

```
1. Beziehe nächstes Textfenster.
       2. Zähle die Worthäufigkeiten der enthaltenen Wörter.
       3. Berechne alle Kookurrenzen im aktuellen Textfenster.
       4. Für alle Kookurrenzen w_i mit w_j :
              5. Zähle die Kookurrenzhäufigkeit
6. Wiederhole ab (1), bis alle Textfenster bearbeitet.
```

Quelltext 2: Zählalgorithmus für Wort- und Kookurrenzhäufigkeiten

Formalisiert: Sei V das Vokabular mit Kardinalität $|V|$ und beliebiger, aber fester (d.h. wohldefinierter) Sortierung. Sei T der Textkorpus beliebiger Länge. Sei k die Fensterlänge. Im Anschluss an die Ausführung des Algorithmus ergibt sich wie folgt:

1. *Die Worthäufigkeit eines Wortes:* Für jedes Wort $w \in V$ gilt: $H(w) \in \mathbb{N}_0$ entspricht der Anzahl der Vorkommen des Wortes w im Textkorpus T.

2. *Der Worthäufigkeitsvektor:* Für alle $H(w)$ mit $w \in V$ gilt:
   $Q := (H(w)) \in \mathbb{N}_0^{|V|}$ ist der Vektor aller Worthäufigkeiten.

3. *Die Kookurrenzhäufigkeit zweier Wörter:*
   Für je zwei Wörter $w_i, w_j \in V$ gilt: $H(w_i \& w_j)$ entspricht der Anzahl der Kookurrenzen[17] der Wörter w_i, w_j im Textkorpus T mit Wortabstand $\leqslant k$. Kein Wort kookurriert mit sich selbst: $H(w \& w) = 0 \ \forall w \in V$.

4. *Die Kookurrenzmatrix:* Für jedes $H(w_i \& w_j)$ mit $v, w \in V$ gilt:
   $P := (H(w_i \& w_j)) \in \mathbb{N}_0^{|V| \times |V|}$ ist die Matrix aller Kookurrenzhäufigkeiten.

Aus der Tatsache, dass definitionsgemäß kein Wort mit sich selbst kookurriert, d.h. $H(w \& w) = 0 \ \forall w \in V$, folgt, dass die Diagonale der quadratischen Kookurrenzmatrix P konstant 0 ist. Weiterhin gilt, dass die obere und die untere Dreiecksmatrix von P spiegelsymmetrisch zueinander sind. Dies basiert auf der vereinfachenden Annahme symmetrischer Assoziationsstärken zwischen zwei Wörtern, also
$H(w_i \& w_j) = H(w_j \& w_i) \ \forall w_i, w_j \in V$.

2.1.3.1 Kookurrenzhäufigkeiten

Tabelle 6 als Beispiel mit den vier Wörtern „Vater“, „Mutter“, „Kind“ und Familie“ anhand eines Beispielkorpus mag dies verdeutlichen (in Anlehnung an [BÖH1997]).

$P=H(w_i,w_j)$	Mutter	Vater	Kind	Familie	$Q=H(w)$
Mutter	0	1860	2901	324	13942
Vater	1860	0	1302	392	12238
Kind	2901	1302	0	703	39377
Familie	324	392	703	0	12900

Tabelle 6: Worthäufigkeiten und Kookurrenzen „Vater“, „Mutter“, „Kind“, „Familie“

Deutlich zu erkennen ist die definitionsgemäße 0-Kookurrenz jedes Wortes mit sich selbst (man nimmt an, dass kein Wort mit sich selbst kookurriert) und die Symmetrie der Kookurrenzmatrix.

Abbildung 7 auf S. 36 visualisiert die Daten aus Tabelle 6 (in Anlehnung an [BÖH1997]). Erwartungsgemäß ist damit jedoch noch nicht viel gewonnen: „Mit bloßem Auge“ ist lediglich zu erkennen, dass eine stärkere Assoziation zwischen „Mutter“ und „Kind“ als zwischen „Vater“ und „Kind“ besteht. Die Werte 2091 und 1302 sind insofern vergleichbar, da die absolute Häufigkeit von „Mutter“ und „Vater“ in etwa identisch ist (13.942 und 12.238).

Keinerlei Aussage lässt sich jedoch z.B. über die Vergleichbarkeit der Assoziationen von „Vater“ und „Kind“ bzw. „Vater“ und „Familie“ fällen: Die absoluten Häufigkeiten von „Kind“ und „Familie“ differieren zu stark.

17 Im Bereich der Computerlinguistik wird statt der Schreibweise „$H(w_i \& w_j)$“ oft „$H(w_i, w_j)$“ verwendet.

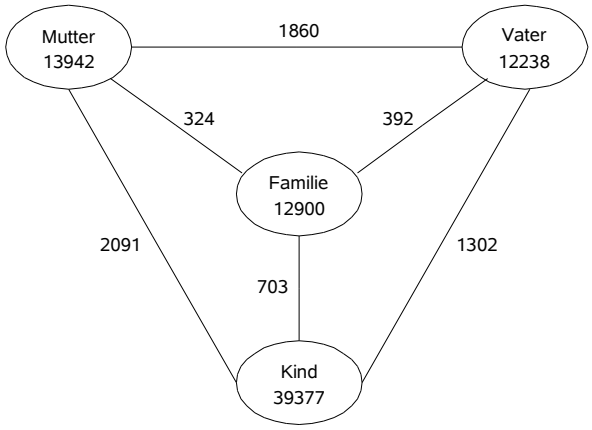

Abbildung 7: Worthäufigkeiten und Kookurrenzen „Vater", „Mutter", „Kind", „Familie"

Zur Berechnung der eigentlich benötigten, miteinander quantitativ vergleichbaren bedingten Wahrscheinlichkeiten $p(w_i|w_j)$ bedarf es einer Normierung der Assoziationshäufigkeiten. Erst dadurch werden sie als Assoziationsstärken vergleichbar. Dabei müssen insbesondere die Knotengewichte d.h. die absoluten Häufigkeiten „verschwinden".

2.1.3.2 Assoziationsnormen

Böhnisch diskutiert in [BÖH1997] verschiedene Methoden, um aus den $H(w_i \& w_j)$ jeweils zugehörige normierte $\psi(w_i \& w_j) = \psi_{ij}$ zu berechnen. Am Ende seiner Versuche legt er dar, dass seine Eigenentwicklung, basierend auf der harmonischen Mittelwertbildung der bedingten Wahrscheinlichkeiten $p(A|B)$ und $P(B|A)$ die besten Ergebnisse im Vergleich mit menschlichen Assoziationen darstellt.

Somit gilt im weiteren Verlauf für alle weiteren Berechnungen:

$$\psi(w_i \& w_j) = \frac{H(w_i \& w_j)}{\sqrt{(H(w_i) \cdot H(w_j))}}$$ ([BÖH1997], S. 16, Formel (5))

Als resultierende Matrix ergibt sich aus Tabelle 6 auf S. 35 bzw. aus Abbildung 7 normiert Tabelle 8: Die 0-Diagonale ist definitionsgemäß erhalten geblieben, ebenso die Symmetrie der oberen und unteren Dreiecksmatrix (Tabelle angelehnt an [BÖH1997], S. 5).

$\psi(w_i,w_j)$	Mutter	Vater	Kind	Familie
Mutter	0,0000	0,1424	0,1238	0,0242
Vater	0,1424	0,0000	0,0593	0,0312
Kind	0,1238	0,0593	0,0000	0,0312
Familie	0,0242	0,0312	0,0312	0,0000

Tabelle 8: Normierte Assoziationsstärken „Vater", „Mutter", „Kind", „Familie"

Graphisch dargestellt in Abbildung 9 auf Seite 37 sind die normierten Assoziationsstärken weitaus anschaulicher (Abbildung angelehnt an [BÖH1997] (S. 4)).

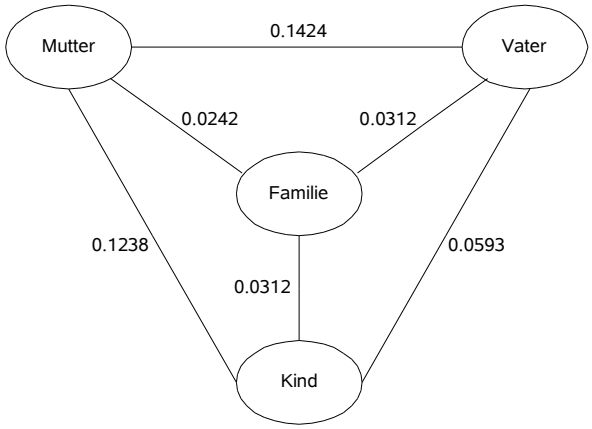

Abbildung 9: Normierte Assoziationsstärken „Vater“, „Mutter“, „Kind“, „Familie“

Die ursprüngliche Deutung bleibt bestehen: Die Assoziation „Mutter“ und „Kind“ ist mehr als doppelt so hoch wie zwischen „Vater“ und „Kind“. Nun sind auch die Assoziationen von „Vater“ und „Kind“ und „Vater“ und „Familie“ beurteilbar.

2.2 Assoziative Wortnetze

Die Wortassoziationen zwischen je zwei Wörtern eines Vokabulars lassen sich netzförmig als ein Graph im informatischen Sinne darstellen, dessen Knoten durch die Wörter des Vokabulars gebildet werden. Der Graph ist vollständig und gewichtet: Das Kantengewicht zwischen zwei Wortknoten entspricht der Stärke der gegenseitigen Assoziation der beiden Wörter: Je stärker die Assoziation zwischen zwei Wörtern, desto stärker das Gewicht.

Wie zuvor sollen symmetrische Assoziationsstärken angenommen sein, d.h. ein Wort w_1 vermag ein anderes Wort w_2 ebenso stark zu assoziieren wie umgekehrt. Dies ist streng genommen nicht ganz richtig: Einige Studien (etwa [CH1989]) legen nahe, dass im Englischen durch die recht klar definierte Satzstellung das Assoziationsgewicht von z.B. „doctor“ zu „nurse“ ein anderes ist als das von „nurse“ zu „doctor”. Obwohl [BÖH1997] beschreibt, dass dies im Deutschen vernachlässigbar sei, weil durch die freiere Satzstellung hier nur geringe Stärkenunterschiede zwischen den Assoziationsrichtungen auftreten, ist das Problem der Asymmetrie der Assoziationsgewichte bis heute nicht gelöst (lt. M. Wettler, pers. Gespräch). Dennoch soll es auch im Rahmen dieser Arbeit bei der angenommenen Symmetrie belassen werden: Diese Vereinfachung führt für alle weiteren Betrachtungen und Anwendungen zu einer Halbierung der Datenmenge und (später) zu beschleunigten Berechnungen.

Da Visualisierungen vollständiger Graphen allzu leicht unübersichtlich werden (denn jeder Knoten hat eine Kante zu jedem anderen Knoten), kann ein solcher in diesem Spezialfall dadurch in einen leichter darstellbaren unvollständigen Graph überführt werden, indem man Kanten unterhalb eines gewissen Schwellwert-Gewichts nicht einzeichnet. Dabei bleiben die starken Assoziationsgewichte unverändert enthalten und der Graph ist ohne nennenswerten Informationsverlust i.d.R. leichter kreuzungsfrei darstellbar. Abbildung 10 zeigt einen gewichteten, ungerichteten Beispiel-Graph mit symmetrischen Assoziationsstärken, der originalgetreu aus [BÖH1997] (S. 3) entnommen ist, seinerseits als Zitat aus [WR1993].

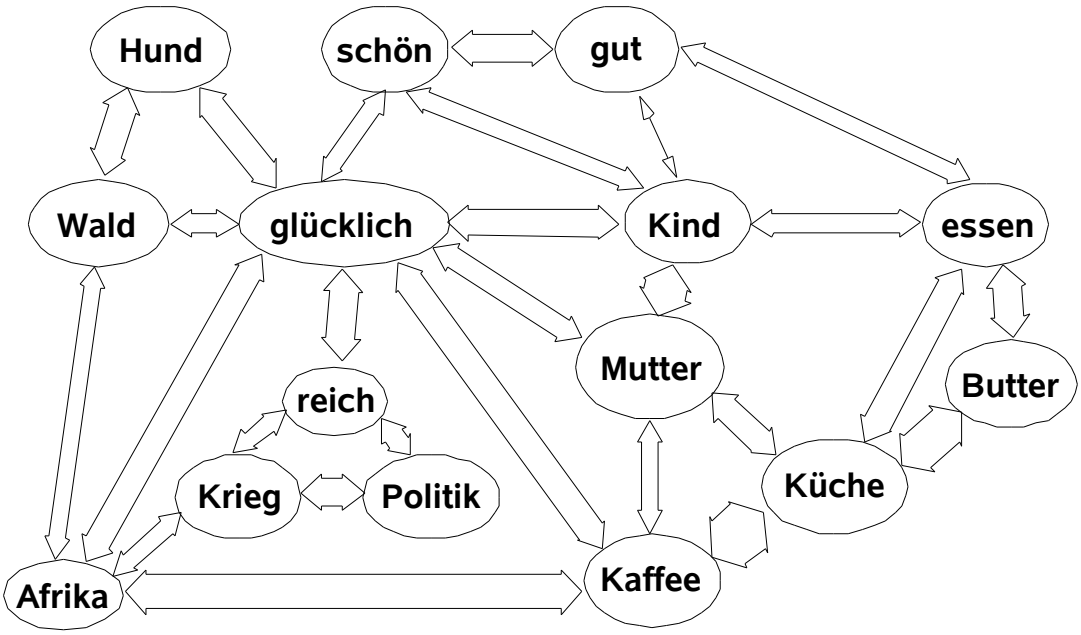

Abbildung 10: Beispiel-Netzwerk (unvollständiger Graph)

Die Breite der Pfeile ist dabei proportional zum Logarithmus der Assoziationsstärke, die Länge der Pfeile ist irrelevant. Gut zu erkennen ist, dass etwa „Kaffee“ stark mit „Küche“ assoziiert ist, ebenso „Kind“ mit „Mutter“. Hingegen ist die Assoziation zwischen „Wald“ und „Afrika“ eher gering, noch schwächer ist die Assoziationsstärke zwischen „Kind“ und „gut“.

2.2.1 Computersimulierte Assoziation

Assoziationsversuche wie der von Galton (vgl. [GAL1880]) zur Erforschung menschlicher Assoziation oder auch allgemein zur Ermittlung der realen Assoziationsstärken der Wörter einer Sprache bzw. zugehöriger Konzepte sind mit großem logistischen Aufwand verbunden – und daher zeitraubend und kostspielig. Wünschenswert ist eine wie auch immer geartete „Simulation“ menschlicher Assoziation.

Rapp und Wettler haben 1991 nachgewiesen (vgl. [RW1991]), dass sich Galtons freie Wortassoziationen gut mit Hilfe von Hebbschen Lernprozessen (vgl. Hebbsches Postulat weiter oben) anhand von korpusbasierten Untersuchungen vorhersagen lassen. Basis dafür war nicht nur die Nachbildung des Lernvorgangs zur Ermittlung der Assoziationsstärken, sondern auch die anschließende Simulation des menschlichen Assoziationsvermögens im Computer.

Werden die Assoziationsstärken empirisch ermittelt, muss für die genaue Bestimmung der durchschnittlichen Assoziationsstärken offensichtlich eine möglichst umfangreiche Studie durchgeführt werden. Im Gegensatz dazu ist bei der Berechnung der Assoziationsstärken eine gute Annäherung der *berechneten* an die *realen* Assoziationsstärken nur mit einem hinreichend großen und vielfältigen Textkorpus möglich.

Während gesprochene Wörter bis heute immer noch große Schwierigkeiten bei der automatischen Bearbeitung im Computer verursachen, sind in heutiger Zeit immer mehr Schriftdokumente digital verfügbar und somit direkt maschinell verarbeitbar. Zu diesem Zweck wurden seit 1989 von der Arbeitsgruppe Kognitive Psychologie der Universität Paderborn deutschsprachige maschinenlesbare Texte in großem Umfang zusammengetragen. Dabei handelt es sich einerseits um mehrere Jahrgänge der Berliner „Tageszeitung“ (TAZ), der „Frankfurter Rundschau“ (FR) und der Münchener „Süddeutschen Zeitung“ (SZ), andererseits in geringerem Umfang auch um Texte aus der klassischen Literatur.

Korpus	MB komprimiert	ca. Mio. Wörter	Anteil
TAZ	220,82	111	**34,5%**
FR	64,39	32	**10,1%**
SZ	317,86	160	**49,7%**
Klassik	36,66	18	**5,7%**
Summe	**639,73**	**322**	**100,0%**

Tabelle 11: Volumina der einzelnen Teilkorpora

Das Gesamtvolumen des digital verfügbaren und für diese Arbeit verwendeten Korpus mit geschätzter Wortanzahl der einzelnen Teilkorpora ist in Tabelle 11 dargestellt.

2.2.2 Künstliche Neuronale Netzwerke

Wird jedes Wort des verwendeten Vokabulars bzw. des zugehörigen assoziativen Wortnetzes wiederum durch ein Neuron modelliert, ergibt sich zusammen mit den berechneten Assoziationsstärken (z.B. mit SST oder Assoziationsnormen) ein Neuronales Netzwerk: Die Assoziationsstärke zwischen je zwei Wörtern entspricht einer synaptischen Verbindung zwischen den zugehörigen Wort-Neuronen. Mit dieser Modellierung lässt sich die Aktivierung bzw. das Aktionspotenzial eines Wort-Neurons ge-

wichtet anhand der Assoziationsstärken über die Synapsen an die anderen Wort-Neuronen weiterleiten. Der Aktivierungszustand nach der Reizausbreitung im Neuronalen Netz entspricht damit einer simulierten Wortassoziation.

Derartige „Künstliche Neuronale Netzwerke" lassen sich rechnergestützt realisieren. Basis dafür ist häufig eine wie bereits in im Abschnitt 2.1.2.2 definierte Assoziationsmatrix mit den Assoziationsstärken zwischen je zwei Wörtern, d.h. zwischen den „Künstlichen Neuronen". In diesem Fall bestehen diese gedachten Neuronen nur in der Modellvorstellung und werden nicht diskret modelliert, also insbesondere nicht einzeln (z.B. objektorientiert) implementiert.

In Michael Böhnischs ATA2 (vgl. [BÖH1997]), das für diese Arbeit Verwendung fand, wird ein Assoziatives Wortnetz mit Hilfe eines solchen Künstlichen Neuronalen Netzwerks implementiert und für die Simulation von Wortassoziationen verwendet. Böhnisch beschreibt dies wie folgt:

> *»Knoten in diesem Netz haben „Aktivitäten", die durch „Stimulation" erhöht werden können. [...] Jeder Knoten überträgt seine Aktivität auf seine benachbarten Knoten, anteilig gewichtet entsprechend den Kantengewichten, und propagiert so einen Anregungszustand durch das gesamte Netz. Dieses dynamische Modell wird deswegen „Spreading Activation Network" genannt. [...] Ein Wortnetz ist damit ein Spezialfall eines einschichtigen Neuronalen Netzwerks.«*
>
> *[BÖH1997], S. 3*

2.2.2.1 Netzrepräsentation

Sei V das Vokabular des Assoziativen Wortnetzes. Die Wörter seien mit Hilfe einer beliebigen, aber im weiteren Verlauf festen Ordnung geordnet, z.B. alphabetisch aufsteigend. Innerhalb dieser Sortierung seien die Wörter von 1 bis n durchnummeriert (o.B.d.A.).

Die Assoziationsstärke eines Stimulus-Wortes $w_i = p$ hinsichtlich eines anderen Wortes $w_j = q$ bezeichne ich fortan mit $\psi(p \& q)$ oder mit ψ_{ij}. Wiederum sei von symmetrischen Assoziationsstärken ausgegangen, d.h. $\psi_{ij} = \psi_{ji} \forall\, w_i, w_j \in V$. Weiterhin gelte definitionsgemäß, dass sich kein Wort selbst assoziiert, d.h. $\psi_{ii} = 0 \forall\, w_i \in V$.

Wie bei der SST in Abschnitt 2.1.2.2 ergibt sich als Darstellung aller Assoziationsstärken zwischen je zwei Wörtern im Netzwerk wiederum eine quadratische Assoziationsmatrix mit spiegelsymmetrischer oberer und unterer Dreiecksmatrix sowie Nulldiagonale.

$$\Psi = \begin{pmatrix} 0 & \psi_{12} & \psi_{13} & \cdots & \cdots & \psi_{1n} \\ \psi_{12} & 0 & \psi_{23} & \cdots & \cdots & \psi_{2n} \\ \psi_{13} & \psi_{23} & 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \psi_{(n-1)n} \\ \psi_{1n} & \psi_{2n} & \cdots & \cdots & \psi_{(n-1)n} & 0 \end{pmatrix}$$

Dies ist zugleich die einfachste Darstellung eines vollständigen gewichteten, ungerichteten Graphen ohne Kanten von einem Knoten auf sich selbst (vgl. Abbildung 10 auf S. 38).

2.2.2.2 Reizstimulierung

Das so definierte Künstliche Neuronale Netz werde nun trainiert, damit es die Assoziationsstärken zwischen den einzelnen Wörtern des Vokabulars widerspiegelt. Dabei bildet das Auszählen der Kookurrenzen mit anschließender Normierung der Kookurrenzhäufigkeiten zu Assoziationsstärken (vgl. Abschnitt 2.1.3) einen Spezialfall des Trainierens, alternativ kann auch SST (vgl. Abschnitt 2.1.2.1) zum Einsatz kommen.

Die Produktion von freien Wortassoziation auf vorgegebene Stimuluswörter kann mit einem trainierten Künstlichen Neuronalen Netz wie folgt simuliert werden: Einer oder mehrere Wortknoten werden angeregt, woraufhin die netzimmanente Reizpropagierung abläuft und schlussendlich eine Reihe von aktivierten Knoten mitsamt ihrer daraus folgenden Aktivierungen liefert. Kognitionswissenschaftlich bedeutet dies: Die Assoziationssimulation erhält eines oder mehrere Reizwörter aus dem Vokabular als Eingabe und ermittelt die daraus resultierenden assoziierten Wörter. Von größtem Interesse ist eine Liste der assoziierten Wörter in absteigender Reihe ihrer Aktivierungsstärke.

Der Stimulus bildet dabei einen Vektor v_0, dessen Komponenten den Wörtern des Vokabulars entsprechen – in der o.a. beliebigen, aber festen Sortierung des Vokabulars. Die Aktivierungsstärke der einzelnen Stimuluswörter lässt sich stetig zwischen 0 und 1 variieren, praktisch wird hier aber für ein zu aktivierendes Wort der Wert 1 (d.h. maximale Aktivierung) verwendet, für ein nicht zu aktivierendes Wort der Wert 0. Bei einem Reiz mit wenigen Stimuluswörtern ergibt sich so ein Vektor, dessen Komponenten nahezu alle 0 sind.

Die Berechnung der resultierenden Aktivierung erfolgt durch Multiplikation der Assoziationsmatrix Ψ mit v_0. Also:

$$\Psi = \begin{pmatrix} 0 & \psi_{12} & \psi_{13} & \cdots & \cdots & \psi_{1n} \\ \psi_{12} & 0 & \psi_{23} & \cdots & \cdots & \psi_{2n} \\ \psi_{13} & \psi_{23} & 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \psi_{(n-1)n} \\ \psi_{1n} & \psi_{2n} & \cdots & \cdots & \psi_{(n-1)n} & 0 \end{pmatrix}, \quad v_0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} . \text{ Damit: } v_1 = \|\Psi \cdot v_0\| . \text{ (1)}$$

Die Normierung erfolgt, wie in [BÖH1997] (S. 8) definiert, mit Hilfe der üblichen 2-Norm, d.h. der Euklidischen Norm $\|a\| := a/|a|$.

Der resultierende normierte Ergebnisvektor v_1 entspricht dem Output des neuronalen Netzes bei einfacher Reizpropagierung des Eingabevektors v_0 . Die berechnete Aktivierung eines Wortes, repräsentiert durch seine Komponente des Ausgabevektors v_1 , ist die Summe der Assoziationsstärken dieses Wortes zu den im Eingabevektor v_0 stimulierten Wörtern. Kognitionswissenschaftlich betrachtet enthält v_1 also die primären Assoziationsstärken nach Präsentation der durch die in v_0 als Reiz definierten Stimuluswörter, im einfachsten Fall sogar nur eines einzigen Stimulusworts. — In aller Regel wird man die Komponenten von v_1 nun in absteigender Folge sortieren, um die zugehörigen assoziierten Wörter ausgeben zu können, die am stärksten assoziierten Wörter zuerst.

2.2.2.3 Multiple Reizpropagierung

An dieser Stelle schließt sich je nach Experimentalsituation ggf. die multiple Reizpropagierung an (vgl. [FWR1995], [BÖH1997]). Kognitionswissenschaftlich betrachtet: Es lassen sich neben direkten auch indirekte Assoziationen zum Stimulus berechnen. Ein Beispiel: Liefert etwa das Reizwort „Hund“ die direkte Assoziation „Katze“, so liefert eine Restimulierung des Netzes auch die Assoziationen auf das Reizwort „Katze“, also etwa die Assoziation „Maus“. Je nach Anwendungsfall kann es durchaus wünschenswert sein, zu „Hund“ auch die indirekte Assoziation „Maus“ liefern zu können.

Darüber hinaus kann die Re-Stimulierung Assoziationen zweiten oder allgemein höheren Grades liefern, die nur durch mehrere Stimuluswörter assoziiert werden können, in diesem Falle durch mehrere (unterschiedlich starke) Assoziationen ersten Grades, die als Input für den zweiten Berechnungsschritt dienen. Ein Beispiel: Liefert etwa die Stimulierung des Wortknotens „Küche“ die beiden Wörter „Wasserhahn“ und „Herd“, so würde eine Sekundär-Stimulierung der letztgenannten Wortknoten ggf. das Wort „Teekessel“ als Ausgabe liefern.

Bezogen auf die Berechnungsvorschrift (1) auf Seite 42 ergibt sich: v_1 wird wiederum als Eingabevektor derselben Berechnung aufgefasst. Je nach Verfahren der multiplen Propagierung kann dabei noch der ursprüngliche Stimulusvektor v_0 erneut aufsummiert werden. Die Berechnungsvorschrift lautet somit:

$v_1' = v_1 + v_0$ mit anschließender Propagierung $v_2 = \|\Psi \cdot v_1'\|$.

Das Verfahren lässt sich über mehrere Schritte iterieren. Allgemein gilt:

$v_{i+1}' = v_{i+1} + v_0$ mit anschließender Propagierung $v_{i+2} = \|\Psi \cdot v_{i+1}'\|$. (1')

Die zugehörige Algebraische Analyse der multiplen Propagierung soll an dieser Stelle nur kurz und ohne jeglichen Beweis angerissen werden, im Detail findet sie sich in [FWR1995]: Die Aktivierung v_0 des Künstlichen Neuronalen Netzes relaxiert graduell und konvergiert für $i \to \infty$ in den meisten Fällen gegen einen Grenzwert v_i .

Ist der Spektralradius[18] von Ψ kleiner als 1, konvergiert v_i in jedem Fall gegen den Nullvektor. Ist der Spektralradius von Ψ dagegen größer als 1, so konvergiert v_i in den meisten Fällen (vgl. [FWR1995], S. 690 unten) gegen den Eigenvektor von Ψ mit dem maximalen Eigenwert – und in dem Fall ist dieses Ergebnis sogar identisch für jeden beliebigen Eingabevektor.

Weiter unten im Kapitel 2.5 über die geänderte Einsatzweise von ATA2 für diese Arbeit lege ich in Abschnitt 2.5.3.1 dar, dass mehrfache Reizpropagierung im vorliegenden Anwendungsfall prinzipbedingt wenig sinnvoll ist und daher erwartungsgemäß keine verwertbaren Ergebnisse liefert. Eine über diese grobe Beschreibung hinaus gehende Erörterung kann daher an dieser Stelle entfallen.

2.3 Gesamt-Vokabular vs. Testvokabular

Der Textkorpus besteht aus vielen unterschiedlichen Einzelwörtern, die als Vereinigungsmenge betrachtet ein *Gesamtvokabular* bzw. den Wortschatz des Korpus darstellen. Theoretisch ließe sich zwischen je zwei Wörtern des Gesamtvokabulars die Kookurrenzhäufigkeit auszählen und mittels Normierung die Assoziationsstärke berechnen. In der Praxis wird man jedoch immer auf ein kleineres *Testvokabular* oder kurz *Vokabular* zurückgreifen, um Speicherplatz und Rechenzeit zu sparen.

18 Der Spektralradius einer Matrix ist der Betrag ihres betragsgrößten Eigenwerts. Anschaulich repräsentiert er ein Maß für das „maximale Wachstum“ eines Vektors, der mit der Matrix multipliziert wird.

Die Reduktion des Gesamtvokabulars zum Testvokabular unterliegt der Sorgfalt des Versuchsleiters. Das Testvokabular muss sowohl die Wörter enthalten, deren stärkste Assoziationen ermittelt werden sollen, als auch die Wörter, die diesen Assoziationen entsprechen. Enthielte beispielsweise das Testvokabular nur das Wort „Mutter“, aber nicht das Wort „Kind“, könnte aus „Mutter“ niemals die offensichtliche Assoziation „Kind“ berechnet werden. Ebenso wenig wäre es möglich, jedwede Assoziation auf das Wort „Kind“ zu berechnen.

Das gesamte Verfahren steht und fällt also mit der Auswahl eines repräsentativen, ggf. anwendungsspezifischen Testvokabulars. Es bietet sich unter anderem an, Testtexte aus dem gerade untersuchten Themenbereich (z.B. Werbetexte) als „Wortlieferanten“ zu verwenden. Diese Texte brauchen nicht zwingend dem Textkorpus zu entstammen: Es geht lediglich darum, eine angemessene Menge Vokabular-Einträge zu ermitteln, die bei allen weiteren Berechnungen Verwendung finden.

Bei allen Fenstertechnik-Algorithmen (also beim schrittweisen Lernen oder beim Auszählen der Wort- und Kookurrenzhäufigkeiten) werden dann nur die Wörter betrachtet, die im Testvokabular enthalten sind. Nichtsdestotrotz verbleiben die Nicht-Vokabular-Wörter im gerade betrachteten Textfenster, d.h. sie bilden sozusagen „Unwörter“, deren Kookurrenzen mit den Wörtern des Testvokabulars nicht gewertet werden.

2.3.1 Morphologie und Lemmatisierung

Die zu verarbeitenden Texte sollten vor der Kookurrenzanalyse vereinfacht werden, indem Wortformen desselben Wortstamms zu einem einzigen Wortknoten des Assoziativen Netzes zusammengefasst werden. Neben den positiven Auswirkungen auf den Speicherbedarf und die Rechenzeit (vgl. [BÖH1997], S. 8ff) hilft dieser Vorgang auch, die durch Wortflexion wie Konjugation von Verben und Deklination von Substantiven und Adjektiven entstehenden „Unwörter“ auf Vokabularwörter zurückzuführen und somit die sonst vorhandene „Zersplitterung“ des Kookurrenzpotenzials durch die verschiedenen Wortformen zu vermeiden.

Ein derartiger Wortstamm heißt auch „Lemma“. Den Vorgang der Rückführung der unterschiedlichen Wortformen auf ihr Lemma bezeichnet man als *Lemmatisierung*.

Beispielsweise wird der Genitiv Singular „Hundes“ (etwa aus dem Satz „... das Fell des Hundes“) auf „Hund“ zurückgeführt, wie auch der Nominativ Plural „Hunde“. Ebenso wird z.B. „ging“ (etwa aus „Der Mann ging“) auf „gehen“ zurückgeführt.

Durch diesen Vorgang werden deutsche Sätze in ein recht merkwürdig aussehendes „Pidgin Deutsch“ übersetzt, siehe Text 3.

```
Der Mann ging, um das Fell des Hundes zu streicheln.
Der Mann gehen . um das Fell des Hund zu streicheln .
```

Text 3: „Pidgin-Deutsch“ durch korrekte Lemmatisierung

Der in Böhnischs ATA2-Programm enthaltene Lemmatisierer verwendet eine vorab berechnete, statische Ersetzungstabelle, in der für (hoffentlich) jedes flektierte Wort eine Grundform enthalten ist. Diese Ersetzungstabelle wurde mit der Windows-Software „Morphy“ generiert, einem integrierten Morphologie- und Tagging-Programmpaket für die Deutsche Sprache. Es wurde von Wolfgang Lezius 1995 in der Arbeitsgruppe Kognitive Psychologie der Universität Paderborn entwickelt und 1996 auf der KONVENS-Konferenz vorgestellt (s. [LRW1996]). Morphy ist als Public Domain-Programm frei erhältlich unter [LEZ1999].

Die von Morphy für ATA2 generierte Lemmatisierungs-Tabelle umfasst knapp 78.000 Wortgrundformen, sowie zu jeder enthaltenen Grundform die zugehörigen flektierten Formen. Der ATA2-Lemmatisierer muss damit auskommen: Mehr als die statische Ersetzungstabelle steht ihm nicht zur Verfügung. Morphy dagegen verfolgt bei seiner morphologischen Analyse einen weitaus eleganteren Ansatz[19]. Leider ist dieser Vorgang rechenintensiv und würde die Funktion von ATA2 stark bremsen. Daher wurden auch die von ATA2 verwendeten Textkorpora vorab mit der Morphy-basierten Lemmatisierungs-Tabelle lemmatisiert. Ebenso müssen auch Testtexte, wenn sie als „Wortlieferanten“ für das Testvokabular dienen sollen, lemmatisiert werden. Gleichermaßen können auch nur lemmatisierte Stimulus-Wörter verwendet werden, um Assoziationen zu berechnen.

2.3.1.1 Homographen

Ein großes Problem einer derartigen wortweisen Lemmatisierung mit einer statischen Ersetzungstabelle sind *Homographen*, d.h. Wörter mit identischer Schreibweise aber unterschiedlicher Bedeutung. Ein prominentes Beispiel ist das Wort „weiß“, das einerseits adjektivisch als Farbname verwendet werden kann, aber auch als 1. bzw. 3. Person, Präsenz, Indikativ von „wissen“ („Ich/er weiß ...“).

Morphy ist in der Lage, diese Mehrdeutigkeit insofern aufzulösen, dass es alle Möglichkeiten nennt (siehe Text 4 auf S. 46).

19 Morphy verwendet ein Lexikon unflektierter Wortstämme. Es schneidet von einem zu analysierenden Wort sämtliche möglichen Anhängsel ab, findet alle dazu passenden unflektierten Wortstämme – und versucht regelbasiert, von diesen zurück zum ursprünglichen Wort zurück zu finden. Dieser Vorgang listet sämtliche (regelbasiert auflösbaren) Mehrdeutigkeiten auf (vgl. [LRW1996], S. 3).

```
weiß
VER wissen NON 1 SIN PRÄ
VER wissen NON 3 SIN PRÄ
ADJ weiß PRD GRU
```

Text 4: Morphy-Ausgabe zu „weiß“

Jedoch muss sich der tabellenbasierte Lemmatisierer für eine Alternative „entscheiden“. Es besteht nicht die Möglichkeit, mehrere mögliche Deutungen zu verwenden: Jedes flektierte Wort wird durch ein unflektiertes ersetzt, das via Tabelle als Grundform ermittelt wurde. Das Ergebnis wird kommentarlos weiterverwendet.

Demnach wäre etwa bei der Verbform „wußte“ (1./3. Person, Imperfekt, Indikativ, „Ich/er wußte ...“) die Rückführung auf den Infinitiv „wissen“ absolut korrekt. Im Gegensatz dazu gäbe es bei o.a. „weiß“ keine Unterscheidungsmöglichkeit, ob „wissen“ oder „weiß“ (als Farbbezeichnung) zu wählen ist. Geradezu grotesk mutet deswegen folgende Lemmatisierung aus Text 5 an.

```
Die Wand ist weiß.
Die winden sein wissen .
```

Text 5: „Pidgin-Deutsch“ durch falsche Lemmatisierung

Der ATA2-Lemmatisierer interpretiert „Wand“ als 1. bzw. 3. Person, Imperfekt, Indikativ von „winden“ und die Farbe „weiß“ wie oben als 1. Person, Präsenz, Indikativ von „wissen“. Das Resultat ist schlicht vollkommen unbrauchbar. Nichtsdestotrotz gilt: Sofern im gleichermaßen vorab lemmatisierten Korpus der Satz „Die Wand ist weiß“ enthalten war, wurde dieser ebenfalls auf dieselbe Weise verstümmelt.

Die Ursache für den offensichtlichen Mangel der unzulässigen willkürlichen Rückführung mehrerer Homographen auf eine gemeinsame (und damit potenziell falsche) Grundform liegt in der wenngleich großen doch weiterhin begrenzten Ersetzungstabelle: Flektierte Wortformen, für die keine Ersetzungsregeln in der Tabelle enthalten sind, kann Morphy immerhin noch mit einer Fehlermeldung quittieren. Der ATA2-Lemmatisierer kennt keine derartige Möglichkeit: Er lässt nicht-lemmatisierbare Wortformen unverändert passieren. Dies gilt wiederum sowohl für den vorab lemmatisierten Korpus, als auch für die zu verwendenden „Wortlieferant“-Testtexte.

2.3.1.2 Komposita

Ein weiteres Problem der deutschen Sprache ist die Möglichkeit, *Komposita*, d.h. zusammengesetzte Substantive mit oder ohne Bindestrich zu bilden. Es bietet sich an, Komposita in deren Grundwörter zu zerlegen, z.B. „Hundenasen“ in „Hunde“ und „Nasen“, die wiederum korrekt in deren Grundformen „Hund“ und „Nase“ lemmatisiert werden können. Morphy gelingt mithilfe seiner Nominalkompositaanalyse sehr gut auch die Zerlegung von Komposita ohne Bindestrich. Der ATA2-Lemmatisierer

hingegen kann Komposita nur dann in Einzelwörter zerlegen und lemmatisieren, wenn sie mit Bindestrich im Text erscheinen. Ist das nicht der Fall, passieren Komposita unzerlegt und unlemmatisiert die Lemmatisierung.

Bei genauer Betrachtung wird folgendes deutlich: Der Textkorpus ist bei der Vorverarbeitung auf dieselbe Art und Weise teilweise fehlerhaft lemmatisiert worden. Das heißt, im lemmatisierten Korpus tauchen durchaus unzerlegte Komposita und potenziell falsche Lemmatisierungen aufgrund von nicht aufgelösten Mehrdeutigkeiten auf. Werden für einen Versuch also Testtexte für die Verwendung als Vokabular-Lieferant lemmatisiert, ist es wenig hilfreich, augenscheinlich fehlerhafte Lemmatisierungen „von Hand" zu korrigieren: Komposita, die beim Testtext nicht korrekt zerlegt werden, wurden auch bei der Korpus-Lemmatisierung nicht korrekt zerlegt, und mehrdeutige Wortformen wurden gleichermaßen potenziell fehlerhaft lemmatisiert übernommen. Das Dilemma liegt auf der Hand: Es hilft zwar der optischen Qualität einer Lemmatisierung, eine augenscheinlich falsche Lemmatisierung wie in Text 5 auf S. 46 von Hand zu korrigieren, führt aber letztlich zu einer verzerrten Korrespondenz des Testvokabulars mit dem Korpus.

2.3.2 Exemplarische Korpusanalyse

Um die Prognosen in Bezug auf die Lemmatisierungsqualität der vorab lemmatisierten Textkorpora zu überprüfen, wurde exemplarisch der TAZ-Korpus hinsichtlich der enthaltenen Worthäufigkeiten genauer untersucht. Es ging dabei nicht um die praktische Verbesserung des Korpus, sondern nur um die Gewinnung eines genaueren Einblicks in die Qualität des Korpusmaterials und daraus resultierender systematischer Probleme bei deren Verwendung.

Der TAZ-Korpus macht mit seinen 111 Mio. Wörtern (unkomprimiert etwa 600 Megabyte ASCII-Text) knapp 35% des verfügbaren Gesamtkorpus aus (vgl. Tabelle 11 auf S. 39). In den Vorversuchen wurde dieser Teilkorpus als Kompromisslösung zwischen Korpusgröße und benötigter Rechenzeit verwendet. Das vollständige Vokabular dieses Teilkorpus wurde hinsichtlich seiner absoluten Worthäufigkeiten ausgezählt. Hierzu wurde in *Borland Delphi* ein kleines Windows-Hilfsprogramm entwickelt, das diese Aufgabe erledigte. Darin kam ein balancierter Binärbaum[20] als Datenstruktur zum Einsatz, der Zugriffe auf die Zähler der einzelnen Wörter in logarithmischer Zugriffszeit realisierte.

20 Balancierte Binärbäume und deren logarithmisches Laufzeitverhalten sind in [WIR1975] ausführlich beschrieben. Um die traditionell fehlerträchtige Entwicklung eines einsatzfähigen Programmmoduls abzukürzen, kam eine auf o.a. Quelle basierende, bereits implementierte und getestete Delphi-Unit von [TSC2000] zum Einsatz, die lediglich für die Verwendung im aktuellen Hilfsprogramm angepasst werden musste.

Der erste Programmlauf zum Auszählen der Worthäufigkeiten benötigte auf einer 1,5 GHz schnellen AMD-CPU mit 512 Megabyte Arbeitsspeicher bei lokal auf der Festplatte vorliegendem TAZ-Korpus etwas weniger als 10 Minuten. Daraus resultierten etwa 1,2 Mio. unterscheidbare Wortformen in einer nunmehr lediglich 22 Megabyte großen Indexdatei, die auf o.a. Windows-PC bequem im Hauptspeicher gehalten werden kann. Bei erneuter Verwendung eines balancierten Binärbaums kann die Wortliste ohne spürbaren Zeitversatz quasi in Echtzeit durchsucht werden.[21] Die Resultate dieser Zusatzuntersuchung sind vielfältig.

2.3.2.1 Resultate

Zu erwarten, aber unbedingt an dieser Stelle erwähnenswert ist die Tatsache, dass der TAZ-Korpus noch auf der alten deutschen Rechtschreibung basiert, z.B. *„übrigbleiben“* statt *„übrig bleiben“*. Daher müssen auch alle Wortlieferant-Testtexte in der alten Rechtschreibung gehalten werden, insbesondere auch in Hinblick auf Doppel-S- / Eszet-Schreibweisen (z.B. „daß“ statt „dass“). Weiterhin ist, wie bereits prognostiziert, der Stand der Lemmatisierung hinsichtlich ungebräuchlicher Wörter und Wortformen aufgrund der begrenzten, statischen Ersetzungstabelle mitunter unbefriedigend.

Der TAZ-Korpus umfasst 1,2 Mio. unterscheidbare Wortformen. Darin sind in recht großer Zahl unbrauchbare d.h. falsch geschriebene Wörter enthalten, z.B. *„öffentlicehn“* statt *„öffentlichen“*. Die Ursache dürfte in der Methode der Digitalisierung zu suchen sein: Ein derartiges Wort deutet auf einen Tippfehler beim „Erfassen von Hand“ d.h. beim Abtippen hin, oder, sofern die Texte bereits digital übernommen worden sind, auf einen echten Druckfehler in der ursprünglichen Ausgabe, d.h. einen Tippfehler des damaligen Redakteurs.

Derart unbrauchbare Wörter sind jedoch nur bedingt problematisch: Sie haben in aller Regel eine Häufigkeit von 1, kommen also nur einmal im gesamten Korpus vor. Sofern Sie nicht in den Eingabetexten, d.h. im Rahmen des Wortnetz-Vokabulars und der Stimuli vorkommen, werden Sie bei der Kookurrenzanalyse schlicht ignoriert und verbrauchen dabei höchstens unnötig Rechenzeit. Andererseits führen solche „Unwörter“ zu etwas inkorrekten Assoziationsstärken, da z.B. *„öffentlicehn“* nicht in den Kookurrenzen mitgezählt wird, in denen das korrekt geschriebene Wort *„öffentlichen“* bzw. dessen lemmatisierte Grundform *„öffentlich“* eigentlich hätte gezählt werden müssen[22].

21 Das zugehörige Hilfsprogramm „CoCou2 – Corpus Count 2“ liegt mitsamt Quelltext und der 22 Megabyte großen, 1,2 Mio. Wortformen enthaltenden Indexdatei auf dem beigefügten Datenträger dieser Arbeit bei.

22 Eine Rückführung auf die korrekt geschriebenen Wortformen und damit eine Verbesserung der Korpus-Qualität wäre ggf. mit Hilfe von Wortähnlichkeits-Untersuchungen möglich, etwa unter Verwendung des Levenshtein-Abstandes (nach Vladimir I. Levenshtein) zweier Wörter, vgl. [LEV1965]. Der zu erwartende Rechenaufwand ist jedoch immens.

Durch Filtern aller Wörter mit einer Häufigkeit von 1 nach der Worthäufigkeitszählung werden nahezu alle fehlerhaften Wörter zuverlässig entfernt. Diese Annahme ist gültig, sofern sich die fehlerhaften Schreibweisen nicht wiederholen. Auf diese Weise reduziert sich die Anzahl der unterscheidbaren Wortformen des TAZ-Korpus von 1,2 Mio. auf lediglich etwas mehr als 526.000 verschiedene Wörter. Natürlich werden dabei neben falsch geschriebenen Fehlwörtern auch korrekt geschriebene, lediglich ungebräuchliche Wörter nicht mehr mitgezählt. Jedoch erscheint ein ungebräuchliches Wort mit der Worthäufigkeit 1 auch nur in je einer Kookurrenz von maximal 22 Wörtern (bei Kookurrenz-Fensterlänge 12 sind das 11 Wörter links und 11 Wörter rechts mit dem betreffenden Wort in der Mitte), bildet also *a priori* keine starke Kookurrenz.

Die Groß-/Kleinschreibung stellt ein Problem für den Korpus dar. Gut zu sehen ist dies an den Wörtern „wie“ und „fremd“ in Text 6.

```
wie: 244893              fremd: 4971
Wie: 46839               Fremd: 100
```

Text 6: Worthäufigkeiten „wie“ / „Wie“ / „fremd“ / „Fremd“

Keine der beiden Unterscheidungen macht hinsichtlich der unterschiedlicher Wortstammformen oder angenommener Homographen Sinn. Natürlich ließen sich derartige Wörter zusammenfassen, allerdings nur durch Eingriff in die verwendete Software bzw. durch Re-Lemmatisierung des Textkorpus ohne Groß-/Kleinschreibung.

Komposita kommen im Korpus niemals mit Bindestrich vor, sondern immer ohne. Ungebräuchliche Komposita in den Testtexten müssen zwangsläufig in Einzelsubstantive aufgespaltet werden, um sie überhaupt untersuchungsfähig zu machen, da sie als zusammengesetzte Wörter nicht im Korpus enthalten sind.

```
Geschmack: 1730          Fremd: 100
Nuance: 289              Sprache: 10478
Geschmacksnuance: 0      Fremdsprache: 130
```

Text 7: Komposita-Häufigkeiten „Geschmacksnuance“ / „Fremdsprache“

Ein Beispiel dafür ist „*Geschmacksnuance*“ (s. Text 7), wobei sowohl „*Geschmack*“ als auch „*Nuance*“ mit akzeptabler Worthäufigkeit im Korpus vorhanden sind, das Kompositum allerdings kein einziges Mal.

Bei so genannten „Schlagwörtern“ dagegen kann es sogar vorkommen, dass das Kompositum häufiger vorkommt als eines der beiden Grundwörter. Am Beispiel „*Fremdsprache*“ in Text 7 ist dies deutlich zu erkennen.

2.3.3 Flaschenhals Lemmatisierung

Zur *a priori*-Verbesserung der zu erwartenden Ergebnisse müsste bereits eine Stufe vorher angesetzt werden: Beim noch nicht lemmatisierten Text. Wie Böhnisch ebenfalls empfiehlt ([BÖH1997], S. 9), sollten ungebräuchliche Wörter entweder entfallen oder durch gebräuchlichere ersetzt werden. Auf diese Weise lässt sich ein noch nicht lemmatisierter Text behutsam vorbereiten, damit der Lemmatisierer ihn nicht verstümmelt, um auf Korpus-Seite auf entsprechend unverstümmelte Pendants zu treffen.

Es steht fest, dass die Lemmatisierungskomponente den sprichwörtlichen Flaschenhals des Verfahrens darstellt, der dessen Leistungsfähigkeit begrenzt. Einerseits sollen verschiedene flektierte Wortformen zu einem Wortknoten zusammengefasst werden, um die Zersplitterung des Kookurrenzpotenzials des zugehörigen Lemmas zu verhindern. Andererseits führt eben diese automatische Zusammenfassung über eine statische, begrenzte Ersetzungstabelle zu großen Problemen mit Homographen, während nicht in der Ersetzungstabelle enthaltene Wortformen trotzdem zu zersplittertem Kookurrenzpotenzial zumindest bei einem Teil des Vokabulars führt. Andererseits ist unklar, ob ungebräuchliche Komposita zerlegt werden sollten, damit zumindest die Teilwörter Kookurrenzen bilden oder ob Schlagwörter doch als Komposita stehen bleiben sollten.

Zwei Ansätze könnten diese Probleme wenn nicht vollständig lösen, so doch zumindest lindern: Zum einen müsste die Ersetzungstabelle des Lemmatisierers stark erweitert und damit fehlerbereinigt werden. Hierbei sind einige Vereinfachungen der neuen deutschen Rechtschreibung überaus hilfreich, z.B. Auftrennung von künstlichen Komposita (z.B. „kennen lernen“ statt „kennenlernen“) oder korrigierte Schreibweisen durch Rückführung auf den eigentlichen Wortstamm (z.B. „aufwändig“ statt „aufwendig“)[23].

Sinnvoll wäre darüber hinaus, insbesondere hinsichtlich der durch Homographen hervorgerufenen Mehrdeutigkeiten, die Verwendung eines syntaktischen Taggers bei der Lemmatisierung. Das heißt, Mehrdeutigkeiten bei der Lemmatisierung würden durch kontextsensitive Satzanalyse korrekt aufgelöst[24]. Dieser Ansatz führt aber nur dann zum Erfolg, wenn auch der Roh-Korpus bei der Vorab-Lemmatisierung getaggt würde. Morphy ([LEZ1999]) würde für 250 Mio. Wörter bei angenommenen 20 Wörtern pro Sekunde (vgl. [LRW1996], S. 8) dafür aber gut 144 Tage d.h. beinahe 5 Monate Rechenzeit benötigen. Dies ist leider derzeit nicht durchführbar.

23 Eine Vermischung von alter und neuer deutscher Rechtschreibung, wie sie in der aktuellen Literatur derzeit allzu häufig anzutreffen ist, würde jedoch voraussichtlich großes Durcheinander anrichten.

24 Der nahezu klassische Beispielsatz „Winde das im Winde flatternde Seil um die Winde“ kann mit Morphy ([LEZ1999]) fehlerfrei getaggt werden, wobei die drei verschiedenen Ableitungen des Wortes „Winde“ korrekt lemmatisiert werden (1. Imperativ Singular des Verbs „winden“; 2. (altertümlicher) Dativ Singular des Substantivs „Wind“; 3. Akkusativ Singular des Substantivs „Winde“).

2.4 ATA2 – Assoziative Text Analyse Version 2

Weiter oben im Rahmen der grundsätzlichen Diskussion von Wortnetzen und deren Training kam schon mehrfach das Softwarepaket „ATA2“, *Assoziative Text Analyse, Version 2*, zur Sprache, das den Autoabstracting-Experimenten dieser Arbeit zu Grunde liegt. Die Theorien, auf denen ATA2 aufbaut, sind bereits in den beiden Abschnitten 2.1 (Assoziatives Lernen), 2.2 (Assoziative Wortnetze) und 2.3 (Gesamt-Vokabular vs. Testvokabular) beschrieben worden. Die schematische Verwendungsweise ist in Abbildung 12 dargestellt.

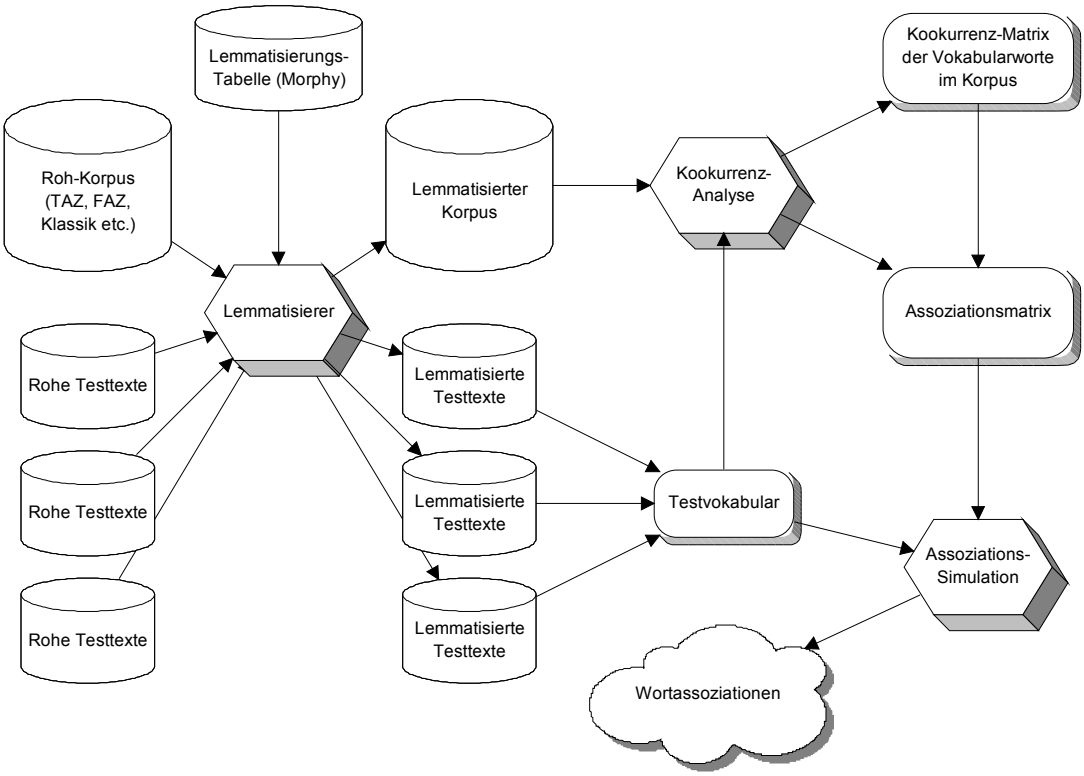

Abbildung 12: Schematische Darstellung der ATA2-Verwendungsweise

ATA2 implementiert ein Assoziatives Wortnetz, indem es ein Künstliches Neuronales Netz mit den Wörtern eines Testvokabulars als Wortknoten aufbaut. Die Assoziationsstärken zwischen den Wörtern werden nicht durch inkrementelles Lernen, sondern vielmehr durch Auszählen von Kookurrenzhäufigkeiten innerhalb eines Textkorpus und anschließender Normierung berechnet.

Der Textkorpus wurde vor der Verwendung mit Hilfe einer mit Morphy ([LEZ1999]) generierten, statischen Ersetzungstabellen lemmatisiert. Wenn Testtexte als Wortlieferanten für das zu verwendende Testvokabular dienen sollen, müssen sie ebenfalls lemmatisiert werden.

Im Anschluss an die Kookurrenzanalyse kann das Künstliche Neuronale Netz mit einem oder mehreren Reizwörtern stimuliert werden. Das Ergebnis der Reizausbreitung liefert die am stärksten mit den Stimuluswörtern assoziierten Wörter mitsamt deren Assoziationsstärken, entsprechend den Ergebnissen eines (simulierten) Assoziationsversuches mit menschlichen Versuchspersonen, deren Wortassoziationen sich allerdings auf das Testvokabular beschränken müssen.

2.4.1 Details zu ATA2

Reinhard Rapp realisierte bereits 1996 eine entsprechende Software (vgl. [RAP1996]), sozusagen „ATA1“. Die wissenschaftliche Basis geht zurück auf einen Artikel von M. Wettler, R. Rapp und R. Ferber [WRF1993] und beruht letztlich auf der Dissertation von R. Rapp von 1994 (Buchausgabe [RAP1996]). ATA2 wurde von Michael Böhnisch 1997 implementiert und im Rahmen seiner Diplomarbeit [BÖH1997] dokumentiert. Es bildet mit seinen über 30 shell-bedienbaren UNIX-Programmen quasi eine Re-Implementation von Rapps Software, die neben der Einbeziehung zusätzlicher wissenschaftlicher Ansätze im Wesentlichen schneller arbeitet und mit weitaus größeren Eingaben effizient umgehen kann.

Die Lemmatisierungs-Komponente von ATA2 verwendet Teile der ebenfalls bereits angesprochenen Morphy-Technologie von W. Lezius ([LRW1996]). Der Korpus liegt bereits lemmatisiert vor, ein Zugriff auf den Roh-Korpus ist im Betrieb von ATA2 nicht vorgesehen und auch unnötig. Der Weg zum Testvokabular durch lemmatisierte Testtexte als Wortlieferanten wird ergänzt durch das manuelle Einfügen oder auch Entfernen von erwünschten bzw. unerwünschten Testwörtern. Böhnischs resultierendes Testvokabular besteht aus insgesamt etwa 5.600 Wörtern (s. [BÖH1997], Fußnote S. 10).

Nach jeder Modifikation des Testvokabulars muss ATA2 die Kookurrenzanalyse der Vokabularwörter über den gewünschten Korpus wiederholen. Die dabei verwendete Fenstertechnik schließt Mehrfachzählungen von „nahen Kookurrenzen“ aus (vgl. Abschnitt 2.1.2.3, S. 32ff). Die CPU-Zeit für die Kookurrenzanalyse verhält sich linear zur Größe des verwendeten Korpus. Ein kompletter Durchlauf bei 5.600 Wörtern im Testvokabular benötigt auf der zur Verfügung stehenden Workstation (Sun Sparc Ultra 1 `psycho1.upb.de` bei 166 MHz CPU-Taktfrequenz (UltraSPARC) mit 128 MB Arbeitsspeicher) knapp eineinhalb Stunden. Eine anschließende Assoziationsberechnung (ohne mehrfache Reizpropagierung) benötigt etwa zweieinhalb Minuten.

2.4.2 Mögliche Anwendungen

In verschiedenen Bereichen kann es sinnvoll sein, die „durchschnittlichen“ Assoziationsstärken einer bestimmten Gruppe von Menschen auf Stimuluswörter zu kennen, und diese nicht durch langwierige Feldstudien zu ermitteln, sondern eine preiswerte und schnelle Computersimulation anzuwenden.

Im Bereich der Werbung (gleichermaßen in der Presse sowie in Funk und Fernsehen) interessiert es den Werbetreibenden, was der potenzielle Kunde „zwischen den Zeilen liest“ (vgl. [WR1993]). Dies kann bereits vorab im Bereich des Produktdesigns einfließen, um die potenzielle Wirkung neuer Produkte oder Werbespots vor der eigentlichen Markteinführung richtig einschätzen zu können. Weiterhin kann die werbeübliche Methode, über Assoziationen unterschwellige, über die eigentlichen Produkteigenschaften hinausgehende Suggestionen zu transportieren, mit simulierter Assoziation auf ihre Effizienz überprüft werden.

Gleichermaßen kann in der Politik beim Verfassen einer Rede Wert auf eine bestimmte Wortwahl gelegt werden, die bei den Zuschauern bzw. Zuhörern bestimmte Aspekte stärker oder schwächer assoziiert – häufig sogar unbewusst. Im weitesten Sinne ähnelt eine politische Rede, bei der die eigene Meinung der Opposition als beste Option suggeriert werden soll, in vielen Punkten einer Produktwerbung. Spätestens im Falle der Wahlwerbung sind die Grenzen zwischen Politik und Werbung fließend.

Weiterhin können Assoziative Wortnetze im Information Retrieval dazu verwendet werden, die Wörter aus Suchanfragen an ein Information Retrieval System basierend auf Wortassoziationen zu erweitern, um verbesserte Suchergebnisse zu erzielen. Ferber, Wettler und Rapp haben dies [FWR1995] anschaulich demonstriert.

Im Zusammenspiel oder auch als Ersatz für einen Thesaurus könnte ein Assoziatives Wortnetz auch in ein Textverarbeitungssystem integriert werden, wo es ähnlich einem Thesaurus Synonyme oder verwandte Begriffe liefern könnte.

2.5 Veränderte Anwendungsweise von ATA2

Das ursprüngliche Einsatzgebiet von ATA2 liegt in der Simulation freier menschlicher Assoziationen über die eingegebenen Stimuluswörter hinaus.

Ein für jeden Menschen mit Werbeerfahrung nachvollziehbares Beispiel für derartige Werbeassoziationen beginnt mit der Vorstellung einer Prärielandschaft, berittenen Männern mit derben Gesichtern und Cowboy-Hüten beim Viehumtrieb, einer Kaffee-

kanne auf dem offenen Lagerfeuer – und assoziiert unvermeidlich eine bestimmte amerikanische Zigarettenmarke mit rot-weißem Emblem, deren Markenname nicht einmal mehr genannt zu werden braucht. Fast fühlt man sich wiederum an Pawlows Versuchshunde bei der klassischen Konditionierung (vgl. z.B. [ZIM1992], S. 231) erinnert, deren Speicheltätigkeit aufgrund der ausgebildeten Assoziation zwischen Glockenklang und Fütterung alsbald auch bei alleinigem Glockenklang einsetzte. – Wobei Assoziation im James'schen Sinne ([JAM1890]) natürlich in beiden Richtungen funktioniert: Das heißt, das Rauchen einer bestimmten Zigarettenmarke soll in Gedanken die Freiheit der Cowboys beim Viehumtrieb wachrufen – und beim Fressen dürften die Pawlows Hunde ebenfalls noch den Klang der Glocke assoziiert haben, obwohl man sie natürlich schwerlich fragen konnte.

All das ist für das hier untersuchte Autoabstracting nicht ausreichend: Gesucht sind keine freien, erlernten Assoziationen als konditionierte Stimuli, sondern die Verwendung des Assoziierens als kognitiver Prozess zur Auswahl der Kernwörter in Texten.

2.5.1 Verbot der freien Assoziationen

Beim Autoabstracting geht es, wie auch beim identischen nicht-computergestützten Vorgang durch Menschen, um das Herausstellen der Kernwörter des Textes zum Verfassen eines Autoabstracts und ggf. zur Einspeisung dieser Worte als Schlagworte in Literaturdatenbanken. Es sollen also im Text enthaltene Wörter extrahiert werden und keine zusätzlichen Schlüsselwörter „hinzu assoziiert" werden.

Der Vorgang des freien Assoziierens gelingt uns Menschen dadurch, dass wir zu einem Stimuluswort Assoziationen zu nicht genannten Wörtern erlernt haben. So gelingt uns mühelos die Assoziation von „Schokolade", „Kuh" und „lila" auf eine bestimmte Schokoladenmarke – und auch ATA2 vollbringt diese Leistung problemlos (vgl. [BÖH1997], S. 37).

Wie beim Menschen liegt die Ursache für das freie Assoziieren auch bei ATA2 in seinem größeren Wortschatz bzw. Vokabular. Das ATA2-Testvokabular bestand zum Zeitpunkt von Böhnischs Diplomarbeit aus etwa 5.600 Wörtern (s. [BÖH1997], Fußnote S. 10). Eine Stärke des ATA2-Systems liegt in der leichten Erweiterbarkeit des Vokabulars, woran sich jedoch in jedem Fall eine erneute Kookurrenzanalyse bzw. ein erneutes Training des Neuronalen Netzes mit einem möglichst diversifizierten Textkorpus anschließen muss. Dabei sollte jedoch tunlichst darauf geachtet werden, dass die hinzugefügten Wörter w' überhaupt im Vokabular des verwendeten Korpus T enthalten sind, d.h. $H(w')_T > 0 \; \forall \, w'$.

Während man einem Menschen beim Lesen eines Textes schwerlich das Assoziieren textfremder Wörter verbieten kann, lässt sich ATA2 durch einen Trick davon überzeugen: Das Testvokabular wird *komplett* geleert, und lediglich die Wörter des Textes werden zugelassen. In der schematischen Abbildung 12 auf S. 51 der ATA-Funktionsweise entspräche das der Modifikation, nur noch den gerade betrachteten Testtext als Wortlieferanten zuzulassen und dessen Wörter in das Testvokabular einzufügen.

So bleibt vom lemmatisierten „Madagaskar"-Text (siehe Anhang bzw. Text 10 auf Seite 60) nach Eliminierung der Wort-Duplikate und alphabetischer Sortierung lediglich das in Text 8 dargestellte Testvokabular übrig, das aus 125 Wörtern besteht.

```
Affe, Asphalt, Auftritt, Baum, Beton, Boot, Dank, Die, Ding, Entwicklung,
Erde, Familie, Feuer, Flugzeug, Gebiet, Gehirn, Hang, Haustier, Insel, Jahr,
Jahrtausend, Kampf, Kanu, Lebensraum, Lemuren, Macheten, Madagaskar, Makis,
Mal, Million, Nachfahre, Primat, Sache, Technologie, Und, Vor,
Weiterentwicklung, Widersacher, Wie, Während, Zeit, Zweig, ab, aber,
aggressiv, all, allein, allerdings, an, auf, aus, begnügen, bleiben,
buddeln, damit, dann, das, dem, den, denselben, der, deren, die, diesmal,
doch, ehrgeizig, ein, eineinhalb, entscheiden, erreicht, erstaunlich,
fertigbringen, fühlen, für, gleich, größere, gut, haben, hauen,
herausfinden, herum, im, in, interessiert, jedoch, kommen, kurz, kämpfen,
können, mit, möglich, nach, nehmen, nicht, nur, oder, sagen, schließlich,
sein, sich, sie, spät, stammen, sterben, stochern, tun, um, und, verfügt,
verschonen, von, vor, vorbeigehen, vorfahren, weiter, wie, wieder, wir,
wohl, zu, zwar, über, überall, überleben, übernehmen
```

Text 8: Vokabular des lemmatisierten „Madagaskar"-Textes

Selbstverständlich muss auch hier dafür Sorge getragen werden, dass die Wörter des Minimal-Vokabulars im Korpus enthalten sind.

Wird das Wortnetz von ATA2 nun mit diesem kleinen Vokabular trainiert, bleibt die quadratische Assoziationsmatrix bzw. das Neuronale Netz aufgrund der nur 125 Zeilen bzw. Spalten geradezu „handlich" klein. Die für das Auszählen der Kookurrenzen benötigte Zeit sinkt von eineinhalb Stunden auf 45 bis 60 Minuten, die Assoziationssimulation benötigt statt einiger Minuten nur noch wenige Sekunden.

2.5.2 Worthäufigkeitsanalyse zur Vokabularreduktion

Das in Text 8 dargestellte Testvokabular besteht offensichtlich aus Wörtern, die den Sinn des Textes ausmachen (z.B. „Madagaskar", „Lemuren", „Affe" etc.), sowie aus Füllwörtern, deren Beitrag zum Textsinn eher gering ist. Es handelt sich dabei hauptsächlich um *Funktionswörter* wie bestimmte und unbestimmte Artikel („der", „die", „das", „ein", „eine", etc.), Konjunktionen („und", „oder", etc.) und Präpositionen („in", „bei", „durch", etc.). – In der Computerlinguistik nennt man diese Wort häufig *Stoppwörter*.

Wie bereits im „klassischen“ Algorithmus von H. P. Luhn (vgl. [LUH1958]) sollen nun auch hier vorab sowohl die häufigen Stoppwörter als auch die seltenen Wörter aus dem Testvokabular entfernt werden. — Zur Erinnerung: Die Häufigkeitsverteilung von Wörtern, die sich mit Hilfe des Zipfschen Gesetzes (etwa in [FER2003], S. 67) beschreiben lässt, impliziert zweierlei: Zum einen, dass die sehr häufigen Wörter, die das meiste Textvolumen ausmachen, allein aufgrund ihrer großen Zahl nur wenig zum Textsinn beitragen. Zum zweiten kann bei extrem seltenen Wörtern angenommen werden, das sie nicht in allen relevanten Texten vorkommen – etwa, weil es sich um seltene Synonyme von ansonsten weitaus besser geeigneten Kernwörtern handelt, die ebenfalls im Text enthalten sind.

Häufige Wörter könnten auf einfache Weise mit Stoppwortlisten entfernt werden, seltene Wörter erfordern jedoch eine andere Vorgehensweise. Sinnvoll wäre ein statistischer Weg, der sowohl die häufigen als auch die seltenen Wörter in einem Arbeitsschritt ausfiltert. Ein geeignetes statistisches Mittel ist die Worthäufigkeitsklassenanalyse im Sinne von [QUA1998]: Die Häufigkeitsklasse bildet ein logarithmisches Maß für ein beliebiges Wort im Korpus – im Vergleich zum häufigsten Wort des Korpus. Das häufigste Wort im Korpus ist das Wort „der“: Es kommt 7.680.846x vor. Die Häufigkeitsklasse $c(w)$ für ein Wort beliebiges w berechnet sich wie folgt[25]:

$$c(w)=\left\lfloor \log_2\left(\frac{h(\text{der})}{h(w)}\right)\right\rfloor=\left\lfloor \log_2\left(\frac{7.680.846}{h(w)}\right)\right\rfloor$$

Die Häufigkeitsklasse des Wortes „der“ ist 0, ebenso die der Wörter „die“, „in“ und „und“. Je seltener ein Wort im Korpus vorkommt, desto höher seine Häufigkeitsklasse. Die in den verwendeten Testtexten maximale Häufigkeitsklasse von 19 wird durch das seltene Wort „Makis“ im „Madagaskar“-Text erreicht mit einer absoluten Worthäufigkeit von 9 Vorkommnissen im Korpus. Das Wort „Madagaskar“ selbst fällt mit der absoluten Häufigkeit von 478 in die Worthäufigkeitsklasse 13. In Tabelle 13 auf Seite 57 sind sämtliche Worthäufigkeiten und Worthäufigkeitsklassen des „Madagaskar“-Textes dargestellt.

```
Affe, Asphalt, Auftritt, Baum, Beton, Boot, Dank, Ding, Entwicklung, Erde,
Familie, Feuer, Flugzeug, Gebiet, Gehirn, Hang, Haustier, Insel,
Jahrtausend, Kampf, Kanu, Lebensraum, Lemuren, Macheten, Madagaskar, Mal,
Nachfahre, Primat, Sache, Technologie, Weiterentwicklung, Widersacher,
Während, Zweig, aggressiv, begnügen, buddeln, denselben, diesmal, ehrgeizig,
eineinhalb, erreicht, erstaunlich, fertigbringen, fühlen, gleich, größere,
hauen, herausfinden, herum, interessiert, kämpfen, schließlich, stammen,
sterben, stochern, verfügt, verschonen, vorbeigehen, vorfahren, überall,
überleben, übernehmen
```

Text 9: Wortklassen-reduziertes Vokabular des lemmatisierten „Madagaskar“-Textes

25 Je nach Definition der Häufigkeitsklasse wird nicht abgerundet sondern „konventionell“ ab- bzw. aufgerundet.

Im weiteren Verlauf der Untersuchungen wird das jeweilige Testvokabular daher auf die Wörter der mittleren Worthäufigkeitsklassen beschränkt. Nach Evaluation der Worthäufigkeitsklassen der in allen Testtexten enthaltenen Wörter fiel die willkürliche Entscheidung „nach Augenmaß“, nur die Wörter der Worthäufigkeitsklassen 7 bis 17 einschließlich ins Testvokabular aufzunehmen. Für den „Madagaskar“-Text verblieben dadurch nur noch die in Text 9 auf Seite 56 dargestellten 63 von ehemals 125 Wörtern (vgl. Text 8 auf Seite 55) im Vokabular.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Makis	9	19	Feuer	14998	9	doch	222674	5
Lemuren	74	16	herum	15377	8	dann	278204	4
Macheten	134	15	Auftritt	16316	8	Und	288634	4
buddeln	210	15	kämpfen	17036	8	wieder	292849	4
stochern	386	14	diesmal	17850	8	wir	314506	4
fertigbringen	453	14	überall	17866	8	gut	351025	4
Madagaskar	478	13	überleben	18240	8	all	368424	4
Nachfahre	745	13	Insel	18919	8	sagen	388138	4
Primat	758	13	Baum	19372	8	kommen	410579	4
Haustier	949	12	stammen	22698	8	oder	474420	4
Kanu	953	12	Gebiet	26104	8	aber	556708	3
Asphalt	1163	12	Ding	31675	7	über	626853	3
Widersacher	1460	12	fühlen	40024	7	nur	631085	3
vorbeigehen	1480	12	gleich	42226	7	vor	660957	3
Weiterentwicklung	1626	12	sterben	43553	7	Jahr	706524	3
Lebensraum	1905	11	Während	44174	7	um	739537	3
vorfahren	2143	11	Entwicklung	45188	7	wie	756071	3
Affe	2297	11	übernehmen	46193	7	nach	845550	3
Beton	2608	11	schließlich	47287	7	können	886588	3
verschonen	2665	11	Kampf	47953	7	aus	939461	3
verfügt	2879	11	Sache	51061	7	an	1110543	2
Zweig	2975	11	Familie	53078	7	sie	1394891	2
herausfinden	3022	11	entscheiden	60037	6	Die	1459367	2
begnügen	3058	11	Vor	64800	6	dem	1577653	2
denselben	3257	11	weiter	67277	6	im	1622051	2
ehrgeizig	3416	11	allein	68029	6	für	1632866	2
Jahrtausend	3696	11	deren	77274	6	auf	1707225	2
Gehirn	4095	10	kurz	80368	6	sich	1845848	2
eineinhalb	4418	10	zwar	90344	6	nicht	1940174	1
Erde	4688	10	allerdings	90833	6	mit	1975245	1
interessiert	4734	10	wohl	91128	6	das	2057789	1
aggressiv	5762	10	möglich	95705	6	zu	2236454	1
Mal	6430	10	spät	101138	6	von	2323003	1
Technologie	7145	10	jedoch	118330	6	haben	2451209	1
erstaunlich	7153	10	tun	136947	5	den	2854838	1
Boot	8086	9	Wie	137699	5	ein	3660134	1
Dank	9231	9	nehmen	155916	5	in	4282554	0
erreicht	10913	9	damit	159852	5	sein	5073512	0
hauen	11055	9	ab	173860	5	und	5523379	0
größere	11380	9	Million	197164	5	die	7106150	0
Flugzeug	12010	9	bleiben	197805	5	der	7680846	0
Hang	12385	9	Zeit	214685	5			

Tabelle 13: Worthäufigkeiten und Worthäufigkeitsklassen im „Madagaskar“-Text

2.5.3 Stimulierung aller Wortknoten

Bei der üblichen Anwendung von ATA2 enthält der Stimulus nur eines oder wenige Wörter, um die daraus resultierenden freien Assoziationsstärken zu ermitteln. Im Gegensatz dazu soll hier das Lesen eines kompletten Textes simuliert werden, mit anschließender Ermittlung der Kernwörter durch Analyse der stärksten Assoziation. Ko-

gnitionswissenschaftlich betrachtet sind damit also die Wörter gemeint, die beim Leser „am eindrücklichsten wahrgenommen wurden“, in diesem Falle bereits reduziert auf die Wörter der mittleren, potenziell bedeutsamen Worthäufigkeitsklassen.

Daher muss ATA2 bei der Assoziationssimulation in der Eingabe *sämtliche* Wörter des Textes aktivieren. Im Vergleich mit Berechnungsvorschrift (1) auf Seite 42 werden also alle Komponenten des Eingabevektors v_0 mit 1 initialisiert. Somit ergibt sich:

$$\Psi = \begin{pmatrix} 0 & \psi_{12} & \psi_{13} & \cdots & \cdots & \psi_{1n} \\ \psi_{12} & 0 & \psi_{23} & \cdots & \cdots & \psi_{2n} \\ \psi_{13} & \psi_{23} & 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 & \psi_{(n-1)n} \\ \psi_{1n} & \psi_{2n} & \cdots & \cdots & \psi_{(n-1)n} & 0 \end{pmatrix}, \quad v_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix} . \text{ Damit: } v_1 = \| \Psi \cdot v_0 \| . \text{ (2)}$$

Die Länge des Eingabevektors v_0 sowie des Ausgabevektors v_1 entspricht wiederum der Vokabulargröße – im Falle des „Madagaskar“-Textes also 63. Die berechnete Aktivierung eines Wortes, repräsentiert durch seine Komponente des Ausgabevektors v_1, ist in diesem Sonderfall also die Summe der Assoziationsstärken dieses Wortes zu *allen* anderen Wörtern des Textes.

2.5.3.1 Assoziationen höheren Grades

ATA2 soll dazu verwendet werden, um mit Hilfe der simulierten Wortassoziation die Kernwörter des Textes bzw. des reduzierten Vokabulars zu ermitteln. Das heißt, gesucht sind die Wortknoten des Neuronalen Netzes, die nach gleichmäßiger Aktivierung aller Wortknoten nach der Reizpropagierung am stärksten aktiviert sind. Dies entspricht kognitionswissenschaftlich einem Menschen, der alle Wörter des Textes einmalig liest und die Wörter memorieren soll, die „am eindrücklichsten wahrgenommen wurden“. Dabei handelt es sich folglich um die Primärassoziationen, die Assoziationen ersten Grades.

Assoziationen höheren Grades implizieren, dass eine wiederholte Reizausbreitung stattfindet: Jede weitere Reizpropagierung entspricht dabei der Berechnungsvorschrift (1') (siehe S. 43), bei dem sich der neue Eingabevektor v_i' durch Aufsummieren des ursprünglichen Eingabevektors v_0 auf den Ausgabevektor v_i errechnet. v_0 entspricht aber laut Berechnungsvorschrift (2) (s.o.) bereits dem „Vollstimulus“, dessen sämtliche Beträge 1 sind – d.h. das Ergebnis v_i des vorherigen Berechnungsschritts wird dem Vollstimulus v_0 quasi „aufgeprägt“.

Bezogen auf den Anwendungsfall entsprächen diese unterschiedlichen Gewichte etwa der unterschiedlich intensiven Präsentation der einzelnen Wörter, z.B. in unterschiedlicher Dauer oder unterschiedlich gut lesbar – einer Art „verzerrtem Vollstimulus“.

Unabhängig davon, wie viele Berechnungsschritte letztlich durchgeführt würden, ergebnisrelevant wäre nur der letzte Eingabevektor, d.h. der letzte verzerrte Vollstimulus. Dadurch, dass für die finale Berechnung nicht alle Vektoreinträge genau 1 sind, ergibt sich für die Assoziation ein verzerrtes Bild, das zu verfälschen Ergebnissen führt. Diese Schlussfolgerung ließ sich empirisch in Vorversuchen bestätigen: Die Resultate der Assoziationen höheren Grades waren im Sinne der gesuchten Kernwörter völlig unbrauchbar.

Aus methodischer Sicht sind somit lediglich die Assoziationen ersten Grades bzw. deren Aktivierungsstärken von Interesse. Auf jegliche multiple Reizpropagierung kann bzw. muss daher verzichtet werden.

2.5.4 Manuelle Vorbereitung der Texte

Die Untersuchungen hinsichtlich der Qualität von ATA2s Lemmatisierungskomponente in Kapitel 2.3 auf Seite 43ff führten zu der Entscheidung, die im Rahmen dieser Arbeit zu verarbeitenden Texte durch Wortersetzungen noch vor der Lemmatisierung marginal zu vereinfachen. Dabei wurden die Resultate der exemplarischen Untersuchung des TAZ-Korpus (vgl. Abschnitt 2.3.2 auf S. 47) bzgl. der Worthäufigkeiten auf den Gesamtkorpus verallgemeinert.

Zu beachten war, dass die Texte in zweierlei Hinsicht benutzbar sein mussten: Einerseits für die Verarbeitung mit ATA2, andererseits für die Wortmarkierungsstudie mit den Versuchsteilnehmern.

Insbesondere Komposita bedurften einer besonderen Betrachtung: Kam das Kompositum in seiner lemmatisierten Form mit mehr als verschwindend geringer Häufigkeit im (TAZ-)Korpus vor, konnte es sowohl im Text der Wortmarkierungsstudie verbleiben, als auch im Vokabular für die ATA2-Kookurrenzanalyse mit anschließender Assoziationsberechnung. War hingegen das Kompositum nicht oder nur selten im Korpus vorhanden, wurde es manuell im Text und somit danach durch die Lemmatisierungskomponente in seine Grundwörter bzw. deren lemmatisierte Grundformen aufgespalten. Anderenfalls wäre es spätestens durch die Filterung der hohen Worthäufigkeitsklassen oberhalb von 17 aus dem Testvokabular entfernt worden.

Generell wurden Wörter mit sehr geringen Häufigkeiten gegen Wörter mit etwas höheren Häufigkeiten ausgetauscht: Etwa enthielt einer der Texte das eher ungebräuchliche Wort „Abfallkästchen“, das durch „Abfallkorb“ ersetzt wurde. Dies geschah wiederum im Originaltext noch vor der Lemmatisierung, damit die Wörter der Wortmarkierungsstudie den lemmatisierten Wörtern der ATA2-Analyse vergleichbar blieben.

1. Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb. Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

2. Die entscheiden Entwicklung die an Madagaskar vorbeigehen sein das Auftritt der Affe sie stammen zwar von den gleich vorfahren wie die Lemuren ab verfügt jedoch über größere Gehirn und sein aggressiv Widersacher im Kampf um denselben Lebensraum Während die Makis sich damit begnügen haben in den Baum herum zu Hang und sich wohl zu fühlen sein die Affe ehrgeizig und interessiert sich für all möglich vor all für die Zweig Wie sie nach kurz Zeit herausfinden können sie damit Ding tun die sie allein nicht fertigbringen nach Sache buddeln in Sache herum stochern auf Sache herum hauen Die Affe übernehmen die Erde und der Lemuren Zweig der Primat Familie sterben überall aus nur auf Madagaskar nicht das für Million Jahr von Affe verschonen bleiben Vor eineinhalb Jahrtausend kommen die Affe dann aber schließlich doch auf Madagaskar an oder gut sagen kommen deren Nachfahre auf Madagaskar an wir Dank erstaunlich Weiterentwicklung auf dem Gebiet der Zweig Technologie erreicht wir die Insel mit Kanu spät mit Boot und schließlich mit Flugzeug und nehmen den Kampf um den Lebensraum ein weiter Mal auf diesmal allerdings mit Feuer und Macheten mit Haustier Asphalt und Beton Und wieder kämpfen die Lemuren um überleben

3. Affe, Asphalt, Auftritt, Baum, Beton, Boot, Dank, Die, Ding, Entwicklung, Erde, Familie, Feuer, Flugzeug, Gebiet, Gehirn, Hang, Haustier, Insel, Jahr, Jahrtausend, Kampf, Kanu, Lebensraum, Lemuren, Macheten, Madagaskar, Makis, Mal, Million, Nachfahre, Primat, Sache, Technologie, Und, Vor, Weiterentwicklung, Widersacher, Wie, Während, Zeit, Zweig, ab, aber, aggressiv, all, allein, allerdings, an, auf, aus, begnügen, bleiben, buddeln, damit, dann, das, dem, den, denselben, der, deren, die, diesmal, doch, ehrgeizig, ein, eineinhalb, entscheiden, erreicht, erstaunlich, fertigbringen, fühlen, für, gleich, größere, gut, haben, hauen, herausfinden, herum, im, in, interessiert, jedoch, kommen, kurz, kämpfen, können, mit, möglich, nach, nehmen, nicht, nur, oder, sagen, schließlich, sein, sich, sie, spät, stammen, sterben, stochern, tun, um, und, verfügt, verschonen, von, vor, vorbeigehen, vorfahren, weiter, wie, wieder, wir, wohl, zu, zwar, über, überall, überleben, übernehmen

4. Affe, Asphalt, Auftritt, Baum, Beton, Boot, Dank, Ding, Entwicklung, Erde, Familie, Feuer, Flugzeug, Gebiet, Gehirn, Hang, Haustier, Insel, Jahrtausend, Kampf, Kanu, Lebensraum, Lemuren, Macheten, Madagaskar, Mal, Nachfahre, Primat, Sache, Technologie, Weiterentwicklung, Widersacher, Während, Zweig, aggressiv, begnügen, buddeln, denselben, diesmal, ehrgeizig, eineinhalb, erreicht, erstaunlich, fertigbringen, fühlen, gleich, größere, hauen, herausfinden, herum, interessiert, kämpfen, schließlich, stammen, sterben, stochern, verfügt, verschonen, vorbeigehen, vorfahren, überall, überleben, übernehmen

Text 10: „Madagaskar“-Text – Lesbare und lemmatisierte Version, komplette und verkürzte Wortliste

Offensichtliche Fehl-Lemmatisierungen aufgrund von Homographen wurden manuell bereinigt, sofern im Korpus die korrekten Wortformen in hinreichender Häufigkeit enthalten waren. Falls jedoch die offensichtlich richtige Wortform bereits zum Lemmatisierungs-Zeitpunkt des Korpus zugunsten der falschen Wortform „weglem-

matisiert“ worden war, so dass die richtige Wortform nicht oder nur mit verschwindender Worthäufigkeit im Korpus enthalten war, verblieb die falsche Wortform im zu untersuchenden Textvokabular. Kognitionswissenschaftlich betrachtet bedeutet dies, dass das Wortnetz die beiden Homographen durch falsches oder fehlendes Vorwissen einfach nicht besser unterscheiden kann – ein mitunter auch bei Menschen anzutreffender Fehler.

2.5.4.1 Finale Textvorbereitung

Am Ende der manuell unterstützten Lemmatisierung lag jeder Text jeweils in zwei Versionen vor, wie in Text 10 auf S. 60 für den „Madagaskar“-Text dargestellt.

1. Als ASCII-Fließtext für die Weiterverarbeitung durch die PHP-Programmierung in der Wortmarkierungsstudie für die Wortmarkierungen der Versuchsteilnehmern.
2. Als lemmatisierte „Pidgin-Deutsch“-Version zur Weiterverarbeitung mit ATA2.

   Die lemmatisierte Version dient als Wortlieferant für das Testvokabular:
3. Die duplikateliminierte Wortliste umfasst 125 unterscheidbare, lemmatisierte Wortformen (vgl. Tabelle 13 auf S. 57).
4. Nach Analyse der Worthäufigkeitsklassen (durch eine vorherige Worthäufigkeits-Zählung mit ATA2) und Reduktion auf die mittleren Worthäufigkeitsklassen 7-17 verkürzt sich das Vokabular auf nur noch 63 Wörter (vgl. Text 9 auch S. 56).

Das verkürzte lemmatisierte Vokabular eines Textes dient bei der Kernwortsuche als einzige Informationsquelle. Diese Vorgehensweise gilt identisch für alle im Rahmen dieser Arbeit untersuchten Texte.

ATA2 benötigt die Eingabe des verkürzten Vokabulars in zweimaliger Ausfertigung: Einerseits zum Aufbau des Testvokabulars, um damit anschließend die Kookurrenzanalyse über den Korpus auszuführen; andererseits als Stimulus-Input für die Assoziationssimulation, die die eigentliche Rangfolge der Wörter hinsichtlich ihrer Wichtigkeit im jeweiligen Text liefern soll. In beiden Fällen ist die vollständige, verkürzte Wortliste des Textes gemeint, denn der spätere Stimulus aktiviert ja sämtliche Knoten des Neuronalen Netzes einheitlich stark, d.h. sämtliche Vokabular-Wörter.

2.5.5 Unerwünschter Worthäufigkeits-Bias

Der Grund für die Worthäufigkeitsklassenanalyse (s. Abschnitt 2.5.2 auf Seite 55) bzw. für die Filterung der extrem häufigen Wörter liegt in der starken Assoziation der Funktions- bzw. Stoppwörter begründet. Ein Vorversuch am Beispiel des „Madagaskar"-Textes mit Stimulierung aller Wortknoten des Neuronalen Netzes lieferte das in Tabelle 14 dargestellte Resultat.

Die berechnete Assoziationsstärke in der zweiten Spalte der Tabelle entspricht dabei der Aktivierung des zugehörigen Wortknotens bei einmaliger Reizpropagierung nach gleichmäßiger Aktivierung sämtlicher Knoten, wie in Formel (2) auf Seite 58 beschrieben.

#	Assoziationsstärke	Wort	Korpus-Worthäufigkeit
1	0,4449483119	Lemuren	74
2	0,3347199619	der	7680846
3	0,2700166229	die	7106150
4	0,2281698363	und	5523379
5	0,2172752146	Die	1459367
6	0,2079269300	den	2854838
7	0,2060544004	aus	939461
8	0,1870635323	in	4282554
9	0,1704912207	sein	5073512
10	0,1605480329	ein	3660134

Tabelle 14: Erster Vortest des „Madagaskar"-Textes mit stark assoziierten Funktionswörtern

Der Grund dafür liegt auf der Hand: In der ursprünglichen Verwendung von ATA2 waren die Funktionswörter nicht im untersuchten Vokabular von ca. 5.600 Wörtern enthalten – gemäß Böhnischs Empfehlung ([BÖH1997], S. 10). Dadurch werden sie bei der Kookurrenzanalyse automatisch komplett ignoriert, was auch überaus sinnvoll ist: Etwa die Artikel „der", „die" und „das" sowie deren Flexionen kookurrieren sprachbedingt mit nahezu jedem Substantiv im Textkorpus. Weiterhin haben derartige Wörter eine entsprechend hohe Worthäufigkeit innerhalb des Korpus: Durch Kookurrenz mit anderen, ebenso häufigen Funktionswörtern führt dies zu extrem hohen Assoziationsstärken untereinander. Die Stimulierung des Netzes mit sämtlichen Eingabewörtern des Textes, d.h. insbesondere auch der Funktionswörter, provoziert somit unabwendbar die starke Aktivierung ebendieser Wörter.

Ursprünglich erhob diese Arbeit den Anspruch, dass ATA2 ohne jegliche „Schützenhilfe" des Versuchsleiters einer Wortvorauswahl auskommen sollte. Nichtsdestotrotz war die Worthäufigkeitsklassenanalyse zur Vokabularreduktion (s. Abschnitt 2.5.2) ein automatisch durchführbarer, adäquater Weg. Ein weiterer Testlauf mit dem auf die Worthäufigkeitsklassen 7-17 reduzierten Vokabular wiederum beim „Madagaskar"-Text ist in Tabelle 15 auf S. 63 dargestellt.

#	Assoziationsstärke	Wort	Korpus-Worthäufigkeit
1	0,7222943471	Lemuren	74
2	0,3673782413	Madagaskar	478
3	0,2751325323	Beton	2608
4	0,1647911881	Baum	19372
5	0,1635945754	herum	15377
6	0,1577602623	Kanu	953
7	0,1511272232	Affe	2297
8	0,1445557849	Primat	758
9	0,1420865896	Insel	18919
10	0,1104831008	Feuer	14998

Tabelle 15: Vortest des „Madagaskar"-Textes mit reduziertem Vokabular

Ohne genauere Untersuchung der Resultate der Wortmarkierungsstudie wirkt dieses Ergebnis bereits weitaus plausibler. Nichtsdestotrotz führten Detailanalysen mit den im Ergebnis-Teil dieser Arbeit entwickelten Analysemethoden zu nicht wirklich befriedigenden Resultaten. Nachfolgend stelle ich den Versuch vor, die Ergebnisse zu verbessern.

2.5.5.1 Korrektur des Worthäufigkeits-Bias

Bei genauer Betrachtung der Ergebnisse der Assoziationsberechnung konnte eine systematische Überbetonung (engl. „*bias*") von Wörtern beobachtet werden, die im Korpus häufig vorkommen. Dem lässt sich mit einer Korrekturrechnung begegnen.

Einerseits sei verallgemeinernd angenommen, dass die Wörter mit größerer Korpushäufigkeit weniger zum Sinn des Textes beitragen. Daher wird die berechnete Assoziationsstärke a_w eines jeden Wortes w durch die prozentuale Worthäufigkeit im Korpus T dividiert, d.h. $H(w)_T / |T|$.

Also: $a'_w = a_w / \frac{H(w)_T}{|T|} = \frac{a_w \cdot |T|}{H(w)_T}$

Weiterhin ist die Texthäufigkeit eines Wortes innerhalb des aktuell untersuchten Textes in Übereinstimmung mit [LUH1958] als Maß für seine Wichtigkeit im Text interpretierbar: Ein Wort, das mehrfach im Text vorkommt, trägt wahrscheinlicher zu dessen Sinn bei, als ein Wort, das einmal im Text vorkommt. Daher wird die soeben berechnete Assoziationsstärke a'_w jedes assoziierten Wortes w zusätzlich mit ihrer prozentualen Worthäufigkeit im aktuell betrachteten Text X multipliziert, d.h. $H(w)_X / |X|$.

Also: $a''_w = a'_w \cdot \frac{H(w)_X}{|X|} = \frac{a_w \cdot |T|}{H(w)_T} \cdot \frac{H(w)_X}{|X|} = \frac{a_w \cdot H(w)_X}{H(w)_T} \cdot \frac{|T|}{|X|}$

Der Quotient $|T| / |X|$, d.h. das inverse Verhältnis des Text-Umfangs zum Korpus-Umfang ist für sämtliche Wörter des Testvokabulars desselben Textes identisch. Er bildet also für alle Wörter lediglich einen konstanten Faktor, der die Rangfolge aber nicht

beeinflusst. Weiterhin verschwindet jeder konstante Faktor bei anschließender Normierung ohnehin. Der Quotient kann daher ersatzlos entfallen. Als endgültige korrigierte Assoziationsstärke b_w jedes assoziierten Wortes w ergibt sich daher aus der ursprünglich berechneten Assoziationsstärke a_w:

$$b_w = \frac{a_w \cdot H(w)_x}{H(w)_T} \quad \forall w \in V.$$

Im Anschluss werden die Resultate mit Hilfe der üblichen 2-Norm normiert, d.h. der Euklidischen Norm $\|b\| := b/|b|$.

Zusammengefasst: Zur Korrektur des Worthäufigkeits-Bias wird die errechnete Assoziationsstärke jedes Wortes mit dessen absoluter Häufigkeit im gerade untersuchten Testtext multipliziert und durch dessen absolute Häufigkeit im Korpus dividiert.

Zur Erinnerung: Dies ähnelt erneut der Einbeziehung sowohl lokaler als auch globaler Gewichtungsfaktoren (vgl. Abschnitt 1.2.3 „Kombinierte lokale und globale Gewichtungseinflüsse“, S. 16). Dennoch entspricht die Division durch die Korpushäufigkeit nur bedingt dem Einfluss der inversen Dokumenthäufigkeit: Hier existiert keine wie auch immer geartete „umgebende Textsammlung“, der globale Gewichtungsfaktor stammt aus dem Textkorpus, der für alle zu untersuchenden Testtexte identisch ist.

Das Resultat dieser Korrekturrechnung ist in Tabelle 16 auf Seite 64 für den „Madagaskar“-Text dargestellt. Derart korrigierte Ergebnisse werden im Ergebnisteil der Arbeit betrachtet.

#	Assoziations-stärke	Wort	Korpus-häufigkeit	Text-häufigkeit	Korrigierte Ass.-stärke	Norm./korrig. Ass.-stärke
1	0,7222943471	Lemuren	74	3	0,0292822033	100,000
2	0,3673782413	Madagaskar	478	4	0,0030742949	10,499
3	0,1511272232	Affe	2297	5	0,0003289665	1,123
4	0,0482809111	buddeln	210	1	0,0002299091	0,785
5	0,1445557849	Primat	758	1	0,0001907068	0,651
6	0,1577602623	Kanu	953	1	0,0001655407	0,565
7	0,2751325323	Beton	2608	1	0,0001054956	0,360
8	0,1098737472	Asphalt	1163	1	0,0000944744	0,323
9	0,0622040191	Haustier	949	1	0,0000655469	0,224
10	0,1045652633	vorfahren	2143	1	0,0000487939	0,167

Tabelle 16: Vortest des „Madagaskar“-Textes mit Korrektur des Worthäufigkeits-Bias

Die Division durch die Korpushäufigkeit und die Multiplikation mit der Texthäufigkeit sind natürlich gegensätzlich zueinander: Auch Wörter, deren Relevanz durch relativ hohe Korpushäufigkeit (wie „Affe“ mit 2297) geschwächt wird, können durch hohe Texthäufigkeit (wie „Affe“ mit 5) wieder gestärkt werden.

2.5.6 Zusammenfassung der Ergebnisberechnung

Das beschriebene Verfahren berechnet die Kernwörter des jeweiligen Testtextes mit Hilfe von drei Komponenten:

- Vokabularreduktion
- Assoziationssimulation
- Worthäufigkeits-Bias-Korrektur

Die Vokabularreduktion mit Hilfe der Worthäufigkeitsklassen-Analyse verhilft der Assoziationsberechnung durch Entfernung der zu häufigen und zu seltenen Wörter zu weitaus besseren Ergebnissen, wie in Tabelle 15 auf Seite 63 dargestellt. Es bleibt die Frage, ob die Vokabularreduktion zu rigide mit den Wörtern umgeht, d.h. ob sie zu viele Wörter entfernt. In hier nicht genauer diskutierten Vortests zeigte sich, dass die Assoziationsberechnung ohne vorherige Vokabularreduktion vergleichbare Ergebnisse erzielte, sofern die Worthäufigkeits-Bias-Korrektur nachgeschaltet wurde.

Die Worthäufigkeits-Bias-Korrektur allein liefert jedoch ebenfalls keine zufrieden stellenden Ergebnisse, weder auf dem vollen Testvokabular des jeweils betrachteten Testtextes noch auf dem durch die Vokabularreduktion verkürzten Wortschatz. Es ist und bleibt also die Assoziationssimulation, die die Wörter entsprechend gewichtet. Die Vokabularreduktion realisiert eine sinnvolle Vorfilterung, während die Worthäufigkeits-Bias-Korrektur die Ergebnisse fein einstellt.

Die genauen Resultate und der Vergleich mit den Ergebnissen der Wortmarkierungsstudie finden sich im Ergebnis-Kapitel dieser Arbeit.

3 Realisation der Studie

Um die Leistungsfähigkeit von ATA2 und des Assoziativen Wortnetzes hinsichtlich der Klassifizierung der jeweils zehn wichtigsten Wörter eines Testtextes zu bewerten, müssen zwangsläufig Vergleichswerte vorliegen. Faktisch heißt das: Man benötigt die zehn Kernwörter der Testtexte, die auch ATA2 berechnen soll, und zwar sollen diese Wörter möglichst gesichert und über alle Zweifel erhaben sein.

Wie ermittelt man derart zweifelsfrei korrekte Kernwörter von Testtexten? Aus der Psychologie kommend lag es nahe, eine Studie mit einer hinreichend großen Zahl Versuchsteilnehmer (Vtn)[26] durchzuführen. Man lege also den Vtn exakt die Testtexte vor und lasse sie die zehn Kernwörter z.B. durch Unterstreichen markieren. Sind die Vtn aufmerksam und sorgfältig vorgegangen, so sollten die zehn am häufigsten markierten Wörter zugleich den Kernwörtern entsprechen.

Die Durchführung einer derartigen Studie auf traditionelle Art und Weise erfordert einen nicht zu unterschätzenden materiellen und organisatorischen Aufwand:

1. Nachdem die Testtexte ausgewählt sind, die später ATA2 als Eingabe dienen sollen, müssen die Testbögen für die Vtn mit exakt diesen Texten gedruckt werden.
2. Die Auswahl der Vtn unterliegt dem Problem, einen Termin zu finden, an dem alle Vtn verfügbar sind – oder mehrere Termine anzusetzen.
3. Weiterhin müssen Räumlichkeiten organisiert werden, in denen die Studie durchgeführt werden soll, die bequem von allen Vtn erreicht werden können.
4. Am Ende des Vorgangs stellt sich der Versuchsleiter (d.h. in diesem Fall der Autor) der Fleißarbeit, einen großen Stapel Papier auszuwerten: Es gilt, die Wortmarkierungen auszuzählen und in die Auswertung der Studie zu übertragen, z.B. anfangs in eine Tabelle. Fehler der Vtn, etwa zu wenige oder zu viele markierte Wörter in einem der Testtexte, träten erst jetzt zu Tage.

Diesen logistischen Overhead galt es zu vermeiden. Schnell wurde klar, dass es einfacher wäre, den Vtn die einzelnen Testtexte nicht in Papierform zu präsentieren, sondern am Computermonitor. Damit ließe sich das Problem der fehlerhaften Versuchsdurchführung bereits im Rahmen der Programmierung der Versuchsanordnung ab-

26 In der vorliegenden Arbeit erscheint vor allem bei der Programmierung der Wortmarkierungsstudie mitunter alternativ zu dem Begriff „Versuchsteilnehmer“ (bzw. abgekürzt „Vtn“) der heute eher ungebräuchliche Begriff „Versuchsperson“ (bzw. abgekürzt „Vpn“). In der gedruckten Arbeit herrscht ersterer Begriff vor, in den Quelltexten bis in die Datenbank-Bezeichner letzterer. Die beiden Begriffe werden völlig synonym verwendet.

fangen. Weiterhin könnten die Versuchsresultate, d.h. die Wortmarkierungen, von vornherein in einer zentralen Datenbank gesammelt werden, was die spätere manuelle und damit fehlerträchtige Übernahme der Ergebnisse unnötig macht.

Die Lösung, um auch noch die terminliche und räumliche Problematik zu entzerren, lag auf der Hand: Die Präsentation der Texte am Bildschirm dürfte nicht an einem bestimmten Tag in einem bestimmten Raum stattfinden, sondern an einem beliebigen Computerterminal zu einer beliebigen Zeit. Daraus folgte: Die Programmierung der Versuchsanordnung würde nicht für ein bestimmtes Betriebssystem zum Zwecke einer zentralistischen Studie erfolgen. Vielmehr sollte das World-Wide-Web (WWW) als technisches Medium und das Internet als Kommunikationsschnittstelle genutzt werden, das im universitären und mittlerweile auch im privaten Bereich nahezu uneingeschränkt zur Verfügung steht. Die Vtn sollten sich zu jeder ihnen günstig erscheinenden Tageszeit an einen beliebigen internetfähigen Computer ihrer Wahl setzen können, um unabhängig von Versuchs-Raum und -Zeit die Studie konzentriert durchführen zu können.

Bereits in der ersten Planungsphase wurde offensichtlich, dass das dem Internet immanente Client-/Server-Paradigma (z.B. Webserver / Webbrowser, Mailserver / Mailprogramm etc.) auch für die Versuchsanordnung zutreffen würde.

1. Die teilnahmewilligen Vtn würden vor beliebigen internetfähigen PCs Platz nehmen, die Verbindung zum Internet herstellen und mit (möglichst beliebigen) Webbrowsern an der Studie teilnehmen.
2. Ein zentraler Webserver würde für jeden Vtn via Internet einen Testtext ausliefern, jeder Vtn würde die seiner Ansicht nach zehn wichtigsten Wörter markieren.
3. Die Wortmarkierungen würden zum Server zurückübermittelt werden, woraufhin dieser sie in einer zentralen Datenbank sammeln und (sofern vorhanden) den nächsten Testtext ausliefern würde.

Für die Vtn unsichtbar sollte der Versuchsleiter zu jedem Zeitpunkt die Möglichkeit haben, sich über den aktuellen Stand der Versuchsdurchführung zu informieren.

3.1 Textauswahl

Als Kompromiss zwischen Text-Vielfältigkeit und Bearbeitungszeit für die Studie fiel die Entscheidung, den Vtn neun Testtexte zur Markierung der ihrer Meinung nach zehn wichtigsten Wörter zu präsentieren.

Bei den Texten handelte es sich um Texte verschiedener Schwierigkeitsgrade: Einerseits wurden Texte aus Lesebüchern für das 2. und 3. Schuljahr ausgewählt, andererseits Texte aus der Erwachsenenliteratur – darin enthalten Auszüge aus Sachbüchern oder aus der Belletristik. Es handelte sich um folgende Testtexte, die im kompletten Wortlaut im Anhang dieser Arbeit enthalten sind.

1. *„Madagaskar"*
   „Die letzten ihrer Art" von Douglas Adams & Mark Carwardine.
   ISBN 3-8077-0256-3, Zweitausendeins, 1992.
   (sonst: ISBN 3-4530-6115-2, Heyne, 1992.)
   Auszug dem Kapitel „Zweig-Technologie", S. 10; 204 Wörter.

2. *„Bushaltestelle"*
   „Deutsches Lesebuch, 2. Schuljahr"
   UBB DDL1371(9) - ISBN 3-425-01125-1, Diesterweg, 1974.
   Auszug aus dem Text „Im Omnibus", S. 18; 167 Wörter.

3. *„Wein"*
   „Wein-Enzyklopädie – Die Weinregionen der Welt", C. Foulkes & M. Broadbent.
   ISBN 3-934519-28-8, ECO, Eltville am Rhein, 2000.
   Auszug aus dem Vorwort, S. 5; 190 Wörter.

4. *„Igel"*
   „Ansichten 3 – Lesebuch Primarstufe 3. Schuljahr"
   UBB DDL1761-3 – ISBN 3-592-20630-3, Kamp, 1972.
   Auszug aus dem Text „Der Igel", S. 78; 169 Wörter.

5. *„Computerspiele"*
   „Wir waren Space Invaders – Geschichten vom Computerspielen",
   Mathias Mertens & Tobias O. Meißner.
   ISBN 3-8218-3920-1, Eichborn, 2002.
   Auszug aus dem Vorwort, S. 7; 206 Wörter.

6. *„Tannenwald"*
   „Deutsches Lesebuch, 2. Schuljahr"
   UBB DDL1371(9) - ISBN 3-425-01125-1, Diesterweg, 1974.
   Auszug aus dem Text „Der Tannenwald in der Stadt", S. 43; 186 Wörter.

7. *„Körpersprache"*
   „Körpersprache", Samy Molcho.
   ISBN 3-442-12667-3, Goldmann, 1996.
   Auszug aus dem Kapitel „Unsere erste Sprache", S. 9; 188 Wörter.

8. *„Autor“*
   „Bunte Lesefolgen – Neue Ausgabe – Lesebuch für das 3. Schuljahr“
   UBB DDL1224-3 – ISBN 3-464-01442-8, Cornelsen-Velhagen & Klasing, 1983.
   Auszug aus dem Text „Der Autor“, S.168; 188 Wörter.

9. *„Wissenschaft“*
   „Das Jahr der Graugans“, Konrad Lorenz.
   ISBN 3-423-01795-3, dtv, 1982/1984.
   Auszug aus dem Vorwort, S. 5; 170 Wörter.

Textnr.	Texttitel	Wortanzahl	Unterscheidbare Wortformen	Reduktion auf
1	Madagaskar	204	130	63,7%
2	Bushaltestelle	167	103	61,7%
3	Wein	190	142	74,7%
4	Igel	169	119	70,4%
5	Computerspiele	206	146	70,9%
6	Tannenwald	186	128	68,8%
7	Körpersprache	188	128	68,1%
8	Autor	188	142	75,5%
9	Wissenschaft	170	127	74,7%

Tabelle 17: Anzahl Wörter und Anzahl unterscheidbare Wortformen pro Testtext

Wichtiger als die bloße Anzahl der Wörter pro Text ist die Anzahl der unterscheidbaren Wortformen, die als einzelne zur Markierung als Kernwort bereit standen, wie in Tabelle 17 dargestellt.

3.2 Client-Programmierung

Die Planung des Clients der Wortmarkierungsstudie beinhaltete die Überlegung, wie die Vtn die zehn Kernwörter des jeweiligen Testtextes markieren sollten. Die Realisation sollte zwei Punkte berücksichtigen: Die Bedienung sollte intuitiv sein. Einfacher gesagt: Die Vtn sollten die gewünschten Wörter eines auf dem Bildschirm dargestellten Textes durch einfaches *„point & click“* („darauf zeigen und anklicken“) mit der Maus auswählen können. Dabei sollten möglichst beliebige (moderne) Webbrowser verschiedener Hersteller (etwa Netscape, Opera, Microsoft oder Apple) und auch unter verschiedenen Betriebssystemen (etwa Windows, Linux/UNIX oder MacOS) verwendbar sein. Den kleinsten gemeinsamen Nenner bildete dabei die Überlegung, keine browserspezifischen Funktionen oder ggf. nicht für alle Plattformen erhältliche Plug-In-Zusatzmodule zu verwenden. Vielmehr sollte sich auf die Schnittmenge der Funktionen beschränkt werden, die in allen gängigen Webbrowsern unter allen gängigen Betriebssystemen identisch implementiert ist.

Seit geraumer Zeit stehen zur umfassenden Textgestaltung und -manipulation die *Cascading Style Sheets* (CSS) zur Verfügung. (Mehr Informationen zu CSS auf den Webseiten des WWW-Konsortiums „*W3C*“ unter [CSS2004].) Obwohl die unterschiedlichen Browserhersteller die CSS-Funktionsvielfalt bis heute nicht vollständig in ihre Webbrowser implementiert haben, genügt für die nachfolgend beschriebene benötigte Funktionalität ein geringer Ausschnitt von CSS, der in allen modernen Browsern verfügbar ist.

Weiterhin kann *JavaScript* in Zusammenhang mit dem *Document Object Model* (DOM) zum Einsatz kommen. JavaScript[27] ist eine clientseitige d.h. im Browser implementierte Scriptsprache, deren Programme in WWW-Seiten eingebettet werden können, die im Webbrowser des WWW-Benutzers d.h. des Vtn laufen. Über das DOM lässt sich via JavaScript (standardisiert als *ECMAScript*) auf die einzelnen Elemente der WWW-Seite zugreifen – lies: auf die einzelnen Objekte des Dokuments. (Mehr Informationen zu DOM auf den Webseiten des W3C unter [DOM2003].) Leider existieren auch für JavaScript/DOM zum Teil gravierende Unterschiede in den Implementationen der einzelnen Browserhersteller. Aber auch hier gilt, dass für die nachfolgend beschriebene benötigte Funktionalität nur ein geringer Ausschnitt von JavaScript/DOM genügt, der in allen moderneren Browsern ohne Kompatiblitätsprobleme verfügbar ist.

3.2.1 Cascading Style Sheets und „point & click“

Cascading Style Sheets (CSS) bilden ein multifunktionales Werkzeug zur HTML-Dokumentformatierung. Jeder wissenschaftlich Arbeitende kennt aus Textverarbeitungssystemen (z.B. *OpenOffice.org*-Weiter oder Microsoft Word) die Möglichkeit, Absatzvorlagen für bestimmte Textabschnitte (z.B. Überschriften, Fußzeilen usw.) zu definieren. Statt danach jedes Textelement von Hand zu formatieren (z.B. Wahl des Schriftgrads, des Zeichensatzes, der Zeichenausrichtung usw.) verwendet man die einmalig definierten Vorlagen immer wieder aufs Neue. Der Vorteil liegt auf der Hand: Alle Textelemente, die auf derselben Vorlage basieren, sehen perfekt identisch aus. Es besteht nicht die Gefahr, ein bestimmtes Attribut bei der Formatierung zu vergessen. Weiterhin gestaltet sich eine ggf. nötige spätere Änderung (z.B. der Textgröße von Überschriften, Zeilenabständen oder Seitenrändern) völlig unproblematisch: Es genügt, die zu Grunde liegende Absatz- oder Seitenvorlage zu verändern, und alle zugehörigen Textelemente verändern ihr Aussehen vollautomatisch.

27 Die Programmiersprache „*JAVA*“ einerseits (mitsamt dem Plattform übergreifenden Konzept einer „JAVA-VM“, in der JAVA-Programme ausgeführt werden) und die Scriptsprache „JavaScript“ andererseits sollte man keinesfalls verwechseln. Beide haben außer eines historisch bedingt ähnlich lautenden Namen wenig miteinander zu tun.

3.2.1.1 Style Sheets als Formatvorlagen

Style Sheets liefern in HTML-Dokumenten dieselbe Funktionalität wie Formatvorlagen in Textverarbeitungssystemen. Natürlich lässt sich auch in HTML jedes Textelement mühsam „von Hand" formatieren. Doch hier bildet ebenso die zentrale Definition und Änderbarkeit einen unschätzbaren Vorteil für längere oder mehrseitige Dokumente.

Das bislang unerklärte erste Wort *Cascading* beschreibt, dass Style Sheet-Vorlagen Stilattribute in einer Vererbungshierarchie an andere Vorlagen weitergeben können, ähnlich wie bei der Objektvererbung in der objektorientierten Programmierung. Durch Verfeinerung und Redefinition der Attribute werden so immer spezifischere Stile instanziiert. – *OpenOffice.org* realisiert für Dokumentformatierung ein ähnliches Konzept.

Erfolgt diese Vererbung bzw. Verfeinerung von Stilattributen (von derselben oder von anderen Vorlagen) mehrmals, bildet sich dadurch förmlich eine „Kaskade" von sich gegenseitig verfeinernden Stil-Definitionen. Die treffende Bezeichnung hierfür lautet konsequenterweise *Cascading Style Sheets* (CSS).

3.2.1.2 Anklickbare Elemente – HTML-Links und CSS

Gesucht war eine Umsetzung einer intuitiv bedienbaren *„point & click"*-Benutzerschnittstelle zum Markieren von Wörtern aus Texten. Soll in HTML-Dateien etwas anklickbar sein, ist im einfachsten und kompatibelsten Fall eine `<a href=...>`-Deklaration[28] das Mittel der Wahl – anders ausgedrückt: Ein *Link*. Im Regelfall enthält ein Link einen Querverweis auf ein anderes HTML-Dokument, aber mit Hilfe von JavaScript lässt sich der Wechsel zu einem anderen Dokument unterbinden und stattdessen beliebiger Programmcode ausführen (siehe Abschnitt 3.2.2).

In allen gängigen Webbrowsern verwandelt sich der Mauszeiger beim Überstreichen eines *verlinkten* d.h. querverbundenen Objekts von einem Pfeil in eine Hand mit ausgestrecktem Zeigefinger. Dieses Verhalten ist auch für die Durchführung der Wortmarkierungsstudie wünschenswert. Um jedes einzelne Wort nacheinander anklicken zu können, muss daher *jedes einzelne Wort* des jeweiligen Texttextes durch einen eigenen `<a href=...>`-Link realisiert werden, wobei die Links jeweils geeignete JavaScript-Funktionen aufrufen sollten.

Im ursprünglichen HTML-Standard, d.h. in der Prä-CSS-Zeit, wurden Links grundsätzlich unterstrichen und in blauer Farbe dargestellt. Diese bildeten unübersehbare Blickfänge für den Leser der WWW-Seite, dass hier ein anklickbares Textstück vorhanden war. Usability-Tests wiesen jedoch nach, dass dauerhaft unterstrichene Links das Lesen von extensiv verlinkten HTML-Texten stark erschwerte: Je mehr unterstrichene Wörter vorkamen, desto stärker störte dies den Lesefluss. Für die Wortmarkierungs-

28 Das HTML-Tag „href" steht abgekürzt für „Hyper-Reference" – eine Referenz auf ein anderes Hypertext-Dokument, i.d.R. eine andere Webseite.

studie kann nicht genug Wert auf gute Lesbarkeit der Testtexte gelegt werden: Einzelne Wörter als Links d.h. als anklickbare Textelemente können nur dann verwendet werden, wenn sie nicht-unterstrichen dargestellt werden können.

Die ursprüngliche HTML-Unzulänglichkeit unveränderlich unterstrichener Links wurde mit CSS behoben: Hier existiert die Möglichkeit, die Voreinstellungen der Formatvorlage, die das Aussehen von Links definiert, durch eine verfeinernde, kaskadierende Definition zu modifizieren. Es genügt dabei, dem `<a...>`-Tag in einer CSS-Definition geänderte Schriftattribute zuzuweisen.

In Hinblick auf die Studie sollen Links

1. dieselbe Farbe haben wie der restliche Text – d.h. schwarz,
2. im Normalfall nicht-unterstrichen sein,
3. als Blickfang fett gedruckt erscheinen und
4. bei Berührung mit dem Mauszeiger zur Verdeutlichung der Link-Funktion passend zum ursprünglichen HTML-Standard unterstrichen erscheinen.[29]

In CSS lassen sich die Eigenschaften für Links, über denen die Maus schwebt, mit Hilfe der „hover"-Spezialisierung (engl. für „schweben"[30]) genau definieren. Die CSS-Definition für das gewünschte Aussehen von Links findet sich in der Datei `da_style.css` und ist in Quelltext 11 dargestellt.

```
a {
  color: #000000;
  font-weight: bold;
}
a, a:active, a:link, a:visited { text-decoration: none; }
a:hover { text-decoration: underline;}
```

Quelltext 11: CSS-Definition des Link-Aussehens

„`a`" steht für das Aussehen von `<a href...>`-Tags (also Links), die grundsätzlich in schwarzer Farbe und fett gedruckt erscheinen sollen. Alle Linkausprägungen, ob besucht („`visited`"), unbesucht („`link`") oder gerade aktiviert („`active`"), erscheinen ohne jegliche Auszeichnung („`text-decoration: none`"), d.h. insbesondere ohne Unterstreichung. Nur die „`hover`"-Links sollen unterstrichen dargestellt werden.

Im Resultat erscheint ein Testtext im Webbrowser des Versuchsteilnehmers als fett gedruckte Textpassage, deren einzelne Wörter bei darüber schwebendem Mauszeiger unterstrichen dargestellt werden, siehe Abbildung 18 auf Seite 73.

29 Das Konzept von Links, die den Lesefluss nicht durch Unterstreichung stören, jedoch durch ein weniger auffälliges Textattribut (z.B. Fettdruck, z.B. Kursivschrift) als solche erkennbar sind, sowie eine bei Berührung mit der Maus eingeblendete Unterstreichung als visuelle Rückmeldung, hat sich in den letzten Jahren weithin etabliert.

30 Gemeint ist, dass der Mauszeiger ohne Tastenaktivität über einem Link „schwebt". Synonym wird häufig der Begriff „rollover" verwendet, engl. für „überrollen" des Links mit der Maus(-kugel).

Abbildung 18: Link mit Hover-Unterstreichung

Zusammengefasst zeichnet den Client im Webbrowser des Versuchsteilnehmers die Verlinkung jedes einzelnen Wortes mit einem eigenen `<a href...>`-Tag aus, sowie die Gestaltung der Link-Optik via Hover-Formatierung mithilfe von CSS.

3.2.2 JavaScript-Links und Formular-Manipulation via DOM

Links, wie sie für jedes einzelne Wort der Testtexte verwendet werden, bilden normalerweise eine Verknüpfung einer Textpassage eines HTML-Dokuments mit einem anderen HTML-Dokument. Eine derartige „Querverlinkung“ bildet den essenziellen Kern eines jeden Hypertextsystems: Im Gegensatz zum linearen Lesen eines Dokuments kann an jeder vom Autor definierten Stelle mit Hilfe einer solchen Querverbindung zu einem beigeordneten weiteren Dokument verzweigt werden[31].

Die für die Benutzerschnittstelle der Wortmarkierungsstudie geforderte Funktionalität verbietet jedoch den Wechsel zu einer anderen WWW-Seite: Der Vtn soll ohne Seitenneuaufbau der Reihe nach exakt zehn Wörter anklicken und erst nach Bewältigung dieser Aufgabe zum nächsten Text wechseln.

Statt in ein `<a href...>`-Tag die URL[32] eines anderen Dokuments einzutragen, lässt sich dort auch eine JavaScript-Anweisung einsetzen. Ein einfacher normaler Link sähe beispielsweise wie folgt aus:

```
<a href="http://www-psycho.upb.de/">Psychologie-Homepage</a>
```

Dagegen sähe eine einfache eingebettete JavaScript-Anweisung z.B. wie folgt aus:

```
<a href="javascript:alert('Hallo Welt!')">Hallo Welt</a>
```

31 Das wiederholte Springen von Dokument zu Dokument mit Hilfe von immer neuen Querverweisen ähnelt den Springen eines Surfers von Welle zu Welle. Durch diese Ähnlichkeit wurde Ende der 1990er Jahre der Begriff des „Websurfens“ geprägt, und der Besucher einer Webseite wurde als „Websurfer“ bezeichnet. Unschönerweise enthält dieses Bild den Aspekt eines letztlich nicht sehr interessierten Besuchers, der sich von immer neuen Links unmotiviert zu immer neuen Seiten treiben lässt.

32 „URL“ steht für „*Uniform Resource Locator*“ und beschreibt den „Ort“ eines vom Browser erreichbaren Objekts, im einfachsten Fall etwa `http://www.server.domain/pfad/dokument.html`

Natürlich ist man nicht darauf beschränkt, die gesamte JavaScript-Programmierung in die `<a href...>`-Tags einzubetten. Ebenfalls lassen sich so selbstgeschriebene JavaScript-Funktionen aufrufen, die an geeigneter anderer Stelle definiert wurden.

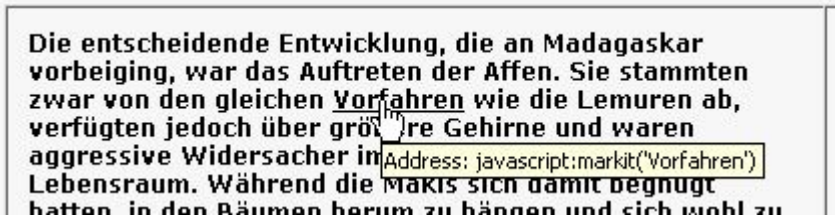

Abbildung 19: Link mit JavaScript-Funktionsaufruf „markit“

Der Webbrowser *Opera* zeigt nach einigen Sekunden des Verweilens des Mauszeigers auf einem Link, wie in Abbildung 19 zu sehen, den Inhalt des `<a href...>`-Tags an – vermeintlich eine Adresse, d.h. eine URL, in diesem Fall jedoch ein JavaScript-Aufruf: In der Abbildung ist zu sehen, dass ein Mausklick auf das Wort „Vorfahren“ die JavaScript-Funktion „`markit(...)`“ mit sich exakt dem Wort „Vorfahren“ als Parameter aufrufen würde. De facto ruft *jedes* Wort des Testtextes die Funktion „`markit (...)`“ mit dem Wortlaut des angeklickten Wortes als Parameter auf.

Die Funktion „`markit(...)`“ ist zusammen mit den anderen benötigten JavaScript-Definitionen in der Datei `da_markwords.js` enthalten, die von den servergenerierten HTML-Dokumenten automatisch eingebunden wird. Sie sorgt dafür, dass maximal zehn Wörter markiert werden, dass keines der Wörter doppelt markiert wird und dass der Vtn den Überblick behält, wie viele und welche Wörter bereits markiert worden sind.

Solche Visualisierungen können mit Hilfe von einzeiligen Formular-Eingabefeldern in HTML realisiert werden. HTML-Formulare bilden den Dreh- und Angelpunkt bei der Kommunikation einer WWW-Seite mit einem Webserver via CGI[33]. Neben einzeiligen Eingabefeldern gibt es Anklickfelder, Auswahllisten und einige weitere Elemente, deren Inhalte auf diesem Wege an den Webserver übermittelt werden.

Neben der üblichen Verwendung von Formular-Eingabefeldern, d.h. zur Eingabe von Texten, lassen sie sich auch *„read-only“* bzw. schreibgeschützt definieren, d.h. ihr Inhalt lässt sich dann nicht mehr vom Computerbenutzer editieren. JavaScript-Programme können den Inhalt eines Eingabefeldes, editierbar oder schreibgeschützt, gleichermaßen auslesen wie auch mit beliebigen Inhalten füllen.

Hierzu sind die in das HTML-Dokument eingebetteten Formulare über das *DOM*, das *Document Object Model* (vgl. [DOM2003]) mit JavaScript-Kommandos ansprechbar: Ein Formular fügt sich wie jedes andere Element des Dokuments in dessen Objekthie-

33 „CGI“ steht für *„Common Gateway Interface“* und beschreibt allgemein die Kommunikation eines WWW-Dokuments in einem Webbrowser mit einem auf dem Server aufrufbaren Programm, das auf geeignete Weise auf die in das Formular eingegebenen Daten reagiert.

rarchie ein. Die hierzu benötigten Teilmengen von JavaScript und des DOM sind trotz aller browserspezifischen Unterschiede seit Einführung von JavaScript unverändert und identisch in allen Browsern enthalten.

Der einfachste Weg zum Zugriff auf den Inhalt eines Formular-Eingabefeldes lautet in JavaScript/DOM-Schreibweise „`document.Formularname.Eingabefeldname.value`“. Eine simple Formular-Deklaration eines Testformulars mit einem darin enthaltenen Test-Eingabefeld lautet also etwa wie im Codefragment in Quelltext 12 dargestellt.

```
<form name="testform">
  <input name="testeingabe" type="text">
</form>
```

Quelltext 12: Beispielhafte Formular-Deklaration

Wenn die Funktion „`markit(...)`“ nun exakt das oben definierte Eingabefeld mit einem Wort füllen soll, lautet die dazu passende Implementation wie in Quelltext 13.

```
function markit(wort) {
  document.testform.testeingabe.value=wort;
}
```

Quelltext 13: Beispielhafte Implementation von „`markit(...)`“

Die Benutzerschnittstelle der Wortmarkierungsstudie benutzt tatsächlich einzeilige Formular-Eingabefelder, wie in Abbildung 20 zu sehen, um die markierten Wörter darzustellen, sowie um sie später an den Server zu übermitteln, der diese dann in der Datenbank der Studie vermerkt.

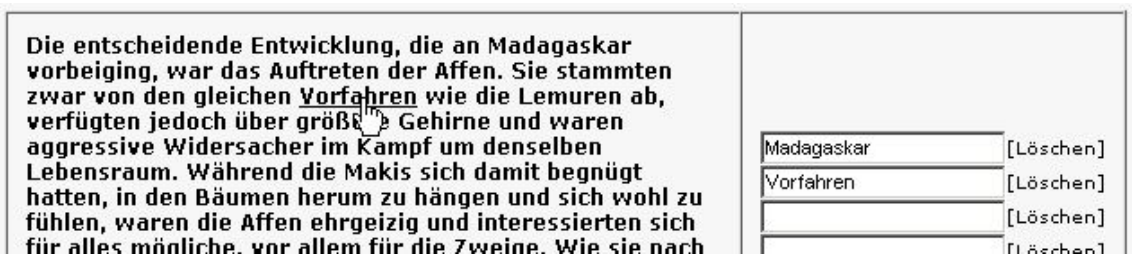

Abbildung 20: „`markit(...)`“ nach dem Eintragen eines angeklickten Wortes

Die reale Implementation von `markit(...)` ist dabei ungleich komplexer:

- Es gibt nicht ein Eingabefeld, sondern zehn.
- Während des Anklickens der Wörter muss Buch geführt werden, wie viele Wörter der Vtn bereits markiert hat.
- Der Vtn darf nicht mehr als zehn Wörter markieren.
- Wörter dürfen nicht doppelt markiert werden.
- Der Versuchsteilnehmer soll die Gelegenheit haben, ein bereits markiertes Wort zu löschen, wenn ein anderes Wort als wichtiger erachtet wird. (Dies übernimmt konsequenterweise nicht „`markit(...)`“ sondern eine zweite JavaScript-Funktion namens „`remove(...)`“.)

Trotz all dieser Anforderungen geht der Kern der Implementation von `markit` im Grunde genommen nicht wesentlich über das soeben Beschriebene hinaus: Nach Anklicken des zu einem Wort gehörenden JavaScript-Links wird das übergebene Wort, sofern möglich, via DOM in ein noch freies Formulareingabefeld übertragen.

3.3 Server-Programmierung

Vollkommen getrennt von der Bedienbarkeit der Client-Seite, die der Vtn zu sehen bekommt, liegt die Serverprogrammierung. Einerseits müssen Basisentscheidungen hinsichtlich der Datenbankstruktur getroffen werden, andererseits muss die Kommunikation zwischen Client und Server definiert und programmiert werden.

Der Autor dieser Arbeit hat mehrere Jahre lang Erfahrung in der Programmierung von Webservern gesammelt. Aufgrund guter Resultate bei verschiedensten Projekten fiel die Wahl auf Server-Infrastrukturen, die plattformübergreifend und kostenlos verfügbar sind:

1. *PHP*. PHP ist ein rekursives Akronym und steht für „PHP: Hypertext Preprocessor“. PHP ist eine als OpenSource-Projekt frei verfügbare Scriptsprache, die für den Einsatz auf Webservern konzipiert und dahingehend optimiert ist. Eine umfassende Dokumentation dieser Sprache findet sich unter [PHP2004].

2. *MySQL*. MySQL ist ein als OpenSource-Projekt frei verfügbares Datenbank-Management-System (DBMS), das insbesondere in Verbindung mit WWW-Servern stark verbreitet ist. Informationen zu MySQL finden sich unter [SQL2004].

3. *Apache* (ein als OpenSource-Projekt frei verfügbarer HTTP-Server, mehr Informationen zu Apache siehe [APA2004]) oder ein anderer geeigneter HTTP-Server mit installiertem PHP-Modul und darin enthaltener MySQL-Anbindung bildet das Fundament einer dynamischen, datenbankgestützten Webanwendung.

PHP, MySQL oder Apache in ihrer vollen Funktionalität auch nur ansatzweise beschreiben zu wollen, würde den Umfang dieser Arbeit bei Weitem sprengen. Für die Wortmarkierungsstudie sind nur einige Kernaspekte von Bedeutung:

- PHP-Code wird auf geeignete Weise direkt in HTML-Dateien eingebettet. Wird eine WWW-Seite vom HTTP-Server abgerufen, die PHP-Code enthält, wird das eingebettete Programm ausgeführt. Das Aussehen der sich im Webbrowser aufbauenden Internetseite wird also durch die PHP-Programmierung bestimmt.

- PHP-Programme können mit Dateien arbeiten, die zusammen mit den HTML-Dateien im Dateisystem des Webservers gespeichert sind. Im vorliegenden Anwendungsfall bedeutet das: PHP ist in der Lage, auf dem Server gespeicherte Testtexte zu lesen, sie zu verarbeiten und daraus WWW-Seiten zu erstellen, die die Vtn zu sehen bekommen, um dort Wörter durch Anklicken auszuwählen.
- MySQL ist ein DBMS, das (von einigen speziellen Eigenschaften abgesehen) wie jedes andere DBMS agiert: Es verwaltet relationale Datenbanken, deren Relationen bzw. „Tabellen“ mit Hilfe der Query-Sprache SQL manipuliert werden können.
- Die enorme Verbreitung von MySQL im WWW-Bereich ist vor allem dadurch zu erklären, dass PHP (wie andere HTML-taugliche Scriptsprachen auch) eine leistungsfähige Schnittstelle zu MySQL-Datenbanken mitbringt. PHP ist somit dazu prädestiniert, bei der Generierung dynamischer WWW-Inhalte auch Datenbankinhalte einzubeziehen oder im Rahmen einer Webanwendung zu manipulieren.

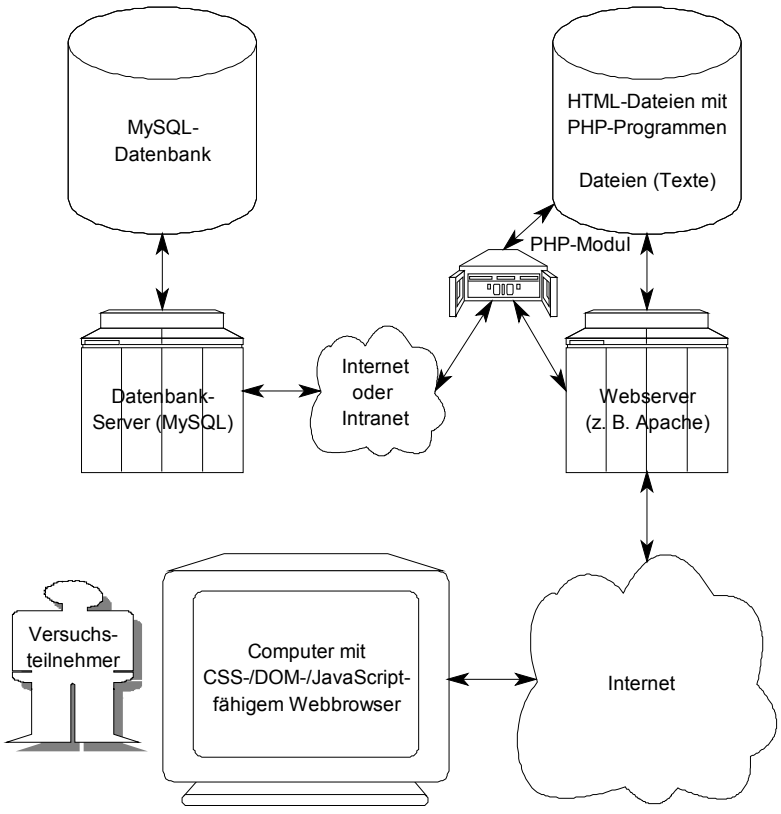

Abbildung 21: Client-Server-Architektur der Wortmarkierungsstudie

Lässt man eine Homepage in heutiger Zeit bei einem der großen Internet-Service-Provider (ISP) hosten[34], und wählt man ein hinreichend großes Webhosting-Paket, ist in der serverseitigen Infrastruktur häufig sowohl PHP- als auch MySQL-Unterstützung enthalten. Der Autor dieser Arbeit hat sich beruflich wie privat jahrelang mit der Programmierung von Online-Präsenzen beschäftigt, die dynamische Seiten mit PHP einschließlich Datenbankanbindung an MySQL enthielten. Es lag also nahe, die für diese Arbeit durchzuführende Studie im Rahmen des Serververbandes durchzuführen, in der auch dessen private Homepage gehostet wird.

Abbildung 21 auf Seite 77 visualisiert die Client-Server-Architektur der Wortmarkierungsstudie: Der Versuchsteilnehmer verwendet einen internetfähigen Computer mit entsprechend ausgestattetem Webbrowser, der die benötigten CSS-, DOM- und JavaScript-Elemente korrekt auswertet. Das Internet dient als Kommunikationsweg zum Webserver des ISP, auf dem z.B. Apache läuft. Mithilfe des installierten PHP-Moduls (das natürlich ein installiertes Softwarepaket und kein Hardware-Modul darstellt), werden die zur Studie gehörigen PHP-Programme ausgeführt, die in den HTML-Dateien eingebettet sind. PHP-Programme können direkt auf Dateien zugreifen, die auf der Festplatte des Webservers gespeichert sind – in diesem Fall auf die dem Versuchsteilnehmer zu präsentierenden Testtexte. Über das Internet bzw. das Intranet des ISP greift das PHP-Modul, kontrolliert durch die in die HTML-Dateien eingebetteten PHP-Programme, auf den Datenbankserver zu. Dieser speichert in der Datenbank die zur Studie gehörigen Daten, d.h. vor allem die Resultate in Form der von den Vtn markierten Wörter.

3.3.1 Basisüberlegungen zur Datenspeicherung

Bei der Durchführung der Wortmarkierungsstudie fallen pro Versuchsteilnehmer und Testtext zehn markierte Wörter an. Im ersten, naiven Ansatz würde der Versuchsaufbau lediglich eine „computergestützte Strichliste“ führen, welche Wörter wie häufig markiert würden. Dies entspricht der absoluten Häufigkeit der Wortmarkierungen.

Zusätzlich ist es wünschenswert, die Resultate einzelner Vtn untereinander in Beziehung setzen zu können – d.h. zu untersuchen, in wieweit sich die markierten Wörter der Vtn paarweise untereinander bzw. von den computerermittelten Lösungen unterscheiden. Dies geht über das reine Zählen von Häufigkeiten hinaus und ordnet den einzelnen Vtn miteinander vergleichbare „Markierungskonstellationen“ zu.

34 „Hosten“ ist ein Anglizismus und leitet sich vom Wort „to host“ = bewirten ab. Ein Host ist grundsätzlich ein Server, der Daten liefert. Im Falle eines ISP spricht man von „Webhosting“, und mit Host ist der Webserver im eigentlichen Sinne gemeint. „Hosten lassen“ bedeutet also, die zu einer Internetpräsenz gehörenden Dateien auf einem Webserver zu deponieren, damit der Server die Dokumente auf Anfrage über das Internet ausliefern kann.

3.3.1.1 Anonymität der Versuchsteilnehmer

Die Anonymität bei der Durchführung der Wortmarkierungsstudie stand im Vordergrund der Überlegungen. Das heißt, ein Vtn sollte nicht dazu gezwungen sein, Namen oder Emailadresse angeben zu müssen, um an der Studie teilnehmen zu können. Einige Rahmenpunkte sollten trotzdem gewährleistet sein:

1. Jeder Vtn bearbeitet jeden der Testtexte exakt einmal. Er kann keinen Text auslassen und keinen Text doppelt bearbeiten.
2. Alle Ergebnisse eines Vtn sollten zu einem Gesamt-Datensatz rekonstruierbar sein.
   - Die textweise Rekonstruierbarkeit dient, wie bereits beschrieben, der textweisen Vergleichbarkeit der Versuchsteilnehmer.
   - Eine studienweite Rekonstruierbarkeit erschließt die Möglichkeit, Unregelmäßigkeiten (z.B. Wortmarkierungen sinnfreier Wörter) studienweit für einen Vtn überprüfen zu können. Sollte ein Vtn durch unsinnige Wortmarkierungen die Studie zu sabotieren versuchen, könnte er komplett aus der Wertung entfernt werden[35].

Der Versuchsaufbau muss daher die einzelnen Vtn, obgleich sie anonym an der Studie teilnehmen sollen, kontinuierlich voneinander unterscheiden können. Die Verwendung einer Versuchsteilnehmer-Identifikation bzw. „Versuchspersonen-Identifikation“ namens „*VPID*“[36] war daher unvermeidlich: Jede Wortmarkierung eines Vtn muss in Relation mit seiner VPID in der Datenbank gespeichert werden.

3.3.1.2 Mehrstufige Versuchsdurchführung und Datenspeicherung

Bei der Wortmarkierungsstudie werden die Testtexte durch die Versuchspersonen textweise bearbeitet, d.h. pro Testtext erfolgt ein Bildschirmneuaufbau.

Letzterer entspricht dem Laden bzw. Neuladen der aktuell dargestellten WWW-Seite, die von der PHP-Programmierung dynamisch erzeugt wird. Die Zwischenergebnisse, d.h. die Wortmarkierungen pro Vtn und Testtext, müssen geeignet zwischengespeichert werden. Grundsätzlich kommen hier zwei Möglichkeiten in Betracht:

1. Speicherung auf dem Client. Die Zwischenergebnisse werden von Text zu Text mitgeführt, akkumulieren sich in einer Art „Eimerkette“ und werden erst am Ende der Studie „in einem Rutsch“ in die Datenbank geschrieben. Dieser Mechanismus ist in Abbildung 22 auf Seite 80 dargestellt.

35 Natürlich ließen sich auch unsinnige Markierungen eines Vtn für einen einzigen Text aus der Studie entfernen, statt den Vtn komplett zu löschen. Dies würde aber zu unterschiedlichen Vtn-Gesamtheiten bei den einzelnen Testtexten führen, was möglichst von vornherein vermieden werden soll.

36 „Versuchsperson“ vs. „Versuchsteilnehmer“ — siehe Fußnote 26 auf S. 66.

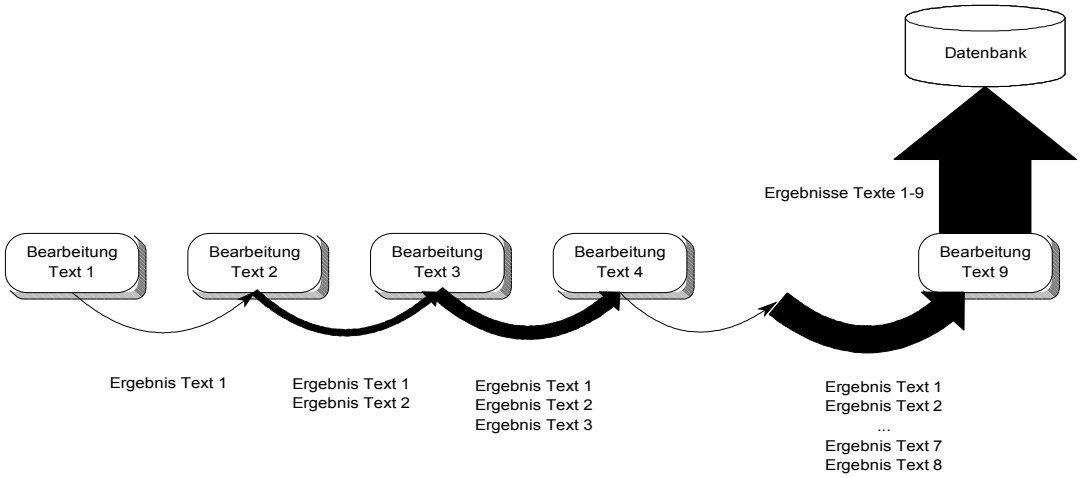

Abbildung 22: Schrittweises Weitergeben von Zwischenergebnissen - „Eimerkette“

2. Speicherung auf dem Server. Die Zwischenergebnisse jeweils eines Textes werden sofort nach Textbearbeitung in der Datenbank zwischengespeichert. Dieser Mechanismus ist in Abbildung 23 dargestellt.

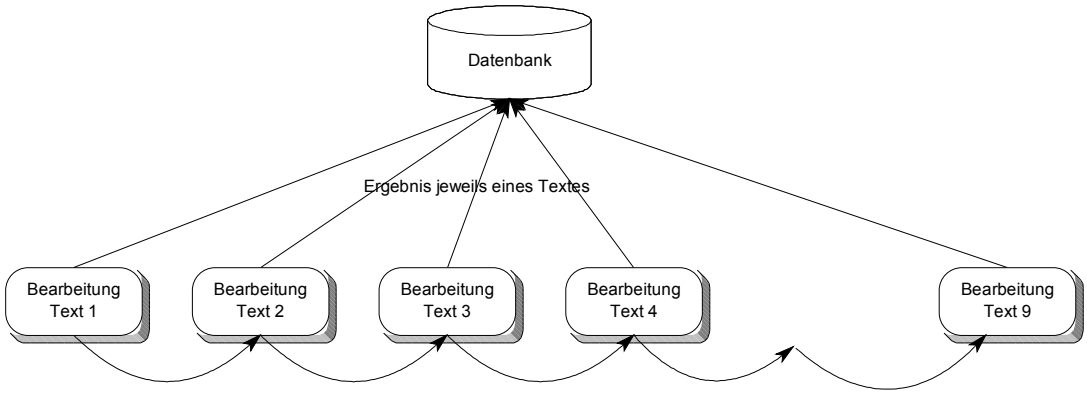

Abbildung 23: Schrittweise Speicherung von Zwischenergebnissen

Der Ansatz mit schrittweiser Weitergabe hat den Vorteil einer vereinfachten Datenbankanbindung, da ja nur ein einziges Mal am Ende der Studie in die Datenbank geschrieben wird. Nichtsdestotrotz müssen die Zwischenergebnisse als unsichtbare Formular-Einträge über alle noch folgenden Bildschirmneuaufbauten gerettet werden, wodurch das Transfervolumen von Text zu Text steigt; der „Eimer“ der „Eimerkette“ wird also von Text zu Text voller. Dies erfordert entsprechende Disziplin bei der Programmierung der Webanwendung.

Das größte Problem ist jedoch ein anderes: Unterbricht ein Vtn die Studie, z.B. durch Schließen des Browserfenster (aus Versehen oder durch Programmabsturz), sind sämtliche Zwischenergebnisse verloren und die Studie muss erneut und vollständig durchgeführt werden, beginnend mit dem ersten Testtext.

Wird alternativ nach jeder Textbearbeitung direkt zwischengespeichert, entfällt das „Eimerketten"-Problem automatisch. Weiterhin führt dies sofort zu dem Vorteil, dass ein Vtn auch absichtlich die Studie unterbrechen kann, um sie später nahtlos an derselben Stelle fortzuführen. Weiterhin ist das Verfahren ungleich flexibler: Es muss streng genommen bei Versuchsbeginn nicht einmal feststehen, wie viele Texte es überhaupt geben wird, da kein statischer Zwischenspeicher programmiert werden muss[37].

Erkauft wird diese Flexibilität durch den Umstand, dass schon zu Beginn der Studiendurchführung die VPID des Vtn feststehen muss, die eindeutige Identifikation eines jeden Versuchsteilnehmers. Nur so kann die Programmierung dafür Sorge tragen, dass die Wortmarkierungen der einzelnen Texte personenbezogen gespeichert werden können, und dass jeder Vtn jeden Text exakt einmal bearbeitet und keinen auslassen kann.

3.3.1.3 Anonymisierte Identifikation der Versuchsteilnehmer

Wenn ein Vtn die Studiendurchführung unterbrechen und später wieder aufnehmen kann, muss ein wie auch immer gearteter Identifikations- bzw. Authentifikationsmechanismus geschaffen werden. Im üblichen Sprachgebrauch einer Anwendung, in der ein System die Zugriffsberechtigung eines Benutzers erkennen soll, sind zwei Stufen nötig: *Identifikation* und *Authentifikation*. Bei der Identifikation gibt der Benutzer sich als eine bestimmte Person zu erkennen, während bei der Authentifikation der Beweis erbracht ist, dass der Benutzer wirklich die genannte Person ist.

Als bekanntestes Beispiel mag die Bankkarte am Geldautomaten dienen: Durch Einschieben der Karte identifiziert sich die Person als Inhaber des Kontos, dessen Kontonummer auf der Karte vermerkt ist. Erst nach Eingabe der Geheimzahl akzeptiert der Geldautomat die Person, deren Identität er nun für bewiesen ansieht. Ein anderes Beispiel ist die Anmeldung an einen gesicherten Computer: Der Computer verlangt zur Identifikation den Benutzernamen und zur Authentifikation das Passwort.

Der Bearbeitung der Studie muss also ein Anmeldevorgang vorausgehen, der in der Datenbank für einen Vtn eine VPID anlegt und mit einer Identifikation mitsamt Authentifikationsmerkmal verankert. Aufgrund des geringen Sicherheitsbedarfs und in Hinblick auf die Usability der Wortmarkierungsstudie wurde auf das Authentifikationsmerkmal verzichtet: Pro Vtn gibt es stattdessen nur eine Identifikation, die bei Studienbeginn vom Server studienweit einmalig generiert wird. Dieses identifizierende Wort, obgleich von kryptischem „passwortartigem" Aussehen, dient zur Identifikation dieses Vtn ohne anschließende Authentifikation. Die Vorgehensweise ist akzeptabel, da die Identifikation, d.h. das „Passwort", zufällig generiert wird; kein Vtn kann durch simples Raten auf die Kennung eines anderen Vtn kommen.

37 Da die Anzahl der Testtexte zum Zeitpunkt der Programmierung des Versuchsaufbaus noch nicht geklärt war und erst kurz vor Studienbeginn festgelegt wurde, war eine möglichst flexible Lösung *per se* wünschenswert.

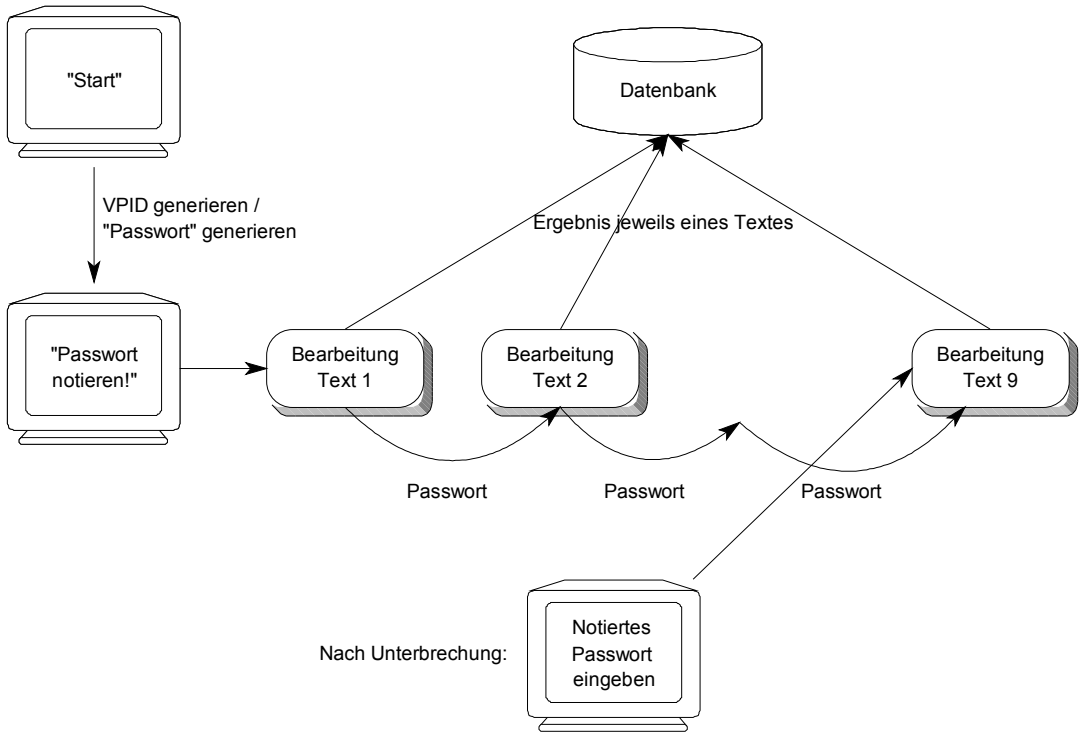

Abbildung 24: Studienstart mit „Passwort" / nach Unterbrechung

Nach dem ersten Mausklick des teilnahmewilligen Vtn wird eine VPID und eine zugehörige kryptische Kennung als „Passwort" generiert und im Klartext auf dem Schirm ausgegeben. Noch vor der eigentlichen Textbearbeitung wird der Vtn dazu aufgefordert, dieses „Passwort" für eine mögliche Unterbrechung der Studie zu notieren. Erst nach einem weiteren bestätigenden Mausklick wird der erste Text zur Wortmarkierung präsentiert. – Ein Schema der Durchführung der Studie ist in Abbildung 24 dargestellt.

Notwendigerweise muss die Identifikation bzw. das „Passwort" nach jeder Textbearbeitung vorliegen, um die in der Datenbank zu speichernden Zwischenergebnisse der zugehörigen VPID zuordnen zu können.

Hinsichtlich der Usability sollte das „Passwort" allerdings nicht vor jedem Seitenneuaufbau erneut abgefragt werden. Üblicherweise würde eine Webanwendung stattdessen Browser-Cookies oder Session-IDs verwenden, wie es bei Webshops oder Diskussionsforen gängige Praxis ist.

Der o.a. simple Identifikationsmechanismus ohne Authentifikation rechtfertigt jedoch keine weitere Steigerung der Programmkomplexität der Versuchsdurchführung. Ein simplerer, wenngleich „schmutzigerer" Weg ist, das identifizierende „Passwort" des Vtn als unsichtbaren Formular-Eintrag im Klartext in den HTML-Code einzubetten und so über den Bildschirmneuaufbau hinweg zu retten.

Da die Klartext-Einbettung jedoch in jedem Fall Ergebnis einer CGI-Anforderung ist, muss und wird jeder Webbrowser nach dem Schließen des Browserfensters jede auf der lokalen Festplatte zwischengespeicherte Kopie des HTML-Textes verwerfen. Der Forderung, das Browserfenster im Anschluss an die Bearbeitung der Studie zu schließen, kann und wird jeder Versuchsteilnehmer problemlos nachkommen.

3.3.2 Datenbankentwurf

Nach der grundsätzlichen Überlegung, zu welchem *Zeitpunkt* die Daten der Zwischenergebnisse gespeichert werden sollen, folgt nunmehr die Definition, *wie* diese Daten gespeichert werden. Im Datenbankentwurf schließen sich die Arbeitsschritte des konzeptionellen und relationalen Datenbankentwurfs an, bevor die relationale Struktur mit Hilfe einer relationalen Anfragesprache (in diesem Falle in SQL) im DBMS angelegt und (später) verwendet werden kann.

3.3.2.1 Konzeptioneller Datenbankentwurf

Im konzeptionellen Datenbankentwurf werden in Hinblick auf das weit verbreitete *ER-Modell* (Entity-Relationship-Modell) so genannte „*Entities*“ und „*Relations*“ spezifiziert: „Gegenstände“ und deren „Beziehungen“ untereinander.

- Der erste „Gegenstand“ ist ein Versuchsteilnehmer. Über ihn ist und bleibt, da die Studie anonym durchgeführt wird, wenig bekannt: Für ihn wird beim Anlegen des Datensatzes vom DBMS eine systemweit eindeutige numerische VPID vergeben und von der Programmierung eine Klartext-Identifikation, die aber ebenfalls systemweit eindeutig ist. Erweitern lässt sich der Datensatz eines Vtn durch statistische Informationen: Bei Anlegen eines Vtn wurde dessen Geschlecht erfasst.
- Als zweites sind die Texte zu nennen. Sie enthalten die zu markierenden Wörter. Da sie aber nicht mit in der Datenbank gespeichert werden sollen, sondern als ASCII-Dateien auf dem Server gespeichert bleiben, lässt sich ein Text auf eine simple Zahl reduzieren: Seine Nummer 1 bis 9, mit der die PHP-Programmierung den zugehörigen Dateinamen rekonstruieren und den Text von der Server-Festplatte nachladen kann.
- Nun fehlen noch die Wortmarkierungen. Jeder Vtn markiert für jeden der neun Texte seine persönliche Auswahl von zehn im Text enthaltenen Wörtern.

Zusammen ergibt sich daraus das erste ER-Modell, wie in Abbildung 25 auf Seite 84 dargestellt.

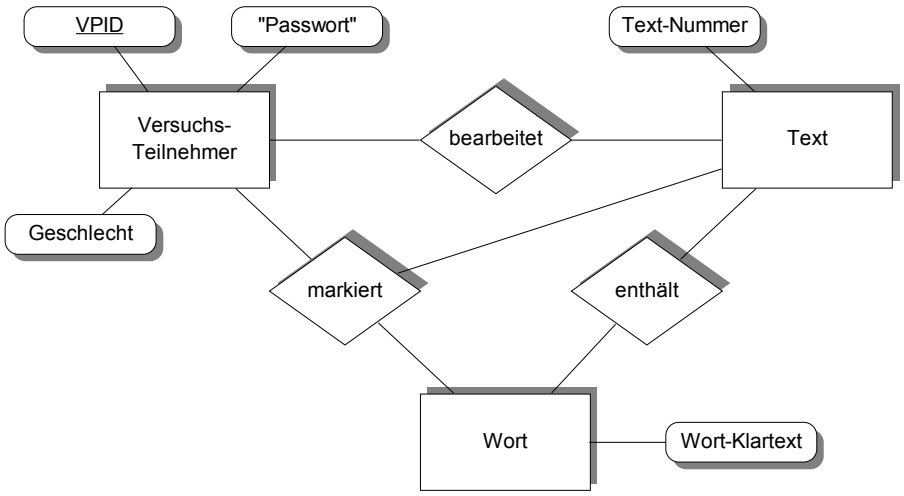

Abbildung 25: ER-Modell / erster konzeptioneller Datenbank-Entwurf

Es wurde bereits herausgestellt, dass ein Text lediglich durch seine Nummer identifiziert bzw. auf der Server-Festplatte referenziert wird. Da die Texte nummeriert sind und in aufsteigender Folge ihrer Nummer von den Vtn bearbeitet werden sollen, muss nicht minutiös gespeichert werden, welcher Vtn bereits welche Texte bearbeitet hat: Es genügt, den „Bearbeitungsfortschritt“ als Anzahl der bereits bearbeiteten Texte pro Vtn zu speichern. Ist er kleiner als 9, wird der nächste Text präsentiert. Ist er gleich 9, hat der Vtn die Studie komplett bearbeitet. Daraus resultiert das vereinfachte ER-Modell in Abbildung 26.

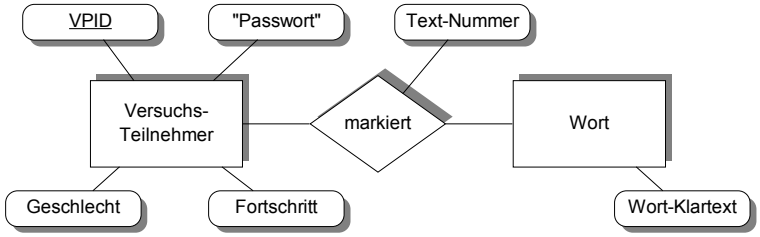

Abbildung 26: ER-Modell ohne Entity „Text“

Diese Modellierung erfüllt die Anforderungen, hat aber eine Schwäche: Die Speicherung der Wort-Klartexte führt zu einer unnötig hohen Datenredundanz. Es liegt nahe, eine zusätzliche Klartext-Wort-Entity einzuführen: Die Speicherung von Wortmarkierungen geschieht in dem Fall nicht über Wort-Klartexte selbst, sondern über Referenzen auf die gespeicherten Klartext-Wörter. Ein nützlicher Nebeneffekt ist, dass ein Klartext-Wort selbst dann nur einmal gespeichert zu werden braucht, wenn es in mehreren Testtexten als Wortmarkierung vorkommt.

Diese Überlegung führt zu abschließendem ER-Modell in Abbildung 27 auf Seite 85.

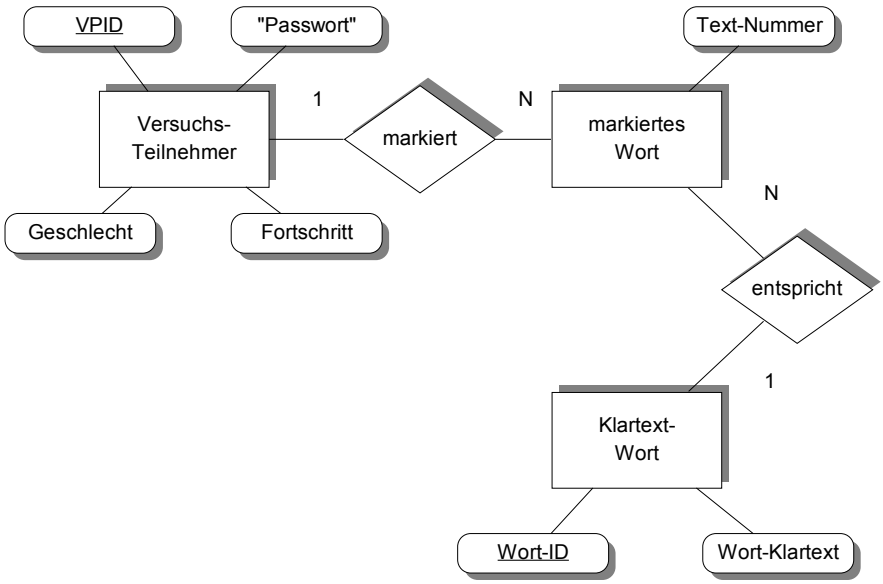

Abbildung 27: ER-Modell mit Entity „Klartext-Wort“

Offensichtlich gilt, dass „markiert“ eine 1:N-Beziehung ist: Ein Versuchsteilnehmer markiert mehrere Texte und in jedem Text noch mehrere Wörter. Ebenso ist „entspricht“ eine 1:N-Relation, da jedes Klartext-Wort beliebig oft referenziert werden kann.

3.3.2.2 Relationaler Datenbankentwurf und SQL-Create-Anweisungen

Der relationale Datenbankentwurf ist nunmehr nur noch eine Formsache, denn das Relationenschema ist recht überschaubar. Ein Großteil der Datenbankoptimierung ist bereits im konzeptionellen Entwurf geschehen, so dass alles weitere „durch scharfes Hinsehen“ ohne langwierige Analyse der Einhaltung von Normalformen-Kriterien erfolgen kann.

```
Versuchspersonen : {[VPID:int, Passwd:string,
       Geschlecht:char, Finished:int]};
Worte-Klartext : {[WortID:int, Wort:string]};
Wortmarkierung : {[TextNr:int, VPID:int, WortID:int]);
```

Text 14: Relationaler Datenbankentwurf

Im relationalen Datenbankentwurf wurde das in Abbildung 27 dargestellte ER-Modell in das in Text 14 dargestellte relationale Modell überführt:

- Die Entity „Versuchsteilnehmer“ wurde in einer Relation namens „`Versuchspersonen`“ inklusive der im ER-Modell angehängten Attribute realisiert. Die VPID bildet wie erwartet den Primärschlüssel, das ausgewählte „Unterscheidungsmerkmal“ zwischen mehreren Versuchsteilnehmern.

- Die Entity „Klartext-Wort“ und die Relation „entspricht“ verschmelzen zur Relation „`Worte-Klartext`“, die jedem Wort eindeutig eine Wort-ID zuweist. Die Wort-ID bildet den Primärschlüssel zur Unterscheidung unterschiedlicher Wörter.
- Die Entity „Markiertes Wort“ und die Relation „markiert“ verschmelzen zur Relation „`Wortmarkierung`“, die als Fremdschlüssel die beiden Referenzen auf den Versuchsteilnehmer einerseits (`versuchspersonen:VPID`) und auf das markierte Wort andererseits (`Worte-Klartext:WortID`) enthält. Zusammen mit der Textnummer bilden alle drei Attribute zusammen den Primärschlüssel: Keines der Attribute kann entfernt werden, ohne die eindeutige Unterscheidbarkeit der einzelnen Wortmarkierungen aufzugeben.

Die SQL-Create-Anweisungen lesen sich in Quelltext 15 demnach wie folgt (entnommen aus der Datei `da_settings.php`):

```
create table da_versuchspersonen (
   vpid integer primary key auto_increment,
   passwd varchar(20) unique not null,
   geschlecht varchar(1) not null,
   finished tinyint not null default 0
 );

create table da_worteklartext (
   wortid integer primary key auto_increment,
   wort varchar(40) unique not null
 );

create table da_wortmarkierungen (
   textnr tinyint not null,
   vpid   integer not null references da_versuchspersonen,
   wortid integer not null references da_worteklartext,
   primary key (textnr,vpid,wortid)
 );
```

Quelltext 15: SQL-Create-Anweisungen der benötigten drei Relationen

Aufgrund der nahezu englischsprachigen Lesbarkeit von SQL sind die drei Create-Anweisungen fast ohne weitere Erklärungen verständlich: Die Schlüsselwörter „`references`“ in der Deklaration von `da_wortmarkierungen` bilden die eigentlichen Relationen d.h. Beziehungen. Die Schlüsseldefinition kann in `da_wortmarkierungen` wegen der drei Komponenten nicht *inline* bei der Attributdeklaration geschehen (wie bei `da_versuchspersonen` und `da_worteklartext`), sondern muss als eigene Zeile deklariert werden.

3.3.3 Die Client-/Server-Programmierung in Beispielen

Die gesamte Programmierung, die HTML-Elemente, PHP-Programmteile, CSS-Definition und JavaScript-Programmierung umfasst, liegt bei 2.816 Zeilen (Stand: 22.7.2004), die sich wie in Tabelle 28 gelistet aufteilen.

1618	da_auswertung.php
670	da_various_functions.php
101	da_settings.php
95	da_markwords.js
72	da_anmeldedialog.php
53	index.php
43	da_informationen.php
39	auswertung.php
29	da_style.css
25	da_informationen2.php
18	da_zeitdauer.php
17	da_studiebeendet.php
16	da_formaction.js
14	da_footer.php
6	da_serversecret.php
2816	**Zeilen gesamt**

Tabelle 28: Anzahl Programmzeilen

Nach dem „Löwenanteil“ der Versuchsauswertung in `da_auswertung` stellt `da_various_functions` den zweitgrößten des Quelltext-Umfangs, der für jeden Bildschirmneuaufbau auf `index.php` die passende Funktion bereithält und dabei die benötigten JavaScript-Funktionen einbindet. `index.php` bildet die zentrale Schaltstelle im Verlauf der Studie: Sämtliche Ausgaben und Textpräsentationen werden über diese zentrale Startseite abgewickelt.

An dieser Stelle kann aus verständlichen Gründen nicht auf jede einzelne Unterroutine eingegangen werden. Daher soll hier lediglich die eine oder andere Besonderheit bedacht werden.

Die Beschreibung der Auswertung mit Hilfe von `auswertung.php` und `da_auswertung.php` folgt in Abschnitt 3.4 ab Seite 88.

3.3.3.1 Automatische Generierung des identifizierenden Passworts

Die Benutzeridentifikation, die von dem Versuchsteilnehmer niedergeschrieben werden soll und bei vorzeitiger Unterbrechung der Studie zur Wiederanmeldung dient, wird von `da_generatepasswd` zufällig generiert. Es besteht aus sechs Zeichen, genauer gesagt abwechselnd aus einem Kleinbuchstaben und einer Ziffer. Die Ziffer „0“ wird dabei nicht generiert, um eine Verwechselungsgefahr mit dem großen „O“ von vornherein auszuschließen.

Das `passwd`-Attribut der Relation `da_versuchspersonen` ist als `unique` definiert, d.h. das MySQL-DBMS prüft diese statische Integritätsbedingung bereits im Datenbank-Handler. Sollte der Zufallsalgorithmus wider Erwarten ein zweites Mal dieselbe Iden-

tifikation generieren[38], scheitert das Anlegen des Vtn-Datensatzes in `da_registervpn`. In dem Fall kann der teilnahmewillige Vtn den Anmeldeversuch wiederholen und erhält im nächsten Anlauf eine eigene, definitiv eindeutige Identifikation.

3.3.3.2 Textpräsentation im Browser des Versuchsteilnehmers

Kern der Textpräsentation mit den per CSS formatierten JavaScript-Links bildet die Routine `da_printtext`. Hier wird der für den aktuell identifizierten Vtn gerade anstehende Testtext 1 bis 9 aus den Dateien `da_text1.txt` bis `da_text9.txt` eingeladen und in entsprechende Einzelwörter zerlegt, die danach in entsprechende `markit(...)`-`<a href...>`-Tags eingepackt werden. PHP ist durch seine mächtigen Zeichenketten-Operationen besonders gut für derartige Aufgaben geeignet, bei der die jeweilige Eingabedatei an Zeilenenden und Leerzeichen in einzelne Wörter zerschnitten wird. Besondere Betrachtung fanden nicht-alphanumerische Zeichen (etwa Satzzeichen, Anführungsstriche etc.) an den Wortanfängen und Wortenden, die keinesfalls mit verlinkt werden durften.

3.3.3.3 Gekapselte Zugriffe auf die Klartext-Worttabelle

Das Handling der Klartext-Wörter ist in der Funktion `da_findorcreatewortid` gekapselt. Wenn die zehn markierten Wörter des Vtn zum gerade bearbeiteten Text nacheinander in der Relation `da_wortmarkierungen` gespeichert werden sollen, versucht `da_findorcreatewortid`, das jeweilige Wort in `da_worteklartext` zu finden und so dessen Wort-ID zu ermitteln. Misslingt diese Suche, wird das Wort als neuer Eintrag in die Relation eingefügt und die so erhaltene Wort-ID zurückgemeldet.

3.4 Ausgabe der Zwischenergebnisse

`auswertung.php` und die Include-Datei `da_auswertung.php` dienten bereits während der Versuchslaufzeit als Übersicht für den Versuchsleiter. Zu jeder Zeit konnte er einsehen, wie viele Vtn die Studie begonnen hatten, jeweils wie viele Texte bearbeitet hatten und die Studie beendet hatten.

Zusätzlich sollten als erster grober Überblick die absoluten Wortmarkierungshäufigkeiten ausgegeben werden, d.h. wie viele Versuchsteilnehmer welche Wörter wie häufig als wichtigste Wörter der neun Testtexte ausgewählt hatten.

38 Die Anzahl möglicher Versuchsteilnehmer-Identifikationen (VPID) ist bei 3 Buchstaben und 3 Ziffern (ohne 0) rein rechnerisch 26^3 multipliziert mit 9^3. Es könnten also theoretisch 12.812.904 Personen am Versuch teilnehmen.

3.4.1 Einfache Statistiken

MySQL verfügt in ihrer Anfragesprache SQL wie alle relationalen Datenbanken über die Aggregatsfunktion „`count(*)`“, um die zu einer „`where`“-Bedingung passenden Datensätze zu zählen. Etwa zählt

- `select count(*) as cnt from da_versuchspersonen where finished=9`
  die Anzahl der Versuchsteilnehmer, die die Studie komplett bearbeitetet haben.

Statistiken dieser Art lassen sich durch entsprechende Gruppierung via „`group by`“ bequem auf unterschiedliche Teilbereiche der Vtn anwenden. Etwa zählt

- ```
  select finished,count(*) as cnt from da_versuchspersonen
      group by finished order by finished desc
  ```
  die Anzahl der Vtn, die jeweils alle 9, 8, 7, ... bzw. keinen Testtext bearbeitet haben, in absteigender Reihenfolge der bearbeiteten Texte.

3.4.2 Worthäufigkeiten via Star-Join

Während die simplen Vtn-Anzahl- und Textbearbeitungs-Statistiken lediglich auf die Versuchsteilnehmer-Relation `da_versuchspersonen` zugreifen, muss für die Auszählung der absoluten Markierungshäufigkeiten primär die Wortmarkierungs-Relation hinzugezogen werden. Da aber nur Vtn einbezogen werden sollen, die alle Texte bearbeitet haben, müssen alle anderen weggefiltert werden, die vorher abgebrochen haben oder sich gerade mitten in der Bearbeitung befinden.

Daher muss die durch die VPID gebildete Referenz von der Wortmarkierungs-Relation zur Versuchsteilnehmer-Relation durch das DBMS innerhalb einer *Join*-Operation[39] hergestellt werden. Da danach nicht nur die nichts sagenden Wort-IDs ausgegeben werden sollen, sondern vielmehr die zugehörigen Klartext-Wörter, muss mittels eines zweiten Joins die Referenz zur Klartext-Wörter-Relation erfolgen.

Die SQL-Anweisung des zugehörigen Star-Joins[40], die die absoluten Worthäufigkeiten konzentriert, lautet wie in Quelltext 16 auf S. 90 dargestellt.

39 Ein „Join“ stellt eine Teilmenge des Kreuzproduktes zweier Relationen dar: Jedes Element der ersten Relation wird mit jedem Element der zweiten Relation kombiniert. Allerdings werden nur bestimmte Elementkombinationen durchgelassen – in diesem Fall die Kombinationen, in denen die VPID des Versuchsteilnehmers mit der VPID der Wortmarkierung übereinstimmt. Diesen „Filter“ nennt man *Join-Bedingung*. In diesem speziellen Fall, in dem die gleichnamigen Attribute zweier Relationen identisch überdeckt werden sollen, nennt man ihn *„natürlichen Join“*. Allgemein nennt man einen Join, bei dem die Gleichheit zweier Attribute geprüft wird, *„Equi-Join“*.

40 In der Datenbankprogrammierung nennt man einen derartigen „multiplen Join“ über mehrere Relationen auch „Star-Join“, da die Daten sternförmig aus verschiedenen Relationen aus allen Richtungen herangeschafft und mit Hilfe mehrerer Join-Bedingungen verknüpft werden. Im Falle der Wortmarkierungsstudie wäre der „Stern“ allerdings lediglich dreizackig.
  ```

```
select textnr, wort, count(m.wortid) as cnt
from da_wortmarkierungen m, da_worteklartext k,
     da_versuchspersonen v
where m.wortid = k.wortid and
      m.vpid = v.vpid and
      v.finished=9
group by m.wortid,textnr
order by textnr asc, cnt desc, wort asc
```

Quelltext 16: Star-Join-SQL-Anweisung für absolute Worthäufigkeiten

Der Join „`m.wortid = k.wortid`“ dereferenziert die Klartext-Wörter an die Wort-IDs, die von der Wortmarkierungs-Relation geliefert werden. Der Join „`m.vpid = v.vpid`“ verknüpft die Wortmarkierungen mit den Vtn, aus denen mit Hilfe von „`v.finished=9`“ nur die einbezogen werden, die alle Texte komplett bearbeitet haben.

Die eigentliche Summierung der absoluten Markierungshäufigkeiten geschieht in „`count(m.wortid)`“, wird allerdings durch die Gruppierung „`group by m.wortid`“ für jedes Wort einzeln durchgeführt. Verknüpft mit der zweiten Gruppierungsbedingung „`group by textnr`“ erfolgt die Bearbeitung zusätzlich noch textweise.

Erst die anschließende Sortierung „`order by textnr asc, cnt desc, wort asc`“ führt zu einer ansehnlichen Ausgabe – d.h. erst nach der Nummer des Testtextes, für jeden Testtext weiterhin absteigend hinsichtlich der absoluten Markierungshäufigkeit und bei Bedarf (d.h. bei gleicher Markierungshäufigkeit mehrerer Wörter) alphabetisch nach den gleich häufig markierten Wörtern.

Ein Ausschnitt der Ausgabe des „Madagaskar“-Testtextes mit Stand vom 10.02.2004 sah bei 55 einbezogenen Versuchsteilnehmern wie in Abbildung 29 aus.

Wort	Abs. Häufigkeit
Affen	51
Madagaskar	49
Lemuren	44
Überleben	41
Lebensraum	40
Weiterentwicklungen	30
Entwicklung	23
Gehirne	19
Kampf	19
Zweige	18
Nachfahren	16
ehrgeizig	15
Primaten	13
starb	13
Technologie	13
Vorfahren	12
Widersacher	12
Erde	11
Makis	10
aggressive	8

Text Nr. 1:

Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb. Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Abbildung 29: „Madagaskar“-Text – Auswertung Stand 10.02.2004

Weitergehende Auswertungen in Form von Detailauswertungen der einzelnen Texte wurden ebenfalls in `auswertungen.php` und `da_auswertungen.php` implementiert. Nachdem die ATA2-Ergebnisse vorlagen, konnten in Echtzeit Auswertungsmethoden entwickelt, getestet und erweitert werden – sogar während die Studie noch lief, d.h.

während zusätzliche Vtn weitere Wortmarkierungen in das System einbrachten. Eine Beschreibung dieser ergänzenden Detailauswertungen findet sich im Kapitel „Ergebnisse“, allerdings größtenteils ohne genaueres Eingehen auf die Programmdetails.

3.5 Zusammenfassung

Die komplette, umfassend kommentierte Implementation der Webanwendung liegt dieser Arbeit auf einem Datenträger bei, mit Ausnahme der Datei `da_serversecret.php`, die die notwendigen Zugangsdetails zum Datenbankserver des Autors enthält.

Zum Zeitpunkt der Studiendurchführung war die Webanwendung auf einem UNIX-basierten Webserver beim ISP „1&1 Webhosting“ (`http://hosting.1und1.de/`) installiert und arbeitete unter PHP Version 4.2.3 (vgl. [PHP2004]) mit angebundener MySQL Version 4.0.17 (vgl. [SQL2004]).

Um die Webanwendung auf einem anderen Server zu installieren, sollte es genügen, in der Datei `da_settings.php` bzw. `da_serversecret.php` die Details der Datenbankverbindung (Datenbank-Name und -Server sowie Benutzername und Passwort) einzutragen. Sofern die Dateien danach unter Beibehaltung der vorhandenen Verzeichnisstruktur auf einen PHP-kompatiblen Webserver abgelegt werden, werden die drei benötigten Relationen automatisch beim ersten Seitenaufruf angelegt.

4 Ergebnisse: ATA2 vs. Versuchsteilnehmer

Die Wortmarkierungsstudie, in der den Versuchsteilnehmern neun Testtexte vorgelegt wurden, startete Anfang Februar 2004, insgesamt haben über einen Zeitraum von etwas mehr als drei Wochen 66 Versuchsteilnehmer (Vtn) daran teilgenommen. Die Basis-Ergebnisse in Form der absoluten Häufigkeiten der Wortmarkierungen ließen sich von der ersten Minute an online abrufen: Die Programmierung in PHP unter Verwendung der MySQL-Datenbank erlaubte Echtzeit-Einblick in den Fortschritt der Studie.

Dieselben Texte wurden davon abgekoppelt zeitgleich mit ATA2 verarbeitet. Nachdem die ATA2-Ergebnisse vorlagen, wurden sie direkt in die Online-Auswertung der Wortmarkierungsstudie integriert. Das heißt, ein Vergleich der berechneten Resultate mit den Wortmarkierungshäufigkeiten der Vtn war ab diesem Zeitpunkt ebenfalls in Echtzeit durchführbar, während noch weitere Versuchsteilnehmer hinzu kamen. Nach und nach wurde die Echtzeit-Auswertung immer weiter verfeinert. Lediglich die Aufbereitung der Daten für diese Arbeit erforderte eine Übernahme in ein anderes Softwarepaket, in aller Regel zum bequemen Tabellenlayout.

Dieses Kapitel beschäftigt sich umfassend mit dem Vergleich der ATA2-Resultate mit den durch die Vtn gewonnen Referenzwerten. Im ersten Teil werden prinzipielle Probleme und deren Lösungsansätze besprochen, bevor darauf aufbauend die Auswertungsmethoden vorgestellt werden. Der zweite Teil stellt in kompakter Form die eigentlichen Resultate dar.

4.1 Textumfang und Beispieltext

Zur Erinnerung: Den Vtn wurden die Testtexte weitgehend unverändert zur Wortmarkierung präsentiert, sofern nicht in Hinblick auf die Verwendbarkeit in ATA2 eine vorliegende „Korpus-Schwäche“ die Aufteilung von selten gebrauchten Komposita in einzelne Substantive erforderte. Das heißt, die Texte wurden nicht-lemmatisiert inklusive sämtlicher Flexionen wie Deklinationen oder Konjugationen zur Wortmarkierung präsentiert.

ATA2 hingegen arbeitete wie üblich in Hinblick auf den lemmatisierten Textkorpus auf lemmatisierten Texten, bei denen sämtliche Flexionen auf ihre Grundformen zurückgeführt worden waren. Zusätzlich war das Textvokabular durch die Worthäufigkeitsklassenanalyse auf die Häufigkeitsklassen 7 bis 17 reduziert worden.

Beide Textformen sind am Beispiel des „Madagaskar"-Textes in Text 17 dargestellt.

```
Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen.
Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über
größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während
die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen,
waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die
Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein
nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum
hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb
überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb.
Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an
oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher
Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit
Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den
Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren,
Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Affe, Asphalt, Auftritt, Baum, Beton, Boot, Dank, Ding, Entwicklung, Erde, Familie, Feuer,
Flugzeug, Gebiet, Gehirn, Hang, Haustier, Insel, Jahrtausend, Kampf, Kanu, Lebensraum,
Lemuren, Macheten, Madagaskar, Mal, Nachfahre, Primat, Sache, Technologie,
Weiterentwicklung, Widersacher, Während, Zweig, aggressiv, begnügen, buddeln, denselben,
diesmal, ehrgeizig, eineinhalb, erreicht, erstaunlich, fertigbringen, fühlen, gleich,
größere, hauen, herausfinden, herum, interessiert, kämpfen, schließlich, stammen, sterben,
stochern, verfügt, verschonen, vorbeigehen, vorfahren, überall, überleben, übernehmen
```

Text 17: „Madagaskar"-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Hier reduziert sich die Textmenge also von ursprünglich 204 Wörtern im Originaltext auf 63 Vokabulareinträge bzw. nur noch 30,9% des ursprünglichen Textvolumens.

Ein ähnliches Verhältnis ergibt sich bei allen neun Testtexten, die in Tabelle 30 abgebildet.

Textnr.	Texttitel	Wortanzahl	Unterscheidbare Wortformen	Reduktion auf	Vokabularreduzierte Anzahl Lemmata	Reduktion auf
1	Madagaskar	204	130	63,7%	63	30,9%
2	Bushaltestelle	167	103	61,7%	43	25,7%
3	Wein	190	142	74,7%	69	36,3%
4	Igel	169	119	70,4%	62	36,7%
5	Computerspiele	206	146	70,9%	55	26,7%
6	Tannenwald	186	128	68,8%	53	28,5%
7	Körpersprache	188	128	68,1%	63	33,5%
8	Autor	188	142	75,5%	64	34,0%
9	Wissenschaft	170	127	74,7%	65	38,2%

Tabelle 30:Anzahl Wörter, Anzahl unterscheidbare Wortformen und Vokabulareinträge pro Testtext

Im weiteren Verlauf beziehe ich mich bei den meisten Beispielen auf den o.a. „Madagaskar"-Text, d.h. Text 17 im Klartext und im reduzierten Vokabular dient zu jeder Zeit als Anschauungsmaterial. Beziehe ich mich auf andere Testtexte, werden sie (sofern nötig) an Ort und Stelle vorgestellt.

Alle Testtexte mit sämtlichen nachfolgend entwickelten Kennwerten finden sich im Anhang.

4.2 Versuchsimmanente Unwägbarkeiten

Aufgrund der eben beschriebenen unterschiedlichen Textbasen für die ATA2-Berechnungen einerseits und für die Wortmarkierungsstudie andererseits ist eine direkte Vergleichbarkeit der so durch das Assoziative Wortnetz berechneten Kernwörter der Testtexte mit den durch die Vtn mehrheitlich als Kernwörter markierten Wörtern nicht unmittelbar möglich. Erst durch die im Folgenden vorgestellten Datenanpassungen kann die Bewertung der Leistungsfähigkeit von ATA2 hinsichtlich der Kernwortsuche im Vergleich mit den Vtn erfolgen.

4.2.1 Notwendige Entlemmatisierung und Wortklassen

Um die mit ATA2 berechneten lemmatisierten Kernwörter mit den nicht-lemmatisierten Wortmarkierungen der Vtn vergleichen zu können, musste im ersten Schritt die Lemmatisierung rückgängig gemacht werden: Sämtliche von ATA2 assoziierten Wort-Grundformen mussten in ihre in den ursprünglichen Testtexten enthaltenen Flexionen zurückgeführt werden.

Die	Die
entscheidende	entscheiden
Entwicklung	Entwicklung
die	die
an	an
Madagaskar	Madagaskar
vorbeiging	vorbeigehen
war	sein
das	das
Auftreten	Auftritt
der	der
Affen	Affe
Sie	sie
stammten	stammen
zwar	zwar
von	von
den	den
gleichen	gleich
Vorfahren	vorfahren
:	:

Text 18: Nicht-lemmatisierte und lemmatisierte Wortformen in der Gegenüberstellung

Um dies in Echtzeit durchführen zu können, wurden „Gegenüberstellungen“ verwendet, die jedes einzelne Wort eines Textes mit den jeweils zugehörigen lemmatisierten Wörtern in Beziehung setzte. Als Beispiel mag der Anfang des „Madagaskar“-Textes in Form seiner Lemmatisierungs-Gegenüberstellung dienen, wie in Text 18 darge-

stellt. Somit lässt sich jedes von ATA2 berechnete Kernwort auf seine ggf. unterschiedliche Original-Wortform zurückführen, beispielsweise „Affe“ auf „Affen“. Letzteres ist in der Datenbank der Wortmarkierungsstudie enthalten, ersteres dagegen nicht.

Ein Problem liegt dabei auf der Hand: Die Abbildung von den nicht-lemmatisierten auf die lemmatisierten Wortformen ist nicht injektiv und somit nicht bijektiv. Das heißt im Detail, dass beispielsweise beide nicht-lemmatisierten Wörter „Zweige“ und „Zweig“ zum selben Wort „Zweig“ lemmatisiert werden, dass aber umgekehrt die Rückführung von „Zweig“ mehrdeutig zu „Zweige“ und „Zweig“ führt.

Wann immer die von ATA2 als mögliche Kernwörter assoziierten lemmatisierten Wortformen mit den Wortmarkierungen der Vtn verglichen werden sollten, war es wie im Falle von „Zweig“ nötig, eine ganze *Wortklasse* von ggf. mehreren nicht-lemmatisierten Wortformen zu beachten – sofern im Testtext mehrere Flexionen desselben Wortes vorkamen und nach der Worthäufigkeitsklassen-Reduktion des Vokabulars als Kernwörter in Frage kamen.

Der Vorgang der Überführung der lemmatisierten Wortformen auf ATA2-Seite hinüber zu nicht-lemmatisierten Wörtern bzw. Wortklassen wird im Folgenden als *Entlemmatisierung* bezeichnet.

Die zwanzig am stärksten von ATA2 assoziierten Wörter (im Folgenden „Top 20“ genannt) zum „Madagaskar“-Texte sind in ihrer entlemmatisierten Form in Text 19 dargestellt.

```
Lemuren, Madagaskar, Affen, buddeln, Primaten, Kanus, Beton,
Asphalt, Haustieren, Vorfahren, [Zweige/Zweig], herum, stochern,
Erde, Lebensraum, Gehirne, Nachfahren, Macheten, denselben, Jahr-
tausenden
```

Text 19: Top 20 des „Madagaskar“-Textes - entlemmatisiert

Faktisch handelt es sich also um 21 Wörter, da das Wort „Zweig“ durch die Wortklasse „[Zweige/Zweig]“ repräsentiert wird.

4.2.2 Flexionskorrektur von Entlemmatisierungs-Wortklassen

Die im vorangegangenen Abschnitt beschriebenen Wortklassen bilden die Ursache für ein weiteres Problem: Die Vtn waren während der Wortmarkierungsstudie nicht darauf festgelegt, ob sie im „Madagaskar“-Text als wichtigstes Wort etwa „Zweig“ oder „Zweige“ markieren sollten (sofern sie die Wortklasse „[Zweig/Zweige]“ als Kernaspekt des Textes ansahen).

Dementsprechend finden sich beide Wörter derselben Wortklasse in der Liste der von den Versuchsteilnehmern markierten Wörter: „Zweige“ wurde 23x als Kernwort markiert, „Zweig“ wurde 2x markiert (originale und flexionskorrigierte Markierungshäufigkeiten siehe Anhang).

Während die Datenbankabfrage zur simplen, automatischen Statistikberechnung mit Hilfe von SQL-Joins und -„group by“s korrekterweise die Wortmarkierungshäufigkeiten beider einzelner Wörter ermittelt, ist es für die weitere Verarbeitung notwendig, die verschiedenen Flexionen einer Wortklasse zusammenzufassen: Die Markierungshäufigkeiten aller Wörter einer Wortklasse werden in einem einzigen Eintrag konzentriert. Am Beispiel bedeutet dies, dass „Zweige“ in der korrigierten Form 25x markiert wurde, während „Zweig“ aus der Liste entfernt wird.

Der Vorgang der Zusammenführung der Wortmarkierungshäufigkeiten mehrerer Flexionen derselben Wortklasse wird im Folgenden als *Flexionskorrektur* bezeichnet.

```
Wein befriedigt unseren Genuß ebenso wie unser sachliches Interesse, er ist also ein
gleichermaßen sinnliches wie intellektuelles Phänomen. Am meisten gibt er uns, wenn diese
beiden Aspekte seiner Persönlichkeit ausgewogen sind, wenn der ganze Charakter eines Weins
zu einem bestimmten Essen oder Anlaß perfekt paßt und wenn seine Geschmacks - Nuancen zum
Vergleich mit anderen Weinen, anderen Jahrgängen, anderen Anlässen anregen. Es gibt
Weintrinker, die sich ausschließlich an der sinnlichen Seite des Weins erfreuen und ihn
gedankenlos hinunter stürzen, bis die Flasche geleert ist. Andere wiederum diskutieren über
seltene Weine und bewerten und analysieren sie so gründlich, daß sie darüber vergessen, die
Signale zu genießen, die ihre Sinne ihnen vermitteln. Beide Typen von Weintrinkern greifen
in ihrer Einseitigkeit zu kurz: Die wahre Befriedigung, die der Wein zu verschaffen vermag,
liegt in der Ausgewogenheit zwischen alltäglichem Genuß und esoterischen Erlebnis. Der echte
Weinliebhaber weiß, daß gewisse Weine zu gewissen Zeiten ein reines Vergnügen sind und daß
das Denken - von der Analyse ganz zu schweigen - dem Vergnügen im Weg steht. Und daß
umgekehrt ein guter Wein viel Vergnügen bereiten kann, aber nur, wenn man dies zuläßt und
ihn bewußt genießt.

Analyse, Anlaß, Anlässen, Aspekt, Ausgewogenheit, Befriedigung, Beide, Charakter,
Einseitigkeit, Erlebnis, Essen, Flasche, Genuß, Geschmack, Interesse, Jahrgang, Nuance,
Persönlichkeit, Phänomen, Signal, Sinn, Sturz, Typ, Vergleich, Wein, Weinliebhaber,
Weintrinker, alltäglich, analysieren, anregen, ausgewogen, ausschließlich, befriedigen,
bereit, bestimmt, bewerten, bewußt, darüber, diskutieren, ebenso, echt, erfreuen,
esoterisch, gedankenlos, genießen, genießt, gewisse, gleichermaßen, greifen, gründlich,
hinunter, intellektuell, leeren, passen, perfekt, rein, sachlich, schweigen, selten,
sinnlich, umkehren, vergessen, vergnügen, vermitteln, vermögen, verschaffen, wahr, wiederum,
zulassen
```

Text 20: „Wein“-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Im Falle von „Zweige“ und „Zweig“ im „Madagaskar“-Text fällt der Effekt der Flexionskorrektur mit ±2 eher gering aus, die Rangfolge der am häufigsten markierten Wörter ändert sich nur leicht. Im Vergleich dazu ist der Unterschied beispielsweise beim „Wein“-Text (190 Wörter / 69 Lemmata, siehe Text 20) bei Aufaddierung von „Weintrinker“ (16x) und „Weintrinkern“ (7x) zu 16+7 = 23x durchaus signifikant und korrigiert die Rangposition der Wortklasse „[Weintrinker/Weintrinkern]“ von Rangposition 16 auf 9 (originale und flexionskorrigierte Markierungshäufigkeiten siehe Anhang).

4.2.3 „Doppeltmarkierungen“

In der im vorherigen Abschnitt beschriebenen Flexionskorrektur liegt ein weiteres, weniger offensichtliches Problem verborgen: Die Wortmarkierungsstudie hinderte die Versuchsteilnehmer gemäß Spezifikation zwar wirksam daran, ein Wort mehrfach zu markieren. Da die Programmierung der Wortmarkierungsstudie jedoch keinen Zugriff auf die unterschiedlichen Flexionsformen einer Wortklasse hatte, war es einem Vtn durchaus möglich, mehrere Flexionen derselben Wortklasse zu markieren – also etwa „Zweig“ *und* „Zweige“. Nach der Flexionskorrektur hätte der betreffende Vtn diese Wortklasse quasi doppelt markiert, dafür aber insgesamt nur neun Wörter ausgewählt.

Eine genaue Analyse der Wortmarkierungen offenbarte, dass der Fall tatsächlich aufgetreten ist, wenngleich äußerst selten. Mit Hilfe der studienweiten Rekonstruierbarkeit der Vtn-Daten wäre es nun möglich gewesen, einen solchen Vtn komplett von der Studie ausschließen. (Ein lediglich textweiser Ausschluss würde zu variierenden Versuchsteilnehmer-Gesamtheiten bei den verschiedenen Testtexten führen.)

Aufgrund der geringen Anzahl derartiger Vorfälle ist eine signifikante Verfälschung der Studienergebnisse allerdings nicht zu befürchten. Um die Gesamtheit der Versuchsteilnehmer nicht unnötig zu dezimieren, wurde daher von der studienweiten Entfernung der betreffenden Versuchsteilnehmer abgesehen.

Im Falle einer Wiederholungsstudie sollte sicher gestellt werden, dass eine derartige „Doppeltmarkierung“ mehrerer Wörter einer Wortklasse vom Versuchsaufbau verhindert wird. Erhielte die Programmierung des Versuchsaufbaus vom Studienbeginn an Zugriff auf die Lemmatisierungs-Gegenüberstellungslisten, wären gleichermaßen die nachträgliche Entlemmatisierung und die Flexionskorrektur von vornherein unnötig sowie die Doppeltmarkierung ausgeschlossen.

4.2.4 ATA2-Ränge und mehrfach belegte reale Rangplätze

ATA2 berechnete pro Text eine Rangfolge der Wörter hinsichtlich ihrer angenommenen Wichtigkeit im jeweils bearbeiteten lemmatisierten Text, gefolgt von der anschließenden Entlemmatisierung der berechneten Wörter: Das am stärksten assoziierte Wort bildet Rang 1 der *ATA2-Rangfolge*.

Parallel dazu produzierten die Versuchsteilnehmer durch Markierung der zehn von ihnen jeweils am wichtigsten erachteten Wörter mehrheitlich die reale Rangfolge der wichtigsten Wörter des Textes.

Das Resultat nach anschließender Flexionskorrektur ist eine Wortliste, absteigend sortiert nach der Markierungshäufigkeit: Das Wort, das am häufigsten von den Vtn markiert wurde, ist offensichtlich das wichtigste Wort des Textes und bildet Rang 1 der *realen Rangfolge*.

Rang	Real	Anzahl
7	Entwicklung	26
8	Kampf	26

Tabelle 31: Zwei Wörter mit gleicher Wortmarkierungshäufigkeit

Problematisch dabei bleibt, dass mitunter Wörter von gleich vielen Vtn als Kernwörter markiert wurden: So wurden, wie in Tabelle 31 dargestellt, im „Madagaskar"-Text die Wörter „Entwicklung" und „Kampf" von je 26 der 66 Vtn markiert.

Eine exakte Rangbewertung, welches der beiden Wörter wichtiger für den Kern des Textes ist, ist daher unmöglich. Eine nachträgliche alphabetische Sortierung benachteiligt eines der Wörter hinsichtlich seiner Wichtigkeit für den Text. Deswegen teilen sich derartige Wörter im weiteren Verlauf der Auswertung *denselben Rangplatz*, während das nächst unwichtigere Wort von unten nachrutscht.

Im weiteren Verlauf soll gleichermaßen sowohl die ATA2-Rangfolge als auch die durch die Wortmarkierungsstudie mehrheitlich ermittelte reale Rangfolge der wichtigsten Wörter (evtl. mit Mehrfachplatzierungen) auf die obersten zehn Rangplätze begrenzt werden. Diese werden im weiteren Verlauf mit *„Top 10"* bezeichnet – einerseits als „ATA2-Top 10", andererseits als „reale Top 10".

Für den Madagaskar-Text ergibt sich die Top 10 wie in Tabelle 32 dargestellt.

Rang	ATA2	Real	Anzahl
1	Lemuren	Affen	61
2	Madagaskar	Madagaskar	59
3	Affen	Lemuren	53
4	buddeln	Überleben	49
5	Primaten	Lebensraum	48
6	Kanus	Weiterentwicklungen	33
7	Beton	Entwicklung Kampf	26
8	Asphalt	[Zweige/Zweig]	25
9	Haustieren	Gehirne	24
10	Vorfahren	Nachfahren	20

Tabelle 32: „Madagaskar"-Top 10 – ATA2 und real

Analog dazu lässt sich die „ATA2-Top 20" und die „reale Top 20" ermitteln: Auch wenn die Versuchsteilnehmer nur jeweils zehn Wörter pro Text markiert haben, lassen sich die weiteren Markierungshäufigkeiten (ggf. mit doppelt vergebenen Rangpositionen) als die nächsten zehn weniger wichtigen Wörter ausmachen. Auch hier finden sich die beiden Wörter „starb" und „Widersacher", die jeweils von 13 Vtn als Kernwort markiert wurden. Sie teilen sich den Rang 15, wie in Tabelle 33 auf Seite 99 dargestellt.

Rang	ATA2	Real	Anzahl
11	[Zweige/Zweig]	ehrgeizig	19
12	herum	Technologie	18
13	stochern	Primaten	16
14	Erde	Vorfahren	15
15	Lebensraum	starb Widersacher	13
16	Gehirne	Makis	12
17	Nachfahren	Erde	11
18	Macheten	übernahmen	10
19	denselben	aggressive	9
20	Jahrtausenden	größere	6

Tabelle 33: „Madagaskar"-Top 20 (11-20) – ATA2 und real

Damit ergeben sich als reale Top 20-Liste insgesamt 22 Wörter, die von den Vtn mehrheitlich zu den Kernwörtern des Textes gewählt wurden. Bei einer größeren Gesamtheit von Vtn hätte eine entsprechende Diversifikation ggf. dazu geführt, dass jedes Wort einen eigenen Rangplatz belegt. Nichtsdestotrotz bliebe theoretisch wie praktisch der mögliche Ausgang, dass zwei Wörter gleich häufig markiert werden.

4.2.5 Ungeordnete Betrachtung der ATA2-Top 10 / -Top 20

Innerhalb der von ATA2 berechneten Top 10 bzw. Top 20 wird im weiteren Verlauf der Auswertung die errechnete Reihenfolge der zehn bzw. zwanzig Wörter ignoriert, sofern nicht im speziellen Fall anders angegeben.

Natürlich berechnet ATA2 durch seine simulierte menschliche Assoziation im Rahmen der beschriebenen Verwendungsweise eine angenommene Rangfolge der Wörter des Testtextes, die sich direkt aus der Assoziationsstärke der einzelnen Wörter ergibt. Anders betrachtet soll ATA2 aber auch als Simulation eines Vtn bei der Durchführung der Wortmarkierungsstudie aufgefasst werden. Daher sollen die ATA2-Resultate auch insbesondere mit den Wortmarkierungen jeder einzelnen Vtn verglichen werden.

Dies ist nicht zu verwechseln mit der realen Rangfolge, die quasi das „mehrheitliche Meinungsbild" aller Vtn widerspiegelt und so die realen Kernwörter gewichtet nutzbar macht: Die reale Rangfolge bildet die Maßgabe, an der sich sowohl die Wortmarkierungen eines einzelnen Vtn und auch die ATA2-Berechnung messen müssen.

Dazu wird für jede der im Folgenden hergeleiteten vergleichenden Analysen pro Testtext *jeder einzelne* Datensatz der 66 Versuchsteilnehmer rekonstruiert und auf dieselbe Weise beurteilt wie die von ATA2 berechnete Top 10 bzw. Top 20. Somit erscheint ATA2 quasi als 67. Vtn der Studie. Insbesondere im Vergleich mit aggregierten Größen wie etwa Mittelwerten über die Gesamtheit der menschlichen Vtn gelingt die Einschätzung der ATA2-Resultate – im qualitativen Vergleich der maschinellen mit einer durchschnittlichen menschlichen Suche nach den Kernwörtern eines Textes.

Im Gegensatz zu ATA2 konnten die menschlichen Vtn im Rahmen der Wortmarkierungsstudie jedoch keine Rangfolge der von ihnen ausgewählten Wörter bestimmen. Es galt lediglich, die zehn ihnen am wichtigsten erscheinenden Wörter in beliebiger Reihenfolge auszuwählen. Die Benutzerschnittstelle der Wortmarkierungsstudie ließ dahingehend auch keine nachträgliche Korrektur der Markierungsreihenfolge zu. Dementsprechend liegen in der Versuchsdatenbank keine Auswertungen hinsichtlich der Markierungsreihenfolge vor.

Eine vergleichende Auswertung der ATA2-Wortrangfolge im Vergleich mit den Wortmarkierungen der einzelnen Vtn ist daher prinzipbedingt unmöglich. Hätte seitens des Versuchsaufbaus für die Vtn die Möglichkeit bestanden, eine Rangfolge der markierten Wörter zu definieren, wäre ein Vergleich mit der ATA2-Rangfolge natürlich möglich und sinnvoll gewesen.

4.3 Trefferquote ATA2

Aus dem Vergleich der ATA2-Rangfolge mit der realen Rangfolge ergibt sich die Kernfrage der Studie: *Wie gut* ist die von ATA2 ermittelte Top 10 bzw. Top 20 verglichen mit den realen Top 10 bzw. Top 20? Da ATA2 augenscheinlich nicht die exakt richtigen Wörter berechnet hat, interessiert uns: Wie schlägt sich ATA2 bei der Prognose der zehn wichtigsten Wörter im Vergleich zu den einzelnen Vtn? Kann ATA2 zumindest die durchschnittliche Wortmarkierung eines Vtn ansatzweise simulieren?

Es lässt sich in Tabelle 32 auf Seite 98 leicht auszählen, dass ATA2 beim „Madagaskar“-Text drei der zehn realen Top 10-Wörter ebenfalls als Top 10-Wörter berechnet hat: „Lemuren“, „Madagaskar“ und „Affen“. ATA2 erzielte also eine Trefferquote von 30%. — Zur Erinnerung: Wenn zwei Wörter gleich häufig von den Vtn markiert wurden, teilen sie sich eine Rangplatzierung, beim „Madagaskar“-Text teilen sich „Entwicklung“ und „Kampf“ mit jeweils 26 Markierungen den Rang 7. Die ATA2-Trefferquote bezieht sich also auf die Anzahl der korrekten Wörter innerhalb der 10 Wörter der ATA2-Top 10, nicht auf die nunmehr 11 Wörter in den realen Top 10.

Natürlich haben ebenfalls nicht alle Vtn durchgängig die realen Top 10-Wörter markiert. In diesem Fall hätte sich keine Worthäufigkeits-Rangfolge ergeben, vielmehr wären die zehn wichtigsten Wörter allesamt mit der Worthäufigkeit 66 auf dem zusammengefassten Rang 1 erschienen, ohne jegliche Abstufungen.

Daher lohnt sich ebenfalls die Untersuchung: Wie viele der realen Top 10-Wörter markierten die einzelnen Vtn beim „Madagaskar"-Text richtig? Wie viele der 66 Vtn erreichten eine Trefferquote von 30%, wie viele eine Trefferquote von 40% und so weiter? Dies ist in Tabelle 34 dargestellt.

Trefferquote (relativ)	30%	40%	50%	60%	70%	80%	90%
Teilnehmer (von 66)	2	4	12	17	12	15	4
Teilnehmer (relativ)	3,00%	6,00%	18,10%	25,70%	18,10%	22,70%	6,00%

Tabelle 34: Top 10-Trefferquote der Versuchsteilnehmer beim „Madagaskar"-Text

Klar zu erkennen ist, dass wie zu erwarten kein einziger der Vtn sämtliche realen Top 10-Wörter markiert hat, d.h. kein Vtn hat eine Trefferquote von 100% erreicht. Vielmehr errechnet sich die durchschnittliche Trefferquote der 66 Vtn beim „Madagaskar"-Text d.h. der Erwartungswert der Trefferquote wie folgt:

$$E(X)=\sum \%(Teilnehmer)\cdot\%(Treffer)=64{,}24\,\%\ .$$

ATA2 gelingt es beim „Madagaskar"-Text mit drei Treffern bzw. 30% Trefferquote also lediglich, etwas weniger als die Hälfte der Kernwörter zu finden, die ein Vtn durchschnittlich markiert hätte.

4.3.1 Lineare ungewichtete Bewertung

Mathematisch entspricht das bloße Auszählen von Treffern gegen eine maximal erreichbare Trefferzahl einem linearen ungewichteten Modell. Anschaulich machen lässt es sich als einzelne Größenachse – oder anders ausgedrückt, als Zahlenstrahl: Gestartet wird im Nullpunkt, das „Ziel", d.h. das optimale Ergebnis, befindet sich im Punkt 10 (im Falle einer Top 10-Wertung). Jeder zusätzliche Treffer, d.h. jede Übereinstimmung, erreicht die nächstgelegene Einheit weiter rechts, siehe Abbildung 35.

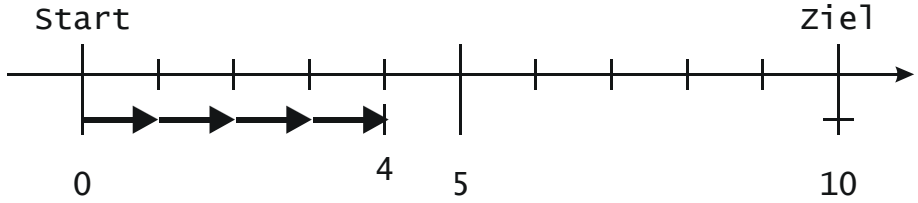

Abbildung 35: Trefferauszählung am Zahlenstrahl – ungewichtetes lineares Modell

Das gewünschte Ziel ist in diesem Beispiel, zehn Schritte nach rechts zurückzulegen, um von Punkt 0 zu Punkt 10 zu gelangen. Es wurden vier Schritte zurückgelegt, entsprechend vier von zehn gefundenen Wörtern. Der Abstand vom Ziel ist in einer derartigen linearen Darstellung logischerweise 6, denn $10-4=6$. Normiert, d.h. auf das Intervall $[0..1]$ bezogen, ist die Trefferquote 0,4 und der Abstand vom Ziel 0,6, entsprechend den prozentualen Angaben 40% Treffer und 60% Fehler.

4.4 Multidimensionale Bewertung

Die Aussagekraft der rein quantitativen Analyse ist indes gering: Wenn es ATA2 beim „Madagaskar"-Text (vgl. Text 17 auf Seite 93) gelingt, in nur drei Treffern die allerwichtigsten Wörter zu finden („Affen" (1), „Madagaskar" (2) und „Lemuren" (3)), so ist dies eine sehr gute Leistung und abstrahiert den Kern des Textes bereits sehr gut.

Wenn ATA2 dagegen im „Autor"-Text (siehe Text 21 - 188 Wörter / 64 Lemmata nach Vokabularreduktion) in sogar vier Treffern nur drei der unteren realen Top 10-Ränge errechnet (siehe Tabelle 36 auf Seite 103): „Urheber" (7), „Konzentration" (8) und „Nachforschungen" (9) und dazu noch „Autor" (1)), so ist das sicherlich trotz höherer prozentualer Trefferquote dennoch ein signifikant schlechteres Ergebnis. Die Treffer-Quantität sagt nichts über die Qualität der dazu führenden Treffer aus.

```
Der Autor steht beim Werdegang eines Buches an erster Stelle, er schreibt den Text. Das
lateinische Wort Autor bedeutet Urheber, ganz gleich ob Mann oder Frau. Allerdings wird der
Begriff im allgemeinen nur auf Schriftsteller und Dichter angewandt. Ihr Werkzeug ist die
Sprache. Mit der müssen sie gut umgehen können. Doch das allein genügt noch nicht! Will ein
Schriftsteller eine Geschichte erfinden und das Geschehen spannend darstellen, muß er Ideen
haben und viel Phantasie entwickeln. Beim Verfassen eines Sachbuches muß er über den
betreffenden Bereich viel Wissen sammeln, das heißt er muß persönliche Erfahrungen machen
und umfangreiche Nachforschungen anstellen. Natürlich hängt das Schreiben auch von der
Personen - Gruppe ab, an die sich ein Autor wendet. Es ist schon ein Unterschied, ob er sich
bei ganz jungen Lesern oder bei Erwachsenen verständlich machen möchte.
Autoren arbeiten sehr unterschiedlich. Manche können diese schwierige Tätigkeit, die so viel
Konzentration und Ausdauer verlangt, nur kurze Zeit am Tag ausüben. Sie bringen vielleicht
nur ein paar Sätze zustande. Andere schreiben die ganze Nacht hindurch und vollenden
Geschichten und Romane in einem Zug. Oft wird dann immer und immer wieder verändert und
verbessert.

Allerdings, Ausdauer, Begriff, Beim, Bereich, Dichter, Erfahrungen, Erwachsenen, Geschehen,
Ideen, Konzentration, Lesern, Manche, Nachforschungen, Nacht, Natürlich, Oft, Personen,
Phantasie, Romane, Sachbuches, Schriftsteller, Sprache, Stelle, Sätze, Text, Tätigkeit,
Unterschied, Urheber, Verfassen, Werdegang, Werkzeug, Zug, Autor, allgemeinen, angewandt,
anstellen, ausüben, bedeutet, betreffenden, darstellen, entwickeln, erfinden, genügt, heißt,
hindurch, hängt, lateinische, paar, persönliche, sammeln, schwierige, spannend,
umfangreiche, umgehen, unterschiedlich, verbessert, verlangt, verständlich, verändert,
vielleicht, vollenden, wendet, zustande
```

Text 21: „Autor"-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Nötig ist daher eine Methode, die neben der bloßen Existenz der Treffer auch deren Platzierung in den realen Top 10 bzw. Top 20 bewertet, um die Qualität der Trefferquote besser einschätzen zu können. Weiterhin soll ein differenzierteres Bewertungsmodell über die lineare „Auszählen"-Bewertung hinaus entwickelt werden.

Als Berechnungsvorschrift hierfür wird in den nachfolgenden Abschnitten zunächst ein mehrdimensionales vektorbasiertes Modell entwickelt, das anfangs ebenfalls nur die Quantität bewertet. Anschließend wird es durch Gewichtung auch die Qualität bewerten. In den nachfolgenden Abschnitten folgt die schrittweise Herleitung des Modells.

Rang	ATA2	Real	Anzahl
1	Ausdauer	Autor	65
2	Sachbuches	Phantasie	52
3	Nachforschungen	Wissen	49
4	Dichter	Sprache	46
5	Konzentration	Buches	38
6	[Autor/Autoren]	Ideen	35
7	Werdegang	Urheber	34
8	anstellen	Konzentration	31
9	Urheber	Nachforschungen Schriftsteller	29
10	lateinische	Erfahrungen	24

Tabelle 36: „Autor"-Top 10 – ATA2 und real

4.4.1 Ungewichtete n-dimensionale Bewertung

Im ersten Schritt soll die lineare Bewertung verlassen werden. Von nun an wird nicht mehr jedes gefundene Wort als ein Schritt „geradeaus zum Ziel" aufgefasst. Vielmehr soll nun jedes Wort als Dimension eines Koordinatensystems dargestellt werden: Statt mit mehreren gefundenen Wörtern immer in dieselbe Richtung zu laufen, führt jedes gefundene Wort auf seiner eigenen Achse vom Punkt 0 zum Punkt 1.

Zwei Dimensionen – zwei zu findende Wörter

Der zweidimensionale Fall, d.h. für zwei zu findende Wörter, ist in Abbildung 37 dargestellt.

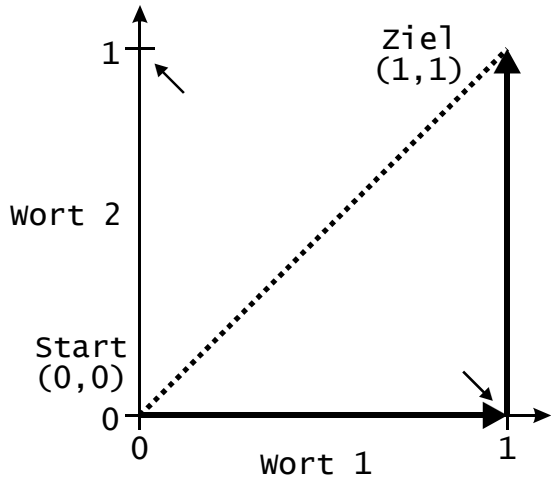

Abbildung 37: 2-Wort-Suche im 2D-Koordinatensystem

Der Startpunkt liegt im Punkt $(0,0)$, d.h. im Ursprung des zweidimensionalen Koordinatensystems. Das zu erreichende Ziel liegt im Punkt $(1,1)$, d.h. in der gegenüberliegenden Ecke des *Einheitsquadrats*, das von $(0,0)$ und $(1,1)$ aufgespannt wird.

Die Horizontale Achse ist dem ersten Wort zugeordnet, die vertikale Achse dem zweiten Wort. Das Auffinden des ersten Wortes entspricht dem horizontalen Pfeil, d.h. dem Vektor $(1,0)$ – der Bewegung auf der horizontalen Achse entlang. Das Auffinden des zweiten Wortes entspricht dem vertikalen Pfeil, d.h. dem Vektor $(0,1)$ – der Be-

wegung parallel zur vertikalen Achse. Daraus resultiert der Zielvektor $(1,1)$, der dem Ortsvektor des Zielpunktes entspricht. Der Zielvektor (1,1) wird also durch Vektoraddition der beiden Teilvektoren $(1,0)$ und $(0,1)$ erzielt.

Es ist zum Erreichen des Zielpunktes gleichgültig, in welcher Reihenfolge die beiden Wörter gefunden werden, d.h. ob erst horizontal oder erst vertikal eine Einheit aufaddiert wird, sofern nur am Ende auf jeder Koordinate eine Einheit aufaddiert wurde.

Das ist nicht der Fall, wenn nur eines der beiden Wörter gefunden wird: Wird nur Wort 1 gefunden, gelangt der gedachte Weg vom Startpunkt zum Zielpunkt nur zur Koordinate $(1,0)$. Wird alternativ nur Wort 2 gefunden, wird Punkt $(0,1)$ erreicht. Die erreichten Ziele entsprechen den freien Ecken des Einheitsquadrats und sind in Abbildung 37 auf Seite 103 mit Pfeilen markiert.

Wird das Ziel nicht erreicht, lässt sich geometrisch die Qualität der Teillösung herleiten: Relevant ist, ähnlich wie im linearen Modell (vgl. Abbildung 35 auf Seite 101), der Abstand des erreichten Punktes vom maximal erreichbaren Zielpunkt. Der Abstand zweier Punkte im „normalen" Euklidischen 2D-Raum lässt sich mathematisch leicht berechnen: Sind die beiden Punkte durch die Koordinaten $(x_{1,}y_1)$ bzw. $(x_{2,}y_2)$ gegeben, lautet die Berechnungsformel für deren Distanz d:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

Klarer gesagt: Der Abstand zweier Punkte entspricht der Quadratwurzel aus der Summe der Differenzen-Quadrate in jeder einzelnen Dimension. – Diese Abstandsberechnung von Punkten bzw. die Längenberechnung von Vektoren mit Hilfe von Quadraten und Quadratwurzeln nennt man auch *2-Norm* bzw. *Euklidische Norm*.

Wird der Zielpunkt (1,1) durch Finden beider Wörter erreicht, ist der Abstand offensichtlich 0 – oder rechnerisch $\sqrt{0^2+0^2}=\sqrt{0}=0$. Wird nur ein Wort gefunden, ist der Abstand 1, denn $\sqrt{0^2+1^2}=\sqrt{1^2+0^2}=\sqrt{1^2}=1$. Wird keines der beiden Wörter gefunden, so ergibt sich der maximale erreichbare Abstand $d_{max}=\sqrt{1^2+1^2}=\sqrt{2}\approx 1.414$.

Abstandswerte verschiedener Modelle sollen gut vergleichbar sein, daher werden die Resultate zweckmäßigerweise normiert – d.h. verhältnismäßig auf das Intervall $[0..1]$ projiziert. Die maximale Entfernung entspricht 1, die minimale Entfernung entspricht 0. Also:

$d_{norm}=\frac{d}{d_{max}}$. Bei nur einem gefundenen Wort ist der Abstand $\frac{1}{\sqrt{2}}\approx 0{,}707$.

Je weiter der Abstand der gefundenen Lösung von der optimalen Lösung ist, desto geringer ist deren Qualität. Aufgrund der $[0..1]$-Normierung lässt sich dies durch simple Invertierung ableiten:

$q_d = 1 - d_{norm}$. Ein einzelnes gefundenes Wort hat also die Qualität $1 - \frac{1}{\sqrt{2}} \approx 0{,}292$.

Zum Vergleich: Beim linearen „Treffer auszählen"-Modell hätte das Finden von einem von zwei Wörtern die Qualität von 0,5 bzw. 50% ergeben.

Drei Dimensionen – drei zu findende Wörter

Wenn statt zwei Wörtern drei gesucht und gefunden werden sollen, ergibt sich daraus ein dreidimensionales Koordinatensystem mit drei den einzelnen Wörtern entsprechenden Achsen und den zugehörigen „Fund"-Vektoren $(1,0,0)$, $(0,1,0)$ und $(0,0,1)$: Nur wenn alle drei Wörter gefunden werden, wird der Zielpunkt $(1,1,1)$ erreicht. Dies verdeutlicht Abbildung 38 auf grafische Weise.

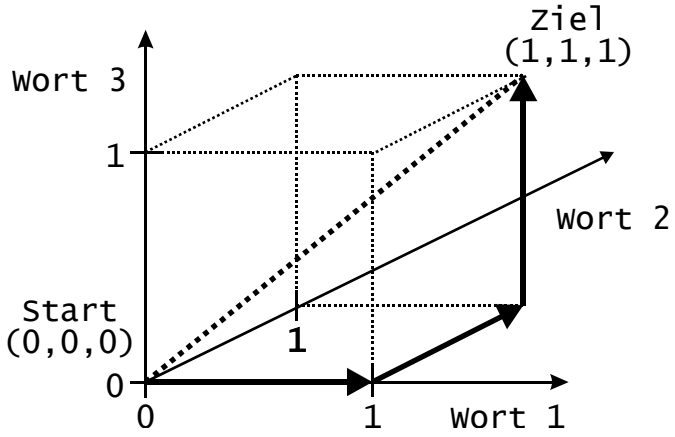

Abbildung 38: 3-Wort-Suche im 3D-Koordinatensystem

Statt des Einheitsquadrates ergibt sich nun der *Einheitswürfel*: Der optimale Zielpunkt $(1,1,1)$ ergibt sich durch Addition der einzelnen Vektoren zum resultierenden Vektor $(1,1,1)$ als Diagonale des Einheitswürfels in drei Dimensionen.

Es gibt nun eine Fülle von Möglichkeiten: Wird kein Wort gefunden, bleibt es beim Ursprung $(0,0,0)$. Wird dagegen ein Wort gefunden, entspricht dies einem der Resultate $(1,0,0)$, $(0,1,0)$ oder $(0,0,1)$ – je nachdem, welches der drei Wörter gefunden wird. Somit entspricht das Finden von zwei Wörtern einem der Resultate $(1,1,0)$, $(1,0,1)$ oder $(0,1,1)$ – je nachdem, welches der drei Wörter *nicht* gefunden wird. Jede dieser nicht-optimalen Lösungen entspricht einem der sieben noch freien Eckpunkte des in Abbildung 38 dargestellten Einheitswürfels.

Mit Hilfe der Euklidischen Norm ergeben sich nun vier mögliche Distanzen: Wird kein Wort gefunden entspricht das der maximalen Distanz $\sqrt{1^2+1^2+1^2}=\sqrt{3}\approx 1{,}732$. Bei einem gefundenen Wort ist der Abstand noch $\sqrt{0^2+1^2+1^2}=\sqrt{2}\approx 1{,}414$. Bei zwei gefundenen Wörtern bleibt nur noch der Abstand $\sqrt{0^2+0^2+1^2}=\sqrt{1}=1$, während der Abstand der optimalen Lösung wie gehabt 0 ist.

Nach Normierung der Distanzen und Invertierung zu Qualitäten ergeben sich die Werte 1, 0,423, 0,184 und 0 für drei, zwei, ein oder kein gefundenes von drei möglichen Wörtern. – Zum Vergleich: Beim linearen „Treffer auszählen"-Modell (vgl. Abbildung 35 auf Seite 101) hätten sich die Werte 1, 0,66, 0,33 und 0 ergeben.

Als grober Trend ist wie in 2D nun auch in 3D deutlich sichtbar, dass das mehrdimensionale Vektormodell gegenüber dem linearen Modell „strenger" hinsichtlich nicht gefundener Wörter bewertet. Das heißt, es „bestraft" fehlende Wörter stärker, als es gefundene Wörter „belohnt".

n Dimensionen – n zu findende Wörter

Sollen statt zwei oder drei Wörtern zehn oder sogar zwanzig gefunden und die Qualität der Lösung beurteilt werden, ergibt sich statt des Koordinatensystems in 2D bzw. 3D eines in 10D bzw. in 20D. Obwohl sich höhere Dimensionen als 3D kein Wesen eines 3D-Universums mehr bildlich vorstellen kann (vgl. [STE2003]), bleibt das Grundprinzip unserer Untersuchungen immer identisch simpel. — Eine verdeutlichende Abbildung muss an dieser Stelle entfallen, da mehr als drei Dimensionen prinzipbedingt nicht in eine zweidimensionale Skizze projizierbar sind.

Nichtsdestotrotz: Für je ein zu findendes Wort wird je eine Achse des Koordinatensystems verwendet. Jedes gefundene Wort entspricht einem Schritt vom Ursprung $(0{,}0{,}..{,}0)$ auf der bzw. parallel zur jeweiligen Achse. Werden alle Wörter gefunden, wird der gegenüberliegenden Eckpunkt des 10- bzw. 20-dimensionalen *„Einheits-Hyperwürfels"* erreicht. Die Distanz der gefundenen Lösung zur bestmöglichen Lösung, beschrieben durch den Ortsvektor $(1{,}1{,}..{,}1)$, lässt sich durch Euklidische Norm berechnen, d.h. wie oben durch die Quadratwurzel aus der Summe der Differenzen-Quadrate in jeder einzelnen Dimension. Die Distanz wird zuerst mit Hilfe der Maximaldistanz zum Punkt $(1{,}1{,}..{,}1)$ der optimalen Lösung normiert und daraufhin invertiert, was zu einem $[0..1]$-Maß für die Qualität der gefundenen Lösung führt.

Diese Lösung funktioniert bei zwei zu findenden Wörter und der Qualitätsbestimmung, bei drei, bei zehn, bei zwanzig – oder allgemein bei n zu findenden Wörtern. Die resultierende Bewertung ist daher allgemein *n-dimensional*.

4.4.2 Gewichtete n-dimensionale Bewertung

Das soeben entwickelte n-dimensionale Bewertungssystem hat einen entscheidenden Schwachpunkt: Obwohl jedes einzelne gefundene oder nicht gefundene Wort eine eigene Dimension darstellt und für sich selbst bewertet wird, ist es bislang dennoch vollkommen gleichgültig, *welche* Wörter gefunden bzw. nicht gefunden werden: Wieder zählt nur die Quantität und nicht die Qualität der gefundenen bzw. nicht gefundenen Wörter.

Bezogen auf das Modell heißt das: Wenn alle Wörter aller Ränge bis auf eines gefunden werden, entspricht der Abstand zwischen dem durch Vektoraddition erreichten Punkt und dem Zielpunkt, der durch den Einheitsvektor $(1,1,..,1)$ repräsentiert wird, immer dem (unnormierten) Abstand $d=\sqrt{0^2+0^2+..+1^2+..+0^2}=\sqrt{1}=1$.

Hinsichtlich des ursprünglichen Problems bedeutet dies: Egal, ob das allerwichtigste Wort nicht gefunden wird oder lediglich das Wort an Platz 20 nicht gefunden wird, sind die Bewertungen der Qualitäten der beiden Lösungen exakt identisch. Somit wurde bislang noch keinerlei Verbesserung zum linearen Modell (vgl. Abbildung 35 auf Seite 101) erzielt.

An dieser Stelle erfolgt eine Modifikation des Verfahrens, zugegebenermaßen eine willkürliche: Statt jedes gefundene Wort durch das Einheitsgewicht 1 auf seiner zugehörigen Achse zu bewerten, werden die Wörter nummeriert und mit entsprechend abgestuften Gewichten in Beziehung gesetzt: Bei 10 zu findenden Wörtern wird dem wichtigsten Wort Nr. 1 das Gewicht 10 zugewiesen, dem weniger wichtigen Wort Nr. 2 das Gewicht 9, dem nächst unwichtigeren Wort Nr. 3 das Gewicht 8 und so weiter, bis am Ende das 10. Wort noch das Gewicht 1 zugewiesen bekommt.

Bei n zu findenden Wörtern ergibt sich allgemein: Wort 1 erhält das Gewicht n, Wort 2 das Gewicht $n-1$, Wort 3 das Gewicht $n-2$ und so weiter, und Wort n erhält das Gewicht 1.

Bezogen auf unser Vektormodell heißt das: Es wird kein Einheits-Hyperwürfel mehr verwendet, denn die Kantenlänge ist nicht mehr konstant 1. Vielmehr ergibt sich eine Art „Hyperquader“, sozusagen ein n-dimensionaler, in je zwei Dimensionen rechteckiger Körper. Zur Verdeutlichung: Im dreidimensionalen Fall ergibt sich ein Quader der Kantenlängen 3/2/1 – eine Art „Schuhkarton“.

Der Ortsvektor des maximal erreichbaren Ziels, dessen Vektor sich immer noch durch Vektoraddition der Teilvektoren ergibt, lautet somit auch nicht mehr $(1,1,..,1)$ sondern $(n,n-1,n-2,..,2,1)$. Im 10D-Fall d.h. bei der Suche innerhalb der Top 10 lautet der Zielvektor also $(10,9,8,7,6,5,4,3,2,1)$.

Wird im 10D-Fall das wichtigste Wort 1 nicht gefunden, lautet der resultierende Vektor $(0,9,8,7,6,5,4,3,2,1)$ und der (unnormierte) Abstand zum Zielvektor ist $\sqrt{10^2+0^2+..+0^2}=\sqrt{10^2}=10$. Wird aber das unwichtigste Wort 10 nicht gefunden, lautet der resultierende Vektor $(10,9,8,7,6,5,4,3,2,0)$ und der Abstand zum Zielvektor ist lediglich $\sqrt{0^2+..+0^2+1^2}=\sqrt{1^2}=1$. Dies jeweils bezogen auf das minimale Ergebnis $(0,0,..,0)$ mit dem Abstand vom Zielvektor $\sqrt{10^2+9^2+...+2^2+1^2}=\sqrt{385}\approx 19{,}62$.

Daraus resultiert, dass im 10D-Fall das Nicht-Finden des wichtigsten Wortes mehr als die Hälfte der zurückzulegenden Strecke zum Zielpunkt ausmacht, während das Nicht-Finden des unwichtigsten Wortes lediglich etwa einem Zwanzigstel entspricht.

Es ist also einerseits möglich, eine gute Trefferqualität durch das Finden weniger sehr wichtiger Kernwörter zu finden, während viele unwichtigere Wörter nicht zwingend gefunden zu werden brauchen. Alternativ ergibt sich ebenfalls eine gute Trefferqualität, wenn wenige der wichtigsten Wörter nicht gefunden werden, dafür aber fast alle der etwas weniger wichtigen.

Aus der Gewichtung der einzelnen Wörter im Vektormodell resultiert also ein funktionierendes Qualitätsmaß für die Güte der von ATA2 berechneten Kernwörter im Vergleich zu den mehrheitlich durch die Vtn ermittelten realen Kernwörtern.

Die hier entwickelte gewichtete Qualitätsberechnung mit n-dimensionaler Euklidischer Distanzberechnung, normierten Abständen und Invertierung zur Qualität bezeichne ich im weiteren Verlauf als *Gewichtete n-dim. Euklidische Qualität* oder kurz als *Gewichtete Qualität*.

4.4.2.1 Gewichtung mehrfach belegter Ränge

Das soeben hergeleitete Modell der Gewichteten Qualität ist damit nahezu komplett. Für die praktische Anwendung muss nur noch ein zusätzliches Detail in Betracht gezogen werden: Bislang wurde davon ausgegangen, dass jedes Wort einen eigenen Rang und somit ein eigenes Gewicht hat. Das heißt, die Qualitätsberechnung für zehn von ATA2 berechnete Wörter gelingt im Vergleich mit den zehn Wörtern der realen Top 10 in zehn Dimensionen mit den Gewichten $10,9,8,...,1$.

Nicht in dieser Betrachtung enthalten sind dabei mehrfach belegte Ränge, wie etwa beim Madagaskar-Text der Rang 7 („Entwicklung“ und „Kampf“, siehe Tabelle 32 auf Seite 98). Es wäre möglich, dass ATA2 keinen, nur einen oder gar alle Vertreter eines mehrfach belegten Rangs ermittelt. Trotz gleicher Wichtigkeit wiegen diese drei Fälle unterschiedlich schwer, d.h. es müssen alle Wörter der Top 10 betrachtet werden – im Falle des Madagaskar-Textes also alle elf Wörter.

Die offensichtliche Lösung ist, die Dimensionalität auf 11 zu erhöhen, dabei aber die Gewichte der mehrfach belegten Ränge identisch zu halten. Als optimaler Zielvektor ergibt sich beim Madagaskar-Text also der 11-dimensionale Vektor $(10,9,8,7,6,5,4,4,3,2,1)$. Die siebtbeste Platzierung „4“ ist doppelt vorhanden.

Gelingt es ATA2, eines der beiden Wörter des doppelt belegten Rangs zu berechnen, ist es hinsichtlich der resultierenden Qualität gleichgültig, welches der beiden Wörter es ermittelt. Berechnet es dagegen beide, wird diese Leistung höher eingestuft als die Berechnung der nachfolgenden Ränge 8-10 (mit den korrespondierenden Gewichten 3 bis 1).

Ein letztes Problem darf nicht verschwiegen werden: Natürlich lassen sich unmöglich mit zehn berechneten bzw. markierten Wörtern elf oder mehr Wörter der realen Top 10 treffen, die sich durch mehrfach belegte Ränge ergeben. Dieses Problem wird um so gravierender, je mehr mehrfach belegte Ränge es gibt. Dementsprechend wird dann (und *nur* dann), wenn zehn Treffer erzielt werden, die niedrig gewichtete Dimension 11 „abgeschnitten“ und geht nicht mehr in die Qualitätsberechnung ein.

Am Beispiel: Wenn es ATA2 oder einem Vtn beim „Madagaskar“-Text (vgl. Tabelle 32 auf Seite 98) gelänge, mit den zehn berechneten bzw. markierten Wörtern die Ränge 1-9 abzudecken, darin enthalten auch beide Wörter „Entwicklung“ und „Kampf“ des doppelt belegten Rang 7, so würde Rang 10 nicht mehr betrachtet. – Dieser ginge mit seinem niedrigen Gewicht allerdings sowieso nur marginal in die Gesamtqualität ein. Aber ATA2 oder ein Vtn muss zumindest „die Chance bekommen“, mit zehn berechneten bzw. markierten Wörtern ein optimales Ergebnis zu erzielen.

Allgemein gilt auch für n Dimensionen und beliebig viele mehrfach belegte Ränge, dass nach Erreichen von n hoch gewichteten Treffern die noch folgenden Dimensionen bzw. Ränge abgeschnitten werden und nicht mehr in die Qualitätsberechnung eingehen. – Faktisch ist dieser Fall aber im Rahmen sämtlicher Qualitätsberechnungen kein einziges Mal eingetreten, da weder ATA2 noch die Vtn durchgängig 100%-Trefferquoten erzielt haben.

4.5 Korrelationsanalyse

Über die reine Berechnung von ungewichteten Trefferquoten oder Gewichteten Qualitäten basierend auf der Top 10 bzw. Top 20 soll analysiert werden, ob ein grundsätzlicher statistischer Zusammenhang zwischen den von ATA2 berechneten hohen Rangplätzen und den realen hohen Rangplätzen besteht.

4.5.1 Korrelationsdiagramm

Ein simpler Versuch, die von ATA2 berechneten Ränge mit den realen Rängen zu vergleichen, ist ein Punktdiagramm, bei dem auf der X-Achse die ATA2-Ränge und auf der Y-Achse die realen Ränge dargestellt sind. Für jedes von ATA2 berechnete Wort wird ein Punkt (x, y) in das Diagramm eingezeichnet, wobei x den von ATA2 berechneten Rang des Wortes bezeichnet und y den Rang, den das Wort auf der realen Rangfolge belegt. Jedes von ATA2 berechnete Wort, das von keinem Versuchsteilnehmer markiert worden ist, soll auf der X-Achse[41] in Punkt $(x,0)$ eingezeichnet werden. Umgekehrt soll jedes Wort, das zwar von mindestens einem Vtn markiert worden ist, aber aufgrund der Worthäufigkeitsklassenanalyse vor der ATA2-Berechnung aus dem Testvokabular entfernt worden ist (weil es nicht in den mittleren Häufigkeitsklassen 7 bis 17 liegt), auf der Y-Achse $(0, y)$ eingezeichnet werden.

Unabhängig davon, dass für die lineare Trefferquote und die Gewichtete Qualität die ATA2-Rangplatzierung ignoriert wurde, wird ein Punktwolke erwartet, die bei Betrachtung der Top 20 um die Diagonale zwischen den Punkten $(1,1)$ und $(20,20)$ besteht – oder allgemein um die Diagonale zwischen den Punkten $(1,1)$ und (x, x).

Formaler ausgedrückt gilt:

- Alle Punkte *oberhalb* der horizontalen Achsenparallele durch die 20 ($y \geq 20$) sind Treffer außerhalb der realen Top 20. Das heißt, ATA2 hat ein Wort auf eine seiner Top 20-Rangpositionen gewählt, das von den Versuchsteilnehmern zwar im Rahmen der Wortmarkierungsstudie als Kernwort markiert worden ist, aber eben nicht mit entsprechender Häufigkeit innerhalb der 20 wichtigsten Wörter.
- Alle Punkte *auf* der *horizontalen* Achse ($y = 0$ bzw. realer Rang = 0) entsprechen absoluten Nicht-Treffern, d.h. *Fehlwörtern*. Das heißt, ATA2 hat ein Wort auf eine seiner Top 20-Rangpositionen gewählt, das von den Versuchsteilnehmern *niemals* als Kernwort markiert worden ist. Diese Wörter bezeichne ich im weiteren Verlauf als *Fehlwörter*.
- Lediglich die Punkte *unterhalb und einschließlich* der horizontalen Achsenparallele durch die 20 ($0 < y \leq 20$) sind Treffer hinsichtlich der realen Top 20. — *Nur diese Treffer* gehen (ohne Beachtung ihrer Rangfolge) in die durchschnittliche Trefferquote von 65.00% bzw. in die durchschnittliche Gewichtete Qualität von 0,391 ein.
- Alle Punkte *auf* der *vertikalen* Achse ($x = 0$ bzw. ATA2-Rang = 0) entsprechen Wörtern, die von einigen Vtn als Kernwörter markiert worden sind, jedoch nicht Teil des Testvokabulars waren.

41 Eine andere Möglichkeit wäre ein beliebiger, ansonsten nicht existenter Wert für *y*, z. B. -1, +99 o.ä. Die Darstellung auf der Y-Achse bildet lediglich eine visuelle Verallgemeinerung.

4.5.2 Fehlwortanalyse der Vtn

Die Fehlwörter, d.h. die ATA2-Top 20-Wörter, die von keinem einzigen Vtn als Kernwort markiert worden sind, lohnen einer weiteren Betrachtung: Andererseits gibt es in der über die Top 20 hinaus verlängerten realen Rangfolge ebenfalls Wörter, die von den Vtn nur ein einziges Mal markiert worden sind.

Das bedeutet im Umkehrschluss, dass *nur ein* Vtn im Gegensatz zu *allen anderen* der Meinung war, dieses Wort sei eines der Kernwörter. Dies entspricht exakt der Definition eines Fehlworts, diesmal nicht aus der Sicht von ATA2, sondern aus der Sicht eines einzelnen Vtn. Dies unterstreicht die Vorstellung, ATA2 als „maschinellen Vtn" zu verstehen und mit den menschlichen Vtn zu vergleichen.

Zur Fehlwortanalyse der Vtn genügt es, nach der Flexionskorrektur die Anzahl derjenigen Wörter zu zählen, die nur noch mit Wortmarkierungs-Häufigkeit „1" erscheinen, und diese Anzahl muss prozentual mit der Anzahl der Vtn (d.h. 66) ins Verhältnis gesetzt werden. Das Resultat ist die Fehlwortquote der Vtn zu einem Text und kann direkt mit dem Prozentsatz der von ATA2 berechneten Fehlwörter verglichen werden.

Realer Rang	ATA2-Rang	Wort	Markierungshäufigkeit
25	------	allein	1
25	------	alles	1
25	8	Asphalt	1
25	43	begnügt	1
25	------	doch	1
25	35	eineinhalb	1
25	60	gleichen	1
25	9	Haustieren	1
25	30	herausfanden	1
25	18	Macheten	1
25	------	nicht	1
25	13	stochern	1
25	------	ums	1
25	------	wohl	1
25	------	Zeit	1

Tabelle 39: „Madagaskar"-Fehlwörter mit Markierungshäufigkeit

Am Beispiel des Madagaskar-Textes gilt (siehe Tabelle 39, vgl. Text 17 auf S. 93): Die fünfzehn Wörter „allein, alles, Asphalt, begnügt, doch, eineinhalb, gleichen, Haustieren, herausfanden, Macheten, nicht, stochern, ums, wohl, Zeit" wurden nur von jeweils einem Vtn markiert. Dies entspricht bei diesem Text einer durchschnittlichen Fehlwortquote von immerhin $15/66 = 22{,}72\ \%$.

In Tabelle 39 sind zusätzlich noch die zugehörigen ATA2-Ränge dargestellt. Wörter ohne Rang gehören zu denjenigen Wörtern, die aufgrund ihrer Worthäufigkeitsklasse <7 oder >17 nicht im Testvokabular enthalten waren und somit nicht Teil der Berechnung waren. Die Reihenfolge ist alphabetisch anhand des Wortes gewählt, jede andere Reihenfolge wäre aufgrund des identischen realen Rangs aber ebenso geeignet.

4.5.3 Durchschnittliche Wortmarkierungshäufigkeiten

Über die Fehlwörter hinaus, d.h. als Kernwörter markierte bzw. berechnete Wörter, die kein (anderer) Vtn als Kernwort markiert hat, soll in Anlehnung an [WRF1993] eine weitere Größe ausgewertet werden: Wenn ATA2 oder ein Vtn ein Wort als Kernwort berechnet bzw. markiert, *wie viele* (andere) Vtn haben das Wort ebenfalls als Kernwort markiert? – Dieser Ansatz bildet sozusagen die Verallgemeinerung des Fehlwort-Begriffs, bei dem kein (anderer) Vtn das Wort als Kernwort markiert hat.

Liegt ein Datensatz von zehn markierten Wörtern vor (d.h. einerseits als ATA2-Berechnung der Top 10, andererseits als Vtn-Wortmarkierung), lässt sich für jedes Wort einzeln mit Hilfe der flexionskorrigierten realen Rangfolge die Anzahl der Vtn-Markierungen ermitteln – also quasi die Anzahl der anderen Vtn, die derselben Meinung waren, das Wort als eines der 10 Kernwörter zu markieren.

Über diese zehn Wortmarkierungshäufigkeiten der Einzelwörter wird der Mittelwert berechnet. Dies ist zulässig, da die Vtn im Rahmen der Wortmarkierungsstudie keine Gewichtung (d.h. Reihenfolge) der Wörter vornehmen konnten. Diese *durchschnittliche Wortmarkierungshäufigkeit* beschreibt sozusagen die „Einigkeit“ eines Vtn mit allen anderen Vtn.

Pro Testtext ergibt sich für die ATA2-Berechnung eine einzige durchschnittliche Wortmarkierungshäufigkeit. Für denselben Text ergibt sich pro Vtn ein Mittelwert, d.h. 66 durchschnittliche Wortmarkierungshäufigkeiten pro Text, aus denen sich wie üblich Minimum, Maximum, Mittelwert und Standardabweichung berechnen lassen.

4.5.4 2x2-Felder-Tafeln und Chi²-Unabhängigkeitstest

Der Chi²-Unabhängigkeitstest ist ein statistisches Mittel zur Überprüfung, ob zwischen zwei Experimenten mit demselben beobachteten Ereignis mit hinreichend niedriger Irrtumswahrscheinlichkeit davon auszugehen ist, dass ein Zusammenhang besteht – oder die Beobachtung wahrscheinlich rein zufällig zu Stande gekommen ist. Ein anschauliches Beispiel ist die Anzahl der zu schnell fahrenden Autofahrer bei zwei verschiedenen Verkehrskontrollen zu unterschiedlichen Tageszeiten.

Das „Ereignis“, das im vorliegenden Anwendungsfall betrachtet wird, ist das Auftreten eines Wortes innerhalb bzw. außerhalb der von ATA2 berechneten Top 20, d.h. das Wort gehört bzw. gehört nicht zu den von ATA2 bestimmten zwanzig wichtigsten Wörtern. Die beiden „Experimente“ ergeben sich durch die Betrachtung derjenigen

Wörter innerhalb bzw. außerhalb der realen Top 20 – d.h. das betrachtete Wort gehört bzw. gehört nicht zu den von den Vtn mehrheitlich bestimmten zwanzig wichtigsten Wörtern.

Klarer ausgedrückt: Es wird ausgezählt, wie viele real wichtige Wörter von ATA2 ebenfalls als wichtig oder im Gegensatz dazu als unwichtig eingestuft wurden. Gleichermaßen wird ausgezählt, wie viele real unwichtige Wörter von ATA2 ebenfalls als unwichtig oder im Gegensatz dazu als wichtig eingestuft wurden.

Beim „Madagaskar"-Text (vgl. Text 17 auf S. 93) ergibt sich die in Tabelle 40 auf Seite 113 dargestellte Verteilung.

Text 1	ATA2<=20	ATA2>20	Summe
Real<=20	10	12	22
Real>20	10	44	54
Summe	20	56	76

Tabelle 40: 2x2-Felder-Tafel „Madagaskar"-Text

Zur Erklärung: Von den 22 Wörtern auf den realen Top 20-Rängen (zur Erinnerung: Ränge sind bei gleicher Markierungshäufigkeit doppelt belegt, vgl. Tabelle 32 und Tabelle 33 auf S. 99) werden von ATA2 10 als wichtig angesehen, zwölf als unwichtig. Entgegengesetzt betrachtet sind von den 56 Wörtern, die ATA2 nicht als Top 20-Wörter einstuft, 44 Wörter tatsächlich auch laut realer Rangfolge unwichtig, während zwölf real wichtig sind und somit von ATA2 falsch klassifiziert wurden.

Die Gesamtsumme von 76 Wörtern ergibt sie wie folgt. Zur Erinnerung: Der „Madagaskar"-Text bestand für die Vtn aus 130 unterscheidbaren Wortformen, für ATA2 aus einem reduzierten Vokabular von 63 Wörtern (vgl. Tabelle 30 auf S. 93). Die $76-63=13$ zusätzlichen Wörter sind in Tabelle 41 dargestellt.

ATA2-Rang	Realer Rang	Wort	Markierungshäufigkeit
------	16	Makis	12
------	21	wir	5
------	22	entscheidende	4
------	22	Millionen	4
------	23	aus	3
------	24	Jahre	2
------	25	allein	1
------	25	alles	1
------	25	doch	1
------	25	nicht	1
------	25	ums	1
------	25	wohl	1
------	25	Zeit	1

Tabelle 41: Aus dem „Madagaskar"-Vokabular entfernte, dennoch von Vtn markierte Wörter

Es handelt sich dabei um diejenigen Wörter aus den 130 Wortformen des Originaltextes, die aufgrund der Worthäufigkeitsanalyse nach der Lemmatisierung aus dem Testvokabular entfernt wurden, aber trotzdem von mindestens einem Vtn als Kernwort markiert worden sind – d.h. als wichtige Wörter angesehen wurden. – Hier zeigt sich, dass die Entfernung von „allein", „alles", „doch", „nicht", „ums", „wohl" und „Zeit" aus dem Vokabular korrekt war, da sie nur von einem Vtn bzw. $1/66=1{,}5$ %

der Vtn zu den Kernwörtern gerechnet wurden. Das Gegenbeispiel „Makis“ mit 12 Vtn bzw. $12/66 = 18{,}2$ % der Vtn zeigt, dass die worthäufigkeits-basierte Vokabularreduktion eine Heuristik darstellt, die meist befriedigend arbeitet, aber mitunter versagt.

Bezogen auf die Verteilung in Tabelle 40 finden sich diese 13 Wörter offensichtlich in der Spalte „ATA2>20“. Nur das Wort „Makis“ gehört mit seinem realen Rang von 16 zu den 12 Wörtern im Feld „ATA2>20“ / „Real<=20“. Alle anderen dargestellten Wörter gehören zu den 44 Wörtern im Feld „ATA2>20“ / „Real>20“.

Zum eigentlichen Unabhängigkeitstest: 2x2-Felder-Tafeln[42] können für jeden einzelnen Text aufgestellt werden. Es ist zu erwarten, dass sich in den Feldern „ATA2<=20 / Real<=20“ und „ATA2>20 / Real>20“ möglichst viele Wörter finden. In dem Fall hätte ATA2 viele der realen Top 20 Wörter korrekt in die Top 20 gewählt und viele der real unwichtigeren Wörter korrekt auf die schlechteren Ränge positioniert – ggf. mit Hilfe der vorgeschalteten Vokabularreduktion. Ebenso sollen in den verbleibenden Feldern, in denen gute reale Ränge mit niedrigen ATA2-Rängen korrespondieren und umgekehrt, möglichst wenige Wörter gezählt werden.

	C	D	Summe
A	C_A	D_A	S_A
B	C_B	D_B	S_B
Summe	S_C	S_D	N

Tabelle 42: Generische 2x2-Felder-Tafel

Eine mathematische Prüfgröße für den eben in Wörtern umschriebenen Zusammenhang lässt sich mit dem Chi²-Unabhängigkeitstest berechnen. In einer generischen 2x2-Felder-Tafel wie in Tabelle 42 berechnet sich die Prüfgröße wie folgt[43]:

$$\chi^2 = \frac{N \cdot (C_A \cdot D_B - C_B \cdot D_A)^2}{S_A \cdot S_B \cdot S_C \cdot S_D}$$

Bezogen auf die konkreten Werte des „Madagaskar“-Textes in Tabelle 40 auf Seite 113 ergibt sich:

$$\chi^2 = \frac{76 \cdot (10 \cdot 44 - 10 \cdot 12)^2}{22 \cdot 54 \cdot 20 \cdot 56} \approx 5{,}85$$

Ein Chi²-Wert von 5,85 entspricht laut Umrechnungstabelle in der Literatur einer Wahrscheinlichkeit von lediglich etwa 1,5%, dass es sich bei der o.a. Wortverteilung um ein Zufallsergebnis handelt. Anders ausgedrückt: Der Zusammenhang zwischen der ATA2-Rangfolge und der realen Rangfolge hinsichtlich der ATA2-Klassifikation als Top 20-Wort oder Nicht-Top 20-Wort ist beim „Madagaskar“-Text *sehr signifikant*.

42 In der Literatur werden 2x2-Felder-Tafeln mitunter auch als 4-Felder-Tafeln bezeichnet.

43 Die Berechnung der Prüfgröße bei Chi²-Unabhängigkeitstests mit mehr als 2x2 Feldern erfolgt über eine Summenformel, die sich im Falle von 2x2-Felder-Tabellen in den dargestellten, simplen Bruch umformen lässt.

4.6 Auswertungsergebnisse

Dieser Abschnitt reiht die numerischen Fakten aneinander, deren Methoden im vorherigen Abschnitt beschrieben wurden. Jeder Unterabschnitt hat ein Pendant in der Methodenbeschreibung.

Vorab sind in Tabelle 43 zur erneuten Verdeutlichung die Wortanzahlen der einzelnen Testtexte mitsamt unterscheidbaren Wortformen und Anzahl der Lemmata im worthäufigkeits-reduzierten Testvokabular dargestellt – nebst weiteren, nachfolgend beschriebenen statistischen Angaben pro Testtext.

Textnr.	Texttitel	Wortanzahl	Unterscheidbare Wortformen	Markierte Wortanzahl	Markierte Ränge	Vokabularreduzierte Anzahl Lemmata	10 Worte Prozent	20 Worte Prozent
1	Madagaskar	204	130	62	25	63	15,9%	31,7%
2	Bushaltestelle	167	103	47	26	43	23,3%	46,5%
3	Wein	190	142	58	27	69	14,5%	29,0%
4	Igel	169	119	46	24	62	16,1%	32,3%
5	Computerspiele	206	146	44	22	55	18,2%	36,4%
6	Tannenwald	186	128	51	26	53	18,9%	37,7%
7	Körpersprache	188	128	55	24	63	15,9%	31,7%
8	Autor	188	142	52	25	64	15,6%	31,3%
9	Wissenschaft	170	127	54	24	65	15,4%	30,8%

Tabelle 43: Wortanzahl- / Markierungsanzahl-Statistik

Von den 204 Wörtern bzw. 130 unterscheidbaren Wortformen im „Madagaskar"-Text (vgl. Text 17 auf S. 93) wurden 62 Wörter von jeweils mindestens einem Vtn markiert. Durch Zusammenfassung gleich häufig markierter Wörter (vgl. Tabelle 32 und Tabelle 33 auf S. 99) ergeben sich 25 unterscheidbare Rangpositionen. Die exakte Rangfolge der einzelnen Wörter findet sich für jeden Text im Anhang.

Weiterhin gilt: Die zehn Kernwörter, die ATA2 aus den 63 verbliebenden Lemmata im reduzierten Vokabular innerhalb seiner Top 10 platziert, entsprechen 15,9% des Textmaterials; bei den zwanzig Top 20-Wörtern sind es schon 31,7% des Textmaterials.

Wann immer also im weiteren Verlauf von ATA2s Top 10- und Top 20-Berechnungen die Rede ist, sollte jeweils beachtet werden, welchen Prozentsatz der Wörter ATA2 als wichtig bzw. unwichtig einzustufen hat.

4.6.1 Lineare ungewichtete Trefferquote

4.6.1.1 ATA2-Top 10 vs. Vtn-Top 10-Erwartungswert

Der Erwartungswert der Vtn-Trefferquote ist nicht für jeden Text gleich. (Zur Erinnerung: Der Erwartungswert ist die durchschnittliche Trefferzahl bzw. -quote der Vtn, d.h. $E(X)=\sum \%(Teilnehmer)\cdot\%(Treffer)$, vgl. Tabelle 34 auf S. 101.) Eine vollständige Übersicht ist in Tabelle 44 auf Seite 116 dargestellt.

Textnr.	1	2	3	4	5	6	7	8	9	Ø
Erwartungswert Vtn	64.24%	64.09%	54.39%	65,00%	74.54%	53.78%	59.84%	65.3%	72.42%	65,00%

Tabelle 44: Erwartungswerte der Vtn-Trefferquote für alle neun Versuchstexte

Die Vtn markieren im Mittel über alle Texte folglich jeweils 65% der mehrheitlich ermittelten Kernwörter.

Im Gegensatz dazu schwankt die ATA2-Top 10-Trefferquote ebenfalls hinsichtlich der realen Top 10 für sämtliche neun Versuchstexte, wie in Tabelle 45 dargestellt.

Textnr.	1	2	3	4	5	6	7	8	9	Ø
Trefferquote Top-10:	30%	60%	70%	40%	50%	60%	50%	40%	40%	48,88%

Tabelle 45: ATA2s Top 10-Trefferquote für alle neun Versuchstexte

ATA2s Top 10-Trefferquote liegt durchschnittlich bei 48,88%. Lediglich bei den drei Testtexten 2, 3 und 6 („Bushaltestelle“, s. Text 22; „Wein“, s. Text 20 auf S. 96; „Tannenwald“, s. Anhang) gibt es positive Ausreißer in Richtung des erwünschten durchschnittlichen Vtn-Erwartungswertes.

Viele Leute stehen an der Haltestelle. Ein junger Herr hat einen großen Koffer bei sich. Er will verreisen. Die Frau mit der Einkaufstasche fährt in die Stadt. Die Schulkinder mit dem Ranzen auf dem Rücken lesen noch den Anschlag an der Plakat - Säule. Ein alter Mann mit einem Stock hat sich einstweilen auf die Bank gesetzt. Alle warten auf den Bus. Jetzt kommt der Wagen. Langsam bremst er. Ein paar Leute steigen aus. Der junge Mann mit dem Koffer hat es besonders eilig. Er drängt sich vor und steigt als erster ein. Die Schulkinder schlüpfen noch rasch in den Wagen hinein. Die Leute lösen ihre Fahrscheine. Die Schulkinder zeigen dem Schaffner ihre Monatskarten. Die alte Frau hat noch keinen Sitzplatz gefunden. Sie kann schlecht stehen, weil der Bus schaukelt. Da steht ein junges Mädchen von seinem Sitzplatz auf. Die alte Frau ist froh und dankbar. Beim Aussteigen stecken die Leute ihre Fahrscheine in den Abfall. Da haben es die Schulkinder eilig. Sie drängen vor zum Ausstieg.

Abfall, Alle, Anschlag, Ausstieg, Beim, Bus, Einkaufstasche, Fahrschein, Haltestelle, Herr, Jetzt, Koffer, Langsam, Lese, Monatskarte, Mädchen, Plakat, Ranzen, Schaffner, Schulkind, Sitzplatz, Stock, Säule, aussteigen, besonders, bremsen, dankbar, drängen, eilig, einstweilen, froh, hinein, lösen, paar, rasch, rücken, schaukeln, schlecht, schlüpfen, stecken, verreisen, wagen, warten

Text 22: „Bushaltestelle“-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Ein Zusammenhang zwischen diesem Ergebnis und dem Schwierigkeitsgrad des jeweiligen Textes lässt sich auf den ersten Blick nicht ausmachen: Die Texte mit den ungeraden Nummern (1, 3, 5, 7, 9) stammen aus der Erwachsenenliteratur, die Texte mit den geraden Nummern (2, 4, 6, 8) wurden aus Lesebüchern für Kinder entnommen.

4.6.1.2 ATA2-Top 20 vs. Vtn-Top 10-Erwartungswert

Die ATA2-Trefferquote lässt sich verbessern, wenn statt der Top 10 die Top 20 betrachtet werden – und zwar beidseitig, d.h. sowohl auf ATA2-Seite als auch bei den realen Kernwörtern werden die 20 höchstplatzierten Wörter betrachtet und miteinander verglichen. – Auch hier werden wiederum die von den Vtn gleich oft markierten Wörter zu gemeinsamen Rängen zusammengefasst, beim „Madagaskar“-Text also neben „Entwicklung“ und „Kampf“ mit jeweils 26 Markierungen auf Rang 7 auch „starb“ und „Widersacher“ mit jeweils 13 Markierungen auf Rang 15 (siehe Tabelle 32/33 auf Seite 98/99). Demnach berechnet sich ATA2s Trefferquote als die Prozentzahl aus seinen 20 Wörtern, die unter den 22 Wörtern der realen Top 20 erscheinen.

In Tabelle 46 sind ATA2s Top 20-Trefferquoten für sämtliche neun Versuchstexte dargestellt.

Textnr.	1	2	3	4	5	6	7	8	9	Ø
Trefferquote Top-20:	50%	80%	75%	55%	70%	55%	75%	65%	60%	65,00%

Tabelle 46: ATA2s Top 20-Trefferquote für alle neun Versuchstexte

Bei Erweiterung auf die Top 20 erreicht ATA2 also exakt die 65,00% Trefferquote, die dem durchschnittlichen Top 10-Erwartungswert der Vtn von 65% (siehe Tabelle 44 auf S. 116) entspricht. Auch das Schwanken der Trefferquoten zwischen 50% und 80% ist mit den schwankenden Top 10-Erwartungswerten der Vtn vergleichbar.

4.6.2 Gewichtete multidimensionale Euklidische Qualität

4.6.2.1 Versuchsteilnehmer – 10-dimensional

Tabelle 47 stellt zunächst die Resultate der Vtn dar. Für jeden der neun Testtexte wurden die Wortmarkierungen von jedem einzelnen der 66 Vtn aus der Datenbank rekonstruiert und mit Hilfe des Vektormodells der Gewichteten Qualität bewertet. Pro Text lagen also 66 Gewichtete Qualitäten vor, über die sich Minimum, Maximum, Mittelwert und Standardabweichung berechnen ließen.

Textnr.	1	2	3	4	5	6	7	8	9
Min. Gew. 10D Qual.:	0,089	0,169	0,219	0,164	0,293	0,118	0,116	0,202	0,167
Max. Gew. 10D Qual.:	0,777	0,825	0,767	0,848	0,769	0,695	0,806	0,899	0,658
Ø	0,546	0,515	0,470	0,547	0,522	0,418	0,487	0,531	0,396
Std.Abw.	0,153	0,151	0,132	0,135	0,112	0,146	0,131	0,164	0,117

Tabelle 47: Mittelwerte der Gewichteten Qualitäten der Vtn über alle Texte

Bei jedem Text gab es also Vtn, deren Wortmarkierungen extrem weit von der optimalen Lösung entfernt lagen – anders ist etwa die minimale Gewichtete Qualität von 0,089 bei Text 1 „Madagaskar“ (s. Tabelle 30 auf S. 93) kaum erklärlich. Andererseits gab es auch Vtn, die die korrekten Top 10 nahezu vollständig markiert hatten – so zu erkennen an der maximalen Gewichteten Qualität von 0,899 bei Text 8 „Autor“ (s. Tabelle 34 auf S. 101).

Die Gewichtete Qualität schwankt stark zwischen den einzelnen Vtn und auch von Text zu Text. Natürlich basiert die Auswahl der Kernwörter auf dem Weltwissen der jeweiligen Versuchsteilnehmer über das im jeweiligen Testtext behandelte Themengebiet. Sogar die geringsten Standardabweichungen von 0,112 und 0,117 bei Text 5 „Computerspiele“ (Volltext siehe Anhang) und Text 9 „Wissenschaft“ (siehe Text 23 auf Seite 118) sind dabei nicht unerheblich und zeigen die „Uneinigkeit“ der Versuchsteilnehmer untereinander.

Die Frage, die der Versuchs - Leiter im Labor an die Natur stellt, hat immer schon zur Voraussetzung, daß er eine Vermutung hegt, die er bestätigt oder widerlegt wissen will. Diese Vermutung aber stammt stets aus einer vorher gehenden Beobachtung, mit anderen Wörtern, aus jenen nicht - rationalen kognitiven Leistungen unserer Sinnes - Organe und unseres Nerven - Systems, die aus Sinnes - Daten Wahrnehmungen formen. Es bedeutet eine gewaltige Überschätzung der menschlichen Ratio, wenn sich ein Wissenschaftler einbildet, er wisse und kenne alle Fragen, die man an die Natur stellen kann. Wenn ein Forscher sich die Frage ausdenkt, während er, in sein Laboratorium gebannt, seiner Arbeit ohne Kontakt mit der unabsehbaren Reichhaltigkeit der lebenden Natur obliegt, so kann es allzu leicht geschehen, daß seine Frage an dem wirklich Wesentlichen vorbeigeht und nur Irrelevantes zutage fördert. So werden dann Untersuchungen angestellt, die trotz größtem Scharfsinn und trotz bester methodischer Berücksichtigung aller Einzelheiten keineswegs das Lebenswichtige betreffen. Der Forscher aber, der sich ganz eben dieser Untersuchung widmet, kann dies unmöglich einsehen.

Beobachtung, Berücksichtigung, Datum, Einzelheit, Form, Forscher, Frage, Kontakt, Labor, Laboratorium, Leistung, Leiter, Natur, Nerv, Organ, Ratio, Reichhaltigkeit, Scharfsinn, Sinn, System, Untersuchung, Vermutung, Versuch, Voraussetzung, Wahrnehmung, Wissenschaftler, allzu, anstellen, ausdenken, bannen, bedeuten, bestätigen, betreffen, eben, einbilden, einsehen, fördern, geschehen, gewaltig, größtem, hegen, irrelevant, jene, keineswegs, kennen, kognitiv, lebenswichtig, leicht, menschlich, methodisch, obliegen, rational, stammen, stets, trotz, unabsehbar, unmöglich, vorbeigehen, vorher, wesentlich, widerlegen, widmen, wirklich, zutage, Überschätzung

Text 23: „Wissenschaft"-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Wie in Tabelle 48 dargestellt errechnet sich die durchschnittliche Gewichtete Qualität aller Vtn bezogen auf alle Texte als Meta-Mittelwert über die in Tabelle 47 auf S. 117 aufgelisteten Mittelwerte für jeden einzelnen Text: Das heißt, die Vtn erreichen durchschnittlich eine Gewichtete Qualität von 0,492. Die Schwankung der Mittelwerte von Text zu Text ist dabei gering, wie die geringe Meta-Standardabweichung 0,052 belegt (d.h. die Standardabweichung der Mittelwerte aller Texte in Tabelle 47).

Ø der Ø	0,492
Std.Abw. der Ø	0,052
Ø der Std.Abw.	0,138

Tabelle 48: Meta-Mittelwert – Gewichtete Qualitäten der Vtn über die gesamte Studie und Mittelwert über die „Uneinigkeit" der Vtn

Nicht zu verwechseln ist diese Meta-Standardabweichung mit den Schwankungen der Gewichteten Qualitäten innerhalb eines Textes, als Maß für die „Uneinigkeit" der Versuchsteilnehmer beim jeweiligen Testtext – d.h. mit den Standardabweichungen der einzelnen Texte, die zwischen 0,112 und 0,164 schwanken, im Mittel 0,138 (d.h. als Mittelwert der Standardabweichungen in Tabelle 47 auf S. 117).

Hinter all den Zahlen verbirgt sich folgendes überraschende Ergebnis: Die durchschnittliche Gewichtete Qualität aller Vtn über alle Texte liegt mitnichten nahe am Maximum. Das heißt, Menschen markieren zehn Kernwörter auch in Texten aus Lesebüchern für die 2. und 3. Schulklasse nicht derart souverän, wie man es erwarten sollte. Ein Zusammenhang der mittleren Gewichteten Qualität mit der Schwere des Textes (Erwachsenenliteratur vs. Lesebuchtexte für Kinder) ist nicht auszumachen.

4.6.2.2 ATA2 – 10-dimensional / 20-dimensional

In Tabelle 49 sind neben den linearen ungewichteten Trefferquoten von ATA2 auch die Gewichteten Qualitäten aufgeführt, jeweils bezogen auf die Top 10 und die Top 20.

Textnr.	1	2	3	4	5	6	7	8	9	Ø	Std.-Abw.
Trefferquote Top-10:	30%	60%	70%	40%	50%	60%	50%	40%	40%	48,88%	11,96%
Gew. 10-dim. Eukl. Qual.:	0,377	0,538	0,536	0,369	0,160	0,404	0,485	0,183	0,190	0,360	0,141
Trefferquote Top-20:	50%	80%	75%	55%	70%	55%	75%	65%	60%	65,00%	10,00%
Gew. 20-dim. Eukl, Qual.:	0,379	0,462	0,501	0,394	0,319	0,432	0,392	0,317	0,332	0,391	0,060

Tabelle 49: ATA2s Top 10- / Top-20-Trefferquote und Gewichtete Qualitäten für alle neun Testtexte

Deutlich zu erkennen ist die Aussagekraft der Gewichteten Qualität beispielsweise im Vergleich der Top 10-Bewertungen von Texttext 1 „Madagaskar“ (s. Text 17 auf S. 93) und Testtext 8 „Autor“ (s. Text 21 auf S. 102): Beim „Madagaskar“-Text führen die hochqualitativen Treffer 1-3 zu einer brauchbaren Gewichteten Qualität von 0,377 (vgl. Tabelle 32 auf Seite 98), während beim „Autor“-Text (vgl. Tabelle 36 auf S. 103) die sogar höhere Trefferquote von 40% bei den niedrigen Treffern 7-9 (und 1) lediglich zu einer Gewichteten Qualität von 0,183 führen.

Rang	ATA2	Real	Anzahl
1	[Weintrinker/ Weintrinkern]	[Wein/Weins/ Weinen/Weine]	62
2	[Wein/Weins/ Weinen/Weine]	Ausgewogenheit	47
3	Einseitigkeit	Genuß	46
4	Genuß	Vergnügen	41
5	Flasche	[sinnliches/ sinnlichen]	39
6	Weinliebhaber	Charakter	31
7	[sinnliches/ sinnlichen]	Befriedigung	25
8	Ausgewogenhei	intellektuelles	24
9	Nuancen	[Weintrinker/ Weintrinkern]	23
10	Befriedigung	Weinliebhaber	22

Tabelle 50: „Wein“-Top 10 – ATA2 und real

Im Gegensatz dazu liefert der 70%-Ausreißer beim Testtext 3 „Wein (s. Text 20 auf S. 96) eine hervorragende Gewichtete Qualität von 0,536 (nur die drei Wörter 4, 6 und 8 („Vergnügen“, „Charakter“ und „intellektuelles“) nicht gefunden, vgl. Tabelle 50).

Diese wird jedoch von der sogar geringeren 60%-Trefferquote beim Testtext 2 „Bushaltestelle“ (s. Text 22 auf Seite 116) mit 0,538 noch leicht überschritten (immerhin vier Wörter 6-9 („Herr“, „eilig“, „Leute“, „warten“) nicht gefunden, vgl. Tabelle 51 auf S. 120).

Das Maß der Gewichteten Qualität ist also valide: Eine höhere Trefferzahl führt nicht notwendigerweise zu einer höheren Gewichteten Qualität.

Zurück zur Übersicht der Gewichteten Qualitäten in Tabelle 49 auf Seite 119: Der Top 10-Mittelwert der Gewichteten Qualitäten über alle Texte von 0,360 ist aufgrund der hohen Standardabweichung von 0,141 jedoch wenig aussagekräftig: Die Resultate sind stark schwankend. Auch hier zeigt sich, dass die Betrachtung der Top 10-Bewer-

Rang	ATA2	Real	Anzahl
1	Ranzen	Haltestelle	61
2	Fahrscheine	Fahrscheine	53
3	Schaffner	Sitzplatz	51
4	Schulkinder	Schulkinder	49
5	Haltestelle	Bus Frau	43
6	Bus	Herr	28
7	Monatskarten	eilig	26
8	schaukelt	Leute	24
9	Sitzplatz	warten	23
10	Koffer	Koffer	22

Tabelle 51: „Bushaltestelle"-Top 10 – ATA2 und real

tungen zugunsten der Top 20-Bewertungen vernachlässigt werden sollte: Im Top 20-Bereich wird ein Mittelwert der Gewichteten Qualitäten von 0,391 erzielt, bei einer Standardabweichung von lediglich 0,060, d.h. bei nur geringer Schwankung.

Auch werden beim Wechsel von 10 auf 20 betrachtete Wörter die schlechten Top 10-Resultate 0,160 und 0,183 bei Testtext 5 „Computerspiele" (siehe Anhang) und Testtext 8 „Autor" (siehe Text 21 auf S. 102) stark verbessert: Beide kommen mit 0,319 bzw. 0,317 nahe an den Mittelwertbereich (d.h. Mittelwert ± Standardabweichung).

4.6.2.3 Vergleich ATA2 / Vtn

Was sagt jedoch eine von ATA2 erzielte mittlere Gewichtete Top 20-Qualität von 0,391 mit hinreichend niedriger Standardabweichung von 0,060 aus (vgl. Tabelle 49 auf S. 119)? Klarer gefragt: Ist dies ein gutes oder ein schlechtes Ergebnis?

Im Gegensatz zu ATA2 erreichen die Vtn im Mittel über alle Teilnehmer und Texte bei Bewertung der Top 10-Wörter (die ja genau ihren zehn Wortmarkierungen entsprechen) eine nur etwas höhere Gewichtete Qualität von 0,492 (vgl. Tabelle 48 auf S. 118), jedoch stark schwankend zwischen sehr guten und sehr schlechten Gewichteten Qualitäten. Setzt man 0,391 mit 0,492 in Bezug, erhält man eine ATA2-Erfolgsquote von

$$\frac{0{,}391}{0{,}492}\cdot 100 = 79{,}47\,\%$$ [44].

So betrachtet lässt sich argumentieren, ATA2 liefere relativ souverän Ergebnisse, die mit einer Annäherung von knapp 80% an durchschnittliche Ergebnisse von Versuchsteilnehmern nun wirklich „gar nicht so schlecht" sind.

Allerdings muss an dieser Stelle deutlich darauf hingewiesen werden, dass hier ein ungerechter Vergleich stattfindet: Einerseits wird Bezug genommen auf die Gewichtete Qualität der ATA2-Ergebnisse unter Einbeziehung der berechneten bzw. realen

44 Dieser Quotient ist insoweit korrekt, dass die Gewichtete Qualität zwar ein nicht-lineares Bewertungsmodell enthält, jedoch bei der anschließenden Qualitätsberechnung Abstände vom optimalen Ergebnis per Euklidischer Norm ermittelt und sie anschließend [0..1]-normiert werden. Insofern lassen sich jeweils normierte Gewichtete Qualitäten durchaus linear miteinander vergleichen – auch solche unterschiedlicher Dimensionalitäten.

Top 20. Auf Seiten der Vtn werden dagegen die Top 10-Ergebnisse einbezogen. Dies ist anders schwerlich möglich, da die Vtn zehn Wörter zu markieren hatten - und nicht zwanzig.

Die Standardabweichung von ATA2s Gewichteter Top 10-Qualität liegt mit 0,141 (vgl. Tabelle 49 auf S. 119) nur geringfügig höher als die durchschnittliche Standardabweichung der Vtn von 0,138 (vgl. Tabelle 48 auf S. 118). Obwohl derart starke Schwankungen kaum valide Abschätzungen zulassen, ergibt sich rechnerisch eine ATA2-Erfolgsquote von immerhin noch

$$\frac{0{,}360}{0{,}492} \cdot 100 = 73{,}17\% \, .$$

Klarer ausgedrückt bedeutet dies, dass ATA2 auch bei der Berechnung von zehn Kernwörtern im Vergleich mit „durchschnittlichen" Versuchsteilnehmern eine qualitative Annäherung von deutlich über 70% erzielt, bei ähnlich starker qualitativer Schwankung wie bei durchschnittlichen Vtn auch.

4.6.2.4 Zusammenfassung Gewichtete Qualität

ATA2 erzielt im Vergleich zu den Vtn bezogen auf die Gewichteten Qualitäten eine Annäherung von 75-80%. Betrachtet aus der Perspektive des Bewertungsmodells bedeutet eine Gewichtete Top 20-Qualität von 0,391 (vgl. Tabelle 49 auf S. 119) jedoch, dass der Abstand vom optimalen Zielpunkt im n-dimensionalen Koordinatensystem immer noch mehr als die „halbe Wegstrecke" entfernt liegt.[45]

Dafür sorgt der Umstand, dass eine gute oder gar lediglich mittelmäßige Trefferquote im Vergleich zur realen Top 10 oder Top 20 auch entsprechend viele Nicht-Treffer liefert. Derartige von ATA2 berechnete Wörter wurden von den Vtn nicht oder nur selten zu den wirklichen Kernwörtern gezählt.

4.6.3 Korrelationsanalyse

4.6.3.1 Korrelationsgrafik

In Abbildung 52 auf S. 122 wird versucht, den von ATA2 berechneten Rang als Punktdiagramm in Beziehung mit dem realen Rang zu bringen, entsprechend der Mehrheitsentscheidung der Vtn. Das Diagramm ist auf die ATA2-Ränge 1-20 beschränkt, zeigt aber alle realen Ränge.

45 Dafür ist in letzter Konsequenz die Dreiecksungleichung verantwortlich: Selbst im ungewichteten 2D-Fall führt ein nicht gefundenes Wort von zwei zu findenden Wörtern schon zu einer Entfernung vom optimalen Zielpunkt von mehr als der Hälfte der Wegstrecke. Genau das sagt die Dreiecksungleichung aus: Bei nicht-entarteten Dreiecken ist die Summe der Längen der beiden Katheten immer länger als die Hypotenuse.

Zur Erinnerung: Durch gleich häufig markierte Wörter sind die Ränge mehrfach belegt. Beispielsweise beim „Madagaskar"-Text (vgl. Text 17 auf S. 93) mit seinen 204 Wörtern wurden von den 130 unterscheidbaren Wortformen 62 Wörter von mindestens einem Vtn markiert, zusammen ergeben sich 25 unterscheidbare Rangpositionen (vgl. Tabelle 43 auf S. 115). Die maximal unterscheidbare Rangzahl von 27 wird z.B. beim Testtext 3 „Wein" (vgl. Text 20 auf S. 96) erreicht.

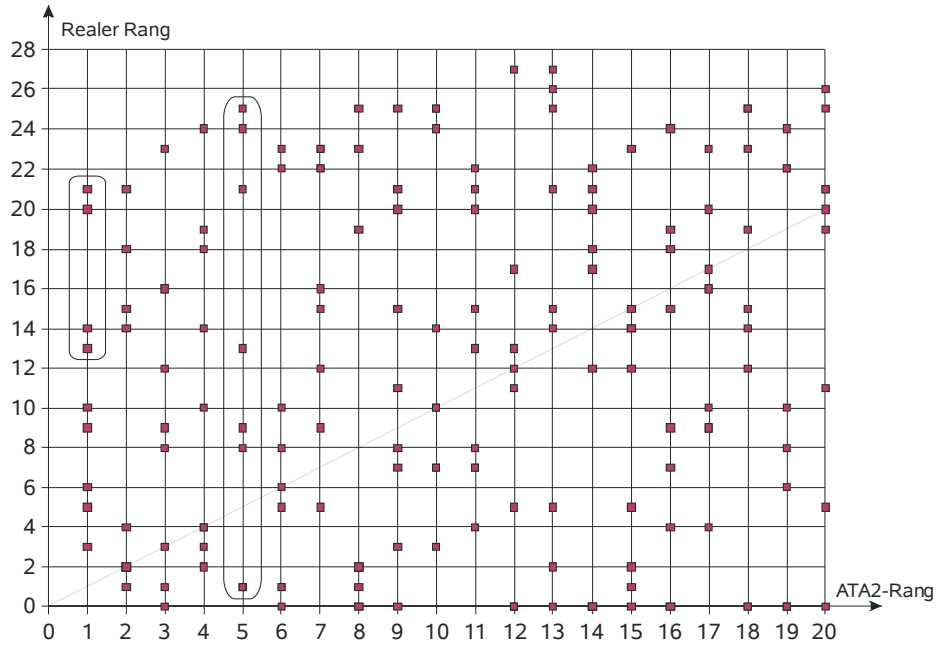

Abbildung 52: Versuch einer Korrelationsgrafik: ATA2-Top 20-Rang vs. Realer Rang

Der Versuch misslingt augenscheinlich: Die gewünschte Punktwolke um die Diagonale $(1{,}1) \rightarrow (20{,}20)$ bzw. allgemein $(1{,}1) \rightarrow (x\,,x)$ ist nicht vorhanden: Die horizontal dargestellten ATA2-Rangpositionen verteilen sich vertikal kunterbunt über die realen Rangpositionen. Beispielsweise ist erkennbar, dass die von ATA2 auf Rang 5 berechneten Kernwörter zwischen Rang 1 und Rang 25 variieren (siehe Markierung).

Wie bereits beschrieben verfügten die Vtn im Rahmen des Versuchsaufbaus der Wortmarkierungsstudie nicht über die Möglichkeit, ihre persönliche Rangfolge innerhalb der von ihnen markierten 10 Wörter zu bestimmen. Angenommen, dass die Bewertung der Wichtigkeit der Kernwörter stark auf dem Vorwissen der Versuchsteilnehmer basiert, so ist es verzeihlich, wenn es keine 1:1-Korrelation zwischen der ATA2-Rangfolge und der realen Rangfolge innerhalb der Top 20 gibt.

Nichtsdestotrotz ist die Menge der stark fehlerhaften Ränge, insbesondere unter den hochrangigen Wörtern der Top 10, unerfreulich hoch. So finden sich auf dem ATA2-Rang 1 sogar Wörter, die von den Vtn lediglich auf Rang 13, 14, 20 und 21 gewählt wurden (siehe Markierung).

Eine Ausweitung von den ATA2-Rängen 1-20 auf sämtliche ATA2-Ränge liefert das in Abbildung 53 abgebildete Diagramm. Wiederum eingezeichnet ist die Diagonale $(1,1) \rightarrow (x, x)$, um die eigentlich eine Punktwolke zu erwarten gewesen war. Einige Beobachtungen fallen sofort ins Auge:

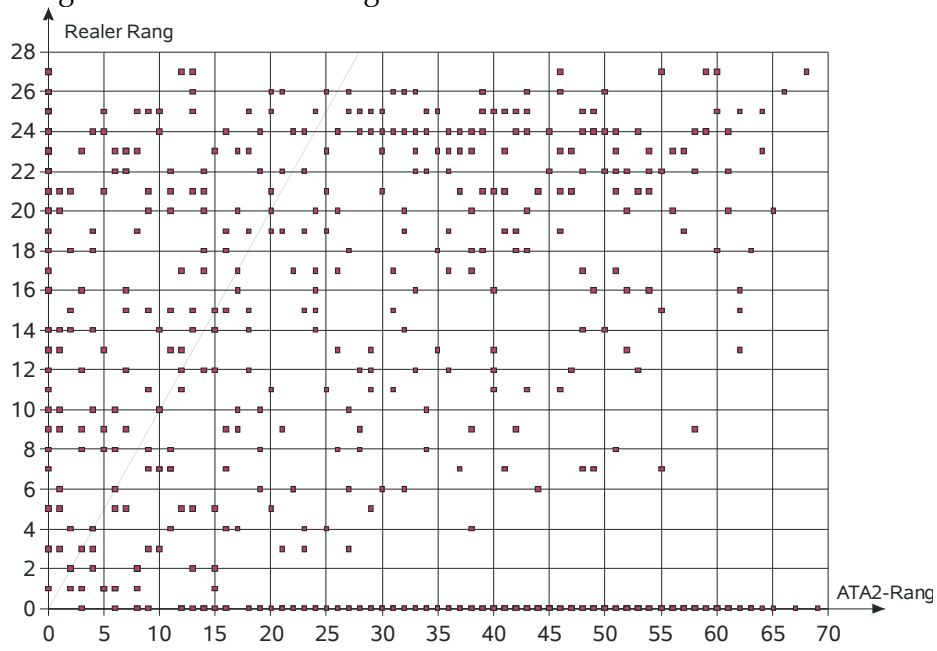

Abbildung 53: Versuch einer Korrelationsgrafik: aller ATA2-Ränge vs. Realer Rang

1. Die Fehlwörter (zur Erinnerung: diejenigen Wörter, die kein Vtn jemals markiert hat), die auf der ATA2-Achse mit realem Rang 0 eingezeichnet sind, häufen sich oberhalb von Rang 20. Das ist insoweit verzeihlich, da ATA2 bei der Berechnung jedem Wort des Testvokabulars einen Rang zuweisen muss – und sei es ein sehr hoher bei extrem ungeeigneten Wörtern. Dies sind exakt die Fehlwörter.

2. Die Wörter, die auf der Real-Achse mit ATA2-Rang 0 eingezeichnet sind, entsprechen den Wörtern, die die Vtn zwar markiert haben, die aber aufgrund der Worthäufigkeitsklassenanalyse aus dem Testvokabular entfernt wurden. Hier ist ein ähnlicher Trend zu beobachten: Sie häufen sich bei höheren Rängen.
   Erfreulich dabei ist, dass in Abbildung 52 auf S. 122, bei Einschränkung auf die hochrangigen ATA-Top 20, kein einziges dieser Wörter vorkommt.

3. Durch scharfes Hinsehen lässt sich ein Trend entdecken, wonach die Punktwolke zumindest bei höheren ATA2-Rängen und niedrigen realen Rängen abnimmt (im Diagramm unten rechts). Das heißt, ATA2 berechnet tendenziell real wichtige Wörter nicht als extrem unwichtig. – Eine genauere Untersuchung dieses Zusammenhangs erfolgt weiter unten im Rahmen der Chi²-Tests ab Seite 126.

Insbesondere letztere Beobachtung lässt sich wie folgt deuten: Aufgrund der vergleichsweise geringen Anzahl von Vtn gibt es eine Fülle von mehrfach belegten Rängen mit geringen Rangnummern von maximal 27, trotz bis zu 62 mindestens einmal markierten Wörtern pro Testtext (vgl. Tabelle 43 auf S. 115).

Bei mehr Vtn hätte sich schon rein statistisch eine Diversifikation der Rangplatzierungen der nun gleichrangigen Wörter ergeben, während die Rang*folge* vermutlich trotzdem im Groben unverändert geblieben wäre. Es steht zu erwarten, dass bei mehr Vtn eine vertikal gestreckte Punktwolke mit ggf. stärkerer Orientierung an der Diagonale $(1,1) \rightarrow (x, x)$ entstanden wäre.

4.6.3.2 ATA2-Ränge und ihre Entsprechungen

Aufgrund der soeben entwickelten, kaum vorhandenen Annäherung der ATA2-Rangfolge an die reale Rangfolge ergibt sich, dass eine durchschnittliche Top 20-Trefferquote von 65,00% bei durchschnittlicher Gewichteter Qualität von immerhin 0,391 mit gewisser Vorsicht betrachtet werden muss: ATA2 hat bei den Top 20-Rängen der neun Testtexte immerhin 21x ein Wort gewählt, das von den Vtn kein einziges Mal als Kernwort markiert worden ist. 42x wählte ATA2 Wörter oberhalb der realen Top 20, die also entsprechend selten markiert worden waren. Insgesamt ergibt sich die in Tabelle 54 dargestellte Verteilung.

Reale Ränge	≤ 20	>20	—
Anzahl	117	42	21
Prozentual	65,00%	23,33%	11,67%

Tabelle 54: ATA2 Top 20-Ränge und Ihre Entsprechungen der realen Ränge

Sollen die ATA2-Top 20-Resultate in der Praxis verwendet werden, muss mit 11,67% Totalfehlern sowie mit 23,33% wenig relevanten Wörtern gerechnet werden. Die Auswertung der von ATA2 gelieferten Rangfolge ist dabei keinesfalls empfehlenswert.

4.6.3.3 Fehlwortquote der Vtn

Textnr.	1	2	3	4	5	6	7	8	9	Ø	Std.-Abw.
Anz. Fehlworte	15	8	11	7	7	12	9	15	12	10,66	2,79
Fehlwortquote	22,72%	12,12%	16,66%	10,60%	10,60%	18,18%	13,63%	22,72%	18,18%	16,16%	4,46%

Tabelle 55: Fehlwortquote der Vtn über alle neun Testtexte

Die Fehlwortquote der Versuchsteilnehmer (zur Erinnerung: Wörter, die nur ein einziger Vtn markiert hat) schwankt über die Versuchstexte, wie in Tabelle 55 dargestellt. Die von den Vtn markierten Wörter entsprechen also zu durchschnittlich 16,16% Fehlwörtern – d.h. Wörtern, die nur sie und kein anderer Vtn als Kernwort ansehen. Verglichen damit ist die von ATA2s erzielte Fehlwortquote von 11,67% hervorragend.

Wiederum muss beachtet werden, dass sich die 16,16% Fehlwortquote der Vtn auf die Top 10 beziehen, da sie ja zehn Wörter zu markieren hatten. ATA2s Fehlwortquote von 11,67% bezieht sich jedoch auf die Top 20 (vgl. Tabelle 54 auf S. 124).

Wir verwenden Zeit und Energie, um neben unserer Mutter - Sprache noch weitere Sprachen zu lernen. Körper - Sprache ist mit der Zeit zu einer Fremd - Sprache geworden. Fremd - Sprachen müssen nicht gelernt werden, aber wir kommen weiter, wenn wir sie beherrschen. Wir vermindern die Gefahr von Mißverständnissen. Es ist mir unerklärlich, warum wir nie die Zeit haben, unsere Primär - Sprache, nämlich die Sprache unseres Körpers, zu verbessern. Da sich niemand des Kommunikations - Mittels Körper - Sprache entziehen oder sie unterdrücken kann, ist es von wesentlichem Nutzen, sie zu lernen - gibt sie uns doch wichtige Informationen über die innere Haltung und Einstellung unserer Mitmenschen. Wenn wir offene Sinne und ein waches Auge für die Signale und Kommentare unserer Körper - Sprache haben, können viele Gespräche und Begegnungen leichter und erfolgreicher verlaufen. Die Kenntnis der Körper - Sprache, des lautlosen Frage- und Antwort- Spiels in unserem körperlichen Verhalten, öffnet direktere Wege zueinander und einen freieren Umgang miteinander. In manchen sprachlosen "Augenblicken" spüren wir das ja auch: Da sagt ein Blick, eine Wendung des Kopfes, eine ergreifende Geste, eine abwehrende Gebärde mehr als tausend Wörter.

Antwort, Auge, Augenblick, Begegnung, Blick, Einstellung, Energie, Frage, Fremd, Gebärde, Gefahr, Geste, Haltung, Information, Kenntnis, Kommentar, Kommunikation, Kopf, Körper, Mitmensch, Mittel, Mißverständnis, Mutter, Primär, Signal, Sinn, Sprache, Umgang, Wendung, abwehren, beherrschen, direkt, entziehen, erfolgreich, ergreifend, frei, inne, körperlich, lautlos, leicht, lernen, manch, miteinander, neben, niemand, nutzen, nämlich, offen, sprachlos, spüren, tausend, unerklärlich, unterdrücken, verbessern, verhalten, verlaufen, vermindern, verwenden, wach, warum, wesentlich, zueinander, öffnen

Text 24: „Körpersprache"-Text – Originaltext und verkürztes, lemmatisiertes Vokabular

Im Vergleich dazu ist ATA2s Top 10-Fehlwortquote sogar noch geringer: ATA2 berechnet im Rahmen des kompletten Versuches lediglich fünf Fehlwörter von 9x10 = 90 „Wortmarkierungen" bzw. Top 10-Platzierungen: Bei Testtext 7 „Körpersprache" (s. Text 24, ATA2-Rangfolge siehe Anhang) die drei Fehlwörter „ergreifende, Wendung, unerklärlich", bei Testtext 8 „Autor" (vgl. Text 21 auf S. 102, ATA2-Rangfolge vgl. Tabelle 36 auf S. 103) das Fehlwort „anstellen" und bei Testtext 9 „Wissenschaft" (siehe Text 23 auf S. 118, ATA2-Rangfolge siehe Anhang) das Wort „größtem".

Damit liegt ATA2s Top 10-Fehlwortquote sogar bei lediglich $5/90 = 5{,}55\%$.

4.6.3.4 Durchschnittliche Wortmarkierungshäufigkeiten

Die durchschnittlichen Wortmarkierungshäufigkeiten für die ATA2-Top 10 einerseits und gemittelt für die 66 Vtn andererseits sind in Tabelle 56 auf S. 126 dargestellt.

Die Mittelwerte der mittleren Wortmarkierungshäufigkeiten der Vtn variieren zwischen 15,54 und 27,78, mit Mittel 23,13. Anschaulich bedeutet dies: Die von einer Vtn markierten 10 Wörter werden durchschnittlich von 23,13 anderen Vtn ebenfalls als Kernwörter markiert. Diese „Einigkeit mit anderen Vtn" schwankt mit einer Standardabweichung von 4,22.

Im Gegensatz dazu werden die von ATA2 berechneten zehn Kernwörter von durchschnittlich sogar 23,72 Personen ebenfalls markiert, mit einer nur leicht stärkeren Schwankung von 4,52.

Textnr.	Min	Max	Mittelw.	Std.Abw.	ATA2
1	8,82	34,5	**26,69**	5,33	21,4
2	13,5	36,7	**25,99**	5,22	32,2
3	11,23	20,21	**15,54**	2,5	29,3
4	11,1	36,5	**27,78**	4,72	22
5	18,27	35,08	**27,72**	3,99	19,5
6	5,67	28,45	**17,83**	4,77	25,9
7	10,73	27,25	**19,91**	3,37	24,9
8	12,67	35,73	**23,88**	5,15	21
9	13,3	33,2	**22,8**	4,03	17,3
		Mittelwert:	**23,13**		**23,72**
		Std.Abw.:	**4,22**		**4,52**

Tabelle 56: Durchschnittliche Wortmarkierungshäufigkeiten Vtn / ATA2

Das bedeutet: Die von ATA2 berechneten Top 10-Wortmarkierungen liefern eine große Übereinstimmung mit den Vtn – bezogen auf die Häufigkeit, mit der Vtn dieselben Wörter als Kernwörter markieren, die ATA2 berechnet.

4.6.3.5 2x2-Felder-Tafeln und Chi²-Unabhängigkeitstest

Die 2x2-Felder-Tafeln sind für alle neun Testtexte erstellt worden, für jeden einzelnen Text wurde der Chi²-Unabhängigkeitstest durchgeführt. In Tabelle 57 sind die neun 2x2-Felder-Tafeln mit Chi²-Berechnung für die Top 20 dargestellt.

Text 1	ATA2<=20	ATA2>20	Summe
Real<=20	10	12	22
Real>20	10	44	54
Summe	20	56	76
		Chi²	**5,85**

Text 2	ATA2<=20	ATA2>20	Summe
Real<=20	16	8	24
Real>20	4	30	34
Summe	20	38	58
		Chi²	**18,77**

Text 3	ATA2<=20	ATA2>20	Summe
Real<=20	15	11	26
Real>20	5	44	49
Summe	20	55	75
		Chi²	**19,59**

Text 4	ATA2<=20	ATA2>20	Summe
Real<=20	11	16	27
Real>20	9	26	35
Summe	20	42	62
		Chi²	**1,57**

Text 5	ATA2<=20	ATA2>20	Summe
Real<=20	14	17	31
Real>20	6	30	36
Summe	20	47	67
		Chi²	**6,46**

Text 6	ATA2<=20	ATA2>20	Summe
Real<=20	11	13	24
Real>20	9	34	43
Summe	20	47	67
		Chi²	**4,56**

Text 7	ATA2<=20	ATA2>20	Summe
Real<=20	15	21	36
Real>20	5	30	35
Summe	20	51	71
		Chi²	**6,58**

Text 8	ATA2<=20	ATA2>20	Summe
Real<=20	13	12	25
Real>20	7	43	50
Summe	20	55	75
		Chi²	**12,31**

Text 9	ATA2<=20	ATA2>20	Summe
Real<=20	12	17	29
Real>20	8	35	43
Summe	20	52	72
		Chi²	**4,48**

Tabelle 57: 2x2-Felder-Tabellen mit Chi²-Prüfgrößen für alle Testtexte – Top 20

Zusätzlich dazu wurden auch 2x2-Felder-Tabellen bezogen auf die Top 10 erstellt und dafür die Chi²-Prüfgrößen berechnet. Dies ist in Tabelle 58 dargestellt.

Text 1	ATA2<=10	ATA2>10	Summe
Real<=10	3	8	11
Real>10	7	58	65
Summe	10	66	76
		Chi²	**2,24**

Text 2	ATA2<=10	ATA2>10	Summe
Real<=10	6	5	11
Real>10	4	43	47
Summe	10	48	58
		Chi²	**13,24**

Text 3	ATA2<=10	ATA2>10	Summe
Real<=10	7	3	10
Real>10	3	62	65
Summe	10	65	75
		Chi²	**32,06**

Text 4	ATA2<=10	ATA2>10	Summe
Real<=10	4	7	11
Real>10	6	45	51
Summe	10	52	62
		Chi²	**4,05**

Text 5	ATA2<=10	ATA2>10	Summe
Real<=10	5	8	13
Real>10	5	49	54
Summe	10	57	67
		Chi²	**7,04**

Text 6	ATA2<=10	ATA2>10	Summe
Real<=10	6	4	10
Real>10	4	53	57
Summe	10	57	67
		Chi²	**18,81**

Text 7	ATA2<=10	ATA2>10	Summe
Real<=10	5	7	12
Real>10	5	54	59
Summe	10	61	71
		Chi²	**9,08**

Text 8	ATA2<=10	ATA2>10	Summe
Real<=10	4	7	11
Real>10	6	58	64
Summe	10	65	75
		Chi²	**5,92**

Text 9	ATA2<=10	ATA2>10	Summe
Real<=10	4	11	15
Real>10	6	51	57
Summe	10	62	72
		Chi²	**2,59**

Tabelle 58: 2x2-Felder-Tabellen mit Chi²-Prüfgrößen für alle Testtexte – Top 10

Zusammenfassend finden sich alle Top 20- und Top 10-Chi²-Prüfgrößen in Tabelle 59 auf Seite 127.

Text-Nr.	1	2	3	4	5	6	7	8	9	Ø	Std.Abw.
Chi² (20)	5,85	18,77	19,59	1,57	6,46	4,56	6,58	12,31	4,48	**8,91**	**6,11**
Chi² (10)	2,24	13,24	32,06	4,05	7,04	18,81	9,08	5,92	2,59	**10,56**	**9,12**

Tabelle 59: Alle Top 20- und Top 10-Chi²-Prüfgrößen

Nur in wenigen Ausnahmefällen erscheinen Prüfgrößen <2,71, entsprechend laut Chi²-Umrechnungstabelle in der Literatur einer Wahrscheinlichkeit von >10% für ein Zufallsergebnis. Alle anderen Ergebnisse sind >3,84, entsprechend einer Zufallswahrscheinlichkeit von <5%. Demgegenüber stehen mehrere Ergebnisse >6,64 oder gar >10,83, entsprechend einer Zufallswahrscheinlichkeit von <1% oder sogar <0,1%.

Erwartungsgemäß schwanken die Prüfgrößen innerhalb der Top 10-Bewertung stärker zwischen extrem signifikanten und nicht signifikanten Prüfgrößen. Beschränkt auf die Top 20-Sicht entspricht eine durchschnittliche Prüfgröße von 8,91 laut Chi²-Umrechnungstabelle einer Zufallswahrscheinlichkeit von lediglich 0,29%.

4.6.3.6 Zusammenfassung Korrelationsanalyse

Der Korrelationskoeffizient zwischen der von ATA2 errechneten Rangfolge und der realen Rangfolge ist nicht sinnvoll berechenbar. Augenscheinlich gelingt ATA2 somit nicht die Bestimmung der Reihenfolge innerhalb der Top 20.

Nichtsdestotrotz liegt die Fehlwortquote (d.h. die Anzahl der von ATA2 unter den Top 20-Wörtern ermittelten Wörtern, die sonst kein einziger Vtn als Kernwort markiert hat) sogar im Vergleich mit der Top 10-Fehlwortquote der Vtn erheblich niedriger. Dieser Vorsprung wächst weiter, wenn man die Fehlwörter innerhalb der ATA2-Top 10 betrachtet. Das heißt: Es ist deutlich unwahrscheinlicher, dass ATA2 im Rahmen der Top 10 ein völlig ungeeignetes Kernwort berechnet, als dass einem durchschnittlichen Versuchsteilnehmer derselbe Fehler unterläuft.

Spätestens im Vergleich der durchschnittlichen Wortmarkierungshäufigkeiten aller zehn Wortmarkierungen pro Text unterscheidet sich die durchschnittliche ATA2-Leistung nicht merklich von der durchschnittlichen Leistung der Vtn: Die ATA2-Top 10-Wörter werden durchschnittlich von gleich vielen Vtn markiert wie die zehn markierten Wörter eines durchschnittlichen Vtn von jeweils anderen Vtn.

Letztendlich liefert die 2x2-Felder-Tafel-Analyse mit anschließender Berechnung von Chi²-Prüfgrößen das Ergebnis mit einer durchschnittlichen Irrtumswahrscheinlichkeit von weit unter 1%, dass ATA2s Top 20-Berechnungen signifikant den realen Top 20 entsprechen.

4.7 Zusammenfassung

ATA2 erreicht bei Erweiterung auf die Top 20-Wörter eine ungewichtete Trefferquote, die der Top 10-Trefferquote der Versuchsteilnehmer ebenbürtig ist. Bezüglich der hergeleiteten Gewichteten Qualität reicht ATA2 zu 75-80% an die Vtn heran. Die berechnete Top 20-Reihenfolge ist zwar wenig brauchbar, aber ATA2s Fehlwortquote unterbietet die Vtn, während ATA2s durchschnittliche Markierungshäufigkeit der Top 10-Wörter nahezu identisch ist mit der der Vtn. Mit Hilfe des Chi²-Tests ist mit großer Signifikanz erwiesen, dass es sich bei den ATA2-Top 20-Resultaten nicht um bloße Zufallstreffer handelt.

ATA2 produziert also alles in allem durchaus akzeptable Resultate. Speziell in Hinblick auf die erweitere Top 20-Auswertung, unter Nichtbeachtung der darin eigentlich vorhandenen Reihenfolge, ist ATA2 von durchschnittlichen menschlichen Vtn kaum zu unterscheiden.

5 Fazit und Ausblick

An dieser Stelle wird ein Erklärungsversuch vorgestellt, warum die ATA2-Ergebnisse sich in der beschriebenen Form ergeben. Dabei wird die angewandte Vorgehensweise von ATA2 vertiefend kognitionswissenschaftlich hinterfragt, indem zusätzliche Versuche mit kleinen Variationen vorgestellt und deren Ergebnisse diskutiert werden.

Darauf aufbauend soll anhand eines „worst case scenario“, eines für Versuchsteilnehmer leichten aber für ATA2 schweren Textes, auf grundsätzliche Probleme aufmerksam gemacht werden. Den Abschluss bildet ein Fazit des Experiments in Hinblick auf eine mögliche Anwendung.

5.1 Kognitionswissenschaftliche Betrachtung

Was genau vollbringt ATA2 intern, wenn es simulierte Wortassoziationen berechnet, die gezwungenermaßen innerhalb eines gegebenen Vokabulars bleiben?

Zur Erinnerung: Als einzige Eingabe für das von ATA2 zu erstellende Assoziative Wortnetz dient das Vokabular, das sich aus den Wörtern des Testtextes zusammensetzt. Es wird nach der Lemmatisierung mit Hilfe der Worthäufigkeitsklassenanalyse auf die Wörter der mittleren Häufigkeitsklassen 7 bis 17 reduziert (vgl. „Vokabularreduktion“, Kapitel 2.5.2, S. 55), wobei die Wahl dieser beiden Grenzen vom Versuchsleiter durch visuelle Evaluation (zu deutsch: „durch scharfes Hinsehen“) definiert wurden und ggf. genauer untersucht werden sollten. Aus jedem der verbleibenden Wörter entsteht ein Wortknoten des Künstlichen Neuronalen Netzes. Letzteres berechnet anhand normierter Kookurrenzhäufigkeiten, welche Wörter am stärksten assoziiert werden, wenn sämtliche Wörter als Reiz präsentiert werden – d.h. wenn alle Wortknoten gleichmäßig stimuliert werden und sich der Reiz im Netz einstufig (d.h. ohne multiple Reizpropagierung) ausbreitet (vgl. Berechnungsvorschrift (2), Kapitel 2.5.3, Seite 58).

Bei der Reizpräsentation nur eines Wortes liefert die ATA2-Berechnung die mit dem Reizwort assoziierten Wörter des (ggf. umfangreichen) Testvokabulars in absteigender Reihenfolge der Assoziationsstärke. Die dabei berechnete Aktivierung eines Wortes ist die Summe der von dem Reizwort ausgehenden Assoziationsstärken zu allen anderen Wörtern des Vokabulars – mit nachfolgender Normierung der stärksten Wortassoziation auf 1,0 bzw. 100%.

Wird das ggf. umfangreiches Testvokabular auf ein Teilvokabular reduziert, bleibt diese Deutung unverändert bestehen – jedoch mit dem Zwang, sich bei der Assoziation auf bestimmte ausgewählte Wörter zu beschränken. Kognitionswissenschaftlich betrachtet entspräche dies einem Assoziationsexperiment (vgl. [GAL1880]), bei dem der jeweilige Versuchsteilnehmer nicht frei über seinen gesamten Wortschatz assoziieren dürfte, sondern lediglich eine Liste von möglichen Assoziationen zur Auswahl vorgelegt bekäme. Die assoziierbaren Wörter könnten dabei vom Versuchsleiter beliebig ausgewählt werden. Insbesondere könnte das präsentierte Reizwort selbst in der Liste der assoziierbaren Wörter enthalten sein – oder auch nicht.

Ausgedehnt auf mehrere Reizwörter bzw. sogar auf das komplette, aus einem Testtext resultierende Testvokabular als Stimulus erscheint die kognitionswissenschaftliche Modellierung mit Hilfe von ATA2s Simulation menschlicher Assoziationen durchaus adäquat: Der simulierte Versuchsteilnehmer erhielte einen Text zum Lesen, aber statt freier Wortassoziation bestünde seine Aufgabe darin, sich aus einer Auswahlliste die zehn Wörter auszusuchen, die am besten auf den Text passen. In diesem Fall bestünde die Liste der wähl- bzw. assoziierbaren Wörter exakt aus den Wörtern des Textes.

Werden für die Assoziationsberechnung die häufigen Stoppwörter der Worthäufigkeitsklassen <7 (d.h. Funktionswörter wie Bindewörter, Artikel oder Präpositionen) im Wortnetz belassen, werden eben diese Wörter von ATA2 erwartungsgemäß stark assoziiert (vgl. „Unerwünschter Worthäufigkeits-Bias“, Kapitel 2.5.5 auf S. 62). Sie kookurrieren durch ihre große Korpushäufigkeit einerseits stark untereinander und andererseits mit potenziell wichtigen Wörtern der Worthäufigkeitsklassen 7 bis 17 (s.o.), vor allem mit Substantiven.

In einem gesonderten, hier nicht genauer dokumentierten Zusatzversuch im Vorfeld der Hauptberechnungen konnte gezeigt werden, dass ein ATA2-Wortnetz zu einem Testtext *einschließlich* der Wortknoten der Worthäufigkeitsklassen <7 diese häufigen Wörter auch dann stark assoziiert, wenn nur die Wortknoten der selteneren Worthäufigkeitsklassen $\geqslant 7$ stimuliert werden. Klarer gesagt: Auch wenn die Stoppwörter gar nicht Teil des Stimulus sind, gelangen Sie dennoch unter die stärksten Assoziationen. Dies unterstreicht die Deutung, dass ihre starke Kookurrenz sie durch ihre große Korpushäufigkeit unzulässig überbetont.

Die Korrekturrechnung der Worthäufigkeits-Bias-Korrektur verhilft den selteneren Wörtern auf die höheren Ränge, indem einerseits die Texthäufigkeit der betreffenden Wörter aufmultipliziert und andererseits die Korpushäufigkeit herausdividiert wird. Das heißt, die im Text häufigen Wörter werden gestärkt, die im Korpus häufigen Wörter werden nochmals abgeschwächt.

Das kognitionswissenschaftliche Pendant dazu ist in der Wahrnehmungspsychologie dokumentiert: Ein menschlicher Vtn, der einen Testtext rezipiert, filtert unwichtige Wörter beim Übergang vom Sensorischen Gedächtnis ins Langzeitgedächtnis und kann somit beim späteren Memorieren die Bedeutung sehr gut, den genauen Wortlaut dagegen sehr viel schlechter wiedergeben. Das zugrunde liegende Konzept wird als *Bedeutungsbezogene Wissensrepräsentation* bezeichnet (siehe [AND2001], Kapitel 5). Dies wird durch die Vokabularreduktion mit Hilfe der Worthäufigkeitsklassenanalyse und insbesondere durch die Worthäufigkeits-Bias-Korrektur nachgebildet.

In der Praxis werden durch diesen Vorgang häufig im Text genannte, d.h. potenziell wichtige, jedoch ungebräuchlichere Wörter gestärkt. Gleichzeitig bleiben dabei exakt die wenig sinnhaltigen Wörter auf der Strecke, d.h. gerade die häufigen Wörter der niedrigen Worthäufigkeitsklassen.

Letztere Schwächung der korpushäufigen Wörter ist zulässig, weil der o.a. Zusatzversuch die ungünstige Auswirkung der Korpushäufigkeit klar herausgestellt hat: Korpushäufige Wörter werden unzulässig überbetont, dieser Umstand betrifft die Wörter sämtlicher Worthäufigkeitsklassen. In einem weiteren, hier ebenfalls nicht genauer dokumentierten Zusatzversuch wurde gezeigt, dass auch mit vollständigem Testtext-Vokabular (d.h. ohne Reduktion auf die Worthäufigkeitsklassen $\geqslant 7$) ähnlich gute Ergebnisse erzielt werden können, sofern die beschriebene Worthäufigkeits-Bias-Korrektur angewandt wird. In diesem Fall werden die hoch assoziierten Stoppwörter korrekt auf die hinteren Ränge verwiesen.

Natürlich arbeiten die beiden Berechnungsschritte der Worthäufigkeits-Bias-Korrektur gegensätzlich zueinander. Insbesondere in dem Fall, dass korpushäufige Wörter auch texthäufige und damit potenziell wichtige Wörter sind, werden eben diese Wörter durch ihre Texthäufigkeit verstärkt, jedoch durch ihre Korpushäufigkeit geschwächt. Dieser Fall soll im nachfolgenden Abschnitt am Beispiel überprüft werden.

5.2 Worst Case Scenario

Mit Hilfe eines konstruierten, aber dennoch sinnvollen Textes kann experimentell gezeigt werden, dass das beschriebene Verfahren exakt bei denjenigen Texten zu ungünstigen Ergebnissen führt, deren Kernwörter zu den korpushäufigen Wörtern der Worthäufigkeitsklassen <7 gehören.

Der Autor dieser Arbeit hat einen derartigen Text verfasst, in Hinblick auf den nicht mehr ganz modernen Textkorpus wiederum in alter deutscher Schreibweise. Inhaltlich behandelt der Text in 181 Wörtern die Unterscheidung der beiden unterschiedlich geschriebenen Wörter „Daß“ und „Das“ sowie die Unterscheidung der beiden identisch geschriebenen Verwendungsformen von „Das“ als Artikel und als Bindewort.

```
Das Wort "Das" hat in der deutschen Sprache zwei unterschiedliche Bedeutungen, was besonders
Ausländer, die Deutsch lernen, als schwierig zu nehmende Hürde empfinden. So ist zum einen
"Das" der bestimmte Artikel für Neutrum-Wörter (das Auto, das Fahrrad und so weiter). Zum
anderen ist "das" aber auch ein Relativ-Pronomen ("Das Auto, das mein Mann mir gekauft
hat, ...", "Das Fahrrad, das im Keller steht, ..."), das aber nicht mit dem Bindewort "daß"
verwechselt werden darf.
Besonders die Unterscheidung zwischen "das" und "daß" fällt oftmals schwer, da beide Wörter
häufig durch ein Komma angefügt werden. Ein Beispiel: "Das Auto, das in der Garage steht,
wird selten benutzt. Darunter, daß es kaum gefahren wird, leidet das Reifenprofil."
Eine Daumenregel, die Kinder oft bereits in der Grundschule lernen, hilft dabei, das
Unterscheiden der unterschiedlich geschriebenen "das" und "daß" zu erleichtern: Wenn es sich
um das Relativ-Pronomen "das" handelt, kann es auch genauso gut durch "dieses", "jenes" oder
"welches" ersetzt werden:
"Das Auto, welches ..."
Bei "daß"-Sätzen wird dieser Austausch nie gelingen, was also ein gutes Indiz dafür
darstellt, daß "daß" und nicht "das" erforderlich ist.

Artikel, Ausländer, Austausch, Auto, Bedeutung, Besonders, Darunter, Fahrrad, Garage,
Grundschule, Hürde, Indiz, Keller, Komma, Neutrum, Pronomen, Relativ, Satz, Sprache,
Unterscheidung, Zum, anfügen, benutzen, besonders, bestimmt, darstellen, empfinden,
erforderlich, erleichtern, ersetzen, gelingen, genauso, geschrieben, handeln
```

Text 25: „Das“-Testtext – Originaltext und reduziertes, lemmatisiertes Vokabular

Der Text, im Wortlaut in Text 25 abgedruckt, ist sicherlich hinreichend verständlich formuliert, so dass Vtn „Daß“ und „Das“ als die Kernwörter des Textes ausmachen können sollten. Eine genaue Untersuchung dazu liegt nicht vor, da der Text nicht Teil der Wortmarkierungsstudie war. Daher ist auch keine mehrheitlich bestimmte reale Top 10- / Top 20-Rangliste vorhanden. Aus diesem Grunde können auch keine objektiven Vergleiche hinsichtlich der Resultate von Vtn durchgeführt werden, um das zugehörige ATA2-Ergebnis qualitativ korrekt einschätzen zu können: Das Instrumentarium der hergeleiteten Analysemethoden steht ohne die reale Wortrangfolge samt und sonders nicht zur Verfügung.

Nichtsdestotrotz ist offensichtlich erwünscht, dass ATA2 gleichermaßen die Wörter „Daß“ und „Das“ liefert, zusätzlich die Bezeichnungen „Bindewort“, „Artikel“, „Komma“, „Relativ“ und „Pronomen“. (Das Kompositum „Relativpronomen“ musste in die beiden Substantive aufgespalten werden, da es als zusammengesetztes Wort nicht im (TAZ-)Korpus enthalten war.) ATA2 sollte aber möglichst nicht die Wörter der im Text genannten Beispiele liefern, also insbesondere nicht „Auto“, „Garage“, „Fahrrad“, „Keller“ oder „Reifenprofil“.

Dieser Testtext ist also in dreierlei Hinsicht eine harte Nuss – ein „worst case scenario“, der schlimmste anzunehmende Fall:

1. Die im Ergebnis erwünschten Kernwörter sind extrem häufige Wörter.

2. Der Text enthält erklärende Beispiele, deren Wörter nicht zum Textsinn gehören.

3. Es gibt äußerst seltene Wörter wie „Bindewort“, „Daumenregel“ und „Reifenprofil“, wobei „Bindewort“ zu den wichtigsten Kernwörtern gehört, „Daumenregel“ nur noch bedingt wichtig ist und „Reifenprofil“ keinesfalls berechnet werden sollte.

Hinzuzufügen ist, dass die Wörter „Daß“ und „Das“ sowie das kleingeschriebene „das“ bereits durch die Worthäufigkeitsklassenanalyse und die anschließende Vokabularreduktion herausgefiltert werden. Das reduzierte Testvokabular enthält nur noch 46 Wörter und ist in Text 25 auf Seite 132 enthalten.

Text 26 zeigt die (nicht entlemmatisierten) ATA2-Top 20-Wörter, in aufsteigender Rangfolge 1-20 bzw. absteigender Reihenfolge ihrer Assoziationsstärke.

```
Pronomen, Neutrum, Komma, Garage, Unterscheidung, Auto, Artikel,
Relativ, Fahrrad, geschrieben, Satz, Grundschule, anfügen,
unterscheiden, Darunter, Sprache, Besonders, Keller, Indiz, Austausch
```

Text 26: „Das“-Testtext – ATA2-Top 20 nach vorheriger Vokabularreduktion

Das Resultat ist weder wirklich gut noch wirklich schlecht. Aufgrund der Vokabularreduktion erscheinen die drei verschiedenen „Das“-Wörter natürlich gar nicht. Die erwarteten Kernwörter „Artikel“, „Komma“, „Relativ“ und „Pronomen“ sowie „Sprache“ und auch „Unterscheidung“ wurden gefunden. „Bindewort“ fehlt: Es liegt in der Worthäufigkeitsklasse 22 extrem seltener Wörter und war somit ebenso wenig im Testvokabular wie das Wort „Daumenregel“, Mitglied der Worthäufigkeitsklasse 20.

Ähnliches gilt für das unerwünschte Wort „Reifenprofil“ (Worthäufigkeitsklasse 18), das somit nicht assoziiert werden konnte. Jedoch bleiben die inhaltlich irrelevanten Wörter „Auto“, „Garage“, „Fahrrad“ und „Keller“ in den Top 20. – Eine vollständige ATA2-Wortliste für diesen Testtext findet sich zusammen mit den anderen Resultaten im Anhang dieser Arbeit.

```
Pronomen, Neutrum, Komma, Auto, Garage, Unterscheidung, Reifenprofil,
Das, Fahrrad, Satz, Grundschule, geschrieben, Relativ, Artikel,
anfügen, das, Daumenregel, Sprache, unterscheiden, Darunter
```

Text 27: „Das“-Testtext – ATA2-Top 20 ohne vorherige Vokabularreduktion

Ein zusätzlicher Versuch zeigte, dass exakt die vorherige Vokabularreduktion dazu führt, dass einerseits die wichtigen „Das“-Wörter nicht unter den Top 20 zu finden sind, jedoch andererseits auch dazu, dass das unerwünschte Wort „Reifenprofil“ entfernt wird.

Die Top 20 der ATA2-Berechnung ohne Vokabularreduktion, in aufsteigender Rangfolge 1-20 bzw. absteigender Reihenfolge ihrer Assoziationsstärke, ist in Text 27 dargestellt: Die „Das“-Schreibweise „daß“ liegt nicht weit entfernt auf Rang 22.

5.2.1 Deutung des Worst Case Scenario

Würde der „worst case scenario“-Testtext über das Wort „Das“ menschlichen Vtn in geeigneter Weise präsentiert, gäbe es sicher keinen Zweifel daran, dass die korrekten Kernwörter markiert und die irrelevanten Wörter ignoriert würden: In geschriebener Form auf Papier oder am Bildschirm helfen bereits die Anführungsstriche um die „Das“-Wörter einerseits und die inhaltlich nicht relevanten Beispieltexte andererseits, Wichtiges von Unwichtigem zu unterscheiden.

ATA2 wertet keine derartigen „*Clues*“ aus: Satzzeichen, ob Anführungszeichen oder Punkte, werden nicht beachtet. Satzstellungen sind irrelevant: Bei der Aufstellung des Testvokabulars für die Kookurrenzanalyse erscheint jedes Wort des (lemmatisierten) Testtextes lediglich einmal, noch reduziert um die sehr häufigen und sehr seltenen Wörter der Worthäufigkeitsklassen <7 und >17. Die Worthäufigkeits-Bias-Korrektur verstärkt häufig im Text vorkommende Wörter und schwächt die korpushäufigen Wörter.

Das Problem liegt in der semantischen Mehrdeutigkeit des Wortes „Das“: Anhand der Schreibweise ohne Einbeziehung zusätzlicher Clues kann nicht unterschieden werden, ob das Wort lediglich als unwichtiges Funktionswort verwendet wird oder ob es ein Kernwort darstellt. Der o.a. Zusatzversuch ohne vorherige Vokabularreduktion verdeutlichte, dass das Wort „Das“ nur aufgrund der Multiplikation mit der Text-Worthäufigkeit in die Top 20 gelangt, siehe Text 27 auf Seite 133.

Daraus folgt: ATA2 „erkennt“ keinesfalls die eigentlich zum Textverständnis semantische Unterscheidung, sondern behilft sich lediglich über eine Heuristik, die für die deutsche Sprache offensichtlich gültig ist.

Problematischer ist die Einbeziehung der Beispielwörter „Auto“, „Garage“, „Keller“ und „Fahrrad“ in die höheren Ränge, obwohl sie definitiv nicht zu den Kernwörtern des Textes gehören: Sie erhalten ihr Gewicht scheinbar durch ihre „Geläufigkeit“, d.h. durchschnittliche Häufigkeit im Korpus, bei zusätzlicher starker Kookurrenz mit verschiedensten Wörtern. (Ohne vorherige Vokabularreduktion kommt auch noch das definitiv irrelevante „Reifenprofil“ hinzu, siehe Text 27 auf Seite 133.) Auch hier gibt es nur eine Möglichkeit, diese von den echten Kernwörtern zu unterscheiden: Eine tiefe Einsicht in die semantische Struktur des Testtextes, d.h. die Kenntnis darüber, dass die in Anführungsstrichen enthaltenen Beispielsätze definitiv keine inhaltlich relevanten Wörter enthalten.

5.2.2 Einfluss von Komposita

Ein weiteres Indiz dafür, dass fehlende semantische Zusammenhänge die Beurteilung erschweren, findet sich im Resultat des Testtextes 5 „Computerspiele“: Das Kompositum „[Computerspiele/Computerspiel]“ wurde *absichtlich* nicht in das 727x im Korpus enthaltene Schlagwort „Computerspiel“ (Worthäufigkeitsklasse 13) lemmatisiert, sondern vielmehr in seine Teilwörter „Computer“ (19.419x, Worthäufigkeitsklasse 8) und „Spiel“ (63.243x, Worthäufigkeitsklasse 6).

In Text 28 ist der Text im Volltext mitsamt reduziertem Vokabular aufgeführt, inklusive der „Computer – Spiele“-Zerlegung. Die reale und die ATA2-Top 10 *mit* Zerlegung von „Computerspiele“ in „Computer – Spiele“ ist in Tabelle 60 auf Seite 136 dargestellt.

```
Warum haben wir ein Buch über Computer - Spiele geschrieben? Die einfachste Antwort lautet:
Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber
geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computer -
Spielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft.
Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der
Ausbreitung und Faszination von Spielen gesetzt werden.
Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit
der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar,
was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann
dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computer - Spiele
erhalten hat, nachzuzeichnen und zu erklären.
Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich
deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei
unvermeidlich subjektiv.
Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu
bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt
begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion.
Das muß sich ändern.

Abhandlung, Analyse, Antwort, Ausbreitung, Auseinandersetzung, Auswahl, Beschreibung,
Bewertung, Computer, Diskussion, Entwicklung, Erfahrung, Erziehung, Faszination, Gegen,
Gewalt, Kapitel, Klassiker, Materie, Recht, Rückbesinnung, Versuch, Vielleicht, Weil,
Wirkung, Zumindest, auswählen, bemängeln, darum, darüber, diejenige, einschränken, einzeln,
endlich, erscheinen, erschöpfen, erwachsen, fangen, genau, hauptsächlich, klar, lauten,
nachzuzeichnen, orientieren, pädagogisch, sicherlich, subjektiv, unbedingt, unvermeidlich,
verstehen, verwunderlich, warum, Über, ändern, überhaupt
```

Text 28: „Computerspiele“-Testtext – Originaltext und reduziertes, lemmatisiertes Vokabular

„Computerspiel“ ist selbstverständlich das Kernwort des Textes und wird demnach in seinen Teilwörtern „Spiel“ und „Computer“ von den Vtn auf die Plätze 1 und 2 der realen Rangfolge gewählt. ATA2 findet im Rahmen seiner Top 10 jedoch lediglich „Computer“. „Spiel“ wird aufgrund der hohen Korpushäufigkeit und daraus resultierender Worthäufigkeitsklasse 6 noch vor der ATA2-Berechnung aus dem Testvokabular entfernt (siehe Text 28). Eine derart niedrige Worthäufigkeitsklasse für dieses Wort ist aufgrund des in großen Teilen zeitungsbasierten Korpus und den darin enthaltenen Sportmeldungen durchaus wenig verwunderlich.

Rang	ATA2	Real	Anzahl
1	Abhandlungen	Spiele	62
2	Materie	Computer	61
3	Rückbesinnung	Buch	49
4	Beschreibung	Diskussion	43
5	nachzuzeichnen	Auseinandersetzung	42
6	Erziehung	Erziehung	39
7	subjektiv	Faszination Gewalt	36
8	Computer	Rückbesinnung	26
9	verwunderlich	Klassikern subjektiv pädagogisch	25
10	Bewertung	Bewertung	24

Tabelle 60: „Computerspiele"-Top 10 (mit Zerlegung „Computer – Spiel") – ATA und real

Ein hier nicht genauer dokumentierter Zusatzversuch hat jedoch gezeigt, dass auch das manuelle Belassen des Wortes „Spiel" im Testvokabular das Wort nicht auf niedrigem Rang assoziierte. Grund dafür ist einerseits die entsprechende Abschwächung eines derart häufigen Wortes im Rahmen der Worthäufigkeits-Bias-Korrektur. Andererseits, und dies wiegt weitaus schwerer, fehlt dem wortbasiert arbeitenden ATA2 die Abstraktion, die beiden Wörter zum Gesamtkonzept „*Computer-Spiel*" zusammenzufassen bzw. „zusammenzuassoziieren". Diese Leistung wurde von den Vtn natürlich souverän gelöst.

```
Abhandlung, Analyse, Antwort, Ausbreitung, Auseinandersetzung, Auswahl, Beschreibung,
Bewertung, Computer, Computerspiel, Diskussion, Entwicklung, Erfahrung, Erziehung,
Faszination, Gegen, Gewalt, Kapitel, Klassiker, Materie, Recht, Rückbesinnung, Versuch,
Vielleicht, Weil, Wirkung, Zumindest, auswählen, bemängeln, darum, darüber, diejenige,
einschränken, einzeln, endlich, erscheinen, erschöpfen, erwachsen, fangen, genau,
hauptsächlich, klar, lauten, nachzuzeichnen, orientieren, pädagogisch, sicherlich,
subjektiv, unbedingt, unvermeidlich, verstehen, verwunderlich, warum, Über, ändern,
überhaupt
```

Text 29: „Computerspiele"-Testtext – Reduziertes, lemmatisiertes Vokabular plus „Computerspiel"

Zum Vergleich: Wird im „Computerspiele"-Vokabular vor der ATA2-Berechnung das Lemma „Computerspiel" ergänzt (wie in Text 29 dargestellt, vgl. Original-Vokabular in Text 28 auf Seite 135), ergibt sich die in Text 30 dargestellte Top 20, in aufsteigender Rangfolge 1-20 bzw. absteigender Reihenfolge der Assoziationsstärke .

```
Computerspiel, Abhandlung, Materie, Rückbesinnung, Faszination,
Beschreibung, nachzuzeichnen, Erziehung, Bewertung, subjektiv,
verwunderlich, Analyse, Auswahl, Computer, erschöpfen, Kapitel,
Klassiker, Wirkung, Auseinandersetzung, Ausbreitung
```

Text 30: ATA2-Top 20 „Computerspiele" - mit Kompositum „Computerspiel"

Auch dieses Ergebnis ist nicht direkt mit den realen Top 10 (s. Tabelle 60) bzw. den realen Top 20 (s. Anhang) vergleichbar, ebenso wenig lassen sich Qualitätsmaße errechnen, da die Vtn die beiden Teilwörter „Computer" und „Spiel" markiert haben und das Kompositum „Computerspiel" demnach nicht in der realen Top 20 enthalten ist. Dennoch entspricht es dem erwarteten Ergebnis, dass das seltenere Wort „Computerspiel" auf Rang 1 der ATA2-Rangfolge zu finden ist. „Computer" bringt es aufgrund der hohen Kookurrenz zu „Computerspiel" immerhin noch auf Rang 14.

5.3 Fazit

Trotz der dargestellten prinzipiellen Schwächen einer rein wortbasierten Arbeitsweise ist es erfreulich, dass die Ergebnisse der für diese Arbeit verwendeten Vorgehensweise mit ATA2, d.h. mit

- Vokabularreduktion auf mittlere Worthäufigkeitsklassen,
- darauf errechneten kookurrenzbasierten Assoziationsstärken und
- anschließender Worthäufigkeits-Bias-Korrektur

mit den durchschnittlichen Ergebnissen der Versuchsteilnehmer vergleichbar ist.

Es liegt in der üblichen Struktur von Sachtexten, dass seltenere, jedoch hinreichend stark kookurrierende Wörter in eine Abhandlung aus häufigeren Wörtern eingehüllt werden – eben in einen leicht verständlichen Erklärungstext. Dies gilt ebenfalls für die meisten Texte, die dieser Studie zugrunde lagen, waren es doch fast allesamt Lehrbuch- oder Sachbuchtexte. Diese Texte bewältigt ATA2 überaus zufrieden stellend, wenngleich mit qualitativen Schwankungen, die jedoch mit der „Uneinigkeit" der Versuchsteilnehmer untereinander vergleichbar ist.

Es zeigt sich, dass die eigentlich zum Textverständnis benötigten semantischen Zusammenhänge in den meisten Fällen nicht vollständig aufgelöst zu werden brauchen, sondern dass Heuristiken eine wirksame Möglichkeit darstellen, sie nachzubilden bzw. hinreichend auszuwerten.

In der Praxis scheitert die verwendete Vorgehensweise jedoch an Komposita-Zerlegungen, durch die semantische Zusammenhänge entfernt (Beispiel: „Computerspiel" vs. „Computer" und „Spiel") oder sogar durch Mehrdeutigkeiten (Beispiel: „Spiel – Sport treiben" vs. „Spiel – spielen zum Zeitvertreib") entstellt wurden. Insbesondere dann, wenn ein wichtiges Wort aufgrund einer inhaltlichen Färbung des verwendeten Korpus durch große Worthäufigkeit in eine niedrige Worthäufigkeitsklasse gerät (Beispiel: „Spiel" und Sportmeldungen in Zeitungen), die im Rahmen der Vokabularreduktion ausgefiltert wird, kann das Verfahren die Wichtigkeit des Wortes prinzipiell nicht mehr bestimmen.

Dies ließe sich jedoch ggf. durch einen besser lemmatisierten, noch umfangreicheren und vielfältigeren Textkorpus inklusive Komposita verbessern oder sogar beheben. Überaus interessant wäre die Verwendung semantisch getaggter Eingaben, d.h. einerseits eines (bislang nicht verfügbaren) hinreichend großen getaggten Korpus und andererseits getaggter Testtexte bzw. Testvokabulare. Dazu ist ATA2 derzeit nicht in der

Lage, da sowohl Korpus als auch Vokabular ohne Tagging-Informationen in simplen ASCII-Dateien vorliegen. Eine Umstellung auf getaggte Eingaben würde umfangreiche Re-Implementationen an ATA2 auf veränderten Dateiformaten erfordern.[46]

Weiterhin geraten ATA2 auch Texte zum „Verhängnis“, in denen semantische Konstrukte (wie etwa Beispielsätze) dazu führen, dass stark kookurrierende Wörter mittlerer Worthäufigkeitsklassen in den Text eingebettet werden, die bei wort- und häufigkeitsbasierter Analyse dann auch irrtümlich als Kernwörter ausgemacht werden.

Die Heuristik der Worthäufigkeits-Bias-Korrektur, bei der korpushäufige Wörter geschwächt und texthäufige Wörter gestärkt werden, funktioniert erfreulich gut; allerdings nur so lange, wie korpushäufige Kernwörter nicht durch die Vokabularreduktion entfernt werden, und solange sie tatsächlich auch hinreichend häufig im Testtext vorkommen – was nicht immer gegeben sein muss: Der „Das“-Text (siehe Text 25 auf S. 132) ließe sich auch so umformulieren, dass die unterschiedlichen Schreibweisen mit nur geringerer Text-Häufigkeit vorkämen. In dem Fall könnte ATA2 sie auch dann nicht als Kernwörter ausmachen, wenn sie durch eine weniger radikale Vokabularreduktion noch im Testvokabular des Testtextes verblieben wären.

Jedoch lässt sich die Klassifizierung, welche korpushäufigen und zugleich textseltenen Wörter trotzdem Kernwörter darstellen, keinesfalls trivial wortbasiert berechnen, da hierzu semantische Clues wie Satzstellungen oder Satzzeichen zwingend erforderlich sind. Ebenso wenig werden aufgespaltene Komposita, deren Bestandteile mehrdeutig sind und demnach unterschiedliche Kookurrenzen bilden, korrekt zusammen assoziiert, um eine hohe Bewertung hervorzurufen. Auch lassen sich Wörter aus eingeschobenen Beispiel-Sätzen ohne tiefere Einsicht in die Satzstruktur nicht als inhaltlich irrelevant einstufen.

5.3.1 Ausblick

Es wurde gezeigt, dass ein computergestütztes Autoabstracting-Verfahren, das wie ATA2 rein wortbasiert und auf ungetaggten d.h. immanent mehrdeutigen Eingaben arbeitet, im Vergleich mit durchschnittlichen Versuchsteilnehmern ohne Bewertung der gefundenen Rangfolge durchaus verwertbare Ergebnisse liefern kann.

46 Als Lemmatisierer käme Morphy ([LRW1996]) in Betracht, jedoch müsste vorab der gesamte Korpus CPU-intensiv getaggt werden. ATA2 müsste darüber hinaus in die Lage versetzt werden, nicht nur auf bloßen ASCII-Wörtern zu arbeiten, sondern vielmehr Morphys syntaktische Zusatzinformationen des Testvokabulars zu berücksichtigen.

Allerdings wird ein derartiges Verfahren schwerlich jemals mit geschulten Redakteuren konkurrieren können, die durch ihre umfassende Texterfahrung und ihren Weitblick hinsichtlich der Textmehrdeutigkeiten weitaus souveräner Kernwörter in Texten ausmachen können.

5.3.2 Unterstützende Verwendung in Suchmaschinen

Bislang ist ungeklärt, in wieweit die dargestellte Vorgehensweise mit ATA2 mit längeren Sachtexten (beispielsweise mehrere Dutzend Seiten lange Veröffentlichungen) umgehen kann. Es steht zu erwarten, dass dabei durch signifikant größeres Vokabular immanent das Problem der semantischen Mehrdeutigkeiten entsteht, und dass variierende Worthäufigkeiten in den unterschiedlichen Dokumentteilen zu falschen Rangbewertungen führen.

Nimmt man aber als zu untersuchende Testtexte die Inhalte von Webseiten an, deren Texte üblicherweise relativ kurz sind, ist das Verfahren mit ATA2 Erfolg versprechend: Der Vorgang ließe sich automatisieren und in den Prozess der Indexierung von neuen HTML-Seiten in eine Suchmaschine integrieren. Unter Zuhilfenahme eines leistungsfähigeren Lemmatisierungsmoduls und eines vielfältigeren, umfangreicheren Korpus ist ein Ergebnis prognostizierbar, dass hinsichtlich der Auswahl der Kernwörter qualitativ einer durchschnittlichen, nicht gesondert ausgebildeten Person entspricht.

Es wäre sogar denkbar, die Texte der HTML-Seiten selbst als Textkorpus für die Kookurrenzanalyse zu verwenden. Das heißt, die HTML-Seiten dienen nicht nur als Lieferanten für das Testvokabular jeweils eines Testtextes, d.h. einer HTML-Seite, sondern vielmehr dient die Gesamtheit des Textmaterials als Basis für die Kookurrenzanalyse. Es gäbe also keine Vokabeln mehr, die nicht im Korpus enthalten wären.

Basis für diese Überlegung ist, die komplette Kookurrenzmatrix über das komplette Korpus-Vokabular auf einem entsprechend großen Sekundärspeicher (sprich: einer großen Festplatte) zu lagern. Wann immer eine weitere HTML-Seite in die Kookurrenzmatrix eingepflegt werden soll, könnte für ein bislang noch nicht enthaltenes Wort je eine zusätzliche Zeile und Spalte hinzugefügt werden, danach könnte die Kookurrenzzählung wie gehabt durchgeführt werden, allerdings auf der externen Datenstruktur. – Die konsistente Vokabular-Sortierung (z.B. alphabetisch) ließe sich über entsprechend leistungsfähige zweidimensionale Indexstrukturen gewährleisten.

Die jeweils aktuelle, „kleine“ Kookurrenzmatrix zu einem der Testtexte, d.h. zu einer der zu untersuchenden Webseiten, ließe sich dann jeweils durch „Slicing“ gewinnen, d.h. durch Ausschneiden der benötigten, jeweils zu den Wörtern gehörigen Zeilen und Spalten der Gesamt-Kookurrenzmatrix. Die Assoziationsberechnung als solche verliefe dann wiederum mit geringem CPU-Bedarf.

Da der Textkorpus mit neu hinzukommenden Dokumenten wächst, ändern sich die Worthäufigkeitsklassen (vgl. [QUA1998]). Der Vorgang der Vokabularreduktion der Textvokabulare mit Hilfe der Worthäufigkeitsklassenanalyse muss also von Zeit zu Zeit nachgeführt und auf das ggf. veränderte Wort mit maximaler Häufigkeit bezogen werden. Ebenso müssen die heuristischen oberen und unteren Grenzklassen für die Reduktion der Vokabulargröße dabei dynamisch angepasst werden.

Weiterhin verändern sich mit einem erweiterten Korpus die Kookurrenzhäufigkeiten bzw. die Assoziationsstärken zwischen den Wörtern. Es bietet sich also an, die in der Datenbank zu den einzelnen Webseiten gespeicherten Kernwörter turnusmäßig neu zu berechnen. Dies könnte in die regelmäßige Prüfung des Seitenbestandes der Suchmaschine integriert werden.

Mit Hilfe der automatisiert und seitenbasiert berechneten Top 20-Kernwörter ließen sich die Fundstellen eines konventionellen, suchwortbasierten booleschen Retrievals weitaus besser in Hinblick auf die thematische Zugehörigkeit zu den eingegebenen Suchwörtern beurteilen. Dies geht weit über Googles PageRank (vgl. [PBMW1998]) hinaus, in dessen Rangliste außer das bloße Vorhandensein der Suchwörter keinerlei inhaltliche Bezüge zu ihnen und ihrer Relevanz bewertet werden, sondern lediglich die „Link-Prominenz“ der gefundenen Webseiten.

5.3.3 Abschließender Anmerkungen

Insgesamt bildet die in dieser Arbeit durchgeführte Untersuchung eine gute Basis für weiterführende, darauf aufbauende Untersuchungen. Die nächsten zu erreichenden Ziele sollten die Automatisierung und signifikante Beschleunigung der Kookurrenz-Analyse bei stark verbesserte Lemmatisierungsqualität sein.

In einem weitergehenden Test sind interessante Ergebnisse zu erwarten, wenn das Verfahren auf eine größere Anzahl Testtexte angewendet würde, als es für diese Arbeit im Vergleich mit den Resultaten der Wortmarkierungsstudie mit realen Versuchsteilnehmern realisierbar war.

Anhang

8+27+4+1 = 40 Seiten. Im Detail:

- Grund-Auswertung Wortmarkierungsstudie (WWW-Ausdruck) – 8 Seiten
    - Teilnehmerstatistik
    - ATA2-Übersicht Trefferquoten / Gewichtete Qualitäten
    - Absolute Markierungshäufigkeiten für alle neun Testtexte
- Für jeden der neun Testtexte der Wortmarkierungsstudie – 3x9 = 27 Seiten
    - Volltext / Wortliste mit Häufigkeitsklassenanalyse / Lemmatisierter Text
    - Detailauswertung
        - Versuchsteilnehmer: Trefferquote / Gewichtete Qualität / Fehlwortquote / Durchschnittliche Markierungshäufigkeit
        - ATA2-Trefferquoten im Detail / Durchschnittliche Markierungsanzahl
    - Wortauswahl-Rang-Gegenüberstellung
        - ATA2-Rangfolge (komplett) vs. reale Rangfolge (komplett)
        - Reale Rangfolge (komplett) vs. ATA2-Rangfolge (komplett)
- Für „Das“-Text / „Computerspiele“ (mit Kompositum) – 2x2 = 4 Seiten
    - Volltext / Wortliste mit Häufigkeitsklassenanalyse / Lemmatisierter Text
    - ATA2-Rangfolge nach Worthäufigkeits-Bias-Korrektur (komplett)
- CDROM
    - Server-Abzug Webanwendung Wormarkierungsstudie
        - Quelltexte (ohne `da_serversecret.php`)
        - Datenbank-Dump (SQL)
    - PDF-Version dieser gedruckten Arbeit (komplett inkl. Anhang)
    - Corpus Count 2 – TAZ-Korpus Hilfsprogramm (inkl. Quelltext)

Studie "Autoabstracting" - Auswertung

- 121 registrierte Versuchspersonen davon Geschlecht M: 46 / W: 75
- 66 vollständig durchgeführte Studien davon Geschlecht M: 27 / W: 39

Anz. bearbeitete Texte:	9	8	7	6	5	4	3	2	1	0
Anz. Teilnehmer:	66	1	1	3	3	3	4	8	5	27

ATA2 vs. Versuchspersonen (flexionskorrigiert)											
Textnr.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mittelwert**	**Std.-Abw.**
Trefferquote Top-10:	30%	60%	70%	40%	50%	60%	50%	40%	40%	**48.88%**	**11.96%**
Gew. 10-dim. Eukl. Qual.:	0.377	0.538	0.536	0.369	0.16	0.404	0.485	0.183	0.19	**0.36**	**0.141**
Trefferquote Top-20:	50%	80%	75%	55%	70%	55%	75%	65%	60%	**65%**	**10%**
Gew. 20-dim. Eukl. Qual.:	0.379	0.462	0.501	0.394	0.319	0.432	0.392	0.317	0.332	**0.391**	**0.06**

Wort	**Abs. Häufigkeit**
Affen	61
Madagaskar	59
Lemuren	53
Überleben	49
Lebensraum	48
Weiterentwicklungen	33
Entwicklung	26
Kampf	26
Gehirne	24
Zweige	23
Nachfahren	20
ehrgeizig	19
Technologie	18
Primaten	16
Vorfahren	15
starb	13
Widersacher	13
Makis	12
Erde	11
übernahmen	10
aggressive	9
größere	6
Auftreten	5
Feuer	5
Jahrtausenden	5
verschont	5
wir	5
Bäumen	4
entscheidende	4
interessierten	4
kämpften	4
Millionen	4
aus	3
Beton	3
Dinge	3
Kanus	3
Booten	2
buddeln	2
erstaunlicher	2
Familie	2
fertigbrachten	2
Flugzeugen	2
hängen	2
Insel	2
Jahre	2
Sachen	2
vorbeiging	2
Zweig	2
allein	1
alles	1
Asphalt	1
begnügt	1
doch	1
eineinhalb	1
gleichen	1

Text Nr. 1 [Detailauswertung] [Ränge ATA2 vs. Real]:

Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb. Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Haustieren	1
herausfanden	1
Macheten	1
nicht	1
stochern	1
ums	1
wohl	1
Zeit	1

Wort	Abs. Häufigkeit
Haltestelle	61
Fahrscheine	53
Sitzplatz	51
Schulkinder	49
Bus	43
Frau	43
Herr	28
eilig	26
Leute	24
warten	23
Koffer	22
Mann	20
Aussteigen	18
Monatskarten	18
Einkaufstasche	16
dankbar	15
Ausstieg	14
Abfall	13
Schaffner	13
Wagen	12
alte	9
Mädchen	9
Ranzen	8
drängen	7
verreisen	7
Stock	5
Viele	5
alter	4
froh	4
schaukelt	4
Stadt	4
drängt	3
kommt	3
steht	3
Alle	2
Bank	2
lösen	2
Plakat	2
rasch	2
steigt	2
Anschlag	1
auf	1
bremst	1
gesetzt	1
Junge	1
junger	1
junges	1
keinen	1
Säule	1
schlecht	1
steigen	1

Text Nr. 2 [Detailauswertung] [Ränge ATA2 vs. Real]:

Viele Leute stehen an der Haltestelle. Ein junger Herr hat einen großen Koffer bei sich. Er will verreisen. Die Frau mit der Einkaufstasche fährt in die Stadt. Die Schulkinder mit dem Ranzen auf dem Rücken lesen noch den Anschlag an der Plakat - Säule. Ein alter Mann mit einem Stock hat sich einstweilen auf die Bank gesetzt. Alle warten auf den Bus. Jetzt kommt der Wagen. Langsam bremst er. Ein paar Leute steigen aus. Der junge Mann mit dem Koffer hat es besonders eilig. Er drängt sich vor und steigt als erster ein. Die Schulkinder schlüpfen noch rasch in den Wagen hinein. Die Leute lösen ihre Fahrscheine. Die Schulkinder zeigen dem Schaffner ihre Monatskarten. Die alte Frau hat noch keinen Sitzplatz gefunden. Sie kann schlecht stehen, weil der Bus schaukelt. Da steht ein junges Mädchen von seinem Sitzplatz auf. Die alte Frau ist froh und dankbar. Beim Aussteigen stecken die Leute ihre Fahrscheine in den Abfall. Da haben es die Schulkinder eilig. Sie drängen vor zum Ausstieg.

Wort	Abs. Häufigkeit
Wein	62
Ausgewogenheit	47
Genuß	46
Vergnügen	41
Charakter	31
Befriedigung	25
sinnliches	25
intellektuelles	24
Weinliebhaber	22
analysieren	19

Text Nr. 3 [Detailauswertung] [Ränge ATA2 vs. Real]:

Wein befriedigt unseren Genuß ebenso wie unser sachliches Interesse, er ist also ein gleichermaßen sinnliches wie intellektuelles Phänomen. Am meisten gibt er uns, wenn diese beiden Aspekte seiner Persönlichkeit ausgewogen sind, wenn der ganze Charakter eines Weins zu einem bestimmten Essen oder Anlaß perfekt paßt und wenn seine Geschmacks - Nuancen zum Vergleich mit anderen Weinen, anderen Jahrgängen, anderen Anlässen anregen. Es gibt Weintrinker, die sich ausschließlich an der sinnlichen Seite des Weins erfreuen und ihn gedankenlos hinunter stürzen, bis die Flasche geleert ist. Andere wiederum diskutieren über seltene

Persönlichkeit	19
Einseitigkeit	18
Interesse	17
Geschmacks	16
Phänomen	16
Weintrinker	16
Erlebnis	15
bewußt	14
sinnlichen	14
Sinne	13
gedankenlos	11
Typen	9
diskutieren	8
esoterischen	8
Jahrgängen	8
Nuancen	8
Signale	8
Essen	7
Weintrinkern	7
Anlaß	6
ausgewogen	6
befriedigt	5
Denken	5
genießt	5
paßt	5
sachliches	5
anregen	4
bewerten	4
genießen	4
Vergleich	4
alltäglichem	3
Analyse	3
Flasche	3
wahre	3
Aspekte	2
Beide	2
guter	2
perfekt	2
zuläßt	2
anderen	1
Anlässen	1
ausschließlich	1
der	1
erfreuen	1
ganze	1
geleert	1
greifen	1
gründlich	1
kurz	1
vergessen	1

Weine und bewerten und analysieren sie so gründlich, daß sie darüber vergessen, die Signale zu genießen, die ihre Sinne ihnen vermitteln. Beide Typen von Weintrinkern greifen in ihrer Einseitigkeit zu kurz: Die wahre Befriedigung, die der Wein zu verschaffen vermag, liegt in der Ausgewogenheit zwischen alltäglichem Genuß und esoterischen Erlebnis. Der echte Weinliebhaber weiß, daß gewisse Weine zu gewissen Zeiten ein reines Vergnügen sind und daß das Denken - von der Analyse ganz zu schweigen - dem Vergnügen im Weg steht. Und daß umgekehrt ein guter Wein viel Vergnügen bereiten kann, aber nur, wenn man dies zuläßt und ihn bewußt genießt.

Wort	Abs. Häufigkeit
Stacheln	65
Igel	64
Feinde	46
Gehölze	41
unangreifbar	39
Kugel	37
Dämmerung	29
Lebensweise	28
einrollen	27
Siedlungen	27
nächtlichen	26
Schutz	25
Aufenthalt	24
Wasser	24
überlisten	18
Tierwelt	17
Pfütze	10
Haare	9
überraschen	9
einmalige	8
heimischen	7
öffnet	7
Erscheinung	6

Text Nr. 4 [Detailauswertung] [Ränge ATA2 vs. Real]:

Mit seinen Stacheln ist der Igel eine einmalige Erscheinung in der heimischen Tierwelt. Der ganze Rücken und die Flanken sind mit festen, spitzen Stacheln versehen. Am Kopf und auf der Unter - Seite schließen sich borstige Haare an. Zum Schutz dieser Körperteile kann sich der Igel so einrollen, daß er eine Kugel bildet, die ringsum Stacheln hat. In dieser Stellung ist der Igel unangreifbar für seine Feinde, von denen vor allem Fuchs, Dachs, Iltis und Hunde zu nennen sind. Es muß ihnen schon gelingen, den Igel in geöffnetem Zustand zu überraschen oder die stachelige Kugel zu überlisten. Dazu rollen sie ihn in eine Pfütze oder sonstwie ins Wasser, wo sich der Igel immer öffnet - und dann ist es um ihn geschehen. Als Aufenthalt liebt der Igel kleine Gehölze mit Unter - Wuchs, Hecken, Wald - Ränder und Gärten. Es zieht ihn oft in die Nähe menschlicher Siedlungen. Daß man ihn nicht häufiger zu sehen bekommt, liegt an seiner nächtlichen Lebensweise. Erst in der Dämmerung verläßt er sein Lager.

Gärten	6
Nähe	6
rollen	6
geöffnetem	5
Fuchs	4
Hunde	4
Iltis	4
menschlicher	4
spitzen	4
borstige	3
Dachs	3
Hecken	3
geschehen	2
Körperteile	2
Rücken	2
stachelige	2
bildet	1
Flanken	1
Lager	1
ringsum	1
verläßt	1
versehen	1
Wald	1

Wort	Abs. Häufigkeit
Computer	61
Spiele	59
Buch	49
Diskussion	43
Auseinandersetzung	42
Erziehung	39
Faszination	36
Gewalt	36
Rückbesinnung	26
Klassikern	25
pädagogischen	25
subjektiv	25
Bewertung	24
Entwicklung	19
Analyse	15
Ausbreitung	15
Auswahl	11
Recht	11
Abhandlungen	9
einschränkende	8
ändern	7
Deutschland	7
erwachsen	7
öffentliche	7
Wirkungen	7
Erfahrung	6
Beschreibung	5
erklären	4
Versuch	4
begonnen	3
geschrieben	3
Spielen	3
endlich	2
Gegen	2
Materie	2
nachzuzeichnen	2
verwunderlich	2
andere	1
diejenigen	1
jetzt	1
Kapitel	1
keine	1
mußte	1
Warum	1
wenig	1
Wirkung	1

Text Nr. 5 [Detailauswertung] [Ränge ATA2 vs. Real]:

Warum haben wir ein Buch über Computer - Spiele geschrieben? Die einfachste Antwort lautet: Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computer - Spielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft. Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der Ausbreitung und Faszination von Spielen gesetzt werden. Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar, was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computer - Spiele erhalten hat, nachzuzeichnen und zu erklären. Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei unvermeidlich subjektiv. Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion. Das muß sich ändern.

Wort	Abs. Häufigkeit
Weihnachtsbaum	55
Marktplatz	43
Beil	40
Tannenbäume	36
Junge	34
kosteten	32
Bindfaden	30
Tannen	28
Hause	27
Schulweg	26
Zweige	22
hübsche	21
Stadt	21
geprüft	19
betrachteten	18
Mauer	18
gefallen	16
gerade	16
Wald	15
Männer	14
tragen	12
bequem	7
Baum	6
Berge	6
Frau	6
Frauen	6
gewachsen	6
Mann	6
Nacht	6
volle	6
wählte	6
abgeladen	5
Handschuhen	5
Leute	5
Tanne	5
Auto	4
Bäume	4
überall	4
nach	3
nicht	3
Augen	2
band	2
Platz	2
dicken	1
drehten	1
fragten	1
Hier	1
ihn	1
Seiten	1
spitzte	1
Tasche	1
traute	1
Über	1
von	1
zurück	1

Text Nr. 6 [Detailauswertung] [Ränge ATA2 vs. Real]:

Als der Junge auf seinem Schulweg am Marktplatz vorbeikam, traute er seinen Augen nicht. Über Nacht war der Tannen - Wald in die Stadt gekommen. An der ganzen langen Mauer entlang standen Tannenbäume, große und kleine, breite und schlanke. Von einem Auto wurden ganze Berge von Tannen abgeladen. Männer mit dicken Handschuhen hoben sie auf und stellten sie an die Mauer, überall dorthin, wo noch Platz war. Hier und da traten Männer und Frauen an die Bäume heran. Sie drehten eine Tanne und betrachteten sie von allen Seiten. Dann stellten sie den Baum wieder an seinen Platz zurück. Er hatte ihnen nicht gefallen. Ein zweiter wurde geprüft, ein dritter und ein vierter. Der Weihnachtsbaum sollte doch auch recht gerade gewachsen sein und volle Zweige haben. Ein Mann trat hinzu. Er hatte ein Beil in der Hand. Aus der Tasche hing ihm ein langer Bindfaden. Die Leute fragten ihn, was die Tannen kosteten. Eine Frau wählte eine hübsche Tanne aus. Der Mann spitzte den Baum mit dem Beil an. Dann band er die Zweige mit dem Bindfaden zusammen, und die Frau konnte ihn bequem nach Hause tragen.

Wort	Abs. Häufigkeit
Körper	57
Sprache	49
Mißverständnissen	42
Kommunikations	39
Fremd	38
Signale	27
Informationen	23
Mitmenschen	22
Mutter	22
Sprachen	20
Energie	19
lernen	19
Zeit	19
Umgang	16
Einstellung	15

Text Nr. 7 [Detailauswertung] [Ränge ATA2 vs. Real]:

Wir verwenden Zeit und Energie, um neben unserer Mutter - Sprache noch weitere Sprachen zu lernen. Körper - Sprache ist mit der Zeit zu einer Fremd - Sprache geworden. Fremd - Sprachen müssen nicht gelernt werden, aber wir kommen weiter, wenn wir sie beherrschen. Wir vermindern die Gefahr von Mißverständnissen. Es ist mir unerklärlich, warum wir nie die Zeit haben, unsere Primär - Sprache, nämlich die Sprache unseres Körpers, zu verbessern. Da sich niemand des Kommunikations - Mittels Körper - Sprache entziehen oder sie unterdrücken kann, ist es von wesentlichem Nutzen, sie zu lernen - gibt sie uns doch wichtige Informationen über die innere Haltung und Einstellung unserer Mitmenschen. Wenn wir offene Sinne und ein waches Auge für die Signale und Kommentare unserer Körper - Sprache haben, können viele Gespräche und Begegnungen leichter und erfolgreicher verlaufen. Die Kenntnis der Körper - Sprache, des lautlosen

Primär	15
Nutzen	13
Sinne	13
Augenblicken	11
erfolgreicher	11
Gespräche	11
Worte	11
Geste	10
sprachlosen	10
freieren	9
Begegnungen	8
direktere	8
Haltung	8
Gebärde	7
innere	7
Verhalten	7
beherrschen	6
Blick	6
Kenntnis	6
Gefahr	5
Mittels	5
verbessern	5
Kommentare	4
Körpers	4
lautlosen	4
spüren	4
leichter	3
Frage	2
miteinander	2
öffnet	2
weitere	2
wesentlichem	2
wichtige	2
Antwort	1
gelernt	1
körperlichen	1
mehr	1
neben	1
sagt	1
vermindern	1
verwenden	1
Wege	1
zueinander	1

Frage- und Antwort- Spiels in unserem körperlichen Verhalten, öffnet direktere Wege zueinander und einen freieren Umgang miteinander. In manchen sprachlosen "Augenblicken" spüren wir das ja auch: Da sagt ein Blick, eine Wendung des Kopfes, eine ergreifende Geste, eine abwehrende Gebärde mehr als tausend Worte.

Wort	Abs. Häufigkeit
Autor	62
Phantasie	52
Wissen	49
Sprache	46
Buches	38
Ideen	35
Urheber	34
Konzentration	31
Nachforschungen	29
Schriftsteller	29
Erfahrungen	24
Werkzeug	22
unterschiedlich	20
Ausdauer	19
Geschichte	19
Text	18
Sachbuches	13
Werdegang	11
verständlich	9
Romane	8
Dichter	7
schreiben	7
spannend	7
Lesern	6
verändert	6
arbeiten	5
Personen	5
verbessert	5
schwierige	4

Text Nr. 8 [Detailauswertung] [Ränge ATA2 vs. Real]:

Der Autor steht beim Werdegang eines Buches an erster Stelle, er schreibt den Text. Das lateinische Wort Autor bedeutet Urheber, ganz gleich ob Mann oder Frau. Allerdings wird der Begriff im allgemeinen nur auf Schriftsteller und Dichter angewandt. Ihr Werkzeug ist die Sprache. Mit der müssen sie gut umgehen können. Doch das allein genügt noch nicht! Will ein Schriftsteller eine Geschichte erfinden und das Geschehen spannend darstellen, muß er Ideen haben und viel Phantasie entwickeln. Beim Verfassen eines Sachbuches muß er über den betreffenden Bereich viel Wissen sammeln, das heißt er muß persönliche Erfahrungen machen und umfangreiche Nachforschungen anstellen. Natürlich hängt das Schreiben auch von der Personen - Gruppe ab, an die sich ein Autor wendet. Es ist schon ein Unterschied, ob er sich bei ganz jungen Lesern oder bei Erwachsenen verständlich machen möchte. Autoren arbeiten sehr unterschiedlich. Manche können diese schwierige Tätigkeit, die so viel Konzentration und Ausdauer verlangt, nur kurze Zeit am Tag ausüben. Sie bringen vielleicht nur ein paar Sätze zustande. Andere schreiben die ganze Nacht hindurch und vollenden Geschichten und Romane in einem Zug. Oft wird dann immer und immer wieder verändert und verbessert.

Verfassen	4
Autoren	3
Gruppe	3
Zeit	3
geschehen	2
schreibt	2
Stelle	2
Tätigkeit	2
Unterschied	2
Wort	2
Begriff	1
ein	1
erfinden	1
erster	1
Erwachsenen	1
kurze	1
lateinische	1
Nacht	1
Oft	1
persönliche	1
sammeln	1
Sätze	1
umgehen	1
vollenden	1
wendet	1

Wort	Abs. Häufigkeit
Beobachtung	53
Vermutung	53
Natur	42
Irrelevantes	38
Labor	34
Frage	33
Wahrnehmungen	30
Forscher	26
Versuchs	26
Lebenswichtige	25
Überschätzung	25
Wesentlichen	23
Wissenschaftler	23
Untersuchungen	17
Scharfsinn	16
Leiter	15
bestätigt	14
widerlegt	13
Laboratorium	12
Ratio	12
Untersuchung	10
Voraussetzung	10
keineswegs	9
kognitiven	9
Sinnes	9
Reichhaltigkeit	8
methodischer	7
vorbeigeht	7
einsehen	6
Fragen	5
Leistungen	5
Arbeit	4
Berücksichtigung	3
Daten	3
Einzelheiten	3
Kontakt	3
unmöglich	3
ausdenkt	2
menschlichen	2
Nerven	2
Organe	2
rationalen	2
vorher	2
wirklich	2
Alle	1
bester	1
betreffen	1

Text Nr. 9 [Detailauswertung] [Ränge ATA2 vs. Real]:

Die Frage, die der Versuchs - Leiter im Labor an die Natur stellt, hat immer schon zur Voraussetzung, daß er eine Vermutung hegt, die er bestätigt oder widerlegt wissen will. Diese Vermutung aber stammt stets aus einer vorher gehenden Beobachtung, mit anderen Worten, aus jenen nicht - rationalen kognitiven Leistungen unserer Sinnes - Organe und unseres Nerven - Systems, die aus Sinnes - Daten Wahrnehmungen formen. Es bedeutet eine gewaltige Überschätzung der menschlichen Ratio, wenn sich ein Wissenschaftler einbildet, er wisse und kenne alle Fragen, die man an die Natur stellen kann. Wenn ein Forscher sich die Frage ausdenkt, während er, in sein Laboratorium gebannt, seiner Arbeit ohne Kontakt mit der unabsehbaren Reichhaltigkeit der lebenden Natur obliegt, so kann es allzu leicht geschehen, daß seine Frage an dem wirklich Wesentlichen vorbeigeht und nur Irrelevantes zutage fördert. So werden dann Untersuchungen angestellt, die trotz größtem Scharfsinn und trotz bester methodischer Berücksichtigung aller Einzelheiten keineswegs das Lebenswichtige betreffen. Der Forscher aber, der sich ganz eben dieser Untersuchung widmet, kann dies unmöglich einsehen.

einbildet	1
gehenden	1
gewaltige	1
kenne	1
lebenden	1
Systems	1
unabsehbaren	1
wisse	1
Wissen	1

1. Madagaskar – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb.
Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Makis	9	19	Feuer	12654	9	Zeit	197886	5
Lemuren	70	16	herum	13189	9	Und	251784	4
Macheten	134	15	Auftritt	14478	8	dann	263034	4
buddeln	210	15	überall	15696	8	wieder	267352	4
stochern	375	14	Baum	16209	8	wir	278795	4
fertigbringen	438	14	kämpfen	16547	8	gut	314979	4
Madagaskar	475	13	diesmal	16889	8	all	334975	4
Nachfahre	743	13	Insel	17965	8	sagen	349278	4
Primat	755	13	überleben	18115	8	kommen	376645	4
Haustier	934	12	stammen	22436	8	oder	443445	4
Kanu	952	12	Gebiet	25555	8	aber	502793	3
vorbeigehen	1153	12	Ding	25678	8	nur	584819	3
Asphalt	1159	12	fühlen	34133	7	über	598179	3
Widersacher	1368	12	sterben	39123	7	vor	630613	3
Weiterentwicklung	1626	12	gleich	39517	7	wie	683018	3
Lebensraum	1904	11	Während	43453	7	Jahr	696675	3
vorfahren	1913	11	Sache	44492	7	um	700687	3
Affe	1959	11	Entwicklung	44887	7	nach	805815	3
Zweig	2076	11	übernehmen	45716	7	können	821919	3
denselben	2194	11	Kampf	46590	7	aus	892600	3
verschonen	2476	11	schließlich	47003	7	an	1039533	2
Beton	2608	11	Familie	51144	7	sie	1210878	2
begnügen	2703	11	entscheiden	59077	6	Die	1421814	2
verfügt	2824	11	allein	59286	6	dem	1472967	2
herausfinden	2968	11	Vor	63021	6	im	1577564	2
ehrgeizig	3355	11	weiter	66773	6	für	1594594	2
Jahrtausend	3499	11	wohl	72525	6	auf	1608537	2
Gehirn	3688	10	deren	72677	6	sich	1718091	2
Erde	3923	10	kurz	75293	6	nicht	1772767	2
eineinhalb	4417	10	zwar	85684	6	mit	1861458	1
interessiert	4638	10	allerdings	89831	6	das	1920541	1
aggressiv	5761	10	möglich	91384	6	zu	2050372	1
Mal	5978	10	spät	97843	6	von	2215196	1
erstaunlich	7008	10	jedoch	116527	5	haben	2295270	1
Technologie	7140	10	tun	117561	5	den	2697266	1
Dank	7808	9	Wie	124941	5	ein	3437822	1
Boot	7906	9	nehmen	139482	5	in	4086574	0
hauen	10106	9	damit	153631	5	sein	4658755	0
erreicht	10499	9	ab	166310	5	und	5064441	0
größere	10608	9	bleiben	185337	5	die	6782081	0
Hang	11655	9	doch	191390	5	der	7367299	0
Flugzeug	12010	9	Million	196686	5			

Die entscheiden Entwicklung die an Madagaskar vorbeigehen sein das Auftritt der Affe sie stammen zwar von den gleich vorfahren wie die Lemuren ab verfügt jedoch über größere Gehirn und sein aggressiv Widersacher im Kampf um denselben Lebensraum Während die Makis sich damit begnügen haben in den Baum herum zu Hang und sich wohl zu fühlen sein die Affe ehrgeizig und interessiert sich für all möglich vor all für die Zweig Wie sie nach kurz Zeit herausfinden können sie damit Ding tun die sie allein nicht fertigbringen nach Sache buddeln in Sache herum stochern auf Sache herum hauen Die Affe übernehmen die Erde und der Lemuren Zweig der Primat Familie sterben überall aus nur auf Madagaskar nicht das für Million Jahr von Affe verschonen bleiben Vor eineinhalb Jahrtausend kommen die Affe dann aber schließlich doch auf Madagaskar an oder gut sagen kommen deren Nachfahre auf Madagaskar an wir Dank erstaunlich Weiterentwicklung auf dem Gebiet der Zweig Technologie erreicht wir die Insel mit Kanu spät mit Boot und schließlich mit Flugzeug und nehmen den Kampf um den Lebensraum ein weiter Mal auf diesmal allerdings mit Feuer und Macheten mit Haustier Asphalt und Beton Und wieder kämpfen die Lemuren um überleben

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 1:

Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb. Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Top 10-Trefferquote der Versuchspersonen:							
Trefferquote (relativ)	30%	40%	50%	60%	70%	80%	90%
Personen (von 66)	2	4	12	17	12	15	4
Personen (relativ)	3%	6%	18.1%	25.7%	18.1%	22.7%	6%
Erwartungswert:	64.24%						
Gewichtete 10-dim. Euklidische Qualität:							
Min / Max / Mittelw. / Std.Abw.	0.089	0.777	**0.546**	0.153			
Durchschnittlichliche Markierungsanzahl anderer Vpn:							
Min / Max / Mittelw. / Std.Abw.	8.818	34.5	**26.688**	5.325			

Top 10 ATA2 - entlemmatisiert: Lemuren, Madagaskar, Affen, buddeln, Primaten, Kanus, Beton, Asphalt, Haustieren, Vorfahren

3 Treffer ATA2 von realen Top 10: Affen (#1), Madagaskar (#2), Lemuren (#3)

7 Treffer ausserhalb Top 10: Primaten (#13), Vorfahren (#14), Beton (#23), Kanus (#23), buddeln (#24), Asphalt (#25), Haustieren (#25)

Text 1 Trefferquote ATA2 Top 10: 30 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.377
Durchschnittlichliche Markierungsanzahl aller 10: 21.4 Vpn

Top 20 ATA2 - entlemmatisiert: Lemuren, Madagaskar, Affen, buddeln, Primaten, Kanus, Beton, Asphalt, Haustieren, Vorfahren, [Zweige/Zweig], herum, stochern, Erde, Lebensraum, Gehirne, Nachfahren, Macheten, denselben, Jahrtausenden

10 Treffer ATA2 von realen Top 20: Affen (#1), Madagaskar (#2), Lemuren (#3), Lebensraum (#5), Zweige (#8), Gehirne (#9), Nachfahren (#10), Primaten (#13), Vorfahren (#14), Erde (#17)

8 Treffer ausserhalb Top 20: Jahrtausenden (#21), Beton (#23), Kanus (#23), buddeln (#24), Asphalt (#25), Haustieren (#25), Macheten (#25), stochern (#25)

2 Wort(e) von Vpn überhaupt nicht markiert: herum, denselben

Text 1 Trefferquote ATA2 Top 20: 50 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.379
Durchschnittlichliche Markierungsanzahl aller 20: 17.45 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Affen	61	**1**	Affen	61
2	Madagaskar	59	**2**	Madagaskar	59
3	Lemuren	53	**3**	Lemuren	53
4	Überleben	49	**4**	Überleben	49
5	Lebensraum	48	**5**	Lebensraum	48
6	Weiterentwicklungen	33	**6**	Weiterentwicklungen	33
7	Entwicklung	26	**7**	Entwicklung	26
7	Kampf	26	**7**	Kampf	26
8	Zweige	25	**8**	Gehirne	24
9	Gehirne	24	**9**	Zweige	23
10	Nachfahren	20	**10**	Nachfahren	20
11	ehrgeizig	19	**11**	ehrgeizig	19
12	Technologie	18	**12**	Technologie	18
13	Primaten	16	**13**	Primaten	16
14	Vorfahren	15	**14**	Vorfahren	15
15	starb	13	**15**	starb	13
15	Widersacher	13	**15**	Widersacher	13
16	Makis	12	**16**	Makis	12
17	Erde	11	**17**	Erde	11
18	übernahmen	10	**18**	übernahmen	10
19	aggressive	9	**19**	aggressive	9
20	größere	6	**20**	größere	6
Anzahl Fehlworte (Hfgk. 1): 15 entsprechend 22.72% (allein, alles, Asphalt, begnügt, doch, eineinhalb, gleichen, Haustieren, herausfanden, Macheten, nicht, stochern, ums, wohl, Zeit)					
[Vollständige Gegenüberstellung ATA2 vs. Real]					

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 1:

Die entscheidende Entwicklung, die an Madagaskar vorbeiging, war das Auftreten der Affen. Sie stammten zwar von den gleichen Vorfahren wie die Lemuren ab, verfügten jedoch über größere Gehirne und waren aggressive Widersacher im Kampf um denselben Lebensraum. Während die Makis sich damit begnügt hatten, in den Bäumen herum zu hängen und sich wohl zu fühlen, waren die Affen ehrgeizig und interessierten sich für alles mögliche, vor allem für die Zweige. Wie sie nach kurzer Zeit herausfanden, konnten sie damit Dinge tun, die sie allein nicht fertigbrachten - nach Sachen buddeln, in Sachen herum stochern, auf Sachen herum hauen. Die Affen übernahmen die Erde, und der Lemuren - Zweig der Primaten - Familie starb überall aus - nur auf Madagaskar nicht, das für Millionen Jahre von Affen verschont blieb. Vor eineinhalb Jahrtausenden kamen die Affen dann aber schließlich doch auf Madagaskar an oder, besser gesagt, kamen deren Nachfahren auf Madagaskar an - wir. Dank erstaunlicher Weiterentwicklungen auf dem Gebiet der Zweig - Technologie erreichten wir die Insel mit Kanus, später mit Booten und schließlich mit Flugzeugen und nahmen den Kampf um den Lebensraum ein weiteres Mal auf, diesmal allerdings mit Feuer und Macheten, mit Haustieren, Asphalt und Beton. Und wieder kämpften die Lemuren ums Überleben.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	3	Lemuren	53	70	3	16
2	2	Madagaskar	59	475	4	13
3	1	Affen	61	1959	5	11
4	24	buddeln	2	210	1	15
5	13	Primaten	16	755	1	13
6	23	Kanus	3	952	1	12
7	23	Beton	3	2608	1	11
8	25	Asphalt	1	1159	1	12
9	25	Haustieren	1	934	1	13
10	14	Vorfahren	15	1913	1	11
11	8	[Zweige/Zweig]	25	2076	3	11
12	------	herum	-	13189	3	9
13	25	stochern	1	375	1	14
14	17	Erde	11	3923	1	10
15	5	Lebensraum	48	1904	2	11
16	9	Gehirne	24	3688	1	11
17	10	Nachfahren	20	743	1	13
18	25	Macheten	1	134	1	15
19	------	denselben	-	2194	1	11
20	21	Jahrtausenden	5	3499	1	11
21	------	Dank	-	7808	1	9
22	24	vorbeiging	2	1153	1	12
23	22	Bäumen	4	16209	1	8
24	15	Widersacher	13	1368	1	12
25	21	Feuer	5	12654	1	9
26	24	Insel	2	17965	1	8
27	6	Weiterentwicklungen	33	1626	1	12
28	24	Booten	2	7906	1	9
29	24	fertigbrachten	2	438	1	14
30	25	herausfanden	1	2968	1	11
31	------	Mal	-	5978	1	10
32	24	Sachen	2	44492	3	7
33	12	Technologie	18	7140	1	10
34	24	Flugzeugen	2	12010	1	9
35	25	eineinhalb	1	4417	1	10
36	22	interessierten	4	4638	1	10
37	24	hängen	2	11655	1	9
38	4	Überleben	49	18115	1	8
39	21	verschont	5	2476	1	11
40	11	ehrgeizig	19	3355	1	11
41	------	verfügten	-	2824	1	11
42	19	aggressive	9	5761	1	10
43	25	begnügt	1	2703	1	11
44	------	hauen	-	10106	1	9
45	24	erstaunlicher	2	7008	1	10
46	23	Dinge	3	25678	1	8
47	------	Gebiet	-	25555	1	8
48	7	Kampf	26	46590	2	7
49	24	Familie	2	51144	1	7
50	------	überall	-	15696	1	8
51	------	stammten	-	22436	1	8
52	20	größere	6	10608	1	9
53	21	Auftreten	5	14478	1	9
54	------	diesmal	-	16889	1	8
55	7	Entwicklung	26	44887	1	7
56	------	erreichten	-	10499	1	9
57	------	schließlich	-	47003	2	7
58	22	kämpften	4	16547	1	8
59	------	fühlen	-	34133	1	7
60	25	gleichen	1	39517	1	7
61	------	Während	-	43453	1	7
62	15	starb	13	39123	1	7
63	18	übernahmen	10	45716	1	7
------	16	Makis	12			
------	21	wir	5			
------	22	entscheidende	4			
------	22	Millionen	4			
------	23	aus	3			
------	24	Jahre	2			
------	25	allein	1			
------	25	alles	1			
------	25	doch	1			
------	25	nicht	1			
------	25	ums	1			
------	25	wohl	1			
------	25	Zeit	1			

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	3	Affen	61	1959	5	11
2	2	Madagaskar	59	475	4	13
3	1	Lemuren	53	70	3	16
4	38	Überleben	49	18115	1	8
5	15	Lebensraum	48	1904	2	11
6	27	Weiterentwicklungen	33	1626	1	12
7	48	Kampf	26	46590	2	7
7	55	Entwicklung	26	44887	1	7
8	11	[Zweige/Zweig]	25	2076	3	11
9	16	Gehirne	24	3688	1	11
10	17	Nachfahren	20	743	1	13
11	40	ehrgeizig	19	3355	1	11
12	33	Technologie	18	7140	1	10
13	5	Primaten	16	755	1	13
14	10	Vorfahren	15	1913	1	11
15	62	starb	13	39123	1	7
15	24	Widersacher	13	1368	1	12
16	------	Makis	12			
17	14	Erde	11	3923	1	10
18	63	übernahmen	10	45716	1	7
19	42	aggressive	9	5761	1	10
20	52	größere	6	10608	1	9
21	53	Auftreten	5	14478	1	9
21	39	verschont	5	2476	1	11
21	20	Jahrtausenden	5	3499	1	11
21	25	Feuer	5	12654	1	9
21	------	wir	5			
22	36	interessierten	4	4638	1	10
22	23	Bäumen	4	16209	1	8
22	58	kämpften	4	16547	1	8
22	------	entscheidende	4			
22	------	Millionen	4			
23	46	Dinge	3	25678	1	8
23	6	Kanus	3	952	1	12
23	------	aus	3			
23	7	Beton	3	2608	1	11
24	------	Jahre	2			
24	49	Familie	2	51144	1	7
24	45	erstaunlicher	2	7008	1	10
24	37	hängen	2	11655	1	9
24	34	Flugzeugen	2	12010	1	9
24	26	Insel	2	17965	1	8
24	32	Sachen	2	44492	3	7
24	22	vorbeiging	2	1153	1	12
24	29	fertigbrachten	2	438	1	14
24	4	buddeln	2	210	1	15
24	28	Booten	2	7906	1	9
25	------	ums	1			
25	18	Macheten	1	134	1	15
25	------	Zeit	1			
25	9	Haustieren	1	934	1	13
25	60	gleichen	1	39517	1	7
25	------	wohl	1			
25	30	herausfanden	1	2968	1	11
25	35	eineinhalb	1	4417	1	10
25	13	stochern	1	375	1	14
25	43	begnügt	1	2703	1	11
25	------	alles	1			
25	------	nicht	1			
25	------	doch	1			
25	------	allein	1			
25	8	Asphalt	1	1159	1	12
------	44	hauen	-	10106	1	9
------	31	Mal	-	5978	1	10
------	41	verfügten	-	2824	1	11
------	21	Dank	-	7808	1	9
------	19	denselben	-	2194	1	11
------	12	herum	-	13189	3	9
------	47	Gebiet	-	25555	1	8
------	50	überall	-	15696	1	8
------	57	schließlich	-	47003	2	7
------	59	fühlen	-	34133	1	7
------	56	erreichten	-	10499	1	9
------	54	diesmal	-	16889	1	8
------	51	stammten	-	22436	1	8
------	61	Während	-	43453	1	7

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

2. Bushaltestelle – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Viele Leute stehen an der Haltestelle. Ein junger Herr hat einen großen Koffer bei sich. Er will verreisen. Die Frau mit der Einkaufstasche fährt in die Stadt. Die Schulkinder mit dem Ranzen auf dem Rücken lesen noch den Anschlag an der Plakat - Säule. Ein alter Mann mit einem Stock hat sich einstweilen auf die Bank gesetzt. Alle warten auf den Omnibus. Jetzt kommt der Wagen. Langsam bremst er. Ein paar Leute steigen aus. Der junge Mann mit dem Koffer hat es besonders eilig. Er drängt sich vor und steigt als erster ein. Die Schulkinder schlüpfen noch rasch in den Wagen hinein. Die Leute lösen ihre Fahrscheine. Die Schulkinder zeigen dem Schaffner ihre Monatskarten. Die alte Frau hat noch keinen Sitzplatz gefunden. Sie kann schlecht stehen, weil der Omnibus schaukelt. Da steht ein junges Mädchen von seinem Sitzplatz auf. Die alte Frau ist froh und dankbar. Beim Aussteigen stecken die Leute ihre Fahrscheine in den Abfall. Da haben es die Schulkinder eilig. Sie drängen vor zum Ausstieg.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Ranzen	80	16	Herr	56581	7	von	2215196	1
Einkaufstasche	228	14	fahren	58014	6	haben	2295270	1
Monatskarte	312	14	Bank	61031	6	den	2697266	1
Fahrschein	433	14	steigen	64655	6	ein	3437822	1
Schaffner	570	13	Leute	72833	6	in	4086574	0
verreisen	601	13	großen	75436	6	sein	4658755	0
einstweilen	896	13	setzen	100768	6	und	5064441	0
schaukeln	1048	12	jung	121233	5	die	6782081	0
Sitzplatz	1255	12	zeigen	124189	5	der	7367299	0
Langsam	1305	12	Stadt	154004	5			
Schulkind	1329	12	finden	165402	5			
schlüpfen	1854	11	weil	167996	5			
Haltestelle	2017	11	Mann	173980	5			
eilig	2583	11	alt	197691	5			
aussteigen	2812	11	Er	201087	5			
Säule	3025	11	Frau	209331	5			
dankbar	3168	11	Ein	220955	5			
Koffer	3454	11	stehen	253583	4			
bremsen	3933	10	da	274291	4			
Ausstieg	4473	10	ihre	300806	4			
Abfall	5763	10	viel	351696	4			
Plakat	7086	10	erst	362491	4			
Stock	7146	10	kommen	376645	4			
froh	7731	9	kein	466949	3			
hinein	8003	9	wollen	492248	3			
rasch	12876	9	zum	574133	3			
Lese	13185	9	vor	630613	3			
Bus	15514	8	noch	672854	3			
Anschlag	18260	8	bei	685320	3			
drängen	18527	8	Der	760069	3			
warten	20853	8	können	821919	3			
rücken	22307	8	aus	892600	3			
stecken	23011	8	er	893675	3			
Beim	24728	8	an	1039533	2			
wagen	25391	8	es	1069022	2			
lösen	26383	8	als	1080482	2			
Alle	27450	8	sie	1210878	2			
Mädchen	29098	7	Die	1421814	2			
Jetzt	36311	7	dem	1472967	2			
paar	48110	7	auf	1608537	2			
schlecht	49025	7	sich	1718091	2			
besonders	49972	7	mit	1861458	1			

viel Leute stehen an der Haltestelle Ein jung Herr haben ein großen Koffer bei sich Er wollen verreisen Die Frau mit der Einkaufstasche fahren in die Stadt Die Schulkind mit dem Ranzen auf dem rücken Lese noch den Anschlag an der Plakat Säule Ein alt Mann mit ein Stock haben sich einstweilen auf die Bank setzen Alle warten auf den Bus Jetzt kommen der wagen Langsam bremsen er Ein paar Leute steigen aus Der jung Mann mit dem Koffer haben es besonders eilig Er drängen sich vor und steigen als erst ein Die Schulkind schlüpfen noch rasch in den wagen hinein Die Leute lösen ihre Fahrschein Die Schulkind zeigen dem Schaffner ihre Monatskarte Die alt Frau haben noch kein Sitzplatz finden sie können schlecht stehen weil der Bus schaukeln da stehen ein jung Mädchen von sein Sitzplatz auf Die alt Frau sein froh und dankbar Beim aussteigen stecken die Leute ihre Fahrschein in den Abfall da haben es die Schulkind eilig sie drängen vor zum Ausstieg

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 2:

Viele Leute stehen an der Haltestelle. Ein junger Herr hat einen großen Koffer bei sich. Er will verreisen. Die Frau mit der Einkaufstasche fährt in die Stadt. Die Schulkinder mit dem Ranzen auf dem Rücken lesen noch den Anschlag an der Plakat - Säule. Ein alter Mann mit einem Stock hat sich einstweilen auf die Bank gesetzt. Alle warten auf den Bus. Jetzt kommt der Wagen. Langsam bremst er. Ein paar Leute steigen aus. Der junge Mann mit dem Koffer hat es besonders eilig. Er drängt sich vor und steigt als erster ein. Die Schulkinder schlüpfen noch rasch in den Wagen hinein. Die Leute lösen ihre Fahrscheine. Die Schulkinder zeigen dem Schaffner ihre Monatskarten. Die alte Frau hat noch keinen Sitzplatz gefunden. Sie kann schlecht stehen, weil der Bus schaukelt. Da steht ein junges Mädchen von seinem Sitzplatz auf. Die alte Frau ist froh und dankbar. Beim Aussteigen stecken die Leute ihre Fahrscheine in den Abfall. Da haben es die Schulkinder eilig. Sie drängen vor zum Ausstieg.

Top 10-Trefferquote der Versuchspersonen:						
Trefferquote (relativ)	40%	50%	60%	70%	80%	90%
Personen (von 66)	2	12	22	18	11	1
Personen (relativ)	3%	18.1%	33.3%	27.2%	16.6%	1.5%
Erwartungswert:	64.09%					

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.169	0.825	**0.515**	0.151

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	13.5	36.7	**25.991**	5.219

Top 10 ATA2 - entlemmatisiert: Ranzen, Fahrscheine, Schaffner, Schulkinder, Haltestelle, Bus, Monatskarten, schaukelt, Sitzplatz, Koffer

6 Treffer ATA2 von realen Top 10: Haltestelle (#1), Fahrscheine (#2), Sitzplatz (#3), Schulkinder (#4), Bus (#5), Koffer (#10)

4 Treffer ausserhalb Top 10: Monatskarten (#12), Schaffner (#16), Ranzen (#20), schaukelt (#23)

Text 2 Trefferquote ATA2 Top 10: 60 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.538
Durchschnittlichliche Markierungsanzahl aller 10: 32.2 Vpn

Top 20 ATA2 - entlemmatisiert: Ranzen, Fahrscheine, Schaffner, Schulkinder, Haltestelle, Bus, Monatskarten, schaukelt, Sitzplatz, Koffer, Einkaufstasche, Aussteigen, verreisen, Langsam, Ausstieg, eilig, Abfall, dankbar, Stock, Mädchen

16 Treffer ATA2 von realen Top 20: Haltestelle (#1), Fahrscheine (#2), Sitzplatz (#3), Schulkinder (#4), Bus (#5), eilig (#7), Koffer (#10), Aussteigen (#12), Monatskarten (#12), Einkaufstasche (#13), dankbar (#14), Ausstieg (#15), Abfall (#16), Schaffner (#16), Mädchen (#19), Ranzen (#20)

3 Treffer ausserhalb Top 20: verreisen (#21), Stock (#22), schaukelt (#23)

1 Wort(e) von Vpn überhaupt nicht markiert: Langsam

Text 2 Trefferquote ATA2 Top 20: 80 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.462
Durchschnittlichliche Markierungsanzahl aller 20: 22.25 Vpn

	Vpn - Top 20 flexionskorrigiert			Vpn - Top 20 Datenbank	
1	Haltestelle	61	**1**	Haltestelle	61
2	Fahrscheine	53	**2**	Fahrscheine	53
3	Sitzplatz	51	**3**	Sitzplatz	51
4	Schulkinder	49	**4**	Schulkinder	49
5	Bus	43	**5**	Bus	43
5	Frau	43	**5**	Frau	43
6	Herr	28	**6**	Herr	28
7	eilig	26	**7**	eilig	26
8	Leute	24	**8**	Leute	24
9	warten	23	**9**	warten	23
10	Koffer	22	**10**	Koffer	22
11	Mann	20	**11**	Mann	20
12	Aussteigen	18	**12**	Aussteigen	18
12	Monatskarten	18	**12**	Monatskarten	18
13	Einkaufstasche	16	**13**	Einkaufstasche	16
14	dankbar	15	**14**	dankbar	15
15	Ausstieg	14	**15**	Ausstieg	14
16	Abfall	13	**16**	Abfall	13
16	alter	13	**16**	Schaffner	13
16	Schaffner	13	**17**	Wagen	12
17	Wagen	12	**18**	alte	9
18	drängt	10	**18**	Mädchen	9
19	Mädchen	9	**19**	Ranzen	8
20	Ranzen	8	**20**	drängen	7
			20	verreisen	7

Anzahl Fehlworte (Hfgk. 1):
8 entsprechend 12.12%
(Anschlag, auf, bremst, gesetzt, Junge, keinen, schlecht, Säule)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 2:

Viele Leute stehen an der Haltestelle. Ein junger Herr hat einen großen Koffer bei sich. Er will verreisen. Die Frau mit der Einkaufstasche fährt in die Stadt. Die Schulkinder mit dem Ranzen auf dem Rücken lesen noch den Anschlag an der Plakat - Säule. Ein alter Mann mit einem Stock hat sich einstweilen auf die Bank gesetzt. Alle warten auf den Bus. Jetzt kommt der Wagen. Langsam bremst er. Ein paar Leute steigen aus. Der junge Mann mit dem Koffer hat es besonders eilig. Er drängt sich vor und steigt als erster ein. Die Schulkinder schlüpfen noch rasch in den Wagen hinein. Die Leute lösen ihre Fahrscheine. Die Schulkinder zeigen dem Schaffner ihre Monatskarten. Die alte Frau hat noch keinen Sitzplatz gefunden. Sie kann schlecht stehen, weil der Bus schaukelt. Da steht ein junges Mädchen von seinem Sitzplatz auf. Die alte Frau ist froh und dankbar. Beim Aussteigen stecken die Leute ihre Fahrscheine in den Abfall. Da haben es die Schulkinder eilig. Sie drängen vor zum Ausstieg.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	20	Ranzen	8	80	1	16
2	2	Fahrscheine	53	433	2	14
3	16	Schaffner	13	570	1	13
4	4	Schulkinder	49	1329	4	12
5	1	Haltestelle	61	2017	1	11
6	5	Bus	43	15514	2	8
7	12	Monatskarten	18	312	1	14
8	23	schaukelt	4	1048	1	12
9	3	Sitzplatz	51	1255	2	12
10	10	Koffer	22	3454	2	11
11	13	Einkaufstasche	16	228	1	15
12	12	Aussteigen	18	2812	1	11
13	21	verreisen	7	601	1	13
14	------	Langsam	-	1305	1	12
15	15	Ausstieg	14	4473	1	10
16	7	eilig	26	2583	2	11
17	16	Abfall	13	5763	1	10
18	14	dankbar	15	3168	1	11
19	22	Stock	5	7146	1	10
20	19	Mädchen	9	29098	1	8
21	------	einstweilen	-	896	1	13
22	------	hinein	-	8003	1	9
23	------	Beim	-	24728	1	8
24	------	lesen	-	13185	1	9
25	------	Rücken	-	22307	1	8
26	17	Wagen	12	25391	2	8
27	18	[drängt/drängen]	10	18527	2	8
28	25	Alle	2	27450	1	8
29	25	lösen	2	26383	1	8
30	------	stecken	-	23011	1	8
31	26	Säule	1	3025	1	11
32	6	Herr	28	56581	1	7
33	26	bremst	1	3933	1	10
34	25	Plakat	2	7086	1	10
35	------	schlüpfen	-	1854	1	12
36	------	Jetzt	-	36311	1	7
37	23	froh	4	7731	1	9
38	9	warten	23	20853	1	8
39	26	Anschlag	1	18260	1	8
40	------	paar	-	48110	1	7
41	25	rasch	2	12876	1	9
42	------	besonders	-	49972	1	7
43	26	schlecht	1	49025	1	7
------	5	Frau	43			
------	8	Leute	24			
------	11	Mann	20			
------	16	alter	13			
------	22	Viele	5			
------	23	Stadt	4			
------	24	kommt	3			
------	24	steht	3			
------	24	steigen	3			
------	25	Bank	2			
------	25	junger	2			
------	26	auf	1			
------	26	gesetzt	1			
------	26	Junge	1			
------	26	keinen	1			

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	5	Haltestelle	61	2017	1	11
2	2	Fahrscheine	53	433	2	14
3	9	Sitzplatz	51	1255	2	12
4	4	Schulkinder	49	1329	4	12
5	6	Bus	43	15514	2	8
5	------	Frau	43			
6	32	Herr	28	56581	1	7
7	16	eilig	26	2583	2	11
8	------	Leute	24			
9	38	warten	23	20853	1	8
10	10	Koffer	22	3454	2	11
11	------	Mann	20			
12	7	Monatskarten	18	312	1	14
12	12	Aussteigen	18	2812	1	11
13	11	Einkaufstasche	16	228	1	15
14	18	dankbar	15	3168	1	11
15	15	Ausstieg	14	4473	1	10
16	------	alter	13			
16	3	Schaffner	13	570	1	13
16	17	Abfall	13	5763	1	10
17	26	Wagen	12	25391	2	8
18	27	[drängt/drängen]	10	18527	2	8
19	20	Mädchen	9	29098	1	8
20	1	Ranzen	8	80	1	16
21	13	verreisen	7	601	1	13
22	19	Stock	5	7146	1	10
22	------	Viele	5			
23	------	Stadt	4			
23	8	schaukelt	4	1048	1	12
23	37	froh	4	7731	1	9
24	------	steht	3			
24	------	steigen	3			
24	------	kommt	3			
25	------	Bank	2			
25	41	rasch	2	12876	1	9
25	------	junger	2			
25	29	lösen	2	26383	1	8
25	34	Plakat	2	7086	1	10
25	28	Alle	2	27450	1	8
26	31	Säule	1	3025	1	11
26	------	gesetzt	1			
26	------	auf	1			
26	39	Anschlag	1	18260	1	8
26	33	bremst	1	3933	1	10
26	------	keinen	1			
26	------	Junge	1			
26	43	schlecht	1	49025	1	7
------	21	einstweilen	-	896	1	13
------	14	Langsam	-	1305	1	12
------	22	hinein	-	8003	1	9
------	25	Rücken	-	22307	1	8
------	42	besonders	-	49972	1	7
------	35	schlüpfen	-	1854	1	12
------	30	stecken	-	23011	1	8
------	36	Jetzt	-	36311	1	7
------	24	lesen	-	13185	1	9
------	40	paar	-	48110	1	7
------	23	Beim	-	24728	1	8

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

3. Wein – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Wein befriedigt unseren Genuß ebenso wie unser sachliches Interesse, er ist also ein gleichermaßen sinnliches wie intellektuelles Phänomen. Am meisten gibt er uns, wenn diese beiden Aspekte seiner Persönlichkeit ausgewogen sind, wenn der ganze Charakter eines Weins zu einem bestimmten Essen oder Anlaß perfekt paßt und wenn seine Geschmacks - Nuancen zum Vergleich mit anderen Weinen, anderen Jahrgängen, anderen Anlässen anregen. Es gibt Weintrinker, die sich ausschließlich an der sinnlichen Seite des Weins erfreuen und ihn gedankenlos hinunter stürzen, bis die Flasche geleert ist. Andere wiederum diskutieren über seltene Weine und bewerten und analysieren sie so gründlich, daß sie darüber vergessen, die Signale zu genießen, die ihre Sinne ihnen vermitteln.
Beide Typen von Weintrinkern greifen in ihrer Einseitigkeit zu kurz: Die wahre Befriedigung, die der Wein zu verschaffen vermag, liegt in der Ausgewogenheit zwischen alltäglichem Genuß und esoterischen Erlebnis. Der echte Weinliebhaber weiß, daß gewisse Weine zu gewissen Zeiten ein reines Vergnügen sind und daß das Denken - von der Analyse ganz zu schweigen - dem Vergnügen im Weg steht. Und daß umgekehrt ein guter Wein viel Vergnügen bereiten kann, aber nur, wenn man dies zuläßt und ihn bewußt genießt.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Weinliebhaber	33	17	zulassen	10089	9	andere	249754	4
Weintrinker	58	16	Typ	10491	9	Und	251784	4
Einseitigkeit	314	14	vermögen	12095	9	stehen	253583	4
Ausgewogenheit	378	14	bewußt	12812	9	Es	263194	4
gedankenlos	419	14	ausschließlich	13869	9	wenn	298994	4
Anlässen	597	13	Beide	14034	9	ihre	300806	4
esoterisch	700	13	wiederum	15271	8	gut	314979	4
Nuance	785	13	Essen	15391	8	viel	351696	4
Befriedigung	1170	12	vermitteln	15677	8	ihr	374106	4
ausgewogen	1498	12	Anlaß	16338	8	geben	411918	4
sinnlich	2234	11	schweigen	16460	8	man	419503	4
hinunter	2483	11	rein	16670	8	oder	443445	4
genießt	2620	11	echt	16992	8	Seite	459776	4
befriedigen	3374	11	diskutieren	18515	8	bis	460372	4
Jahrgang	3615	10	Sinn	19580	8	aber	502793	3
analysieren	3757	10	passen	20591	8	so	559696	3
gleichermaßen	3925	10	gewisse	21761	8	zum	574133	3
anregen	4037	10	wahr	22371	8	nur	584819	3
erfreuen	4128	10	selten	23350	8	über	598179	3
Genuß	4226	10	bestimmt	23736	8	wie	683018	3
sachlich	4457	10	vergessen	24320	8	Der	760069	3
alltäglich	4802	10	Vergleich	25184	8	können	821919	3
Geschmack	4938	10	greifen	27332	8	dies	827459	3
vergnügen	5129	10	bereit	32330	7	er	893675	3
genießen	5190	10	ebenso	37033	7	daß	936255	2
Erlebnis	5257	10	Interesse	43450	7	an	1039533	2
gründlich	5297	10	darüber	46878	7	sie	1210878	2
Phänomen	5804	10	denken	70807	6	Die	1421814	2
Flasche	5965	10	kurz	75293	6	dem	1472967	2
bewerten	6325	10	also	87320	6	im	1577564	2
Persönlichkeit	6483	10	ihnen	96969	6	des	1607888	2
intellektuell	7092	10	Weg	98020	6	sich	1718091	2
umkehren	7744	9	Am	105405	6	mit	1861458	1
leeren	7959	9	liegen	121992	5	das	1920541	1
verschaffen	8027	9	unser	130011	5	zu	2050372	1
perfekt	8636	9	ihn	142987	5	von	2215196	1
Aspekt	8683	9	beide	145017	5	ein	3437822	1
Sturz	8742	9	wissen	152287	5	in	4086574	0
Wein	8780	9	uns	153162	5	sein	4658755	0
Signal	8989	9	Zeit	197886	5	und	5064441	0
Analyse	9007	9	zwischen	198199	5	die	6782081	0
Charakter	9535	9	ganz	223423	5	der	7367299	0

Wein befriedigen unser Genuß ebenso wie unser sachlich Interesse er sein also ein gleichermaßen sinnlich wie intellektuell Phänomen Am viel geben er uns wenn dies beide Aspekt sein Persönlichkeit ausgewogen sein wenn der ganz Charakter ein Wein zu ein bestimmt Essen oder Anlaß perfekt passen und wenn sein Geschmack Nuance zum Vergleich mit andere Wein andere Jahrgang andere Anlässen anregen Es geben Weintrinker die sich ausschließlich an der sinnlich Seite des Wein erfreuen und ihn gedankenlos hinunter Sturz bis die Flasche leeren sein andere wiederum diskutieren über selten Wein und bewerten und analysieren sie so gründlich daß sie darüber vergessen die Signal zu genießen die ihre Sinn ihnen vermitteln Beide Typ von Weintrinker greifen in ihr Einseitigkeit zu kurz Die wahr Befriedigung die der Wein zu verschaffen vermögen liegen in der Ausgewogenheit zwischen alltäglich Genuß und esoterisch Erlebnis Der echt Weinliebhaber wissen daß gewisse Wein zu gewisse Zeit ein rein vergnügen sein und daß das denken von der Analyse ganz zu schweigen dem vergnügen im Weg stehen Und daß umkehren ein gut Wein viel vergnügen bereit können aber nur wenn man dies zulassen und ihn bewußt genießt

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 3:

Wein befriedigt unseren Genuß ebenso wie unser sachliches Interesse, er ist also ein gleichermaßen sinnliches wie intellektuelles Phänomen. Am meisten gibt er uns, wenn diese beiden Aspekte seiner Persönlichkeit ausgewogen sind, wenn der ganze Charakter eines Weins zu einem bestimmten Essen oder Anlaß perfekt paßt und wenn seine Geschmacks - Nuancen zum Vergleich mit anderen Weinen, anderen Jahrgängen, anderen Anlässen anregen. Es gibt Weintrinker, die sich ausschließlich an der sinnlichen Seite des Weins erfreuen und ihn gedankenlos hinunter stürzen, bis die Flasche geleert ist. Andere wiederum diskutieren über seltene Weine und bewerten und analysieren sie so gründlich, daß sie darüber vergessen, die Signale zu genießen, die ihre Sinne ihnen vermitteln. Beide Typen von Weintrinkern greifen in ihrer Einseitigkeit zu kurz: Die wahre Befriedigung, die der Wein zu verschaffen vermag, liegt in der Ausgewogenheit zwischen alltäglichem Genuß und esoterischen Erlebnis. Der echte Weinliebhaber weiß, daß gewisse Weine zu gewissen Zeiten ein reines Vergnügen sind und daß das Denken - von der Analyse ganz zu schweigen - dem Vergnügen im Weg steht. Und daß umgekehrt ein guter Wein viel Vergnügen bereiten kann, aber nur, wenn man dies zuläßt und ihn bewußt genießt.

Top 10-Trefferquote der Versuchspersonen:						
Trefferquote (relativ)	30%	40%	50%	60%	70%	80%
Personen (von 66)	3	12	21	15	13	2
Personen (relativ)	4.5%	18.1%	31.8%	22.7%	19.6%	3%
Erwartungswert:	54.39%					

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.219	0.767	**0.47**	0.132

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	11.23	20.214	**15.541**	2.498

Top 10 ATA2 - entlemmatisiert: [Weintrinker/Weintrinkern], [Wein/Weins/Weinen/Weine], Einseitigkeit, Genuß, Flasche, Weinliebhaber, [sinnliches/sinnlichen], Ausgewogenheit, Nuancen, Befriedigung

7 Treffer ATA2 von realen Top 10: Wein (#1), Ausgewogenheit (#2), Genuß (#3), sinnliches (#5), Befriedigung (#7), Weintrinker (#9), Weinliebhaber (#10)

3 Treffer ausserhalb Top 10: Einseitigkeit (#12), Nuancen (#20), Flasche (#25)

Text 3 Trefferquote ATA2 Top 10: 70 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.536
Durchschnittlichliche Markierungsanzahl aller 10: 29.3 Vpn

Top 20 ATA2 - entlemmatisiert: [Weintrinker/Weintrinkern], [Wein/Weins/Weinen/Weine], Einseitigkeit, Genuß, Flasche, Weinliebhaber, [sinnliches/sinnlichen], Ausgewogenheit, Nuancen, Befriedigung, ausgewogen, Anlässen, erfreuen, esoterischen, Geschmacks, gedankenlos, Vergnügen, Erlebnis, genießen, Jahrgängen

15 Treffer ATA2 von realen Top 20: Wein (#1), Ausgewogenheit (#2), Genuß (#3), Vergnügen (#4), sinnliches (#5), Befriedigung (#7), Weintrinker (#9), Weinliebhaber (#10), Einseitigkeit (#12), Geschmacks (#14), Erlebnis (#15), gedankenlos (#18), esoterischen (#20), Jahrgängen (#20), Nuancen (#20)

5 Treffer ausserhalb Top 20: ausgewogen (#22), genießen (#24), Flasche (#25), Anlässen (#27), erfreuen (#27)

Text 3 Trefferquote ATA2 Top 20: 75 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.501
Durchschnittlichliche Markierungsanzahl aller 20: 20.2 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Wein	62	**1**	Wein	62
2	Ausgewogenheit	47	**2**	Ausgewogenheit	47
3	Genuß	46	**3**	Genuß	46
4	Vergnügen	41	**4**	Vergnügen	41
5	sinnliches	39	**5**	Charakter	31
6	Charakter	31	**6**	Befriedigung	25
7	Befriedigung	25	**6**	sinnliches	25
8	intellektuelles	24	**7**	intellektuelles	24
9	Weintrinker	23	**8**	Weinliebhaber	22
10	Weinliebhaber	22	**9**	analysieren	19
11	analysieren	19	**9**	Persönlichkeit	19
11	Persönlichkeit	19	**10**	Einseitigkeit	18
12	Einseitigkeit	18	**11**	Interesse	17
13	Interesse	17	**12**	Geschmacks	16
14	Geschmacks	16	**12**	Phänomen	16
14	Phänomen	16	**12**	Weintrinker	16
15	Erlebnis	15	**13**	Erlebnis	15
16	bewußt	14	**14**	bewußt	14
17	Sinne	13	**14**	sinnlichen	14
18	gedankenlos	11	**15**	Sinne	13
19	Typen	9	**16**	gedankenlos	11
20	diskutieren	8	**17**	Typen	9
20	esoterischen	8	**18**	diskutieren	8
20	Jahrgängen	8	**18**	esoterischen	8
20	Nuancen	8	**18**	Jahrgängen	8
20	Signale	8	**18**	Nuancen	8
			18	Signale	8
			19	Essen	7
			19	Weintrinkern	7
			20	Anlaß	6
			20	ausgewogen	6

Anzahl Fehlworte (Hfgk. 1):
11 entsprechend 16.66%
(anderen, Anlässen, ausschließlich, der, erfreuen, ganze, geleert, greifen, gründlich, kurz, vergessen)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 3:

Wein befriedigt unseren Genuß ebenso wie unser sachliches Interesse, er ist also ein gleichermaßen sinnliches wie intellektuelles Phänomen. Am meisten gibt er uns, wenn diese beiden Aspekte seiner Persönlichkeit ausgewogen sind, wenn der ganze Charakter eines Weins zu einem bestimmten Essen oder Anlaß perfekt paßt und wenn seine Geschmacks - Nuancen zum Vergleich mit anderen Weinen, anderen Jahrgängen, anderen Anlässen anregen. Es gibt Weintrinker, die sich ausschließlich an der sinnlichen Seite des Weins erfreuen und ihn gedankenlos hinunter stürzen, bis die Flasche geleert ist. Andere wiederum diskutieren über seltene Weine und bewerten und analysieren sie so gründlich, daß sie darüber vergessen, die Signale zu genießen, die ihre Sinne ihnen vermitteln. Beide Typen von Weintrinkern greifen in ihrer Einseitigkeit zu kurz: Die wahre Befriedigung, die der Wein zu verschaffen vermag, liegt in der Ausgewogenheit zwischen alltäglichem Genuß und esoterischen Erlebnis. Der echte Weinliebhaber weiß, daß gewisse Weine zu gewissen Zeiten ein reines Vergnügen sind und daß das Denken - von der Analyse ganz zu schweigen - dem Vergnügen im Weg steht. Und daß umgekehrt ein guter Wein viel Vergnügen bereiten kann, aber nur, wenn man dies zuläßt und ihn bewußt genießt.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	9	[Weintrinker/Weintrinkern]	23	58	2	17
2	1	[Wein/Weins/Weinen/Weine]	62	8780	8	9
3	12	Einseitigkeit	18	314	1	14
4	3	Genuß	46	4226	2	10
5	25	Flasche	3	5965	1	10
6	10	Weinliebhaber	22	33	1	17
7	5	[sinnliches/sinnlichen]	39	2234	2	11
8	2	Ausgewogenheit	47	378	1	14
9	20	Nuancen	8	785	1	13
10	7	Befriedigung	25	1170	1	12
11	22	ausgewogen	6	1498	1	12
12	27	Anlässen	1	597	1	13
13	27	erfreuen	1	4128	1	10
14	20	esoterischen	8	700	1	13
15	14	Geschmacks	16	4938	1	10
16	18	gedankenlos	11	419	1	14
17	4	Vergnügen	41	5129	3	10
18	15	Erlebnis	15	5257	1	10
19	24	genießen	4	5190	1	10
20	20	Jahrgängen	8	3615	1	11
21	26	Aspekte	2	8683	1	9
22	6	Charakter	31	9535	1	9
23	24	anregen	4	4037	1	10
24	14	Phänomen	16	5804	1	10
25	23	befriedigt	5	3374	1	11
26	------	[gewisse/gewissen]	-	21761	2	8
27	25	Analyse	3	9007	1	9
28	------	gleichermaßen	-	3925	1	10
29	11	Persönlichkeit	19	6483	1	10
30	23	genießt	5	2620	1	11
31	11	analysieren	19	3757	1	10
32	26	Beide	2	14034	2	9
33	------	echte	-	16992	1	8
34	8	intellektuelles	24	7092	1	10
35	23	sachliches	5	4457	1	10
36	24	Vergleich	4	25184	1	8
37	21	Essen	7	15391	1	8
38	17	Sinne	13	19580	1	8
39	25	alltäglichem	3	4802	1	10
40	------	verschaffen	-	8027	1	9
41	19	Typen	9	10491	1	9
42	24	bewerten	4	6325	1	10
43	20	Signale	8	8989	1	9
44	------	hinunter	-	2483	1	11
45	------	wiederum	-	15271	1	8
46	27	gründlich	1	5297	1	10
47	------	vermitteln	-	15677	1	8
48	------	bestimmten	-	23736	1	8
49	16	bewußt	14	12812	1	9
50	26	perfekt	2	8636	1	9
51	23	paßt	5	20591	1	8
52	13	Interesse	17	43450	1	7
53	------	ebenso	-	37033	1	7
54	22	Anlaß	6	16338	1	8
55	27	geleert	1	7959	1	9
56	------	umgekehrt	-	7744	1	9
57	------	stürzen	-	8742	1	9
58	------	reines	-	16670	1	8
59	27	ausschließlich	1	13869	1	9
60	27	greifen	1	27332	1	8
61	------	darüber	-	46878	1	7
62	------	vermag	-	12095	1	9
63	------	seltene	-	23350	1	8
64	25	wahre	3	22371	1	8
65	20	diskutieren	8	18515	1	8
66	26	zuläßt	2	10089	1	9
67	------	schweigen	-	16460	1	8
68	27	vergessen	1	24320	1	8
69	------	bereiten	-	32330	1	7
------	23	Denken	5			
------	26	guter	2			
------	27	anderen	1			
------	27	der	1			
------	27	ganze	1			
------	27	kurz	1			

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	2	[Wein/Weins/Weinen/Weine]	62	8780	8	9
2	8	Ausgewogenheit	47	378	1	14
3	4	Genuß	46	4226	2	10
4	17	Vergnügen	41	5129	3	10
5	7	[sinnliches/sinnlichen]	39	2234	2	11
6	22	Charakter	31	9535	1	9
7	10	Befriedigung	25	1170	1	12
8	34	intellektuelles	24	7092	1	10
9	1	[Weintrinker/Weintrinkern]	23	58	2	17
10	6	Weinliebhaber	22	33	1	17
11	29	Persönlichkeit	19	6483	1	10
11	31	analysieren	19	3757	1	10
12	3	Einseitigkeit	18	314	1	14
13	52	Interesse	17	43450	1	7
14	24	Phänomen	16	5804	1	10
14	15	Geschmacks	16	4938	1	10
15	18	Erlebnis	15	5257	1	10
16	49	bewußt	14	12812	1	9
17	38	Sinne	13	19580	1	8
18	16	gedankenlos	11	419	1	14
19	41	Typen	9	10491	1	9
20	43	Signale	8	8989	1	9
20	65	diskutieren	8	18515	1	8
20	14	esoterischen	8	700	1	13
20	9	Nuancen	8	785	1	13
20	20	Jahrgängen	8	3615	1	11
21	37	Essen	7	15391	1	8
22	11	ausgewogen	6	1498	1	12
22	54	Anlaß	6	16338	1	8
23	51	paßt	5	20591	1	8
23	------	Denken	5			
23	35	sachliches	5	4457	1	10
23	30	genießt	5	2620	1	11
23	25	befriedigt	5	3374	1	11
24	19	genießen	4	5190	1	10
24	23	anregen	4	4037	1	10
24	42	bewerten	4	6325	1	10
24	36	Vergleich	4	25184	1	8
25	5	Flasche	3	5965	1	10
25	64	wahre	3	22371	1	8
25	39	alltäglichem	3	4802	1	10
25	27	Analyse	3	9007	1	9
26	50	perfekt	2	8636	1	9
26	21	Aspekte	2	8683	1	9
26	66	zuläßt	2	10089	1	9
26	------	guter	2			
26	32	Beide	2	14034	2	9
27	55	geleert	1	7959	1	9
27	60	greifen	1	27332	1	8
27	59	ausschließlich	1	13869	1	9
27	68	vergessen	1	24320	1	8
27	------	anderen	1			
27	13	erfreuen	1	4128	1	10
27	------	ganze	1			
27	------	kurz	1			
27	------	der	1			
27	12	Anlässen	1	597	1	13
27	46	gründlich	1	5297	1	10
------	63	seltene	-	23350	1	8
------	67	schweigen	-	16460	1	8
------	69	bereiten	-	32330	1	7
------	47	vermitteln	-	15677	1	8
------	40	verschaffen	-	8027	1	9
------	44	hinunter	-	2483	1	11
------	33	echte	-	16992	1	8
------	28	gleichermaßen	-	3925	1	10
------	26	[gewisse/gewissen]	-	21761	2	8
------	45	wiederum	-	15271	1	8
------	48	bestimmten	-	23736	1	8
------	58	reines	-	16670	1	8
------	61	darüber	-	46878	1	7
------	57	stürzen	-	8742	1	9
------	56	umgekehrt	-	7744	1	9
------	53	ebenso	-	37033	1	7
------	62	vermag	-	12095	1	9

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

4. Igel – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Mit seinen Stacheln ist der Igel eine einmalige Erscheinung in der heimischen Tierwelt. Der ganze Rücken und die Flanken sind mit festen, spitzen Stacheln versehen. Am Kopf und auf der Unter - Seite schließen sich borstige Haare an. Zum Schutz dieser Körperteile kann sich der Igel so einrollen, daß er eine Kugel bildet, die ringsum Stacheln hat. In dieser Stellung ist der Igel unangreifbar für seine Feinde, von denen vor allem Fuchs, Dachs, Iltis und Hunde zu nennen sind. Es muß ihnen schon gelingen, den Igel in geöffnetem Zustand zu überraschen oder die stachelige Kugel zu überlisten. Dazu rollen sie ihn in eine Pfütze oder sonstwie ins Wasser, wo sich der Igel immer öffnet - und dann ist es um ihn geschehen.
Als Aufenthalt liebt der Igel kleine Gehölze mit Unter - Wuchs, Hecken, Wald - Ränder und Gärten. Es zieht ihn oft in die Nähe menschlicher Siedlungen. Daß man ihn nicht häufiger zu sehen bekommt, liegt an seiner nächtlichen Lebensweise. Erst in der Dämmerung verläßt er sein Lager.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
borstig	31	17	menschlich	18452	8	Seite	459776	4
Iltis	44	17	Schutz	19302	8	sehen	500329	3
stachelig	103	16	lieben	20407	8	müssen	520472	3
einrollen	133	15	rücken	22307	8	so	559696	3
Gehölz	266	14	Dazu	23365	8	vor	630613	3
unangreifbar	277	14	häufig	24165	8	um	700687	3
überlisten	340	14	öffnen	25283	8	Der	760069	3
sonstwie	364	14	geschehen	27669	8	können	821919	3
Flanke	463	13	bilden	28474	8	dies	827459	3
Pfütze	478	13	Erst	29261	7	er	893675	3
Tierwelt	480	13	Wasser	31282	7	daß	936255	2
Körperteil	618	13	Unter	34448	7	an	1039533	2
ringsum	623	13	verlassen	36739	7	es	1069022	2
Stachel	629	13	wachsen	39190	7	einen	1121613	2
Igel	755	13	rollen	39600	7	sie	1210878	2
Dämmerung	763	13	gelingen	41801	7	für	1594594	2
Hecke	765	13	Kopf	42623	7	auf	1608537	2
geöffnet	933	12	Zum	49252	7	sich	1718091	2
Lebensweise	1233	12	oft	49423	7	nicht	1772767	2
nächtlich	3870	10	fest	52165	7	mit	1861458	1
Garten	4047	10	Daß	59616	6	zu	2050372	1
Kugel	4099	10	bekommen	74446	6	von	2215196	1
Erscheinung	4160	10	nennen	76740	6	haben	2295270	1
Zustand	4789	10	ziehen	79805	6	den	2697266	1
Aufenthalt	5620	10	denen	93694	6	in	4086574	0
Fuchs	5947	10	ihnen	96969	6	sein	4658755	0
heimisch	6392	10	Am	105405	6	und	5064441	0
versehen	6394	10	wo	107113	6	die	6782081	0
Siedlung	6479	10	Als	117369	5	der	7367299	0
einmalig	7273	9	liegen	121992	5			
Feind	7715	9	ihn	142987	5			
Stellung	10855	9	klein	143128	5			
Haar	12189	9	Mit	167025	5			
Dach	12196	9	ganz	223423	5			
Hund	12360	9	immer	247688	4			
schließen	15877	8	dann	263034	4			
Wald	15890	8	Es	263194	4			
überraschen	15990	8	schon	303960	4			
Lager	16058	8	all	334975	4			
Rand	16488	8	In	370333	4			
spitz	18191	8	man	419503	4			
nähen	18399	8	oder	443445	4			

Mit sein Stachel sein der Igel einen einmalig Erscheinung in der heimisch Tierwelt Der ganz rücken und die Flanke sein mit fest spitz Stachel versehen Am Kopf und auf der Unter Seite schließen sich borstig Haar an Zum Schutz dies Körperteil können sich der Igel so einrollen daß er einen Kugel bilden die ringsum Stachel haben In dies Stellung sein der Igel unangreifbar für sein Feind von denen vor all Fuchs Dach Iltis und Hund zu nennen sein Es müssen ihnen schon gelingen den Igel in geöffnet Zustand zu überraschen oder die stachelig Kugel zu überlisten Dazu rollen sie ihn in einen Pfütze oder sonstwie in Wasser wo sich der Igel immer öffnen und dann sein es um ihn geschehen Als Aufenthalt lieben der Igel klein Gehölz mit Unter wachsen Hecke Wald Rand und Garten Es ziehen ihn oft in die nähen menschlich Siedlung Daß man ihn nicht häufig zu sehen bekommen liegen an sein nächtlich Lebensweise Erst in der Dämmerung verlassen er sein Lager

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 4:

Mit seinen Stacheln ist der Igel eine einmalige Erscheinung in der heimischen Tierwelt. Der ganze Rücken und die Flanken sind mit festen, spitzen Stacheln versehen. Am Kopf und auf der Unter - Seite schließen sich borstige Haare an. Zum Schutz dieser Körperteile kann sich der Igel so einrollen, daß er eine Kugel bildet, die ringsum Stacheln hat. In dieser Stellung ist der Igel unangreifbar für seine Feinde, von denen vor allem Fuchs, Dachs, Iltis und Hunde zu nennen sind. Es muß ihnen schon gelingen, den Igel in geöffnetem Zustand zu überraschen oder die stachelige Kugel zu überlisten. Dazu rollen sie ihn in eine Pfütze oder sonstwie ins Wasser, wo sich der Igel immer öffnet - und dann ist es um ihn geschehen. Als Aufenthalt liebt der Igel kleine Gehölze mit Unter - Wuchs, Hecken, Wald - Ränder und Gärten. Es zieht ihn oft in die Nähe menschlicher Siedlungen. Daß man ihn nicht häufiger zu sehen bekommt, liegt an seiner nächtlichen Lebensweise. Erst in der Dämmerung verläßt er sein Lager.

Top 10-Trefferquote der Versuchspersonen:							
Trefferquote (relativ)	30%	40%	50%	60%	70%	80%	90%
Personen (von 66)	1	4	9	19	16	14	3
Personen (relativ)	1.5%	6%	13.6%	28.7%	24.2%	21.2%	4.5%
Erwartungswert:	65%						

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.164	0.848	**0.547**	0.135

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	11.1	36.5	**27.784**	4.717

Top 10 ATA2 - entlemmatisiert: Iltis, Igel, stachelige, Gehölze, einrollen, borstige, Hecken, Stacheln, Pfütze, Flanken

4 Treffer ATA2 von realen Top 10: Stacheln (#1), Igel (#2), Gehölze (#4), einrollen (#9)

6 Treffer ausserhalb Top 10: Pfütze (#15), Iltis (#21), borstige (#22), Hecken (#22), stachelige (#23), Flanken (#24)

Text 4 Trefferquote ATA2 Top 10: 40 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.369
Durchschnittlichliche Markierungsanzahl aller 10: 22 Vpn

Top 20 ATA2 - entlemmatisiert: Iltis, Igel, stachelige, Gehölze, einrollen, borstige, Hecken, Stacheln, Pfütze, Flanken, Dämmerung, unangreifbar, Tierwelt, Fuchs, Körperteile, Gärten, Haare, sonstwie, Kugel, Hunde

11 Treffer ATA2 von realen Top 20: Stacheln (#1), Igel (#2), Gehölze (#4), unangreifbar (#5), Kugel (#6), Dämmerung (#7), einrollen (#9), Tierwelt (#14), Pfütze (#15), Haare (#16), Gärten (#19)

8 Treffer ausserhalb Top 20: Fuchs (#21), Hunde (#21), Iltis (#21), borstige (#22), Hecken (#22), Körperteile (#23), stachelige (#23), Flanken (#24)

1 Wort(e) von Vpn überhaupt nicht markiert: sonstwie

Text 4 Trefferquote ATA2 Top 20: 55 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.394
Durchschnittlichliche Markierungsanzahl aller 20: 18.35 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Stacheln	65	**1**	Stacheln	65
2	Igel	64	**2**	Igel	64
3	Feinde	46	**3**	Feinde	46
4	Gehölze	41	**4**	Gehölze	41
5	unangreifbar	39	**5**	unangreifbar	39
6	Kugel	37	**6**	Kugel	37
7	Dämmerung	29	**7**	Dämmerung	29
8	Lebensweise	28	**8**	Lebensweise	28
9	einrollen	27	**9**	einrollen	27
9	Siedlungen	27	**9**	Siedlungen	27
10	nächtlichen	26	**10**	nächtlichen	26
11	Schutz	25	**11**	Schutz	25
12	Aufenthalt	24	**12**	Aufenthalt	24
12	Wasser	24	**12**	Wasser	24
13	überlisten	18	**13**	überlisten	18
14	Tierwelt	17	**14**	Tierwelt	17
15	Pfütze	10	**15**	Pfütze	10
16	Haare	9	**16**	Haare	9
16	überraschen	9	**16**	überraschen	9
17	einmalige	8	**17**	einmalige	8
18	heimischen	7	**18**	heimischen	7
18	öffnet	7	**18**	öffnet	7
19	Erscheinung	6	**19**	Erscheinung	6
19	Gärten	6	**19**	Gärten	6
19	Nähe	6	**19**	Nähe	6
19	rollen	6	**19**	rollen	6
20	geöffnetem	5	**20**	geöffnetem	5

Anzahl Fehlworte (Hfgk. 1):
7 entsprechend 10.6%
(bildet, Flanken, Lager, ringsum, verläßt, versehen, Wald)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 4:

Mit seinen Stacheln ist der Igel eine einmalige Erscheinung in der heimischen Tierwelt. Der ganze Rücken und die Flanken sind mit festen, spitzen Stacheln versehen. Am Kopf und auf der Unter - Seite schließen sich borstige Haare an. Zum Schutz dieser Körperteile kann sich der Igel so einrollen, daß er eine Kugel bildet, die ringsum Stacheln hat. In dieser Stellung ist der Igel unangreifbar für seine Feinde, von denen vor allem Fuchs, Dachs, Iltis und Hunde zu nennen sind. Es muß ihnen schon gelingen, den Igel in geöffnetem Zustand zu überraschen oder die stachelige Kugel zu überlisten. Dazu rollen sie ihn in eine Pfütze oder sonstwie ins Wasser, wo sich der Igel immer öffnet - und dann ist es um ihn geschehen. Als Aufenthalt liebt der Igel kleine Gehölze mit Unter - Wuchs, Hecken, Wald - Ränder und Gärten. Es zieht ihn oft in die Nähe menschlicher Siedlungen. Daß man ihn nicht häufiger zu sehen bekommt, liegt an seiner nächtlichen Lebensweise. Erst in der Dämmerung verläßt er sein Lager.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	21	Iltis	4	44	1	17
2	2	Igel	64	755	6	13
3	23	stachelige	2	103	1	16
4	4	Gehölze	41	266	1	14
5	9	einrollen	27	133	1	15
6	22	borstige	3	31	1	17
7	22	Hecken	3	765	1	13
8	1	Stacheln	65	629	3	13
9	15	Pfütze	10	478	1	13
10	24	Flanken	1	463	1	14
11	7	Dämmerung	29	763	1	13
12	5	unangreifbar	39	277	1	14
13	14	Tierwelt	17	480	1	13
14	21	Fuchs	4	5947	1	10
15	23	Körperteile	2	618	1	13
16	19	Gärten	6	4047	1	10
17	16	Haare	9	12189	1	9
18	------	sonstwie	-	364	1	14
19	6	Kugel	37	4099	2	10
20	21	Hunde	4	12360	1	9
21	3	Feinde	46	7715	1	9
22	24	ringsum	1	623	1	13
23	22	Dachs	3	12196	1	9
24	20	geöffnetem	5	933	1	13
25	19	Erscheinung	6	4160	1	10
26	24	Wald	1	15890	1	8
27	18	heimischen	7	6392	1	10
28	8	Lebensweise	28	1233	1	12
29	------	Kopf	-	42623	1	7
30	------	Ränder	-	16488	1	8
31	------	Zustand	-	4789	1	10
32	------	Unter	-	34448	2	7
33	23	Rücken	2	22307	1	8
34	10	nächtlichen	26	3870	1	10
35	13	überlisten	18	340	1	14
36	12	Aufenthalt	24	5620	1	10
37	21	menschlicher	4	18452	1	8
38	24	bildet	1	28474	1	8
39	24	versehen	1	6394	1	10
40	12	Wasser	24	31282	1	7
41	------	Erst	-	29261	1	8
42	9	Siedlungen	27	6479	1	10
43	11	Schutz	25	19302	1	8
44	------	liebt	-	20407	1	8
45	------	Stellung	-	10855	1	9
46	19	Nähe	6	18399	1	8
47	21	spitzen	4	18191	1	8
48	24	Lager	1	16058	1	8
49	------	gelingen	-	41801	1	7
50	------	Dazu	-	23365	1	8
51	17	einmalige	8	7273	1	10
52	------	Wuchs	-	39190	1	7
53	------	häufiger	-	24165	1	8
54	23	geschehen	2	27669	1	8
55	------	festen	-	52165	1	7
56	------	schließen	-	15877	1	8
57	19	rollen	6	39600	1	7
58	------	oft	-	49423	1	7
59	24	verläßt	1	36739	1	7
60	18	öffnet	7	25283	1	8
61	------	Zum	-	49252	1	7
62	16	überraschen	9	15990	1	8

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	8	Stacheln	65	629	3	13
2	2	Igel	64	755	6	13
3	21	Feinde	46	7715	1	9
4	4	Gehölze	41	266	1	14
5	12	unangreifbar	39	277	1	14
6	19	Kugel	37	4099	2	10
7	11	Dämmerung	29	763	1	13
8	28	Lebensweise	28	1233	1	12
9	5	einrollen	27	133	1	15
9	42	Siedlungen	27	6479	1	10
10	34	nächtlichen	26	3870	1	10
11	43	Schutz	25	19302	1	8
12	36	Aufenthalt	24	5620	1	10
12	40	Wasser	24	31282	1	7
13	35	überlisten	18	340	1	14
14	13	Tierwelt	17	480	1	13
15	9	Pfütze	10	478	1	13
16	62	überraschen	9	15990	1	8
16	17	Haare	9	12189	1	9
17	51	einmalige	8	7273	1	10
18	27	heimischen	7	6392	1	10
18	60	öffnet	7	25283	1	8
19	46	Nähe	6	18399	1	8
19	25	Erscheinung	6	4160	1	10
19	57	rollen	6	39600	1	7
19	16	Gärten	6	4047	1	10
20	24	geöffnetem	5	933	1	13
21	1	Iltis	4	44	1	17
21	37	menschlicher	4	18452	1	8
21	20	Hunde	4	12360	1	9
21	47	spitzen	4	18191	1	8
21	14	Fuchs	4	5947	1	10
22	7	Hecken	3	765	1	13
22	6	borstige	3	31	1	17
22	23	Dachs	3	12196	1	9
23	3	stachelige	2	103	1	16
23	54	geschehen	2	27669	1	8
23	15	Körperteile	2	618	1	13
23	33	Rücken	2	22307	1	8
24	48	Lager	1	16058	1	8
24	26	Wald	1	15890	1	8
24	10	Flanken	1	463	1	14
24	22	ringsum	1	623	1	13
24	59	verläßt	1	36739	1	7
24	38	bildet	1	28474	1	8
24	39	versehen	1	6394	1	10
------	55	festen	-	52165	1	7
------	53	häufiger	-	24165	1	8
------	58	oft	-	49423	1	7
------	61	Zum	-	49252	1	7
------	56	schließen	-	15877	1	8
------	31	Zustand	-	4789	1	10
------	32	Unter	-	34448	2	7
------	30	Ränder	-	16488	1	8
------	29	Kopf	-	42623	1	7
------	18	sonstwie	-	364	1	14
------	41	Erst	-	29261	1	8
------	44	liebt	-	20407	1	8
------	50	Dazu	-	23365	1	8
------	49	gelingen	-	41801	1	7
------	45	Stellung	-	10855	1	9
------	52	Wuchs	-	39190	1	7

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

5. Computer-Spiele – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Warum haben wir ein Buch über Computer - Spiele geschrieben? Die einfachste Antwort lautet: Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computer - Spielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft. Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der Ausbreitung und Faszination von Spielen gesetzt werden.
Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar, was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computer - Spiele erhalten hat, nachzuzeichnen und zu erklären.
Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei unvermeidlich subjektiv. Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion. Das muß sich ändern.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
nachzuzeichnen	140	15	Versuch	44271	7	ihre	300806	4
Rückbesinnung	461	13	Entwicklung	44887	7	schon	303960	4
Abhandlung	585	13	überhaupt	45241	7	all	334975	4
Ausbreitung	1056	12	verstehen	46128	7	erst	362491	4
verwunderlich	1399	12	darüber	46878	7	durch	373263	4
bemängeln	1637	12	erscheinen	48772	7	kommen	376645	4
Materie	1813	11	warum	50190	7	geben	411918	4
Faszination	2301	11	klar	50991	7	man	419503	4
subjektiv	2543	11	genau	52117	7	kein	466949	3
unvermeidlich	2780	11	lauten	53000	7	aber	502793	3
erschöpfen	2849	11	einige	58153	6	müssen	520472	3
auswählen	3370	11	erklären	60843	6	sollen	555827	3
Beschreibung	3397	11	deshalb	61820	6	so	559696	3
Klassiker	4087	10	Vor	63021	6	Das	576030	3
Bewertung	4443	10	Spiel	63243	6	über	598179	3
Zumindest	4729	10	Buch	63404	6	noch	672854	3
pädagogisch	4754	10	welche	64194	6	wie	683018	3
Erziehung	4872	10	Man	66495	6	am	717511	3
einschränken	5704	10	einfach	70540	6	können	821919	3
hauptsächlich	6503	10	öffentlich	78606	6	dies	827459	3
Kapitel	6629	10	schreiben	84727	6	daß	936255	2
orientieren	6941	10	beginnen	85374	6	an	1039533	2
sicherlich	6949	10	zwar	85684	6	es	1069022	2
Recht	7444	9	erhalten	86780	6	als	1080482	2
diejenige	8428	9	ihrem	87415	6	einen	1121613	2
Analyse	9007	9	dabei	91446	6	sie	1210878	2
Auswahl	9046	9	früh	95413	6	auch	1301212	2
fangen	11442	9	setzen	100768	6	Die	1421814	2
erwachsen	11916	9	ander	105336	6	auf	1608537	2
unbedingt	13260	9	solch	106046	6	sich	1718091	2
Wirkung	14257	9	wo	107113	6	mit	1861458	1
Computer	19419	8	tun	117561	5	das	1920541	1
Vielleicht	20492	8	Aber	143388	5	zu	2050372	1
darum	20503	8	nun	157859	5	von	2215196	1
Gegen	20710	8	eigen	166150	5	haben	2295270	1
Auseinandersetzung	21127	8	weil	167996	5	werden	2520779	1
Weil	25554	8	lang	182707	5	den	2697266	1
Antwort	25629	8	jetzt	187084	5	ein	3437822	1
Gewalt	27786	8	wenig	205941	5	in	4086574	0
Erfahrung	29115	7	Deutschland	211631	5	sein	4658755	0
ändern	32469	7	was	230040	5	und	5064441	0
endlich	33589	7	andere	249754	4	die	6782081	0
einzeln	34283	7	Und	251784	4	der	7367299	0
Über	36621	7	wir	278795	4			
Diskussion	38280	7	gehen	291282	4			

warum haben wir ein Buch über Computer Spiel schreiben Die einfach Antwort lauten Weil es sie geben Und zwar schon so lang daß es verwunderlich sein wie wenig darüber schreiben werden Zumindest in Deutschland wo sich die Auseinandersetzung mit Computer Spiel in pädagogisch Abhandlung über die Wirkung von Gewalt am Computer erschöpfen Gegen dies einschränken Diskussion müssen unbedingt einen Beschreibung und Analyse der Ausbreitung und Faszination von Spiel setzen werden Vielleicht können ein solch Buch erst jetzt erscheinen weil diejenige die Erfahrung mit der Materie haben erst erwachsen werden müssen Die Rückbesinnung fangen an nun werden klar was man früh tun haben und welche Wirkung das auf die eigen Entwicklung haben Man können dies Buch auch als den Versuch verstehen die Erziehung die man durch Computer Spiel erhalten haben nachzuzeichnen und zu erklären Vor all aber sollen die Spiel endlich zu ihrem Recht kommen Die Kapitel orientieren sich deshalb hauptsächlich an einzeln Klassiker Die Auswahl der Spiel sein dabei unvermeidlich subjektiv andere haben sicherlich ander Spiel auswählen und werden an dies Auswahl einige zu bemängeln haben Aber genau darum gehen es einen solch Auseinandersetzung müssen überhaupt beginnen werden Über Spiel und ihre Bewertung geben es noch kein öffentlich Diskussion Das müssen sich ändern

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 5:

Warum haben wir ein Buch über Computer - Spiele geschrieben? Die einfachste Antwort lautet: Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computer - Spielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft. Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der Ausbreitung und Faszination von Spielen gesetzt werden. Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar, was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computer - Spiele erhalten hat, nachzuzeichnen und zu erklären. Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei unvermeidlich subjektiv. Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion. Das muß sich ändern.

Top 10-Trefferquote der Versuchspersonen:							
Trefferquote (relativ)	40%	50%	60%	70%	80%	90%	100%
Personen (von 66)	1	3	11	13	27	10	1
Personen (relativ)	1.5%	4.5%	16.6%	19.6%	40.9%	15.1%	1.5%
Erwartungswert:	74.54%						

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.293	0.769	**0.522**	0.112

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	18.272	35.083	**27.715**	3.993

Top 10 ATA2 - entlemmatisiert: Abhandlungen, Materie, Rückbesinnung, Beschreibung, nachzuzeichnen, Erziehung, subjektiv, Computer, verwunderlich, Bewertung

5 Treffer ATA2 von realen Top 10: Computer (#2), Erziehung (#6), Rückbesinnung (#8), subjektiv (#9), Bewertung (#10)

5 Treffer ausserhalb Top 10: Abhandlungen (#14), Beschreibung (#18), Materie (#21), nachzuzeichnen (#21), verwunderlich (#21)

Text 5 Trefferquote ATA2 Top 10: 50 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.16
Durchschnittlichliche Markierungsanzahl aller 10: 19.5 Vpn

Top 20 ATA2 - entlemmatisiert: Abhandlungen, Materie, Rückbesinnung, Beschreibung, nachzuzeichnen, Erziehung, subjektiv, Computer, verwunderlich, Bewertung, Faszination, Auswahl, erschöpft, Kapitel, Analyse, [Wirkung/Wirkungen], Klassikern, Ausbreitung, bemängeln, Auseinandersetzung

14 Treffer ATA2 von realen Top 20: Computer (#2), Auseinandersetzung (#5), Erziehung (#6), Faszination (#7), Rückbesinnung (#8), Klassikern (#9), subjektiv (#9), Bewertung (#10), Analyse (#12), Ausbreitung (#12), Auswahl (#13), Abhandlungen (#14), Wirkung (#15), Beschreibung (#18)

4 Treffer ausserhalb Top 20: Materie (#21), nachzuzeichnen (#21), verwunderlich (#21), Kapitel (#22)

2 Wort(e) von Vpn überhaupt nicht markiert: erschöpft, bemängeln

Text 5 Trefferquote ATA2 Top 20: 70 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.319
Durchschnittlichliche Markierungsanzahl aller 20: 17.4 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Spiele	62	**1**	Computer	61
2	Computer	61	**2**	Spiele	59
3	Buch	49	**3**	Buch	49
4	Diskussion	43	**4**	Diskussion	43
5	Auseinandersetzung	42	**5**	Auseinandersetzung	42
6	Erziehung	39	**6**	Erziehung	39
7	Faszination	36	**7**	Faszination	36
7	Gewalt	36	**7**	Gewalt	36
8	Rückbesinnung	26	**8**	Rückbesinnung	26
9	Klassikern	25	**9**	Klassikern	25
9	pädagogischen	25	**9**	pädagogischen	25
9	subjektiv	25	**9**	subjektiv	25
10	Bewertung	24	**10**	Bewertung	24
11	Entwicklung	19	**11**	Entwicklung	19
12	Analyse	15	**12**	Analyse	15
12	Ausbreitung	15	**12**	Ausbreitung	15
13	Auswahl	11	**13**	Auswahl	11
13	Recht	11	**13**	Recht	11
14	Abhandlungen	9	**14**	Abhandlungen	9
15	einschränkende	8	**15**	einschränkende	8
15	Wirkung	8	**16**	ändern	7
16	Deutschland	7	**16**	Deutschland	7
16	erwachsen	7	**16**	erwachsen	7
16	ändern	7	**16**	öffentliche	7
16	öffentliche	7	**16**	Wirkungen	7
17	Erfahrung	6	**17**	Erfahrung	6
18	Beschreibung	5	**18**	Beschreibung	5
19	erklären	4	**19**	erklären	4
19	Versuch	4	**19**	Versuch	4
20	begonnen	3	**20**	begonnen	3
20	geschrieben	3	**20**	geschrieben	3
			20	Spielen	3

Anzahl Fehlworte (Hfgk. 1):
8 entsprechend 12.12%
(andere, diejenigen, jetzt, Kapitel, keine, mußte, Warum, wenig)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 5:

Warum haben wir ein Buch über Computer - Spiele geschrieben? Die einfachste Antwort lautet: Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computer - Spielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft. Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der Ausbreitung und Faszination von Spielen gesetzt werden. Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar, was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computer - Spiele erhalten hat, nachzuzeichnen und zu erklären. Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei unvermeidlich subjektiv. Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion. Das muß sich ändern.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	14	Abhandlungen	9	585	1	13
2	21	Materie	2	1813	1	12
3	8	Rückbesinnung	26	461	1	14
4	18	Beschreibung	5	3397	1	11
5	21	nachzuzeichnen	2	140	1	15
6	6	Erziehung	39	4872	1	10
7	9	subjektiv	25	2543	1	11
8	2	Computer	61	19419	4	8
9	21	verwunderlich	2	1399	1	12
10	10	Bewertung	24	4443	1	10
11	7	Faszination	36	2301	1	11
12	13	Auswahl	11	9046	2	9
13	------	erschöpft	-	2849	1	11
14	22	Kapitel	1	6629	1	10
15	12	Analyse	15	9007	1	9
16	15	[Wirkung/Wirkungen]	8	14257	2	9
17	9	Klassikern	25	4087	1	10
18	12	Ausbreitung	15	1056	1	12
19	------	bemängeln	-	1637	1	12
20	5	Auseinandersetzung	42	21127	2	8
21	9	pädagogischen	25	4754	1	10
22	------	unvermeidlich	-	2780	1	11
23	4	Diskussion	43	38280	2	7
24	------	ausgewählt	-	3370	1	11
25	11	Entwicklung	19	44887	1	7
26	13	Recht	11	7444	1	10
27	------	Über	-	36621	3	7
28	------	Zumindest	-	4729	1	10
29	------	orientieren	-	6941	1	10
30	------	darum	-	20503	1	8
31	15	einschränkende	8	5704	1	10
32	19	Versuch	4	44271	1	7
33	22	diejenigen	1	8428	1	9
34	------	unbedingt	-	13260	1	9
35	------	Antwort	-	25629	1	8
36	17	Erfahrung	6	29115	1	8
37	7	Gewalt	36	27786	1	8
38	------	einzelnen	-	34283	1	7
39	------	Weil	-	25554	2	8
40	16	erwachsen	7	11916	1	9
41	------	sicherlich	-	6949	1	10
42	------	Vielleicht	-	20492	1	8
43	------	hauptsächlich	-	6503	1	10
44	21	Gegen	2	20710	1	8
45	------	fängt	-	11442	1	9
46	------	erscheinen	-	48772	1	7
47	------	darüber	-	46878	1	7
48	------	genau	-	52117	1	7
49	------	verstehen	-	46128	1	7
50	22	Warum	1	50190	1	7
51	21	endlich	2	33589	1	7
52	------	überhaupt	-	45241	1	7
53	------	klar	-	50991	1	7
54	16	ändern	7	32469	1	7
55	------	lautet	-	53000	1	7
------	1	Spiele	62			
------	3	Buch	49			
------	16	Deutschland	7			
------	16	öffentliche	7			
------	19	erklären	4			
------	20	begonnen	3			
------	20	geschrieben	3			
------	22	andere	1			
------	22	jetzt	1			
------	22	keine	1			
------	22	mußte	1			
------	22	wenig	1			

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	------	Spiele	62			
2	8	Computer	61	19419	4	8
3	------	Buch	49			
4	23	Diskussion	43	38280	2	7
5	20	Auseinandersetzung	42	21127	2	8
6	6	Erziehung	39	4872	1	10
7	11	Faszination	36	2301	1	11
7	37	Gewalt	36	27786	1	8
8	3	Rückbesinnung	26	461	1	14
9	17	Klassikern	25	4087	1	10
9	7	subjektiv	25	2543	1	11
9	21	pädagogischen	25	4754	1	10
10	10	Bewertung	24	4443	1	10
11	25	Entwicklung	19	44887	1	7
12	18	Ausbreitung	15	1056	1	12
12	15	Analyse	15	9007	1	9
13	12	Auswahl	11	9046	2	9
13	26	Recht	11	7444	1	10
14	1	Abhandlungen	9	585	1	13
15	31	einschränkende	8	5704	1	10
15	16	[Wirkung/Wirkungen]	8	14257	2	9
16	------	öffentliche	7			
16	40	erwachsen	7	11916	1	9
16	------	Deutschland	7			
16	54	ändern	7	32469	1	7
17	36	Erfahrung	6	29115	1	8
18	4	Beschreibung	5	3397	1	11
19	32	Versuch	4	44271	1	7
19	------	erklären	4			
20	------	begonnen	3			
20	------	geschrieben	3			
21	44	Gegen	2	20710	1	8
21	2	Materie	2	1813	1	12
21	5	nachzuzeichnen	2	140	1	15
21	9	verwunderlich	2	1399	1	12
21	51	endlich	2	33589	1	7
22	50	Warum	1	50190	1	7
22	------	jetzt	1			
22	------	mußte	1			
22	14	Kapitel	1	6629	1	10
22	------	wenig	1			
22	33	diejenigen	1	8428	1	9
22	------	keine	1			
22	------	andere	1			
------	49	verstehen	-	46128	1	7
------	52	überhaupt	-	45241	1	7
------	53	klar	-	50991	1	7
------	55	lautet	-	53000	1	7
------	34	unbedingt	-	13260	1	9
------	28	Zumindest	-	4729	1	10
------	29	orientieren	-	6941	1	10
------	30	darum	-	20503	1	8
------	27	Über	-	36621	3	7
------	24	ausgewählt	-	3370	1	11
------	13	erschöpft	-	2849	1	11
------	19	bemängeln	-	1637	1	12
------	22	unvermeidlich	-	2780	1	11
------	35	Antwort	-	25629	1	8
------	38	einzelnen	-	34283	1	7
------	45	fängt	-	11442	1	9
------	46	erscheinen	-	48772	1	7
------	47	darüber	-	46878	1	7
------	43	hauptsächlich	-	6503	1	10
------	42	Vielleicht	-	20492	1	8
------	39	Weil	-	25554	2	8
------	41	sicherlich	-	6949	1	10
------	48	genau	-	52117	1	7

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

6. Tannenwald – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Als der Junge auf seinem Schulweg am Marktplatz vorbeikam, traute er seinen Augen nicht. Über Nacht war der Tannen - Wald in die Stadt gekommen.
An der ganzen langen Mauer entlang standen Tannenbäume, große und kleine, breite und schlanke. Von einem Auto wurden ganze Berge von Tannen abgeladen. Männer mit dicken Handschuhen hoben sie auf und stellten sie an die Mauer, überall dorthin, wo noch Platz war.
Hier und da traten Männer und Frauen an die Bäume heran. Sie drehten eine Tanne und betrachteten sie von allen Seiten. Dann stellten sie den Baum wieder an seinen Platz zurück. Er hatte ihnen nicht gefallen. Ein zweiter wurde geprüft, ein dritter und ein vierter. Der Weihnachtsbaum sollte doch auch recht gerade gewachsen sein und volle Zweige haben.
Ein Mann trat hinzu. Er hatte ein Beil in der Hand. Aus der Tasche hing ihm ein langer Bindfaden. Die Leute fragten ihn, was die Tannen kosteten.
Eine Frau wählte eine hübsche Tanne aus. Der Mann spitzte den Baum mit dem Beil an. Dann band er die Zweige mit dem Bindfaden zusammen, und die Frau konnte ihn bequem nach Hause tragen.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Bindfaden	78	16	wachsen	39190	7	all	334975	4
Tannenbaum	368	14	Nacht	41553	7	kommen	376645	4
abladen	503	13	tragen	41644	7	Seite	459776	4
Beil	583	13	Dann	43040	7	sollen	555827	3
Tanne	623	13	Auto	45112	7	noch	672854	3
Schulweg	744	13	dritt	45841	7	am	717511	3
Weihnachtsbaum	834	13	Auge	46019	7	Der	760069	3
vorbeikommen	1006	12	voll	48462	7	nach	805815	3
Handschuh	1197	12	An	49364	7	können	821919	3
betrachtet	1270	12	treten	49641	7	aus	892600	3
spitzen	1911	11	zusammen	55037	7	er	893675	3
Zweig	2076	11	große	59552	6	an	1039533	2
schlank	3225	11	Hand	62490	6	einen	1121613	2
Marktplatz	3381	11	Aus	65281	6	sie	1210878	2
heran	3622	10	zurück	69872	6	auch	1301212	2
dorthin	4054	10	kosten	71543	6	Die	1421814	2
bequem	4201	10	Leute	72833	6	dem	1472967	2
hübsch	5182	10	gerade	87635	6	auf	1608537	2
binden	7159	10	ihnen	96969	6	nicht	1772767	2
trauen	7217	9	wo	107113	6	mit	1861458	1
entlang	7322	9	recht	108854	6	von	2215196	1
gefallen	7463	9	fragen	117326	5	haben	2295270	1
Tasche	7747	9	Als	117369	5	werden	2520779	1
dick	9806	9	Haus	121812	5	den	2697266	1
hinzu	10357	9	Von	136156	5	ein	3437822	1
Junge	12803	9	ihn	142987	5	in	4086574	0
Mauer	13418	9	klein	143128	5	sein	4658755	0
heben	14846	8	Eine	147799	5	und	5064441	0
überall	15696	8	ihm	152449	5	die	6782081	0
zweit	15844	8	stellen	153146	5	der	7367299	0
Wald	15890	8	Stadt	154004	5			
Baum	16209	8	Mann	173980	5			
viert	16274	8	lang	182707	5			
prüfen	16970	8	doch	191390	5			
drehen	17397	8	Er	201087	5			
hängen	20761	8	Frau	209331	5			
breit	24095	8	Ein	220955	5			
Berg	25736	8	ganz	223423	5			
Platz	25763	8	was	230040	5			
wählen	33360	7	stehen	253583	4			
Hier	35279	7	wieder	267352	4			
Über	36621	7	da	274291	4			

Als der Junge auf sein Schulweg am Marktplatz vorbeikommen trauen er sein Auge nicht Über Nacht sein der Tanne Wald in die Stadt kommen An der ganz lang Mauer entlang stehen Tannenbaum große und klein breit und schlank Von ein Auto werden ganz Berg von Tanne abladen Mann mit dick Handschuh heben sie auf und stellen sie an die Mauer überall dorthin wo noch Platz sein Hier und da treten Mann und Frau an die Baum heran sie drehen einen Tanne und betrachtet sie von all Seite Dann stellen sie den Baum wieder an sein Platz zurück Er haben ihnen nicht gefallen Ein zweit werden prüfen ein dritt und ein viert Der Weihnachtsbaum sollen doch auch recht gerade wachsen sein und voll Zweig haben Ein Mann treten hinzu Er haben ein Beil in der Hand Aus der Tasche hängen ihm ein lang Bindfaden Die Leute fragen ihn was die Tanne kosten Eine Frau wählen einen hübsch Tanne aus Der Mann spitzen den Baum mit dem Beil an Dann binden er die Zweig mit dem Bindfaden zusammen und die Frau können ihn bequem nach Haus tragen

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 6:

Als der Junge auf seinem Schulweg am Marktplatz vorbeikam, traute er seinen Augen nicht. Über Nacht war der Tannen - Wald in die Stadt gekommen. An der ganzen langen Mauer entlang standen Tannenbäume, große und kleine, breite und schlanke. Von einem Auto wurden ganze Berge von Tannen abgeladen. Männer mit dicken Handschuhen hoben sie auf und stellten sie an die Mauer, überall dorthin, wo noch Platz war. Hier und da traten Männer und Frauen an die Bäume heran. Sie drehten eine Tanne und betrachteten sie von allen Seiten. Dann stellten sie den Baum wieder an seinen Platz zurück. Er hatte ihnen nicht gefallen. Ein zweiter wurde geprüft, ein dritter und ein vierter. Der Weihnachtsbaum sollte doch auch recht gerade gewachsen sein und volle Zweige haben. Ein Mann trat hinzu. Er hatte ein Beil in der Hand. Aus der Tasche hing ihm ein langer Bindfaden. Die Leute fragten ihn, was die Tannen kosteten. Eine Frau wählte eine hübsche Tanne aus. Der Mann spitzte den Baum mit dem Beil an. Dann band er die Zweige mit dem Bindfaden zusammen, und die Frau konnte ihn bequem nach Hause tragen.

Top 10-Trefferquote der Versuchspersonen:								
Trefferquote (relativ)	20%	30%	40%	50%	60%	70%	80%	90%
Personen (von 66)	2	5	10	16	19	11	2	1
Personen (relativ)	3%	7.5%	15.1%	24.2%	28.7%	16.6%	3%	1.5%
Erwartungswert:	53.78%							

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.118	0.695	**0.418**	0.146

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	5.666	28.454	**17.828**	4.772

Top 10 ATA2 - entlemmatisiert: [Tannen/Tanne], Tannenbäume, Beil, Schulweg, Weihnachtsbaum, Bindfaden, Handschuhen, [Bäume/Baum], Zweige, band

6 Treffer ATA2 von realen Top 10: Weihnachtsbaum (#1), Beil (#3), Tannenbäume (#4), Tannen (#6), Bindfaden (#8), Schulweg (#10)

4 Treffer ausserhalb Top 10: Zweige (#11), Bäume (#19), Handschuhen (#22), band (#25)

Text 6 Trefferquote ATA2 Top 10: 60 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.404
Durchschnittlichliche Markierungsanzahl aller 10: 25.9 Vpn

Top 20 ATA2 - entlemmatisiert: [Tannen/Tanne], Tannenbäume, Beil, Schulweg, Weihnachtsbaum, Bindfaden, Handschuhen, [Bäume/Baum], Zweige, band, betrachteten, An, Tasche, vorbeikam, Marktplatz, schlanke, Wald, Dann, dorthin, dicken

11 Treffer ATA2 von realen Top 20: Weihnachtsbaum (#1), Marktplatz (#2), Beil (#3), Tannenbäume (#4), Tannen (#6), Bindfaden (#8), Schulweg (#10), Zweige (#11), betrachteten (#15), Wald (#17), Bäume (#19)

4 Treffer ausserhalb Top 20: Handschuhen (#22), band (#25), dicken (#26), Tasche (#26)

5 Wort(e) von Vpn überhaupt nicht markiert: An, vorbeikam, schlanke, Dann, dorthin

Text 6 Trefferquote ATA2 Top 20: 55 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.432
Durchschnittlichliche Markierungsanzahl aller 20: 16.85 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Weihnachtsbaum	55	**1**	Weihnachtsbaum	55
2	Marktplatz	43	**2**	Marktplatz	43
3	Beil	40	**3**	Beil	40
4	Tannenbäume	36	**4**	Tannenbäume	36
5	Junge	34	**5**	Junge	34
6	Tannen	33	**6**	kosteten	32
7	kosteten	32	**7**	Bindfaden	30
8	Bindfaden	30	**8**	Tannen	28
9	Hause	27	**9**	Hause	27
10	Schulweg	26	**10**	Schulweg	26
11	Zweige	22	**11**	Zweige	22
12	hübsche	21	**12**	hübsche	21
12	Stadt	21	**12**	Stadt	21
13	Männer	20	**13**	geprüft	19
14	geprüft	19	**14**	betrachteten	18
15	betrachteten	18	**14**	Mauer	18
15	Mauer	18	**15**	gefallen	16
16	gefallen	16	**15**	gerade	16
16	gerade	16	**16**	Wald	15
17	Wald	15	**17**	Männer	14
18	Frauen	12	**18**	tragen	12
18	tragen	12	**19**	bequem	7
19	Bäume	10	**20**	Baum	6
20	bequem	7	**20**	Berge	6
			20	Frau	6
			20	Frauen	6
			20	gewachsen	6
			20	Mann	6
			20	Nacht	6
			20	volle	6
			20	wählte	6

Anzahl Fehlworte (Hfgk. 1):
12 entsprechend 18.18%
(dicken, drehten, fragten, Hier, ihn, Seiten, spitzte, Tasche, traute, von, zurück, Über)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 6:

Als der Junge auf seinem Schulweg am Marktplatz vorbeikam, traute er seinen Augen nicht. Über Nacht war der Tannen - Wald in die Stadt gekommen. An der ganzen langen Mauer entlang standen Tannenbäume, große und kleine, breite und schlanke. Von einem Auto wurden ganze Berge von Tannen abgeladen. Männer mit dicken Handschuhen hoben sie auf und stellten sie an die Mauer, überall dorthin, wo noch Platz war. Hier und da traten Männer und Frauen an die Bäume heran. Sie drehten eine Tanne und betrachteten sie von allen Seiten. Dann stellten sie den Baum wieder an seinen Platz zurück. Er hatte ihnen nicht gefallen. Ein zweiter wurde geprüft, ein dritter und ein vierter. Der Weihnachtsbaum sollte doch auch recht gerade gewachsen sein und volle Zweige haben. Ein Mann trat hinzu. Er hatte ein Beil in der Hand. Aus der Tasche hing ihm ein langer Bindfaden. Die Leute fragten ihn, was die Tannen kosteten. Eine Frau wählte eine hübsche Tanne aus. Der Mann spitzte den Baum mit dem Beil an. Dann band er die Zweige mit dem Bindfaden zusammen, und die Frau konnte ihn bequem nach Hause tragen.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	6	[Tannen/Tanne]	33	623	5	13
2	4	Tannenbäume	36	368	1	14
3	3	Beil	40	583	2	13
4	10	Schulweg	26	744	1	13
5	1	Weihnachtsbaum	55	834	1	13
6	8	Bindfaden	30	78	2	16
7	22	Handschuhen	5	1197	1	12
8	19	[Bäume/Baum]	10	16209	3	8
9	11	Zweige	22	2076	2	11
10	25	band	2	7159	1	10
11	15	betrachteten	18	1270	1	12
12	------	An	-	49364	5	7
13	26	Tasche	1	7747	1	9
14	------	vorbeikam	-	1006	1	12
15	2	Marktplatz	43	3381	1	11
16	------	schlanke	-	3225	1	11
17	17	Wald	15	15890	1	8
18	------	Dann	-	43040	2	7
19	------	dorthin	-	4054	1	10
20	26	dicken	1	9806	1	9
21	22	abgeladen	5	503	1	13
22	------	heran	-	3622	1	11
23	15	Mauer	18	13418	2	9
24	------	entlang	-	7322	1	10
25	26	drehten	1	17397	1	8
26	20	bequem	7	4201	1	10
27	26	Hier	1	35279	1	7
28	12	hübsche	21	5182	1	10
29	5	Junge	34	12803	1	9
30	21	Berge	6	25736	1	8
31	------	hoben	-	14846	1	9
32	26	spitzte	1	1911	1	11
33	16	gefallen	16	7463	1	10
34	25	Platz	2	25763	2	8
35	25	Augen	2	46019	1	7
36	------	breite	-	24095	1	8
37	------	hing	-	20761	1	8
38	23	Auto	4	45112	1	7
39	26	traute	1	7217	1	10
40	21	Nacht	6	41553	1	7
41	23	überall	4	15696	1	8
42	------	[traten/trat]	-	49641	2	7
43	18	tragen	12	41644	1	7
44	21	gewachsen	6	39190	1	7
45	------	hinzu	-	10357	1	9
46	26	Über	1	36621	1	7
47	21	volle	6	48462	1	7
48	14	geprüft	19	16970	1	8
49	------	zweiter	-	15844	1	8
50	------	vierter	-	16274	1	8
51	------	zusammen	-	55037	1	7
52	------	dritter	-	45841	1	7
53	21	wählte	6	33360	1	7
------	7	kosteten	32			
------	9	Hause	27			
------	12	Stadt	21			
------	13	Männer	20			
------	16	gerade	16			
------	18	Frauen	12			
------	22	Leute	5			
------	24	nach	3			
------	24	nicht	3			
------	26	fragten	1			
------	26	ihn	1			
------	26	Seiten	1			
------	26	von	1			
------	26	zurück	1			

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	5	Weihnachtsbaum	55	834	1	13
2	15	Marktplatz	43	3381	1	11
3	3	Beil	40	583	2	13
4	2	Tannenbäume	36	368	1	14
5	29	Junge	34	12803	1	9
6	1	[Tannen/Tanne]	33	623	5	13
7	------	kosteten	32			
8	6	Bindfaden	30	78	2	16
9	------	Hause	27			
10	4	Schulweg	26	744	1	13
11	9	Zweige	22	2076	2	11
12	------	Stadt	21			
12	28	hübsche	21	5182	1	10
13	------	Männer	20			
14	48	geprüft	19	16970	1	8
15	11	betrachteten	18	1270	1	12
15	23	Mauer	18	13418	2	9
16	33	gefallen	16	7463	1	10
16	------	gerade	16			
17	17	Wald	15	15890	1	8
18	43	tragen	12	41644	1	7
18	------	Frauen	12			
19	8	[Bäume/Baum]	10	16209	3	8
20	26	bequem	7	4201	1	10
21	44	gewachsen	6	39190	1	7
21	40	Nacht	6	41553	1	7
21	53	wählte	6	33360	1	7
21	30	Berge	6	25736	1	8
21	47	volle	6	48462	1	7
22	21	abgeladen	5	503	1	13
22	7	Handschuhen	5	1197	1	12
22	------	Leute	5			
23	41	überall	4	15696	1	8
23	38	Auto	4	45112	1	7
24	------	nicht	3			
24	------	nach	3			
25	34	Platz	2	25763	2	8
25	10	band	2	7159	1	10
25	35	Augen	2	46019	1	7
26	25	drehten	1	17397	1	8
26	------	fragten	1			
26	20	dicken	1	9806	1	9
26	27	Hier	1	35279	1	7
26	13	Tasche	1	7747	1	9
26	46	Über	1	36621	1	7
26	------	zurück	1			
26	39	traute	1	7217	1	10
26	32	spitzte	1	1911	1	11
26	------	von	1			
26	------	ihn	1			
26	------	Seiten	1			
------	12	An	-	49364	5	7
------	16	schlanke	-	3225	1	11
------	14	vorbeikam	-	1006	1	12
------	18	Dann	-	43040	2	7
------	52	dritter	-	45841	1	7
------	45	hinzu	-	10357	1	9
------	31	hoben	-	14846	1	9
------	42	[traten/trat]	-	49641	2	7
------	36	breite	-	24095	1	8
------	49	zweiter	-	15844	1	8
------	50	vierter	-	16274	1	8
------	22	heran	-	3622	1	11
------	24	entlang	-	7322	1	10
------	37	hing	-	20761	1	8
------	51	zusammen	-	55037	1	7
------	19	dorthin	-	4054	1	10

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

7. Körpersprache – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Wir verwenden Zeit und Energie, um neben unserer Mutter - Sprache noch weitere Sprachen zu lernen. Körper - Sprache ist mit der Zeit zu einer Fremd - Sprache geworden. Fremd - Sprachen müssen nicht gelernt werden, aber wir kommen weiter, wenn wir sie beherrschen. Wir vermindern die Gefahr von Mißverständnissen. Es ist mir unerklärlich, warum wir nie die Zeit haben, unsere Primär - Sprache, nämlich die Sprache unseres Körpers, zu verbessern. Da sich niemand des Kommunikations - Mittels Körper - Sprache entziehen oder sie unterdrücken kann, ist es von wesentlichem Nutzen, sie zu lernen - gibt sie uns doch wichtige Informationen über die innere Haltung und Einstellung unserer Mitmenschen.
Wenn wir offene Sinne und ein waches Auge für die Signale und Kommentare unserer Körper - Sprache haben, können viele Gespräche und Begegnungen leichter und erfolgreicher verlaufen. Die Kenntnis der Körper - Sprache, des lautlosen Frage- und Antwort- Spiels in unserem körperlichen Verhalten, öffnet direktere Wege zueinander und einen freieren Umgang miteinander. In manchen sprachlosen "Augenblicken" spüren wir das ja auch: Da sagt ein Blick, eine Wendung des Kopfes, eine ergreifende Geste, eine abwehrende Gebärde mehr als tausend Worte.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Primär	116	15	Antwort	25629	8	In	370333	4
Fremd	277	14	wesentlich	27667	8	kommen	376645	4
ergreifend	375	14	Sprache	27686	8	geben	411918	4
Gebärde	478	13	erfolgreich	29603	7	mehr	429567	4
lautlos	745	13	direkt	34037	7	oder	443445	4
unerklärlich	812	13	lernen	35090	7	aber	502793	3
sprachlos	835	13	niemand	35140	7	müssen	520472	3
abwehren	841	13	Information	38691	7	über	598179	3
Mitmensch	1205	12	nutzen	38994	7	noch	672854	3
Wendung	1552	12	frei	39647	7	um	700687	3
Mißverständnis	1640	12	manch	41771	7	können	821919	3
zueinander	1761	12	Mutter	42564	7	es	1069022	2
vermindern	1794	12	Kopf	42623	7	als	1080482	2
unterdrücken	2748	11	Blick	42967	7	einen	1121613	2
wach	3645	10	nämlich	45310	7	sie	1210878	2
Kommunikation	5606	10	neben	45721	7	auch	1301212	2
Geste	5890	10	Auge	46019	7	Die	1421814	2
körperlich	6695	10	Frage	48250	7	für	1594594	2
Gefahr	7280	9	offen	50008	7	des	1607888	2
beherrschen	7667	9	warum	50190	7	sich	1718091	2
Einstellung	8784	9	leicht	51397	7	nicht	1772767	2
Signal	8989	9	Gespräch	59926	6	mit	1861458	1
entziehen	9105	9	nie	61754	6	das	1920541	1
Begegnung	9849	9	Spiel	63243	6	zu	2050372	1
Kenntnis	10017	9	Wort	73951	6	von	2215196	1
tausend	10720	9	wichtig	78048	6	haben	2295270	1
Kommentar	10815	9	mir	91049	6	werden	2520779	1
Mittel	11090	9	Weg	98020	6	ein	3437822	1
Augenblick	11297	9	ja	103000	6	in	4086574	0
spüren	12316	9	Wenn	122208	5	sein	4658755	0
miteinander	12391	9	unser	130011	5	und	5064441	0
Umgang	13219	9	Wir	146565	5	die	6782081	0
verwenden	13725	9	uns	153162	5	der	7367299	0
Energie	13987	9	doch	191390	5			
inne	14343	9	Zeit	197886	5			
Haltung	15876	8	weit	211429	5			
Körper	16320	8	Es	263194	4			
verlaufen	17427	8	da	274291	4			
Sinn	19580	8	wir	278795	4			
verbessern	19921	8	wenn	298994	4			
verhalten	22591	8	sagen	349278	4			
öffnen	25283	8	viel	351696	4			

Wir verwenden Zeit und Energie um neben unser Mutter Sprache noch weit Sprache zu lernen Körper Sprache sein mit der Zeit zu ein Fremd Sprache werden Fremd Sprache müssen nicht lernen werden aber wir kommen weit wenn wir sie beherrschen Wir vermindern die Gefahr von Mißverständnis Es sein mir unerklärlich warum wir nie die Zeit haben unser Primär Sprache nämlich die Sprache unser Körper zu verbessern da sich niemand des Kommunikation Mittel Körper Sprache entziehen oder sie unterdrücken können sein es von wesentlich nutzen sie zu lernen geben sie uns doch wichtig Information über die inne Haltung und Einstellung unser Mitmensch Wenn wir offen Sinn und ein wach Auge für die Signal und Kommentar unser Körper Sprache haben können viel Gespräch und Begegnung leicht und erfolgreich verlaufen Die Kenntnis der Körper Sprache des lautlos Frage und Antwort Spiel in unser körperlich verhalten öffnen direkt Weg zueinander und ein frei Umgang miteinander In manch sprachlos Augenblick spüren wir das ja auch da sagen ein Blick einen Wendung des Kopf einen ergreifend Geste einen abwehren Gebärde mehr als tausend Wort

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 7:

Wir verwenden Zeit und Energie, um neben unserer Mutter - Sprache noch weitere Sprachen zu lernen. Körper - Sprache ist mit der Zeit zu einer Fremd - Sprache geworden. Fremd - Sprachen müssen nicht gelernt werden, aber wir kommen weiter, wenn wir sie beherrschen. Wir vermindern die Gefahr von Mißverständnissen. Es ist mir unerklärlich, warum wir nie die Zeit haben, unsere Primär - Sprache, nämlich die Sprache unseres Körpers, zu verbessern. Da sich niemand des Kommunikations - Mittels Körper - Sprache entziehen oder sie unterdrücken kann, ist es von wesentlichem Nutzen, sie zu lernen - gibt sie uns doch wichtige Informationen über die innere Haltung und Einstellung unserer Mitmenschen. Wenn wir offene Sinne und ein waches Auge für die Signale und Kommentare unserer Körper - Sprache haben, können viele Gespräche und Begegnungen leichter und erfolgreicher verlaufen. Die Kenntnis der Körper - Sprache, des lautlosen Frage- und Antwort- Spiels in unserem körperlichen Verhalten, öffnet direktere Wege zueinander und einen freieren Umgang miteinander. In manchen sprachlosen "Augenblicken" spüren wir das ja auch: Da sagt ein Blick, eine Wendung des Kopfes, eine ergreifende Geste, eine abwehrende Gebärde mehr als tausend Worte.

Top 10-Trefferquote der Versuchspersonen:						
Trefferquote (relativ)	40%	50%	60%	70%	80%	90%
Personen (von 66)	7	22	16	11	6	4
Personen (relativ)	10.6%	33.3%	24.2%	16.6%	9%	6%
Erwartungswert:	59.84%					
Gewichtete 10-dim. Euklidische Qualität:						
Min / Max / Mittelw. / Std.Abw.	0.116	0.806	**0.487**	0.131		
Durchschnittlichliche Markierungsanzahl anderer Vpn:						
Min / Max / Mittelw. / Std.Abw.	10.727	27.25	**19.907**	3.366		

Top 10 ATA2 - entlemmatisiert: Fremd, Gebärde, ergreifende, [Körper/Körpers], [Sprache/Sprachen], Wendung, Geste, unerklärlich, Mitmenschen, Mißverständnissen

5 Treffer ATA2 von realen Top 10: Sprache (#1), Körper (#2), Mißverständnissen (#3), Fremd (#5), Mitmenschen (#8)

2 Treffer ausserhalb Top 10: Geste (#15), Gebärde (#18)

3 Wort(e) von Vpn überhaupt nicht markiert: ergreifende, Wendung, unerklärlich

Text 7 Trefferquote ATA2 Top 10: 50 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.485
Durchschnittlichliche Markierungsanzahl aller 10: 24.9 Vpn

Top 20 ATA2 - entlemmatisiert: Fremd, Gebärde, ergreifende, [Körper/Körpers], [Sprache/Sprachen], Wendung, Geste, unerklärlich, Mitmenschen, Mißverständnissen, lautlosen, abwehrende, sprachlosen, Primär, Augenblicken, Kommunikations, Gefahr, beherrschen, Energie, Umgang

15 Treffer ATA2 von realen Top 20: Sprache (#1), Körper (#2), Mißverständnissen (#3), Kommunikations (#4), Fremd (#5), Mitmenschen (#8), Energie (#10), Umgang (#11), Primär (#12), Augenblicken (#14), Geste (#15), sprachlosen (#15), Gebärde (#18), beherrschen (#19), Gefahr (#20)

1 Treffer ausserhalb Top 20: lautlosen (#21)

4 Wort(e) von Vpn überhaupt nicht markiert: ergreifende, Wendung, unerklärlich, abwehrende

Text 7 Trefferquote ATA2 Top 20: 75 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.392
Durchschnittlichliche Markierungsanzahl aller 20: 18.7 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Sprache	69	**1**	Körper	57
2	Körper	61	**2**	Sprache	49
3	Mißverständnissen	42	**3**	Mißverständnissen	42
4	Kommunikations	39	**4**	Kommunikations	39
5	Fremd	38	**5**	Fremd	38
6	Signale	27	**6**	Signale	27
7	Informationen	23	**7**	Informationen	23
8	Mitmenschen	22	**8**	Mitmenschen	22
8	Mutter	22	**8**	Mutter	22
9	lernen	20	**9**	Sprachen	20
10	Energie	19	**10**	Energie	19
10	Zeit	19	**10**	lernen	19
11	Umgang	16	**10**	Zeit	19
12	Einstellung	15	**11**	Umgang	16
12	Primär	15	**12**	Einstellung	15
13	Nutzen	13	**12**	Primär	15
13	Sinne	13	**13**	Nutzen	13
14	Augenblicken	11	**13**	Sinne	13
14	erfolgreicher	11	**14**	Augenblicken	11
14	Gespräche	11	**14**	erfolgreicher	11
14	Worte	11	**14**	Gespräche	11
15	Geste	10	**14**	Worte	11
15	sprachlosen	10	**15**	Geste	10
16	freieren	9	**15**	sprachlosen	10
17	Begegnungen	8	**16**	freieren	9
17	direktere	8	**17**	Begegnungen	8
17	Haltung	8	**17**	direktere	8
18	Gebärde	7	**17**	Haltung	8
18	innere	7	**18**	Gebärde	7
18	Verhalten	7	**18**	innere	7
19	beherrschen	6	**18**	Verhalten	7
19	Blick	6	**19**	beherrschen	6
19	Kenntnis	6	**19**	Blick	6
20	Gefahr	5	**19**	Kenntnis	6
20	Mittels	5	**20**	Gefahr	5
20	verbessern	5	**20**	Mittels	5
			20	verbessern	5

Anzahl Fehlworte (Hfgk. 1):
9 entsprechend 13.63%
(Antwort, körperlichen, mehr, neben, sagt, vermindern, verwenden, Wege, zueinander)

[Vollständige Gegenüberstellung ATA2 vs. Real]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 7:

Wir verwenden Zeit und Energie, um neben unserer Mutter - Sprache noch weitere Sprachen zu lernen. Körper - Sprache ist mit der Zeit zu einer Fremd - Sprache geworden. Fremd - Sprachen müssen nicht gelernt werden, aber wir kommen weiter, wenn wir sie beherrschen. Wir vermindern die Gefahr von Mißverständnissen. Es ist mir unerklärlich, warum wir nie die Zeit haben, unsere Primär - Sprache, nämlich die Sprache unseres Körpers, zu verbessern. Da sich niemand des Kommunikations - Mittels Körper - Sprache entziehen oder sie unterdrücken kann, ist es von wesentlichem Nutzen, sie zu lernen - gibt sie uns doch wichtige Informationen über die innere Haltung und Einstellung unserer Mitmenschen. Wenn wir offene Sinne und ein waches Auge für die Signale und Kommentare unserer Körper - Sprache haben, können viele Gespräche und Begegnungen leichter und erfolgreicher verlaufen. Die Kenntnis der Körper - Sprache, des lautlosen Frage- und Antwort- Spiels in unserem körperlichen Verhalten, öffnet direktere Wege zueinander und einen freieren Umgang miteinander. In manchen sprachlosen "Augenblicken" spüren wir das ja auch: Da sagt ein Blick, eine Wendung des Kopfes, eine ergreifende Geste, eine abwehrende Gebärde mehr als tausend Worte.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	5	Fremd	38	277	2	14
2	18	Gebärde	7	478	1	13
3	------	ergreifende	-	375	1	14
4	2	[Körper/Körpers]	61	16320	5	8
5	1	[Sprache/Sprachen]	69	27686	10	8
6	------	Wendung	-	1552	1	12
7	15	Geste	10	5890	1	10
8	------	unerklärlich	-	812	1	13
9	8	Mitmenschen	22	1205	1	12
10	3	Mißverständnissen	42	1640	1	12
11	21	lautlosen	4	745	1	13
12	------	abwehrende	-	841	1	13
13	15	sprachlosen	10	835	1	13
14	12	Primär	15	116	1	16
15	14	Augenblicken	11	11297	1	9
16	4	Kommunikations	39	5606	1	10
17	20	Gefahr	5	7280	1	10
18	19	beherrschen	6	7667	1	9
19	10	Energie	19	13987	1	9
20	11	Umgang	16	13219	1	9
21	------	unterdrücken	-	2748	1	11
22	17	Begegnungen	8	9849	1	9
23	19	Kenntnis	6	10017	1	9
24	17	Haltung	8	15876	1	8
25	19	Blick	6	42967	1	7
26	24	zueinander	1	1761	1	12
27	------	Auge	-	46019	1	7
28	9	[lernen/gelernt]	20	35090	3	7
29	12	Einstellung	15	8784	1	9
30	24	körperlichen	1	6695	1	10
31	24	Antwort	1	25629	1	8
32	20	Mittels	5	11090	1	9
33	24	vermindern	1	1794	1	12
34	------	waches	-	3645	1	11
35	------	entziehen	-	9105	1	9
36	23	Frage	2	48250	1	7
37	23	miteinander	2	12391	1	9
38	18	Verhalten	7	22591	1	8
39	18	innere	7	14343	1	9
40	13	Sinne	13	19580	1	8
41	21	spüren	4	12316	1	9
42	------	verlaufen	-	17427	1	8
43	------	Kopfes	-	42623	1	7
44	6	Signale	27	8989	1	9
45	------	tausend	-	10720	1	9
46	21	Kommentare	4	10815	1	9
47	23	öffnet	2	25283	1	8
48	17	direktere	8	34037	1	7
49	7	Informationen	23	38691	1	7
50	14	erfolgreicher	11	29603	1	8
51	8	Mutter	22	42564	1	7
52	16	freieren	9	39647	1	7
53	24	verwenden	1	13725	1	9
54	------	manchen	-	41771	1	7
55	22	leichter	3	51397	1	7
56	23	wesentlichem	2	27667	1	8
57	------	warum	-	50190	1	7
58	------	nämlich	-	45310	1	7
59	24	neben	1	45721	1	7
60	------	offene	-	50008	1	7
61	20	verbessern	5	19921	1	8
62	13	Nutzen	13	38994	1	7
63	------	niemand	-	35140	1	7
------	10	Zeit	19			
------	14	Gespräche	11			
------	14	Worte	11			
------	23	weitere	2			
------	23	wichtige	2			
------	24	mehr	1			
------	24	sagt	1			
------	24	Wege	1			

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	5	[Sprache/Sprachen]	69	27686	10	8
2	4	[Körper/Körpers]	61	16320	5	8
3	10	Mißverständnissen	42	1640	1	12
4	16	Kommunikations	39	5606	1	10
5	1	Fremd	38	277	2	14
6	44	Signale	27	8989	1	9
7	49	Informationen	23	38691	1	7
8	51	Mutter	22	42564	1	7
8	9	Mitmenschen	22	1205	1	12
9	28	[lernen/gelernt]	20	35090	3	7
10	19	Energie	19	13987	1	9
10	------	Zeit	19			
11	20	Umgang	16	13219	1	9
12	14	Primär	15	116	1	16
12	29	Einstellung	15	8784	1	9
13	40	Sinne	13	19580	1	8
13	62	Nutzen	13	38994	1	7
14	------	Gespräche	11			
14	50	erfolgreicher	11	29603	1	8
14	------	Worte	11			
14	15	Augenblicken	11	11297	1	9
15	13	sprachlosen	10	835	1	13
15	7	Geste	10	5890	1	10
16	52	freieren	9	39647	1	7
17	48	direktere	8	34037	1	7
17	22	Begegnungen	8	9849	1	9
17	24	Haltung	8	15876	1	8
18	39	innere	7	14343	1	9
18	2	Gebärde	7	478	1	13
18	38	Verhalten	7	22591	1	8
19	23	Kenntnis	6	10017	1	9
19	18	beherrschen	6	7667	1	9
19	25	Blick	6	42967	1	7
20	32	Mittels	5	11090	1	9
20	17	Gefahr	5	7280	1	10
20	61	verbessern	5	19921	1	8
21	46	Kommentare	4	10815	1	9
21	11	lautlosen	4	745	1	13
21	41	spüren	4	12316	1	9
22	55	leichter	3	51397	1	7
23	56	wesentlichem	2	27667	1	8
23	47	öffnet	2	25283	1	8
23	------	wichtige	2			
23	------	weitere	2			
23	36	Frage	2	48250	1	7
23	37	miteinander	2	12391	1	9
24	31	Antwort	1	25629	1	8
24	26	zueinander	1	1761	1	12
24	30	körperlichen	1	6695	1	10
24	59	neben	1	45721	1	7
24	33	vermindern	1	1794	1	12
24	53	verwenden	1	13725	1	9
24	------	sagt	1			
24	------	Wege	1			
24	------	mehr	1			
------	21	unterdrücken	-	2748	1	11
------	63	niemand	-	35140	1	7
------	8	unerklärlich	-	812	1	13
------	6	Wendung	-	1552	1	12
------	3	ergreifende	-	375	1	14
------	12	abwehrende	-	841	1	13
------	58	nämlich	-	45310	1	7
------	35	entziehen	-	9105	1	9
------	42	verlaufen	-	17427	1	8
------	43	Kopfes	-	42623	1	7
------	34	waches	-	3645	1	11
------	54	manchen	-	41771	1	7
------	60	offene	-	50008	1	7
------	45	tausend	-	10720	1	9
------	57	warum	-	50190	1	7
------	27	Auge	-	46019	1	7

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

8. Autor – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Der Autor steht beim Werdegang eines Buches an erster Stelle, er schreibt den Text. Das lateinische Wort Autor bedeutet Urheber, ganz gleich ob Mann oder Frau. Allerdings wird der Begriff im allgemeinen nur auf Schriftsteller und Dichter angewandt. Ihr Werkzeug ist die Sprache. Mit der müssen sie gut umgehen können. Doch das allein genügt noch nicht! Will ein Schriftsteller eine Geschichte erfinden und das Geschehen spannend darstellen, muß er Ideen haben und viel Phantasie entwickeln. Beim Verfassen eines Sachbuches muß er über den betreffenden Bereich viel Wissen sammeln, das heißt er muß persönliche Erfahrungen machen und umfangreiche Nachforschungen anstellen. Natürlich hängt das Schreiben auch von der Personen - Gruppe ab, an die sich ein Autor wendet. Es ist schon ein Unterschied, ob er sich bei ganz jungen Lesern oder bei Erwachsenen verständlich machen möchte. Autoren arbeiten sehr unterschiedlich. Manche können diese schwierige Tätigkeit, die so viel Konzentration und Ausdauer verlangt, nur kurze Zeit am Tag ausüben. Sie bringen vielleicht nur ein paar Sätze zustande. Andere schreiben die ganze Nacht hindurch und vollenden Geschichten und Romane in einem Zug. Oft wird dann immer und immer wieder verändert und verbessert.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Sachbuch	519	13	geschehen	27669	8	stehen	253583	4
Werdegang	688	13	Sprache	27686	8	dann	263034	4
Ausdauer	766	13	Zug	27788	8	Es	263194	4
Nachforschung	863	13	Text	28000	8	wieder	267352	4
Urheber	1282	12	verlangen	28187	8	schon	303960	4
lateinisch	1341	12	Erfahrung	29115	7	gut	314979	4
vollenden	1928	11	allgemein	29635	7	viel	351696	4
Dichter	1977	11	entwickeln	30972	7	machen	352516	4
hindurch	1984	11	persönlich	31039	7	erst	362491	4
betreffend	2193	11	Autor	32498	7	ihr	374106	4
Werkzeug	2529	11	bedeuten	32853	7	oder	443445	4
angewandt	3340	11	Idee	34500	7	wollen	492248	3
Oft	4058	10	Person	39275	7	müssen	520472	3
verfassen	4811	10	Nacht	41553	7	so	559696	3
spannend	5001	10	manch	41771	7	Das	576030	3
ausüben	5385	10	Stelle	43302	7	nur	584819	3
verständlich	5465	10	Bereich	45715	7	über	598179	3
anstellen	5647	10	vielleicht	46107	7	noch	672854	3
Konzentration	5896	10	paar	48110	7	bei	685320	3
zustande	6055	10	gleichen	58009	6	am	717511	3
erfinden	6493	10	allein	59286	6	Der	760069	3
umfangreich	8300	9	Buch	63404	6	können	821919	3
Phantasie	8334	9	Wort	73951	6	dies	827459	3
Tätigkeit	10238	9	kurz	75293	6	er	893675	3
erwachsen	11916	9	Gruppe	79910	6	an	1039533	2
umgehen	13354	9	Geschichte	81897	6	einen	1121613	2
Leser	13965	9	mögen	82554	6	sie	1210878	2
genügen	14691	8	schreiben	84727	6	auch	1301212	2
Schriftsteller	15123	8	Tag	108257	6	im	1577564	2
Natürlich	15912	8	Arbeit	112044	6	auf	1608537	2
Begriff	16143	8	ob	112614	6	sich	1718091	2
verändern	16679	8	jung	121233	5	nicht	1772767	2
sammeln	16937	8	sehr	127462	5	das	1920541	1
darstellen	17012	8	bringen	129676	5	von	2215196	1
verbessern	19921	8	Doch	131111	5	haben	2295270	1
unterscheiden	20176	8	wissen	152287	5	werden	2520779	1
hängen	20761	8	beim	163963	5	den	2697266	1
heißen	20967	8	ab	166310	5	ein	3437822	1
Allerdings	23241	8	Mit	167025	5	in	4086574	0
unterschiedlich	23768	8	Mann	173980	5	sein	4658755	0
Roman	23977	8	Zeit	197886	5	und	5064441	0
Satz	24536	8	Frau	209331	5	die	6782081	0
schwierig	24683	8	ganz	223423	5	der	7367299	0
Beim	24728	8	immer	247688	4			
wenden	26211	8	andere	249754	4			

Der Autor stehen beim Werdegang ein Buch an erst Stelle er schreiben den Text Das lateinisch Wort Autor bedeuten Urheber ganz gleichen ob Mann oder Frau Allerdings werden der Begriff im allgemein nur auf Schriftsteller und Dichter angewandt ihr Werkzeug sein die Sprache Mit der müssen sie gut umgehen können Doch das allein genügen noch nicht wollen ein Schriftsteller einen Geschichte erfinden und das geschehen spannend darstellen müssen er Idee haben und viel Phantasie entwickeln Beim verfassen ein Sachbuch müssen er über den betreffend Bereich viel wissen sammeln das heißen er müssen persönlich Erfahrung machen und umfangreich Nachforschung anstellen Natürlich hängen das schreiben auch von der Person Gruppe ab an die sich ein Autor wenden Es sein schon ein unterscheiden ob er sich bei ganz jung Leser oder bei erwachsen verständlich machen mögen Autor Arbeit sehr unterschiedlich manch können dies schwierig Tätigkeit die so viel Konzentration und Ausdauer verlangen nur kurz Zeit am Tag ausüben sie bringen vielleicht nur ein paar Satz zustande andere schreiben die ganz Nacht hindurch und vollenden Geschichte und Roman in ein Zug Oft werden dann immer und immer wieder verändern und verbessern

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 8:

Der Autor steht beim Werdegang eines Buches an erster Stelle, er schreibt den Text. Das lateinische Wort Autor bedeutet Urheber, ganz gleich ob Mann oder Frau. Allerdings wird der Begriff im allgemeinen nur auf Schriftsteller und Dichter angewandt. Ihr Werkzeug ist die Sprache. Mit der müssen sie gut umgehen können. Doch das allein genügt noch nicht! Will ein Schriftsteller eine Geschichte erfinden und das Geschehen spannend darstellen, muß er Ideen haben und viel Phantasie entwickeln. Beim Verfassen eines Sachbuches muß er über den betreffenden Bereich viel Wissen sammeln, das heißt er muß persönliche Erfahrungen machen und umfangreiche Nachforschungen anstellen. Natürlich hängt das Schreiben auch von der Personen - Gruppe ab, an die sich ein Autor wendet. Es ist schon ein Unterschied, ob er sich bei ganz jungen Lesern oder bei Erwachsenen verständlich machen möchte. Autoren arbeiten sehr unterschiedlich. Manche können diese schwierige Tätigkeit, die so viel Konzentration und Ausdauer verlangt, nur kurze Zeit am Tag ausüben. Sie bringen vielleicht nur ein paar Sätze zustande. Andere schreiben die ganze Nacht hindurch und vollenden Geschichten und Romane in einem Zug. Oft wird dann immer und immer wieder verändert und verbessert.

Top 10-Trefferquote der Versuchspersonen:							
Trefferquote (relativ)	40%	50%	60%	70%	80%	90%	100%
Personen (von 66)	4	10	21	16	9	5	1
Personen (relativ)	6%	15.1%	31.8%	24.2%	13.6%	7.5%	1.5%
Erwartungswert:	65.3%						

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.202	0.899	**0.531**	0.164

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	12.666	35.727	**23.883**	5.152

Top 10 ATA2 - entlemmatisiert: Ausdauer, Sachbuches, Nachforschungen, Dichter, Konzentration, [Autor/Autoren], Werdegang, anstellen, Urheber, lateinische

4 Treffer ATA2 von realen Top 10: Autor (#1), Urheber (#7), Konzentration (#8), Nachforschungen (#9)

5 Treffer ausserhalb Top 10: Ausdauer (#13), Sachbuches (#15), Werdegang (#16), Dichter (#19), lateinische (#25)

1 Wort(e) von Vpn überhaupt nicht markiert: anstellen

Text 8 Trefferquote ATA2 Top 10: 40 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.183
Durchschnittlichliche Markierungsanzahl aller 10: 21 Vpn

Top 20 ATA2 - entlemmatisiert: Ausdauer, Sachbuches, Nachforschungen, Dichter, Konzentration, [Autor/Autoren], Werdegang, anstellen, Urheber, lateinische, Lesern, Werkzeug, Phantasie, Romane, hindurch, betreffenden, Schriftsteller, Oft, Verfassen, vollenden

13 Treffer ATA2 von realen Top 20: Autor (#1), Phantasie (#2), Urheber (#7), Konzentration (#8), Nachforschungen (#9), Schriftsteller (#9), Werkzeug (#11), Ausdauer (#13), Sachbuches (#15), Werdegang (#16), Romane (#18), Dichter (#19), Lesern (#20)

4 Treffer ausserhalb Top 20: Verfassen (#22), lateinische (#25), Oft (#25), vollenden (#25)

3 Wort(e) von Vpn überhaupt nicht markiert: anstellen, hindurch, betreffenden

Text 8 Trefferquote ATA2 Top 20: 65 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.317
Durchschnittlichliche Markierungsanzahl aller 20: 16.65 Vpn

	Vpn - Top 20 flexionskorrigiert			**Vpn - Top 20 Datenbank**	
1	Autor	65	**1**	Autor	62
2	Phantasie	52	**2**	Phantasie	52
3	Wissen	49	**3**	Wissen	49
4	Sprache	46	**4**	Sprache	46
5	Buches	38	**5**	Buches	38
6	Ideen	35	**6**	Ideen	35
7	Urheber	34	**7**	Urheber	34
8	Konzentration	31	**8**	Konzentration	31
9	Nachforschungen	29	**9**	Nachforschungen	29
9	Schriftsteller	29	**9**	Schriftsteller	29
10	Erfahrungen	24	**10**	Erfahrungen	24
11	Werkzeug	22	**11**	Werkzeug	22
12	unterschiedlich	20	**12**	unterschiedlich	20
13	Ausdauer	19	**13**	Ausdauer	19
13	Geschichte	19	**13**	Geschichte	19
14	Text	18	**14**	Text	18
15	Sachbuches	13	**15**	Sachbuches	13
16	Werdegang	11	**16**	Werdegang	11
17	schreibt	9	**17**	verständlich	9
17	verständlich	9	**18**	Romane	8
18	Romane	8	**19**	Dichter	7
19	Dichter	7	**19**	schreiben	7
19	spannend	7	**19**	spannend	7
20	Lesern	6	**20**	Lesern	6
20	verändert	6	**20**	verändert	6

Anzahl Fehlworte (Hfgk. 1):
15 entsprechend 22.72%
(Begriff, ein, erfinden, erster, Erwachsenen, kurze, lateinische, Nacht, Oft, persönliche, sammeln, Sätze, umgehen, vollenden, wendet)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 8:

Der Autor steht beim Werdegang eines Buches an erster Stelle, er schreibt den Text. Das lateinische Wort Autor bedeutet Urheber, ganz gleich ob Mann oder Frau. Allerdings wird der Begriff im allgemeinen nur auf Schriftsteller und Dichter angewandt. Ihr Werkzeug ist die Sprache. Mit der müssen sie gut umgehen können. Doch das allein genügt noch nicht! Will ein Schriftsteller eine Geschichte erfinden und das Geschehen spannend darstellen, muß er Ideen haben und viel Phantasie entwickeln. Beim Verfassen eines Sachbuches muß er über den betreffenden Bereich viel Wissen sammeln, das heißt er muß persönliche Erfahrungen machen und umfangreiche Nachforschungen anstellen. Natürlich hängt das Schreiben auch von der Personen - Gruppe ab, an die sich ein Autor wendet. Es ist schon ein Unterschied, ob er sich bei ganz jungen Lesern oder bei Erwachsenen verständlich machen möchte. Autoren arbeiten sehr unterschiedlich. Manche können diese schwierige Tätigkeit, die so viel Konzentration und Ausdauer verlangt, nur kurze Zeit am Tag ausüben. Sie bringen vielleicht nur ein paar Sätze zustande. Andere schreiben die ganze Nacht hindurch und vollenden Geschichten und Romane in einem Zug. Oft wird dann immer und immer wieder verändert und verbessert.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	13	Ausdauer	19	766	1	13
2	15	Sachbuches	13	519	1	13
3	9	Nachforschungen	29	863	1	13
4	19	Dichter	7	1977	1	11
5	8	Konzentration	31	5896	1	10
6	1	[Autor/Autoren]	65	32498	4	7
7	16	Werdegang	11	688	1	13
8	------	anstellen	-	5647	1	10
9	7	Urheber	34	1282	1	12
10	25	lateinische	1	1341	1	12
11	20	Lesern	6	13965	1	9
12	11	Werkzeug	22	2529	1	11
13	2	Phantasie	52	8334	1	9
14	18	Romane	8	23977	1	8
15	------	hindurch	-	1984	1	11
16	------	betreffenden	-	2193	1	11
17	9	Schriftsteller	29	15123	2	8
18	25	Oft	1	4058	1	10
19	22	Verfassen	4	4811	1	10
20	25	vollenden	1	1928	1	11
21	19	spannend	7	5001	1	10
22	24	Tätigkeit	2	10238	1	9
23	------	Beim	-	24728	2	8
24	25	Begriff	1	16143	1	8
25	4	Sprache	46	27686	1	8
26	------	angewandt	-	3340	1	11
27	10	Erfahrungen	24	29115	1	8
28	25	erfinden	1	6493	1	10
29	------	ausüben	-	5385	1	10
30	6	Ideen	35	34500	1	7
31	17	verständlich	9	5465	1	10
32	14	Text	18	28000	1	8
33	------	darstellen	-	17012	1	8
34	------	genügt	-	14691	1	9
35	25	Erwachsenen	1	11916	1	9
36	------	Allerdings	-	23241	1	8
37	------	umfangreiche	-	8300	1	9
38	------	Bereich	-	45715	1	7
39	25	Sätze	1	24536	1	8
40	25	Nacht	1	41553	1	7
41	21	Personen	5	39275	1	7
42	25	sammeln	1	16937	1	8
43	------	Natürlich	-	15912	1	8
44	------	allgemeinen	-	29635	1	8
45	------	hängt	-	20761	1	8
46	------	entwickeln	-	30972	1	7
47	12	unterschiedlich	20	23768	1	8
48	25	persönliche	1	31039	1	7
49	25	umgehen	1	13354	1	9
50	------	zustande	-	6055	1	10
51	24	Unterschied	2	20176	1	8
52	22	schwierige	4	24683	1	8
53	------	heißt	-	20967	1	8
54	21	verbessert	5	19921	1	8
55	------	Zug	-	27788	1	8
56	20	verändert	6	16679	1	8
57	------	verlangt	-	28187	1	8
58	24	Geschehen	2	27669	1	8
59	------	Manche	-	41771	1	7
60	------	bedeutet	-	32853	1	7
61	24	Stelle	2	43302	1	7
62	25	wendet	1	26211	1	8
63	------	vielleicht	-	46107	1	7
64	------	paar	-	48110	1	7
------	3	Wissen	49			
------	5	Buches	38			
------	13	Geschichte	19			
------	17	schreibt	9			
------	21	arbeiten	5			
------	23	Gruppe	3			
------	23	Zeit	3			
------	24	Wort	2			
------	25	ein	1			
------	25	erster	1			
------	25	kurze	1			

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	6	[Autor/Autoren]	65	32498	4	7
2	13	Phantasie	52	8334	1	9
3	------	Wissen	49			
4	25	Sprache	46	27686	1	8
5	------	Buches	38			
6	30	Ideen	35	34500	1	7
7	9	Urheber	34	1282	1	12
8	5	Konzentration	31	5896	1	10
9	17	Schriftsteller	29	15123	2	8
9	3	Nachforschungen	29	863	1	13
10	27	Erfahrungen	24	29115	1	8
11	12	Werkzeug	22	2529	1	11
12	47	unterschiedlich	20	23768	1	8
13	------	Geschichte	19			
13	1	Ausdauer	19	766	1	13
14	32	Text	18	28000	1	8
15	2	Sachbuches	13	519	1	13
16	7	Werdegang	11	688	1	13
17	31	verständlich	9	5465	1	10
17	------	schreibt	9			
18	14	Romane	8	23977	1	8
19	21	spannend	7	5001	1	10
19	4	Dichter	7	1977	1	11
20	56	verändert	6	16679	1	8
20	11	Lesern	6	13965	1	9
21	------	arbeiten	5			
21	54	verbessert	5	19921	1	8
21	41	Personen	5	39275	1	7
22	52	schwierige	4	24683	1	8
22	19	Verfassen	4	4811	1	10
23	------	Zeit	3			
23	------	Gruppe	3			
24	51	Unterschied	2	20176	1	8
24	61	Stelle	2	43302	1	7
24	58	Geschehen	2	27669	1	8
24	22	Tätigkeit	2	10238	1	9
24	------	Wort	2			
25	48	persönliche	1	31039	1	7
25	42	sammeln	1	16937	1	8
25	49	umgehen	1	13354	1	9
25	62	wendet	1	26211	1	8
25	28	erfinden	1	6493	1	10
25	------	ein	1			
25	18	Oft	1	4058	1	10
25	------	erster	1			
25	10	lateinische	1	1341	1	12
25	------	kurze	1			
25	40	Nacht	1	41553	1	7
25	20	vollenden	1	1928	1	11
25	24	Begriff	1	16143	1	8
25	39	Sätze	1	24536	1	8
25	35	Erwachsenen	1	11916	1	9
------	64	paar	-	48110	1	7
------	59	Manche	-	41771	1	7
------	60	bedeutet	-	32853	1	7
------	63	vielleicht	-	46107	1	7
------	38	Bereich	-	45715	1	7
------	29	ausüben	-	5385	1	10
------	33	darstellen	-	17012	1	8
------	34	genügt	-	14691	1	9
------	26	angewandt	-	3340	1	11
------	23	Beim	-	24728	2	8
------	8	anstellen	-	5647	1	10
------	15	hindurch	-	1984	1	11
------	16	betreffenden	-	2193	1	11
------	36	Allerdings	-	23241	1	8
------	37	umfangreiche	-	8300	1	9
------	50	zustande	-	6055	1	10
------	53	heißt	-	20967	1	8
------	55	Zug	-	27788	1	8
------	46	entwickeln	-	30972	1	7
------	45	hängt	-	20761	1	8
------	43	Natürlich	-	15912	1	8
------	44	allgemeinen	-	29635	1	8
------	57	verlangt	-	28187	1	8

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

9. Wissenschaft – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Die Frage, die der Versuchs - Leiter im Labor an die Natur stellt, hat immer schon zur Voraussetzung, daß er eine Vermutung hegt, die er bestätigt oder widerlegt wissen will. Diese Vermutung aber stammt stets aus einer vorher gehenden Beobachtung, mit anderen Worten, aus jenen nicht - rationalen kognitiven Leistungen unserer Sinnes - Organe und unseres Nerven - Systems, die aus Sinnes - Daten Wahrnehmungen formen. Es bedeutet eine gewaltige Überschätzung der menschlichen Ratio, wenn sich ein Wissenschaftler einbildet, er wisse und kenne alle Fragen, die man an die Natur stellen kann. Wenn ein Forscher sich die Frage ausdenkt, während er, in sein Laboratorium gebannt, seiner Arbeit ohne Kontakt mit der unabsehbaren Reichhaltigkeit der lebenden Natur obliegt, so kann es allzu leicht geschehen, daß seine Frage an dem wirklich Wesentlichen vorbeigeht und nur Irrelevantes zutage fördert. So werden dann Untersuchungen angestellt, die trotz größtem Scharfsinn und trotz bester methodischer Berücksichtigung aller Einzelheiten keineswegs das Lebenswichtige betreffen. Der Forscher aber, der sich ganz eben dieser Untersuchung widmet, kann dies unmöglich einsehen.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Reichhaltigkeit	31	17	stets	18559	8	schon	303960	4
Überschätzung	140	15	Untersuchung	19538	8	gut	314979	4
Scharfsinn	160	15	Sinn	19580	8	man	419503	4
Ratio	213	15	fördern	19846	8	oder	443445	4
kognitiv	233	14	betreffen	20364	8	zur	444282	4
irrelevant	359	14	Natur	20558	8	wollen	492248	3
einbilden	369	14	stammen	22436	8	aber	502793	3
größtem	620	13	Leiter	23269	8	so	559696	3
unabsehbar	631	13	Kontakt	23972	8	nur	584819	3
Laboratorium	727	13	Leistung	24158	8	Der	760069	3
methodisch	799	13	bestätigen	24859	8	können	821919	3
lebenswichtig	846	13	wesentlich	27667	8	dies	827459	3
obliegen	940	12	geschehen	27669	8	aus	892600	3
vorbeigehen	1153	12	System	30829	7	er	893675	3
zutage	1835	11	bedeuten	32853	7	daß	936255	2
rational	1937	11	trotz	36598	7	an	1039533	2
ausdenken	1945	11	kennen	42060	7	es	1069022	2
widerlegen	1954	11	Versuch	44271	7	einen	1121613	2
Berücksichtigung	2166	11	Frage	48250	7	Die	1421814	2
bannen	2218	11	wirklich	50386	7	dem	1472967	2
einsehen	2584	11	leicht	51397	7	im	1577564	2
Wahrnehmung	3396	11	eben	55867	7	sich	1718091	2
hegen	3484	11	jene	55875	7	nicht	1772767	2
Labor	3794	10	Diese	70953	6	mit	1861458	1
Vermutung	3966	10	Wort	73951	6	das	1920541	1
Nerv	4081	10	während	74329	6	haben	2295270	1
Einzelheit	4248	10	Arbeit	112044	6	werden	2520779	1
Organ	4513	10	Wenn	122208	5	ein	3437822	1
Beobachtung	5100	10	unser	130011	5	in	4086574	0
anstellen	5647	10	wissen	152287	5	sein	4658755	0
unmöglich	8719	9	stellen	153146	5	und	5064441	0
Forscher	8907	9	ohne	156677	5	die	6782081	0
gewaltig	9356	9	So	159635	5	der	7367299	0
allzu	9471	9	leben	171897	5			
widmen	10072	9	ganz	223423	5			
Form	12139	9	immer	247688	4			
keineswegs	13098	9	alle	248569	4			
Datum	15043	8	andere	249754	4			
vorher	15800	8	dann	263034	4			
Wissenschaftler	16222	8	Es	263194	4			
Voraussetzung	16876	8	gehen	291282	4			
menschlich	18452	8	wenn	298994	4			

Die Frage die der Versuch Leiter im Labor an die Natur stellen haben immer schon zur Voraussetzung daß er einen Vermutung hegen die er bestätigen oder widerlegen wissen wollen Diese Vermutung aber stammen stets aus ein vorher gehen Beobachtung mit andere Wort aus jene nicht rational kognitiv Leistung unser Sinn Organ und unser Nerv System die aus Sinn Datum Wahrnehmung Form Es bedeuten einen gewaltig Überschätzung der menschlich Ratio wenn sich ein Wissenschaftler einbilden er wissen und kennen alle Frage die man an die Natur stellen können Wenn ein Forscher sich die Frage ausdenken während er in sein Laboratorium bannen sein Arbeit ohne Kontakt mit der unabsehbar Reichhaltigkeit der leben Natur obliegen so können es allzu leicht geschehen daß sein Frage an dem wirklich wesentlich vorbeigehen und nur irrelevant zutage fördern So werden dann Untersuchung anstellen die trotz größtem Scharfsinn und trotz gut methodisch Berücksichtigung alle Einzelheit keineswegs das lebenswichtig betreffen Der Forscher aber der sich ganz eben dies Untersuchung widmen können dies unmöglich einsehen

Studie "Autoabstracting" - Auswertung

Detailauswertung

Text Nr. 9:

Die Frage, die der Versuchs - Leiter im Labor an die Natur stellt, hat immer schon zur Voraussetzung, daß er eine Vermutung hegt, die er bestätigt oder widerlegt wissen will. Diese Vermutung aber stammt stets aus einer vorher gehenden Beobachtung, mit anderen Worten, aus jenen nicht - rationalen kognitiven Leistungen unserer Sinnes - Organe und unseres Nerven - Systems, die aus Sinnes - Daten Wahrnehmungen formen. Es bedeutet eine gewaltige Überschätzung der menschlichen Ratio, wenn sich ein Wissenschaftler einbildet, er wisse und kenne alle Fragen, die man an die Natur stellen kann. Wenn ein Forscher sich die Frage ausdenkt, während er, in sein Laboratorium gebannt, seiner Arbeit ohne Kontakt mit der unabsehbaren Reichhaltigkeit der lebenden Natur obliegt, so kann es allzu leicht geschehen, daß seine Frage an dem wirklich Wesentlichen vorbeigeht und nur Irrelevantes zutage fördert. So werden dann Untersuchungen angestellt, die trotz größtem Scharfsinn und trotz bester methodischer Berücksichtigung aller Einzelheiten keineswegs das Lebenswichtige betreffen. Der Forscher aber, der sich ganz eben dieser Untersuchung widmet, kann dies unmöglich einsehen.

Top 10-Trefferquote der Versuchspersonen:								
Trefferquote (relativ)	30%	40%	50%	60%	70%	80%	90%	100%
Personen (von 66)	1	1	3	13	19	19	7	3
Personen (relativ)	1.5%	1.5%	4.5%	19.6%	28.7%	28.7%	10.6%	4.5%
Erwartungswert:	72.42%							

Gewichtete 10-dim. Euklidische Qualität:				
Min / Max / Mittelw. / Std.Abw.	0.167	0.658	**0.396**	0.117

Durchschnittlichliche Markierungsanzahl anderer Vpn:				
Min / Max / Mittelw. / Std.Abw.	13.3	33.2	**22.795**	4.032

Top 10 ATA2 - entlemmatisiert: Scharfsinn, Ratio, kognitiven, Laboratorium, einbildet, Vermutung, rationalen, Natur, größtem, Forscher

4 Treffer ATA2 von realen Top 10: Vermutung (#1), Natur (#2), Forscher (#7), Scharfsinn (#10)

5 Treffer ausserhalb Top 10: Laboratorium (#14), Ratio (#14), kognitiven (#16), rationalen (#23), einbildet (#24)

1 Wort(e) von Vpn überhaupt nicht markiert: größtem

Text 9 Trefferquote ATA2 Top 10: 40 Prozent
Gewichtete 10-dim. Euklidische Qualität: 0.19
Durchschnittlichliche Markierungsanzahl aller 10: 17.3 Vpn

Top 20 ATA2 - entlemmatisiert: Scharfsinn, Ratio, kognitiven, Laboratorium, einbildet, Vermutung, rationalen, Natur, größtem, Forscher, Labor, Reichhaltigkeit, Wahrnehmungen, formen, Beobachtung, unabsehbaren, ausdenkt, Organe, Überschätzung, obliegt

12 Treffer ATA2 von realen Top 20: Beobachtung (#1), Vermutung (#1), Natur (#2), Labor (#4), Wahrnehmungen (#5), Forscher (#7), Überschätzung (#8), Scharfsinn (#10), Laboratorium (#14), Ratio (#14), kognitiven (#16), Reichhaltigkeit (#17)

5 Treffer ausserhalb Top 20: ausdenkt (#23), Organe (#23), rationalen (#23), einbildet (#24), unabsehbaren (#24)

3 Wort(e) von Vpn überhaupt nicht markiert: größtem, formen, obliegt

Text 9 Trefferquote ATA2 Top 20: 60 Prozent
Gewichtete 20-dim. Euklidische Qualität: 0.332
Durchschnittlichliche Markierungsanzahl aller 20: 16.4 Vpn

	Vpn - Top 20 flexionskorrigiert			Vpn - Top 20 Datenbank	
1	Beobachtung	53	**1**	Beobachtung	53
1	Vermutung	53	**1**	Vermutung	53
2	Natur	42	**2**	Natur	42
3	Frage	38	**3**	Irrelevantes	38
3	Irrelevantes	38	**4**	Labor	34
4	Labor	34	**5**	Frage	33
5	Wahrnehmungen	30	**6**	Wahrnehmungen	30
6	Untersuchungen	27	**7**	Forscher	26
7	Forscher	26	**7**	Versuchs	26
7	Versuchs	26	**8**	Lebenswichtige	25
8	Lebenswichtige	25	**8**	Überschätzung	25
8	Überschätzung	25	**9**	Wesentlichen	23
9	Wesentlichen	23	**9**	Wissenschaftler	23
9	Wissenschaftler	23	**10**	Untersuchungen	17
10	Scharfsinn	16	**11**	Scharfsinn	16
11	Leiter	15	**12**	Leiter	15
12	bestätigt	14	**13**	bestätigt	14
13	widerlegt	13	**14**	widerlegt	13
14	Laboratorium	12	**15**	Laboratorium	12
14	Ratio	12	**15**	Ratio	12
15	Voraussetzung	10	**16**	Untersuchung	10
16	keineswegs	9	**16**	Voraussetzung	10
16	kognitiven	9	**17**	keineswegs	9
16	Sinnes	9	**17**	kognitiven	9
17	Reichhaltigkeit	8	**17**	Sinnes	9
18	methodischer	7	**18**	Reichhaltigkeit	8
18	vorbeigeht	7	**19**	methodischer	7
19	einsehen	6	**19**	vorbeigeht	7
20	Leistungen	5	**20**	einsehen	6

Anzahl Fehlworte (Hfgk. 1):
12 entsprechend 18.18%
(Alle, bester, betreffen, einbildet, gehenden, gewaltige, kenne, lebenden, Systems, unabsehbaren, wisse, Wissen)

[Vollständige Gegenüberstellung ATA2 vs. Real]

[Zurück ...]

Studie "Autoabstracting" - Auswertung

Wortauswahl-Rang-Gegenüberstellung

Text Nr. 9:

Die Frage, die der Versuchs - Leiter im Labor an die Natur stellt, hat immer schon zur Voraussetzung, daß er eine Vermutung hegt, die er bestätigt oder widerlegt wissen will. Diese Vermutung aber stammt stets aus einer vorher gehenden Beobachtung, mit anderen Worten, aus jenen nicht - rationalen kognitiven Leistungen unserer Sinnes - Organe und unseres Nerven - Systems, die aus Sinnes - Daten Wahrnehmungen formen. Es bedeutet eine gewaltige Überschätzung der menschlichen Ratio, wenn sich ein Wissenschaftler einbildet, er wisse und kenne alle Fragen, die man an die Natur stellen kann. Wenn ein Forscher sich die Frage ausdenkt, während er, in sein Laboratorium gebannt, seiner Arbeit ohne Kontakt mit der unabsehbaren Reichhaltigkeit der lebenden Natur obliegt, so kann es allzu leicht geschehen, daß seine Frage an dem wirklich Wesentlichen vorbeigeht und nur Irrelevantes zutage fördert. So werden dann Untersuchungen angestellt, die trotz größtem Scharfsinn und trotz bester methodischer Berücksichtigung aller Einzelheiten keineswegs das Lebenswichtige betreffen. Der Forscher aber, der sich ganz eben dieser Untersuchung widmet, kann dies unmöglich einsehen.

Rang 1 = von ATA2 am stärksten assoziiert / von der Mehrzahl der Versuchspersonen markiert

Flexionskorrigierte (d.h. zusammengefasste, aufsummierte) reale Ränge

ATA2-Rang	Realer Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	10	Scharfsinn	16	160	1	15
2	14	Ratio	12	213	1	15
3	16	kognitiven	9	233	1	15
4	14	Laboratorium	12	727	1	13
5	24	einbildet	1	369	1	14
6	1	Vermutung	53	3966	2	10
7	23	rationalen	2	1937	1	11
8	2	Natur	42	20558	3	8
9	------	größtem	-	620	1	13
10	7	Forscher	26	8907	2	9
11	4	Labor	34	3794	1	10
12	17	Reichhaltigkeit	8	31	1	17
13	5	Wahrnehmungen	30	3396	1	11
14	------	formen	-	12139	1	9
15	1	Beobachtung	53	5100	1	10
16	24	unabsehbaren	1	631	1	13
17	23	ausdenkt	2	1945	1	11
18	23	Organe	2	4513	1	10
19	8	Überschätzung	25	140	1	15
20	------	obliegt	-	940	1	12
21	------	zutage	-	1835	1	12
22	6	[Untersuchungen/Untersuchung]	27	19538	2	8
23	3	Irrelevantes	38	359	1	14
24	16	Sinnes	9	19580	2	8
25	23	Nerven	2	4081	1	10
26	8	Lebenswichtige	25	846	1	13
27	3	[Frage/Fragen]	38	48250	4	7
28	9	Wissenschaftler	23	16222	1	8
29	13	widerlegt	13	1954	1	11
30	23	menschlichen	2	18452	1	8
31	------	fördert	-	19846	1	8
32	------	trotz	-	36598	2	7
33	------	angestellt	-	5647	1	10
34	22	Berücksichtigung	3	2166	1	11
35	18	methodischer	7	799	1	13
36	19	einsehen	6	2584	1	11
37	------	gebannt	-	2218	1	11
38	20	Leistungen	5	24158	1	8
39	------	leicht	-	51397	1	7
40	------	allzu	-	9471	1	9
41	7	Versuchs	26	44271	1	7
42	18	vorbeigeht	7	1153	1	12
43	24	gewaltige	1	9356	1	9
44	------	hegt	-	3484	1	11
45	22	Daten	3	15043	1	8
46	11	Leiter	15	23269	1	8
47	------	stets	-	18559	1	8
48	22	Einzelheiten	3	4248	1	10
49	24	Systems	1	30829	1	7
50	24	betreffen	1	20364	1	8
51	22	unmöglich	3	8719	1	9
52	------	widmet	-	10072	1	9
53	12	bestätigt	14	24859	1	8
54	16	keineswegs	9	13098	1	9
55	15	Voraussetzung	10	16876	1	8
56	------	geschehen	-	27669	1	8
57	23	vorher	2	15800	1	8
58	9	Wesentlichen	23	27667	1	8
59	24	kenne	1	42060	1	7
60	------	eben	-	55867	1	7
61	22	Kontakt	3	23972	1	8
62	------	jenen	-	55875	1	7
63	------	bedeutet	-	32853	1	7
64	23	wirklich	2	50386	1	7
65	------	stammt	-	22436	1	8
------	21	Arbeit	4			
------	24	Alle	1			
------	24	bester	1			
------	24	gehenden	1			
------	24	lebenden	1			
------	24	wisse	1			
------	24	Wissen	1			

Realer Rang	ATA2-Rang	Wort	#Vpn	H(w) Korpus	H(w) Text	Wort-Häuf.kl.
1	15	Beobachtung	53	5100	1	10
1	6	Vermutung	53	3966	2	10
2	8	Natur	42	20558	3	8
3	23	Irrelevantes	38	359	1	14
3	27	[Frage/Fragen]	38	48250	4	7
4	11	Labor	34	3794	1	10
5	13	Wahrnehmungen	30	3396	1	11
6	22	[Untersuchungen/Untersuchung]	27	19538	2	8
7	10	Forscher	26	8907	2	9
7	41	Versuchs	26	44271	1	7
8	19	Überschätzung	25	140	1	15
8	26	Lebenswichtige	25	846	1	13
9	58	Wesentlichen	23	27667	1	8
9	28	Wissenschaftler	23	16222	1	8
10	1	Scharfsinn	16	160	1	15
11	46	Leiter	15	23269	1	8
12	53	bestätigt	14	24859	1	8
13	29	widerlegt	13	1954	1	11
14	2	Ratio	12	213	1	15
14	4	Laboratorium	12	727	1	13
15	55	Voraussetzung	10	16876	1	8
16	3	kognitiven	9	233	1	15
16	54	keineswegs	9	13098	1	9
16	24	Sinnes	9	19580	2	8
17	12	Reichhaltigkeit	8	31	1	17
18	35	methodischer	7	799	1	13
18	42	vorbeigeht	7	1153	1	12
19	36	einsehen	6	2584	1	11
20	38	Leistungen	5	24158	1	8
21	------	Arbeit	4			
22	61	Kontakt	3	23972	1	8
22	51	unmöglich	3	8719	1	9
22	45	Daten	3	15043	1	8
22	48	Einzelheiten	3	4248	1	10
22	34	Berücksichtigung	3	2166	1	11
23	17	ausdenkt	2	1945	1	11
23	25	Nerven	2	4081	1	10
23	7	rationalen	2	1937	1	11
23	18	Organe	2	4513	1	10
23	57	vorher	2	15800	1	8
23	64	wirklich	2	50386	1	7
23	30	menschlichen	2	18452	1	8
24	------	gehenden	1			
24	------	lebenden	1			
24	------	wisse	1			
24	------	bester	1			
24	5	einbildet	1	369	1	14
24	59	kenne	1	42060	1	7
24	------	Alle	1			
24	------	Wissen	1			
24	50	betreffen	1	20364	1	8
24	16	unabsehbaren	1	631	1	13
24	43	gewaltige	1	9356	1	9
24	49	Systems	1	30829	1	7
------	65	stammt	-	22436	1	8
------	37	gebannt	-	2218	1	11
------	39	leicht	-	51397	1	7
------	20	obliegt	-	940	1	12
------	21	zutage	-	1835	1	12
------	31	fördert	-	19846	1	8
------	32	trotz	-	36598	2	7
------	33	angestellt	-	5647	1	10
------	40	allzu	-	9471	1	9
------	62	jenen	-	55875	1	7
------	56	geschehen	-	27669	1	8
------	9	größtem	-	620	1	13
------	14	formen	-	12139	1	9
------	47	stets	-	18559	1	8
------	44	hegt	-	3484	1	11
------	52	widmet	-	10072	1	9
------	60	eben	-	55867	1	7
------	63	bedeutet	-	32853	1	7

Ein ATA2-Rang "------" erscheint bei Worten, die von Vpn als Kernworte markiert wurden, aber aufgrund der Worthäufigkeitsklassen-Vokabularreduktion aus dem Vokabular entfernt wurden und somit keinen von ATA2 berechneten Rang haben.

Ein realer Rang "------" erscheint bei Worten, die von keiner Vpn als Kernwort markiert wurden und somit keinen realen Rang haben.

[Zurück ...]

Daß vs. Das – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Das Wort "Das" hat in der deutschen Sprache zwei unterschiedliche Bedeutungen, was besonders Ausländer, die Deutsch lernen, als schwierig zu nehmende Hürde empfinden. So ist zum einen "Das" der bestimmte Artikel für Neutrum-Wörter (das Auto, das Fahrrad und so weiter). Zum anderen ist "das" aber auch ein Relativ-Pronomen ("Das Auto, das mein Mann mir gekauft hat, ..", "Das Fahrrad, das im Keller steht, .."), das aber nicht mit dem Bindewort "daß" verwechselt werden darf. Besonders die Unterscheidung zwischen "das" und "daß" fällt oftmals schwer, da beide Worte häufig durch ein Komma angefügt werden. Ein Beispiel: "Das Auto, das in der Garage steht, wird selten benutzt. Darunter, daß es kaum gefahren wird, leidet das Reifenprofil." Eine Daumenregel, die Kinder oft bereits in der Grundschule lernen, hilft dabei, das Unterscheiden der unterschiedlich geschriebenen "das" und "daß" zu erleichtern: Wenn es sich um das Relativ-Pronomen "das" handelt, kann es auch genauso gut durch "dieses", "jenes" oder "welches" ersetzt werden: "Das Auto, welches .." Bei "daß"-Sätzen wird dieser Austausch nie gelingen, was also ein gutes Indiz dafür darstellt, daß "daß" und nicht "das" erforderlich ist.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
Bindewort	1	22	helfen	39484	7	durch	373263	4
Daumenregel	7	20	gelingen	41801	7	oder	443445	4
Reifenprofil	18	18	Auto	45112	7	aber	502793	3
Pronomen	35	17	Zum	49252	7	so	559696	3
Neutrum	55	17	oft	49423	7	zum	574133	3
anfügen	224	15	besonders	49972	7	Das	576030	3
Relativ	400	14	handeln	53666	7	um	700687	3
Komma	1024	12	jene	55875	7	können	821919	3
Unterscheidung	1249	12	fahren	58014	6	dies	827459	3
geschrieben	1353	12	nie	61754	6	daß	936255	2
Garage	1937	11	welche	64194	6	es	1069022	2
verwechseln	2295	11	kaum	69273	6	als	1080482	2
Darunter	2526	11	Wort	73951	6	auch	1301212	2
oftmals	2610	11	Beispiel	80092	6	dem	1472967	2
Indiz	3347	11	schwer	80704	6	im	1577564	2
Grundschule	4927	10	dafür	86637	6	für	1594594	2
Austausch	5062	10	also	87320	6	sich	1718091	2
Hürde	5107	10	mir	91049	6	nicht	1772767	2
Fahrrad	6554	10	dabei	91446	6	mit	1861458	1
erleichtern	7122	10	Deutsche	100044	6	das	1920541	1
empfinden	10311	9	Bei	119050	5	zu	2050372	1
Keller	10670	9	dürfen	120127	5	haben	2295270	1
erforderlich	12011	9	Wenn	122208	5	werden	2520779	1
benutzen	12133	9	mein	123100	5	ein	3437822	1
ersetzen	12185	9	fallen	125955	5	in	4086574	0
Besonders	12687	9	Kind	128830	5	sein	4658755	0
genauso	13345	9	nehmen	139482	5	und	5064441	0
darstellen	17012	8	bereits	144362	5	die	6782081	0
leiden	18031	8	beide	145017	5	der	7367299	0
unterscheiden	20176	8	Eine	147799	5			
Artikel	21307	8	So	159635	5			
Bedeutung	21767	8	Mann	173980	5			
selten	23350	8	zwischen	198199	5			
bestimmt	23736	8	weit	211429	5			
unterschiedlich	23768	8	deutsch	211557	5			
Ausländer	24038	8	Ein	220955	5			
häufig	24165	8	was	230040	5			
Satz	24536	8	andere	249754	4			
schwierig	24683	8	stehen	253583	4			
Sprache	27686	8	zweien	262643	4			
lernen	35090	7	da	274291	4			
kaufen	36380	7	gut	314979	4			

Das Wort Das haben in der Deutsche Sprache zweien unterschiedlich Bedeutung was besonders Ausländer die deutsch lernen als schwierig zu nehmen Hürde empfinden So sein zum ein Das der bestimmt Artikel für Neutrum Wort das Auto das Fahrrad und so weit Zum andere sein das aber auch ein Relativ Pronomen Das Auto das mein Mann mir kaufen haben Das Fahrrad das im Keller stehen das aber nicht mit dem Bindewort daß verwechseln werden dürfen Besonders die Unterscheidung zwischen das und daß fallen oftmals schwer da beide Wort häufig durch ein Komma anfügen werden Ein Beispiel Das Auto das in der Garage stehen werden selten benutzen Darunter daß es kaum fahren werden leiden das Reifenprofil Eine Daumenregel die Kind oft bereits in der Grundschule lernen helfen dabei das unterscheiden der unterschiedlich geschrieben das und daß zu erleichtern Wenn es sich um das Relativ Pronomen das handeln können es auch genauso gut durch dies jene oder welche ersetzen werden Das Auto welche Bei daß Satz werden dies Austausch nie gelingen was also ein gut Indiz dafür darstellen daß daß und nicht das erforderlich sein

Daß vs. Das – ATA2-Rangfolge

Ursprünglicher Rang	Ursprüngliche Assoziationsstärke	Wort	Korpushäufigkeit	Texthäufigkeit	Worthäufigkeitsklasse	Neuer Rang	Korrigierte Assoziationsstärke
2	0,5257156169	Pronomen	35	2	17	1	0,0300408924
1	0,8236142431	Neutrum	55	1	17	2	0,0149748044
6	0,0445278583	Komma	1024	1	12	3	0,0000434842
10	0,0231311497	Garage	1937	1	11	4	0,0000119417
19	0,0108786601	Unterscheidung	1249	1	12	5	0,0000087099
4	0,0896036975	Auto	45112	4	7	6	0,0000079450
3	0,1599859181	Artikel	21307	1	8	7	0,0000075086
46	0,0013195541	Relativ	400	2	14	8	0,0000065978
11	0,0154122569	Fahrrad	6554	2	10	9	0,0000047032
34	0,0042849985	geschrieben	1353	1	12	10	0,0000031670
5	0,0681435262	Satz	24536	1	8	11	0,0000027773
15	0,0125843894	Grundschule	4927	1	10	12	0,0000025542
47	0,0003404462	anfügen	224	1	15	13	0,0000015198
8	0,0262624961	unterscheiden	20176	1	8	14	0,0000013017
40	0,0029809563	Darunter	2526	1	11	15	0,0000011801
7	0,0317798295	Sprache	27686	1	8	16	0,0000011479
24	0,0070152218	Besonders	12687	2	9	17	0,0000011059
18	0,0110257632	Keller	10670	1	9	18	0,0000010333
39	0,0031250786	Indiz	3347	1	11	19	0,0000009337
35	0,0042462308	Austausch	5062	1	10	20	0,0000008388
44	0,0018163020	verwechseln	2295	1	11	21	0,0000007914
36	0,0039822627	Hürde	5107	1	10	22	0,0000007798
45	0,0015987227	oftmals	2610	1	11	23	0,0000006125
13	0,0133788345	besonders	49972	2	7	24	0,0000005355
16	0,0119633341	Ausländer	24038	1	8	25	0,0000004977
9	0,0253345291	jene	55875	1	7	26	0,0000004534
26	0,0052167245	ersetzen	12185	1	9	27	0,0000004281
27	0,0051698974	benutzen	12133	1	9	28	0,0000004261
29	0,0050127510	unterschiedlich	23768	2	8	29	0,0000004218
41	0,0028810615	erleichtern	7122	1	10	30	0,0000004045
23	0,0071939106	leiden	18031	1	8	31	0,0000003990
22	0,0082352234	Bedeutung	21767	1	8	32	0,0000003783
20	0,0090570836	häufig	24165	1	8	33	0,0000003748
12	0,0144813889	gelingen	41801	1	7	34	0,0000003464
21	0,0085282435	Zum	49252	2	7	35	0,0000003463
37	0,0039692728	genauso	13345	1	9	36	0,0000002974
30	0,0049472392	darstellen	17012	1	8	37	0,0000002908
43	0,0026717328	empfinden	10311	1	9	38	0,0000002591
33	0,0044009543	lernen	35090	2	7	39	0,0000002508
25	0,0058339721	bestimmt	23736	1	8	40	0,0000002458
14	0,0131095110	handeln	53666	1	7	41	0,0000002443
17	0,0114380919	oft	49423	1	7	42	0,0000002314
42	0,0026859313	erforderlich	12011	1	9	43	0,0000002236
32	0,0047208125	selten	23350	1	8	44	0,0000002022
31	0,0049339996	schwierig	24683	1	8	45	0,0000001999
28	0,0051125479	kaufen	36380	1	7	46	0,0000001405
38	0,0038600065	helfen	39484	1	7	47	0,0000000978

Computerspiele – Originaltext / Worthäufigkeitsklassen / Lemmatisiert

Warum haben wir ein Buch über Computerspiele geschrieben? Die einfachste Antwort lautet: Weil es sie gibt. Und zwar schon so lange, daß es verwunderlich ist, wie wenig darüber geschrieben wurde. Zumindest in Deutschland, wo sich die Auseinandersetzung mit Computerspielen in pädagogischen Abhandlungen über die Wirkung von Gewalt am Computer erschöpft. Gegen diese einschränkende Diskussion mußte unbedingt eine Beschreibung und Analyse der Ausbreitung und Faszination von Spielen gesetzt werden.
Vielleicht kann ein solches Buch erst jetzt erscheinen, weil diejenigen, die Erfahrung mit der Materie haben, erst erwachsen werden mußten. Die Rückbesinnung fängt an, nun wird klar, was man früher getan hat und welche Wirkungen das auf die eigene Entwicklung hatte. Man kann dieses Buch auch als den Versuch verstehen, die Erziehung, die man durch Computerspiele erhalten hat, nachzuzeichnen und zu erklären.
Vor allem aber sollen die Spiele endlich zu ihrem Recht kommen. Die Kapitel orientieren sich deshalb hauptsächlich an einzelnen Klassikern. Die Auswahl der Spiele ist dabei unvermeidlich subjektiv. Andere hätten sicherlich andere Spiele ausgewählt und werden an dieser Auswahl einiges zu bemängeln haben. Aber genau darum geht es, eine solche Auseinandersetzung muß überhaupt begonnen werden. Über Spiele und ihre Bewertung gibt es noch keine öffentliche Diskussion. Das muß sich ändern.

Wort	h(w)	c(w)	Wort	h(w)	c(w)	Wort	h(w)	c(w)
nachzuzeichnen	140	15	Diskussion	38280	7	gehen	291282	4
Rückbesinnung	461	13	Versuch	44271	7	ihre	300806	4
Abhandlung	585	13	Entwicklung	44887	7	schon	303960	4
Computerspiel	727	13	überhaupt	45241	7	all	334975	4
Ausbreitung	1056	12	verstehen	46128	7	erst	362491	4
verwunderlich	1399	12	darüber	46878	7	durch	373263	4
bemängeln	1637	12	erscheinen	48772	7	kommen	376645	4
Materie	1813	11	warum	50190	7	geben	411918	4
Faszination	2301	11	klar	50991	7	man	419503	4
subjektiv	2543	11	genau	52117	7	kein	466949	3
unvermeidlich	2780	11	lauten	53000	7	aber	502793	3
erschöpfen	2849	11	einige	58153	6	müssen	520472	3
auswählen	3370	11	erklären	60843	6	sollen	555827	3
Beschreibung	3397	11	deshalb	61820	6	so	559696	3
Klassiker	4087	10	Vor	63021	6	Das	576030	3
Bewertung	4443	10	Spiel	63243	6	über	598179	3
Zumindest	4729	10	Buch	63404	6	noch	672854	3
pädagogisch	4754	10	welche	64194	6	wie	683018	3
Erziehung	4872	10	Man	66495	6	am	717511	3
einschränken	5704	10	einfach	70540	6	können	821919	3
hauptsächlich	6503	10	öffentlich	78606	6	dies	827459	3
Kapitel	6629	10	schreiben	84727	6	daß	936255	2
orientieren	6941	10	beginnen	85374	6	an	1039533	2
sicherlich	6949	10	zwar	85684	6	es	1069022	2
Recht	7444	9	erhalten	86780	6	als	1080482	2
diejenige	8428	9	ihrem	87415	6	einen	1121613	2
Analyse	9007	9	dabei	91446	6	sie	1210878	2
Auswahl	9046	9	früh	95413	6	auch	1301212	2
fangen	11442	9	setzen	100768	6	Die	1421814	2
erwachsen	11916	9	ander	105336	6	auf	1608537	2
unbedingt	13260	9	solch	106046	6	sich	1718091	2
Wirkung	14257	9	wo	107113	6	mit	1861458	1
Computer	19419	8	tun	117561	5	das	1920541	1
Vielleicht	20492	8	Aber	143388	5	zu	2050372	1
darum	20503	8	nun	157859	5	von	2215196	1
Gegen	20710	8	eigen	166150	5	haben	2295270	1
Auseinandersetzung	21127	8	weil	167996	5	werden	2520779	1
Weil	25554	8	lang	182707	5	den	2697266	1
Antwort	25629	8	jetzt	187084	5	ein	3437822	1
Gewalt	27786	8	wenig	205941	5	in	4086574	0
Erfahrung	29115	7	Deutschland	211631	5	sein	4658755	0
ändern	32469	7	was	230040	5	und	5064441	0
endlich	33589	7	andere	249754	4	die	6782081	0
einzeln	34283	7	Und	251784	4	der	7367299	0
Über	36621	7	wir	278795	4			

warum haben wir ein Buch über Computerspiel schreiben Die einfach Antwort lauten Weil es sie geben Und zwar schon so lang daß es verwunderlich sein wie wenig darüber schreiben werden Zumindest in Deutschland wo sich die Auseinandersetzung mit Computerspiel in pädagogisch Abhandlung über die Wirkung von Gewalt am Computer erschöpfen Gegen dies einschränken Diskussion müssen unbedingt einen Beschreibung und Analyse der Ausbreitung und Faszination von Spiel setzen werden Vielleicht können ein solch Buch erst jetzt erscheinen weil diejenige die Erfahrung mit der Materie haben erst erwachsen werden müssen Die Rückbesinnung fangen an nun werden klar was man früh tun haben und welche Wirkung das auf die eigen Entwicklung haben Man können dies Buch auch als den Versuch verstehen die Erziehung die man durch Computerspiel erhalten haben nachzuzeichnen und zu erklären Vor all aber sollen die Spiel endlich zu ihrem Recht kommen Die Kapitel orientieren sich deshalb hauptsächlich an einzeln Klassiker Die Auswahl der Spiel sein dabei unvermeidlich subjektiv andere haben sicherlich ander Spiel auswählen und werden an dies Auswahl einige zu bemängeln haben Aber genau darum gehen es einen solch Auseinandersetzung müssen überhaupt beginnen werden Über Spiel und ihre Bewertung geben es noch kein öffentlich Diskussion Das müssen sich ändern

Computerspiele – ATA2-Rangfolge

Ursprünglicher Rang	Ursprüngliche Assoziationsstär	Wort	Korpushäufigkeit		Texthäufigkeit	Worthäufigkeitsklasse	Neuer Rang	Korrigierte Assoziationsstärke
1	0,6304462227	Computerspiel	727	3	13	1		0,0026015663
32	0,0568423531	Abhandlung	585	1	13	2		0,0000971664
8	0,1383662743	Materie	1813	1	11	3		0,0000763190
49	0,0235074241	Rückbesinnung	461	1	13	4		0,0000509922
16	0,1080384050	Faszination	2301	1	11	5		0,0000469528
7	0,1396840435	Beschreibung	3397	1	11	6		0,0000411198
56	0,0048213095	nachzuzeichnen	140	1	15	7		0,0000344379
6	0,1451365110	Erziehung	4872	1	10	8		0,0000297899
13	0,1257479229	Bewertung	4443	1	10	9		0,0000283025
27	0,0679361600	subjektiv	2543	1	11	10		0,0000267150
44	0,0347157625	verwunderlich	1399	1	12	11		0,0000248147
5	0,1744315219	Analyse	9007	1	9	12		0,0000193662
23	0,0855026884	Auswahl	9046	2	9	13		0,0000189040
2	0,3642641669	Computer	19419	1	8	14		0,0000187581
36	0,0509102684	erschöpfen	2849	1	11	15		0,0000178695
14	0,1172131155	Kapitel	6629	1	10	16		0,0000176819
34	0,0549231391	Klassiker	4087	1	10	17		0,0000134385
22	0,0909049804	Wirkung	14257	2	9	18		0,0000127523
12	0,1298517175	Auseinandersetzung	21127	2	8	19		0,0000122925
55	0,0112986711	Ausbreitung	1056	1	12	20		0,0000106995
37	0,0502822512	pädagogisch	4754	1	10	21		0,0000105768
48	0,0294751399	auswählen	3370	1	11	22		0,0000087463
54	0,0137843929	bemängeln	1637	1	12	23		0,0000084205
52	0,0206771598	unvermeidlich	2780	1	11	24		0,0000074378
46	0,0332099616	Zumindest	4729	1	10	25		0,0000070226
10	0,1330647551	Diskussion	38280	2	7	26		0,0000069522
3	0,2785101274	Entwicklung	44887	1	7	27		0,0000062047
45	0,0342814990	einschränken	5704	1	10	28		0,0000060101
28	0,0655413049	Über	36621	3	7	29		0,0000053692
41	0,0387332906	Recht	7444	1	9	30		0,0000052033
20	0,0993828837	darum	20503	1	8	31		0,0000048472
11	0,1329435573	Gewalt	27786	1	8	32		0,0000047846
47	0,0321646736	orientieren	6941	1	10	33		0,0000046340
4	0,1908253524	Versuch	44271	1	7	34		0,0000043104
43	0,0357389930	diejenige	8428	1	9	35		0,0000042405
33	0,0559846573	unbedingt	13260	1	9	36		0,0000042221
18	0,1050085114	Antwort	25629	1	8	37		0,0000040973
15	0,1157403426	Erfahrung	29115	1	7	38		0,0000039753
9	0,1337988375	einzeln	34283	1	7	39		0,0000039028
39	0,0401121419	erwachsen	11916	1	9	40		0,0000033662
51	0,0209874897	hauptsächlich	6503	1	10	41		0,0000032274
40	0,0400139301	Weil	25554	2	8	42		0,0000031317
29	0,0621433901	Vielleicht	20492	1	8	43		0,0000030326
53	0,0195164713	sicherlich	6949	1	10	44		0,0000028085
38	0,0499963459	Gegen	20710	1	8	45		0,0000024141
19	0,1036441939	erscheinen	48772	1	7	46		0,0000021251
50	0,0231040186	fangen	11442	1	9	47		0,0000020192
17	0,1050587144	genau	52117	1	7	48		0,0000020158
21	0,0918685279	darüber	46878	1	7	49		0,0000019597
25	0,0789802995	verstehen	46128	1	7	50		0,0000017122
24	0,0852454077	warum	50190	1	7	51		0,0000016985
26	0,0736181291	überhaupt	45241	1	7	52		0,0000016272
35	0,0539126561	endlich	33589	1	7	53		0,0000016051
31	0,0601784307	klar	50991	1	7	54		0,0000011802
30	0,0603993733	lauten	53000	1	7	55		0,0000011396
42	0,0365091984	ändern	32469	1	7	56		0,0000011244

Abbildungs- / Tabellenverzeichnis

Text- / Quelltextverzeichnis

Literaturverzeichnis

[PBMW1998]: Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 1998, http://www-db.stanford.edu/~backrub/pageranksub.ps

[LUH1958]: H.P. Luhn, The Automatic Creation of Literatur Abstracts, April 1958

[FER2003]: Reginald Ferber, Information Retrieval, Heidelberg, 2003

[FIIS2003]: Fraunhofer Institut für Integrierte Schaltungen, Query by Humming, 2003, http://www.emt.iis.fraunhofer.de/projekte_themen/q

[SMG1983]: Gerard Salton, Michael J. MacGill, Introduction to modern information retrieval, New York, 1983

[FWR1995]: Reginald Ferber, Manfred Wettler und Reinhard Rapp, An Associative Model of Word Selection in the Generation of Search Queries, 1995

[SHA1951]: Claude E. Shannon, Prediction and Entropy of Printed English, 1951

[GAL1880]: F. Galton, Psychometric experiments, 1880

[ZIM1992]: Philip G. Zimbardo, Psychologie, Berlin, Heidelberg, 1992

[JAM1890]: William James, The principles of psychology, New York, 1890

[FOP1965]: Klaus Foppa, Lernen, Gedächtnis, Verhalten, Köln, 1965

[HEB1949]: Donald Olding Hebb, The Organization of Behavior, New York, 1949

[BÖH1997]: Michael Böhnisch, Vorhersage menschlicher Wortassoziationen, Dezember 1997

[WRF1993]: Manfred Wettler, Reinhard Rapp und Reginald Ferber, Freie Assoziationen und Kontiguitäten von Wörtern und Texten, 1993

[SEI2003]: Petra Seidensticker, Stimulus Sampling Theorie, 2003, http://psycho1.uni-paderborn.de/karas/

[RW1991]: Reinhard Rapp und Manfred Wettler, Prediction of Free Word Associations Based on Hebbian Learning, 1991

[RAP1996]: Reinhard Rapp, Die Berechnung von Assoziationen: ein [...], Hildesheim; Zürich; New York, 1996

[CH1989]: Kenneth War Church und Patrick Hanks, Word association norms, mutual information, and lexicography, Philadelphia, 1989

[WR1993]: Manfred Wettler und Reinhard Rapp, Associative text analysis of advertisements, 1993

[LRW1996]: Wolfgang Lezius, Reinhard Rapp & Manfred Wettler, A Morphology-System and Part-of-Speech Tagger for German, Berlin, 1996

[LEZ1999]: Wolfgang Lezius, Morphy - Morphologie und Tagging für das Deutsche, 1999, http://www.lezius.de/wolfgang/morphy/

[WIR1975]: Niklaus Wirth, Algorithmen und Datenstrukturen, 1975-199

[TSC2000]: Ing.Büro R.Tschaggelar & Giacomo Policicchio, Balanced Binary Trees, 2000, http://www.ibrtses.com/delphi/binarytree.html

[LEV1965]: V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, 1965

[QUA1998]: U. Quasthoff, Deutscher Wortschatz im Internet, 1998, http://wortchatz.uni-leipzig.de/

[CSS2004]: Bert Bos, CSS contact, Cascading Style Sheets, 1997-2004,

http://www.w3.org/Style/CSS/
[DOM2003]: Philippe Le Hégaret, Document Object Model (DOM), 1997-2003,
http://www.w3.org/DOM/
[PHP2004]: Stig Sæther Bakken et al., PHP Handbuch (Deutsche Übersetzung), 1997-2004,
http://www.php.net/manual/de/
[SQL2004]: MySQL AB, MySQL - The World's Most Popular Open Source Database, 1995-2004,
http://www.mysql.com/
[APA2004]: The Apache Software Foundation, Apache HTTP Server Project, 1999-2004,
http://httpd.apache.org/
[STE2003]: Ian Stewart, Flacherland, München, 2003
[AND2001]: John R. Anderson, Kognitive Psychologie, Heldelberg; Berlin, 2001

Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig angefertigt und mich fremder Hilfe nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß veröffentlichtem oder unveröffentlichtem Schrifttum entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Paderborn, 30.07.2004

Gero Zahn