Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

DEPARTMENT INFORMATION

Masterarbeit

Suchmaschinen und Sprache: Eine Studie über den Umgang von Google und BING mit den Besonderheiten der deutschen Sprache

vorgelegt von
Alexandra Gather

Studiengang Informationswissenschaft und -management

erster Prüfer: Prof. Dr. Dirk Lewandowski
zweite Prüferin: Prof. Dr. Ulrike Spree

Hamburg, Oktober 2013

Abstract

In dieser Studie über „Suchmaschinen und Sprache“ wird untersucht, wie die beiden Suchmaschinen Google und BING mit den Besonderheiten „Synonymie, Morphologie und Homographie“ in der deutschen Sprache umgehen. Synonyme Begriffe und morphologische Varianten einer Suchanfrage sollten im Idealfall zu denselben Ergebnissen führen. Für Homographie sollten Suchmaschinen dem Nutzer Hilfestellungen für die richtige Bedeutungsfindung, bzw. ihre Ergebnisse gemäß den Nutzerinteressen anbieten.

Mindestens 50 Testanfragen wurden für jede der untersuchten sprachlichen Besonderheiten an Google und BING gestellt. Für „Morphologie“ und „Synonymie“ wurden die Übereinstimmungen der Ergebnisse bei Eingabe zweier zusammengehöriger Anfragen erhoben. Im Bereich „Homographie“ wurden die Bedeutungsverhältnisse innerhalb der organischen Treffer mit denen der organischen Klickdaten der T-Online-Suche korreliert.

Die durchgeführte Studie zeigt, dass Google und BING die deutsche Sprache im Bereich Synonymie und Morphologie nicht technisch verarbeiten, da die entsprechenden Übereinstimmungen gering und nicht signifikant sind. Bei Suchanfragen mit Homographen korrelieren die Bedeutungsverhältnisse in den Ergebnislisten durchaus positiv mit den Klickpräferenzen der Nutzer. Allerdings ist die Korrelation nicht besonders hoch, und es gibt einige Ausreißer. Google bietet im Gegensatz zu BING einige Hilfestellungen zur Bedeutungsklärung. Keine der beiden meistgenutzten Suchmaschinen Deutschlands kann als eindeutiger Vorreiter auf irgendeinem der untersuchten Gebiete bezeichnet werden.

Schlagworte: Information Retrieval, Retrievalstudie, Retrievaltest, Suchmaschinen, BING, Google, Sprache, Deutsch, Homographie, Morphologie, Synonymie

Inhaltsverzeichnis

Anhangsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

HAW	Hochschule für Angewandte Wissenschaften
HTML	Hypertext Markup Language
IP	Internet Protokoll
MSN	Microsoft Network
RAT	Relevance Assessment Tool
SEO	Suchmaschinenoptimierung
URL	Uniform Resource Locater

1. Einleitung

Die meisten Internetnutzer, auch die deutschsprachigen, beginnen ihre Online-Zeit mit der Nutzung einer Suchmaschine (vgl. EIMEREN 2012). Sie halten dieses Werkzeug offensichtlich für nützlich. Die Qualität von Suchmaschinenergebnissen wird daher viel diskutiert und untersucht (vgl. LEWANDOWSKI 2011a). Suchmaschinen sind aufgrund ihres Entstehungslandes nicht für die Nutzung in deutscher Sprache entwickelt worden (vgl. GOOGLE 2013a, DETTWEILER 2009). Die deutsche Sprache wird allerdings häufig im Internet verwendet, und zwar nicht nur von Muttersprachlern (vgl. EU 2011). Ziel dieser Arbeit ist es, die Qualität der Suchfunktion von Google und BING hinsichtlich ihres Umgangs mit den Besonderheiten der deutschen Sprache anhand von Synonymie (Begriffe bedeuten dasselbe), Morphologie (Grammatikalische Variationen eines Wortes) und Homographie (ein Begriff hat mehrere Bedeutungen) zu untersuchen. Daher wird nach einer Schilderung der Problemstellung und Einordnung der vorliegenden Studie in den Stand der Forschung das methodische Design erläutert, bevor die Beschreibung und Interpretation der Ergebnisse folgt.

2. Problemstellung

Das Kernproblem bei der technischen Verarbeitung von Sprache liegt darin, dass eine Nichtberücksichtigung bestimmter sprachlicher Faktoren in der Regel zu Informationsverlust und Mehraufwand für den Nutzer führt (vgl. STOCK 2007). Die zentrale Diskussion dieser Arbeit adressiert daher die Fragestellung, wie Suchmaschinen mit Synonymie, Morphologie und Homographie in der deutschen Sprache umgehen und wie die aktuelle Lage zu bewerten ist. Suchmaschinen gehören zu den TOP-Online-Anwendungen (vgl. EIMEREN 2012), wobei Google mit 95 % Marktanteil der klare Favorit der Deutschen ist und Microsofts Suchmaschine BING gerade einmal 2 % Marktanteil besitzt, aber immerhin noch vor Yahoo mit 1 % Marktanteil liegt (vgl. SCHMIDT 2012). Google und BING liefern in etwa ähnlich relevante Ergebnisse (vgl. GÜNTHER 2012), daher lässt sich ein starker Markeneffekt bei den Marktanteilen vermuten. Aufgrund der technischen Gleichwertigkeit und der Tatsache, dass sie die meist

benutzten Suchmaschinen in Deutschland sind, werden BING und Google in dieser Arbeit hinsichtlich ihrer Verarbeitung der deutschen Sprache miteinander verglichen. Die Suchmaschine Google gibt es seit 1998, sie ist seit einigen Jahren weltweiter Marktführer (vgl. LEVENE 2010). BING ist Microsofts Nachfolger von Live Search, bzw. MSN, die Suchmaschine ging 2009 online (vgl. LEVENE 2010). BING unterhält eine Kooperation mit Yahoo, in der eine gemeinsame Werbeplattform angeboten wird, durch die mit einem Kauf Anzeigen in BING und Yahoo geschaltet werden können (vgl. MICROSOFT 2013a). Durch die Kooperation verantwortet Yahoo die Weiterentwicklung der Werbeplattform, während Microsoft die Yahoo-Suche übernimmt (vgl. BBC 2009). Dies ist ein weiterer Grund, warum der Vergleich von Google mit Yahoo nicht unternommen wurde, da hinter Yahoos Suchtechnologie letztlich Microsofts Expertise steht.

3. Begriffe und Hintergrund

Als Retrievalsystem können unterschiedlichste Systeme bezeichnet werden, die dazu dienen, Informationen zu suchen und zu finden, etwa Bibliotheken, Datenbanken und Suchmaschinen (vgl. STOCK 2007). Die Präzision (Retrievaleffektivität) gibt Auskunft darüber, wie gut ein Retrievalsystem relevante Treffer gemäß dem jeweiligen Nutzerinteresse ausgibt (vgl. LEWANDOWSKI 2011a) und ist ein technischer Begriff, der im Folgenden in dieser Bedeutung verwendet wird. Um die Präzision eines Systems auf einer Skala zu erfassen, werden in der Regel menschliche Juroren eingesetzt, die die Relevanz einer bestimmten Trefferanzahl bewerten (vgl. LEWANDOWSKI 2011a). Für die durchgeführte Studie spielt die Relevanz eine untergeordnete Rolle, da der Fokus auf der Verarbeitung von sprachlichen Besonderheiten liegt (vgl. Kapitel 5 „Methodischer Rahmen“, S. 10).

Suchmaschinen nutzen sogenannte „Crawler“, die ausgehend von bekannten URLs (Uniform Resource Locator, Weblink*)* allen weiteren Verlinkungen folgen und auf diese Weise neue Websites finden, bzw. die Informationen zu bestehenden aktualisieren. Darauf werden die Websites inhaltlich und strukturell erschlossen, also indiziert. Bei einer Abfrage durchsucht ein entsprechendes technisches Modul der Suchmaschine nicht

das Internet, sondern den eigenen Index und gibt dem Ranking-Algorithmus entsprechend eine Liste mit Ergebnissen aus. Diese Ergebnisse spiegeln meist nicht einmal den gesamten Index der Suchmaschine wieder, denn die Suche nach passenden Dokumenten wird abgebrochen, wenn genügend Treffer erzielt worden sind.
(vgl. GRIESBAUM 2009, BAEZA-YATES 2011, LEWANDOWSKI 2005)

Der Algorithmus von Suchmaschinen berücksichtigt vielerlei Kriterien mit unterschiedlicher Gewichtung. Nach GRIESBAUM lassen sich die Rankingfaktoren in folgende Kategorien einteilen (vgl. GRIESBAUM 2009):

- On-Page-Faktoren: Häufigkeit, Format und Position des Inhalts der Suchanfragen auf der Website, HTML-Struktur
- On-Site-Faktoren: Alter und Größe einer Website, Domain etc.
- Link-Faktoren: Links auf eine Website gelten als Empfehlung
- Nutzer-Faktoren: Geografische Zuordnung der IP-Adresse, Suchsessions, Suchhistorie

Kommerziell orientierte Website-Betreiber manipulieren möglichst viele dieser Faktoren (vgl. SEW 2013), diese Beeinflussung wird unter dem Begriff Suchmaschinenoptimierung zusammengefasst. Alle diese Faktoren beeinflussen die vorliegende Studie, da das Ranking der Ergebnisse eben nicht nur auf rein sprachlichen Kriterien beruht.

Für das Verständnis der weiteren Ausführungen werden folgende Begrifflichkeiten geklärt (vgl. STOCK 2007):

- Worte, die in einer Sprache häufig vorkommen, aber selbst keinen Sinn tragen, bezeichnet man als Stoppworte (z.B.: und, der, die, das, oder).
- Morphologie ist die Wissenschaft, die sich mit unterschiedlichen Wortformen beschäftigt, die durch Zusammensetzung (z.B. Tanne & Zapfen = Tannenzapfen), Beugung (z.B. der Mann, die Männer) oder Ableitung (z.B. Freiheit von frei) entstehen. Die Zusammensetzung (Komposition) kommt besonders in der deutschen Sprache häufig vor (z.B. Kapitänsmütze). Für die Berechnung der Relevanz eines Dokumentes ist eine Zerlegung der Komposita oft zwar

sinnvoll, kann aber auch zu Informationsballast führen. Für die Zerlegung existieren verschiedene Verfahren: zum einen wörterbuchbasierte, die einen sehr hohen Aufwand mit sich bringen, und zum anderen statistische Verfahren, die das Problem von unsinnigen Zerlegungen aufwerfen (z.B. Wellensittich = Welle & Sittich).

- Das Thema Wortsegmentierung ist vor allem für Sprachen schwierig, in denen Wörter nicht durch Leerzeichen voneinander getrennt werden, wie z.B. im Chinesischen (vgl. LIU 2012).
- Ein Thesaurus ist ein Nachschlagewerk zu Begrifflichkeiten und den Beziehungen dieser. Er verweist z.B. auf Ober-, Unterbegriffe und Synonyme. Thesauri können für den Nutzer versteckt im Hintergrund einer Suche mitlaufen und dadurch auch Ergebnisse für synonyme Begriffe berücksichtigen.
- Die Semantik beschäftigt sich mit der Bedeutung von Begriffen. Semantische Wortfelder können also bedeutungsähnliche Begriffe aufzeigen.

4. Stand der Forschung

Mit dem Umgang von Suchmaschinen mit der deutschen Sprache haben sich bisher nur wenige Studien beschäftigt (vgl. GUGGENHEIM 2005, SCHMETZKE 1998). SCHMETZKE untersuchte den Umgang von Suchmaschinen mit den Umlauten der deutschen Sprache und stellte gravierende Mängel fest (vgl. SCHMETZKE 1998). BAR-ILAN und GUGGENHEIM befanden einige Jahre später, dass Sonderzeichen keine besondere Herausforderung mehr für Suchmaschinen darstellen (vgl. GUGGENHEIM 2005). Diese vorangegangenen Studien zur deutschen Sprache nutzten Abfragen, die ein bestimmtes sprachliches Problem provozieren, und bewerteten dann die Trefferanzahlen, sowie teilweise die Ähnlichkeit der ersten zehn Treffer bei Eingabe zweier Varianten.

Auf internationaler Ebene gibt es einige Studien, die verschiedene Suchmaschinen im Hinblick auf die technische Verarbeitung von typischen, sprachlichen Problemen untersuchten. Mit der Tauglichkeit von Suchmaschinen bei der Handhabung anderer Schriftsysteme beschäftigten sich LAZARINIS und EFTHIMIADIS für Griechisch (vgl. LAZARINIS 2007a,

EFTHIMIADIS 2008), MOUKDAD für Arabisch (vgl. MOUKDAD 2001, MOUKDAD 2004), sowie MOUKDAD und LIU für Chinesisch (vgl. MOUKDAD 2005, LIU 2012). Weitere Retrievaltests bezüglich sprachlicher Besonderheiten wie Morphologie und Sonderzeichen bzw. Umlauten finden sich für Polnisch (vgl. SROKA 2000, CHORO 2005), Finnisch (vgl. KETTUNEN 2012), Türkisch (vgl. DEMIRCI 2007), Malaysisch (vgl. HALIM 2006), Ungarisch (vgl. BAR-ILAN 2004, TOTH 2006), sowie Hebräisch, Russisch und Französisch (vgl. BAR-ILAN 2004). Einen Überblick über allgemeine Studien zum Thema Sprache und „Information Retrieval" gibt LAZARINIS (vgl. LAZARINIS 2009).

In den genannten internationalen Studien lässt sich eine gewisse Übereinstimmung in der methodischen Vorgehensweise feststellen. Abgefragt werden stets Begriffe, die ein spezielles linguistisches Problem beinhalten. Am häufigsten wird die Trefferanzahl bei unterschiedlichen Schreibweisen verglichen, sowie die Präzision (Retrievaleffektivität). Teilweise wird die Übereinstimmung innerhalb einer gewissen Trefferanzahl betrachtet. Vereinzelt werden weitere Faktoren bewertet, wie die gelungene Wortsegmentierung bei chinesischen Suchanfragen und Ergebnissen (vgl. MOUKDAD 2005). BAR-ILAN betrachtete zusätzlich zur Trefferanzahl das Auftauchen morphologisch veränderter Formen etc. auf den jeweils ersten 10 Websites der Suchanfrage (vgl. BAR-ILAN 2004) und verzichtete bewusst auf eine Relevanzbewertung. TOTH untersuchte in ihrer Studie, inwieweit Synonyme in Trefferbeschreibungen auftauchen oder Stoppworte markiert sind (vgl. TOTH 2006).

Wie LAZARINIS in seinem Überblick zu fremdsprachigen Retrievalstudien feststellt, mangelt es internationalen Suchmaschinen technisch häufig am Verständnis der jeweiligen Morphologie der Sprache (vgl. LAZARINIS 2009). Lokale Suchmaschinen sind im Retrieval dennoch größtenteils schlechter als internationale. Diese Ergebnisse decken sich mit den zuvor genannten, internationalen Studien. Ausnahmen stellen chinesische und ungarische Suchmaschinen dar (vgl. LAZARINIS 2009, TOTH 2006), bei denen lokale Suchmaschinen besser abschneiden. Global gesehen sind Googles Marktanteile in China und Russland vergleichsweise klein (vgl. SCHMIDT 2012), was die Vermutung nahelegt, dass Google auch mit der russischen

Sprache Schwierigkeiten hat. Google schneidet insgesamt in den Retrievalstudien im Vergleich zu anderen Suchmaschinen meist gut und häufig am besten ab, zu BING liegen bisher keine Erkenntnisse vor.

Methodisch ist an den bisher durchgeführten Studien zu kritisieren, dass sich die Studien überwiegend auf die Messung von Präzision und Trefferanzahl beschränkt haben, anstatt ihren Fokus auf den tatsächlichen Umgang mit den sprachlichen Besonderheiten zu legen. Zudem ist die Anzahl der Suchanfragen in den genannten Studien – wenn überhaupt erwähnt – in fast allen Fällen sehr gering (unter fünfzig), entgegen der gängigen Praxis (vgl. LEWANDOWSKI 2011a).

Unabhängig von dem Umgang mit sprachlichen Hürden gibt es für den deutschsprachigen Suchmaschinenraum einige Studien, die die Retrievaleffektivität von Suchmaschinen für deutsche Suchanfragen unter vielerlei Gesichtspunkten untersuchen und generell zu zufriedenstellenden Ergebnissen kommen (vgl. GRIESBAUM 2002, GRIESBAUM 2004, LEWANDOWSKI 2008b, LEWANDOWSKI 2009, LEWANDOWSKI 2011b, GÜNTHER 2012). Diesen Studien mangelt es aber an der sprachlichen Perspektive ihrer Suchanfragen. So wird in ihnen nicht reflektiert, wie sich z.B. unterschiedliche morphologische Varianten einer Anfrage auf Trefferlisten auswirken oder Ähnliches. Google schneidet durchweg sehr positiv im Vergleich ab. BING zeigt in GÜNTHER 2012 eine durchaus mit Google gleichwertige Präzisionsleistung, und auch der Vorgänger MSN ist technisch in etwa gleichwertig zu Google (vgl. LEWANDOWSKI 2009, LEWANDOWSKI 2011b), in LEWANDOWSKI 2008b enttäuschte MSN jedoch. Aufgrund der zeitlichen Abstände zwischen diesen Studien ist eine aktuelle technische Gleichwertigkeit von Google und BING dadurch nicht widerlegt. Zudem können Studien, die mehr als zwei Jahre zurückliegen, durch den „stetigen Wandel“ (LEWANDOWSKI 2011c, S. 1) in der Suchmaschinenbranche als veraltet angesehen werden.

Über den technischen Hintergrund von Suchmaschinen und ihren Umgang mit Sprachproblemen lässt sich einiges aus dem Information Retrieval lernen. Information Retrieval befasst sich mit dem technischen Prozess, wie Informationen auf eine bestimmte Anfrage hin an den Nutzer

ausgeliefert werden können (vgl. NOHR 2003). Folgende sprachliche Schwierigkeiten sind im Information Retrieval bekannt (vgl. STOCK 2007):

- **Schriftsystem erkennen**, sowie Schreib- und Leserichtung
- **Sprache erkennen**, z.B. durch die Häufigkeiten von sprachspezifischen Buchstabenfolgen
- **Stoppworte**: Stoppworte sollten beim Retrieval nicht beachtet werden, weil ihre Häufigkeit die Rankingergebnisse verschlechtert.
- **Morphologie**: Techniken, die das Retrieval morphologischer Formen verbessern, sind Lemmatisierung und Stemming. Bei der Lemmatisierung werden Worte linguistisch analysiert und auf ihren linguistischen Stamm zurückgeführt. Beim Stemming werden automatische Regeln benutzt (z.B. Abtrennung häufiger Endungen), um zu Stammformen zu gelangen. Eine perfekte Lösung für den Umgang mit Morphologie gibt es allerdings nicht.
- **Homographie**: Homographe sind Worte, die mehrere Bedeutungen haben (z.B. Laster = LKW oder Charakterfehler). Werden Homographe von einem Retrievalsystem nicht aufgelöst, muss der Nutzer selbst seine Suchanfrage präzisieren, um die gewünschten Ergebnisse zu erhalten. Hinterlegte Thesauri oder eine automatische Analyse der umgebenden Worte (z.B. Laster – Fahrzeug – Autobahn bzw. Laster – Schwäche – Ethik) können der Auflösung von Homographen dienen.
- **Synonymie**: Als Synonyme bezeichnet man unterschiedliche Worte, die dasselbe bedeuten (z.B. Orange & Apfelsine). Werden Synonyme beim Information Retrieval nicht beachtet, gehen Informationen verloren, da beispielsweise nur Dokumente ausgegeben werden, die das Stichwort „Orange“ enthalten und keines von denen, die „Apfelsine“ enthalten. Werden allerdings zu viele Synonyme im Retrieval mit einander gleichgesetzt, kommt es zu Informationsballast. Ähnlich wie bei Homographie können Synonyme durch hinterlegte Thesauri oder die Bildung von Wortfeldern aus umgebenden Worten erkannt und aufgelöst werden.

- **Phrasen**: Phrasen setzen sich aus mehreren einzelnen Worten zusammen. Wenn Phrasen von einem Retrievalsystem nicht erkannt werden, besteht die Gefahr des Informationsverlusts. So könnte die Suche nach „Deutsche Demokratische Republik“ hauptsächlich Treffer ausgeben, in denen die Worte „deutsch“, „demokratisch“ und „Republik“ als solches vorkommen, und folglich wären Informationen über die DDR allenfalls rein zufällig zu finden.
- **Begriffsrelationen**: Begriffsrelationen bezeichnen semantische Felder von Begriffen (z.B. Ober- und Unterbegriffe). Hinweise auf solche Begriffsbeziehungen ermöglichen es dem Nutzer, die Suchanfrage zu präzisieren oder zu erweitern, bzw. auf weitere Suchmöglichkeiten aufmerksam gemacht zu werden. Auch dafür können hinterlegte Thesauri dienen.
- **Namen**: Eigennamen sind wegen unterschiedlicher Schreibweisen problematisch. Auch hier kann dem Informationsverlust nur begegnet werden, wenn dem System diese unterschiedlichen Schreibweisen bekannt sind und somit nicht nur nach genau dem eingegebenen Namen gesucht wird.

Es gibt, wie bereits angedeutet, eine Reihe von Techniken im Information Retrieval, die sich mit diesen sprachlichen Problemen auseinandersetzen. Verschiedene Studien beschäftigen sich in diversen Sprachen mit einer Bewertung bzw. Verbesserung dieser Techniken (vgl. u.a. TURNEY 2001, KARANIKOLAS 2009, SPIEGEL 2011, LAZARINIS 2007b, LAZARINIS 2008, ALPKOÇAK 2012, SAVOY 2008, SHATNAWI 2012, LIU 2012, HAMMO 2009, WANG 2007, ALOUFI 2010, LETURIA 2012).

Sogenannte „Semantische Suchmaschinen“ wurden dazu entwickelt, sprachliche Herausforderungen zu meistern, aber existierende Semantische Suchmaschinen weisen große Mängel auf und sind kommerziellen Anbietern bisher unterlegen (vgl. SPREE 2011), weswegen von einer Bewertung Semantischer Suchmaschinen in dieser Studie abgesehen wird.

Suchmaschinen scheinen einigen dieser sprachlichen Probleme mit ihren Techniken nicht gewachsen zu sein. So bemerkt GRIESBAUM:

> „Morphologische und syntaktische Verfahren der Textanalyse, wie z.B. Grund- und Stammformreduktion, Kompositazerlegung oder die Erkennung von Mehrwortbegriffen, finden derzeit meist keine Anwendung.“ (GRIESBAUM 2009, S. 32-33)

Auf Googles Hilfeseiten wird gesagt, dass Stoppwort-Listen existieren und bis zu 40 Sprachen unterstützt werden (vgl. GOOGLE 2013b, GOOGLE 2013c). Auch verspricht Google, dass Stemming gemacht wird und Synonyme in die Suche einbezogen werden (vgl. GOOGLE 2013d). BINGs Hilfeseiten geben weniger Aufschluss über diese technischen Funktionen. Zumindest erklärt MICROSOFT, dass es Stoppwort-Listen gibt und dass man am besten selbst nach Synonymen sucht, um passende Informationen zu finden (vgl. MICROSOFT 2013b). Letzteres lässt vermuten, dass eine generelle Einbeziehung von Synonymen bei BING eher nicht stattfindet. Ähnlich wie Google bietet BING die Suchfunktion in bis zu 40 Sprachen an (vgl. MICROSOFT 2013c). Generell ist es wegen der Manipulationsgefahr unter Suchmaschinenbetreibern üblich, sich in Bezug auf angewendete Verfahren bedeckt zu halten (vgl. LEWANDOWSKI 2005). Gerade bezüglich technischer Verfahren zur Reduktion morphologischer Varianten stellt sich die Frage, inwiefern solche Reduktionen im Web Retrieval förderlich sind. Diese Techniken müssen sprachspezifisch entwickelt werden, was internationale Suchmaschinen höchstens für die wichtigsten Sprachen tun können (vgl. LEWANDOWSKI 2005). Zudem ist die Erhöhung der Präzision im Web fraglich, da in der Regel zu viele Informationen zu einem Thema existieren und somit relevante Informationen gefunden werden, egal, welche morphologische Form eines Wortes eingegeben wurde (vgl. LEWANDOWSKI 2005).

Der Stand der Forschung zeigt, dass sich bisher nur wenige, bereits etwas ältere Studien mit dem Umgang von Suchmaschinen mit der deutschen Sprache beschäftigt haben. Aus dem Information Retrieval ist zudem eine Vielzahl sprachlicher Hürden für Retrievalsysteme bekannt. Während Google von sich selbst eine gewisse sprachliche Verarbeitung verschiedener Sprachprobleme behauptet, hält BING sich in dieser Hinsicht mehr zurück.

5. Methodischer Rahmen

Methodisch orientiert sich die durchgeführte Studie nur leicht an den vorherigen Studien in diesem Bereich. Übernommen wird der Gedanke, Testanfragen zu konzipieren, die ein bestimmtes sprachliches Problem provozieren, sowie die ansatzweise geschehene Bewertung der Ähnlichkeit von Treffern bei Eingabe eines Testwortpaares (vgl. GUGGENHEIM 2005). Anders als bisherige Studien zum Umgang von Suchmaschinen mit der deutschen Sprache wird die aktuelle Studie sich nicht mit der Frage von Umlauten, alter und neuer Rechtschreibung oder scharfem und Doppel-S beschäftigen, da GUGGENHEIM bereits feststellte, dass die damit verbundenen Probleme größtenteils behoben sind (vgl. GUGGENHEIM 2005). Von einer Betrachtung des Umgangs mit Begriffsrelationen wird abgesehen, da diese zu starke Mess- und Bewertungsschwierigkeiten hervorrufen.

Stattdessen ist die durchgeführte Studie auf die Untersuchung folgender sprachlicher Besonderheiten ausgerichtet (vgl. STOCK 2007):

- **Synonymie**: Mehrere Worte meinen dasselbe. Beispiel: Sauna/Dampfbad
- **Morphologie**: Grammatikalische Veränderungen eines Wortes. Beispiel: Lamm/Lämmer, Glaskanne/Kanne aus Glas
- **Homographie** (Polysemie): Ein Wort hat mehrere Bedeutungen. Beispiel: Jaguar (Auto/Raubkatze)

In Lehrbüchern zum Information Retrieval ist die technische Handhabung dieser sprachlichen Schwierigkeiten ein bekanntes und viel behandeltes Thema (vgl. STOCK 2007, NOHR 2003, BAEZA-YATES 2011, MANNING 2010, CROFT 2010, BÜTTCHER 2010, FERBER 2003, CHOWDHURY 2008). Auf der Suche nach Standards, wie eine optimale Ergebnisliste auszusehen hat, in der diese sprachlichen Probleme gelöst sind, stößt man in den genannten Lehrbüchern lediglich auf indirekte Hinweise.

Aus diesen Hinweisen lassen sich Hypothesen ableiten:

- Wenn eine Suchmaschine Synonyme technisch berücksichtigt, dann ergeben sich bei Eingabe der synonymen Wörter dieselben Treffer.
- Wenn eine Suchmaschine morphologische Varianten eines Wortes technisch zusammenführt, dann ergeben sich dieselben Treffer bei der Eingabe zweier unterschiedlicher morphologischer Formen.
- Wenn das Problem „Homographie“ von der Suchmaschine behandelt wird, gibt es direkte Hinweise auf verschiedene Bedeutungsmöglichkeiten (Disambiguierung). Zudem orientiert sich das Verhältnis der Bedeutungen in den Treffern an den gemittelten Nutzerinteressen. Im Idealfall wäre es denkbar, dass eine Suchmaschine aufgrund des Suchprofils die Treffer hinsichtlich einer bestimmten Bedeutung anpasst, Datenschutzbedenken einmal außer Acht gelassen. Bisher versuchen Suchmaschinen anscheinend, ihre Trefferliste bei mehrdeutigen Suchanfragen zu durchmischen, ob bewusst in einem bestimmten Maß, ist allerdings unklar (vgl. LEWANDOWSKI 2007).

Es gibt auch Gegenstimmen zu diesen Hypothesen. So bemerkt u.a. LEVENE zur Handhabung von Synonymie, dass im Web einiges dafür spricht, eine Suche nach Synonymen nur auf ausdrücklichen Wunsch der Nutzer zu ermöglichen, da die Präzision darunter leidet, wenn zu viele Synonyme von der Suchmaschine berücksichtigt werden (vgl. LEVENE 2010, SMITH 2010, STACEY 2004). Google bietet mit der „Tilde“ einen Suchoperator, der speziell auch Synonyme in die Suche einschließen soll (vgl. GOOGLE 2013e), während BING keinen Operator für die Synonymsuche bietet (vgl. MICROSOFT 2013d). Das Vorhandensein dieses Operators bei Google ist verwunderlich, da Google wie bereits erwähnt behauptet, generell Synonyme in Suchanfragen einzubeziehen (vgl. GOOGLE 2013d). Ein kurzer Test mit fünf Begriffen der in der weiteren Analyse verwendeten Synonym-Testworte ergab kein eindeutiges Ergebnis bzgl. der Funktionsweise der „Tilde“. Bei vier der getesteten Begriffe ergaben sich durch die Nutzung der Tilde durchaus Unterschiede (O), bei

einem blieben die untersuchten Treffer unverändert, gleichgültig ob mit oder ohne Eingabe der Tilde vor dem Suchbegriff (X) (vgl. Tabelle 1):

Tabelle 1: Wirkung des Operators "Tilde" (eig. Darst.)

Begriff	Treffer 1-10	Treffer 91-100
Gatte	O	O
Couch	O	O
Vorgesetzter	X	X
Feind	O	O
Dessert	O	O

Eine speziell auf dieses Thema ausgerichtete Studie zur Überprüfung angebotener Operatoren könnte u.a. genaueren Aufschluss über die Funktionstüchtigkeit der Tilde geben. Für die durchgeführte Studie wurde der Operator „Tilde“ nicht verwendet, da Googles Aussage, Synonyme generell mit einzubeziehen, getestet werden sollte und ein Vergleich mit einem ähnlichen Operator bei BING nicht möglich war. Zudem gibt einem die Eingabe der „Tilde“ keine Kontrolle darüber, welche Synonyme in die Suche einbezogen werden, deswegen wird generell empfohlen, Synonyme einzeln manuell abzufragen (vgl. STACEY 2004).

Auch in Bezug auf Morphologie ist es fraglich, ob ein Einbeziehen mehrerer morphologischer Varianten im Internet die Präzision überhaupt erhöht (vgl. LEWANDOWSKI 2005). Diese sehr viel grundsätzlichere Frage, inwieweit Suchmaschinen den Anforderungen des Information Retrievals bezüglich des Umgangs mit Morphologie und Synonymie entsprechen können, oder ob aufgrund der Struktur des Internets andere Maßstäbe gelten sollten, ist nicht Gegenstand der hier bearbeiteten Thematik. Als Bewertungsgrundlage werden die abgeleiteten Hypothesen der Lehrbücher des Information Retrievals herangezogen.

5.1 Kennzahlen

Die auf den Lehrbüchern zum Information Retrieval basierenden Hypothesen (siehe S. 11) bestimmen die Kennzahlen, die in der Studie erhoben wurden.

Die Kennzahlen für die Auswertung des Umgangs mit „Synonymie“ und „Morphologie“ basieren auf einem URL-und Domain Vergleich; die Ermittlung wurde im ersten Durchgang abhängig von der Trefferposition, im zweiten, manuellen Durchgang auch unabhängig von der Trefferposition durchgeführt. So wurden die URLs und Domains für jedes Testwortpaar (z.B. „Baum“, „Bäume“) verglichen, und die Anteile gleicher bzw. unterschiedlicher URLs und Domains auf derselben Position, bzw. auf anderen Positionen festgehalten. Es handelt sich daher bei den erhobenen Kennzahlen um verschiedene Übereinstimmungsraten. Um diese in ihrer Bedeutung einschätzen zu können, wird zusätzlich überprüft, inwieweit Google und BING unabhängig voneinander arbeiten. Dazu werden die Übereinstimmungen der Treffer zwischen Google und BING erhoben. Zeigt sich eine geringe Übereinstimmung, handeln die Algorithmen unabhängig und Unterschiede in der Handhabung von Synonymie werden signifikanter. Es wird keine hohe Übereinstimmung der Treffer zwischen Google und BING erwartet, da belegt ist, dass es zwischen großen Suchmaschinen im Allgemeinen wenige Übereinstimmungen und somit Unterschiede in den Algorithmen gibt (vgl. SPINK 2006, RATHER 2008), auch wenn diese Studien bereits veraltet sind.

Für die Bewertung des Umgangs der Suchmaschinen mit „Homographie“ wurden die Bedeutungsverhältnisse der organischen Treffer mit den Klickpräferenzen der Nutzer korreliert. Dafür wurden zunächst die Treffer für eine Suchanfrage (z.B. „Hamburger“) in diverse Kategorien eingeteilt, die mit einem Code betitelt wurden. Für den weiteren Verlauf am wichtigsten sind die Kategorie „1“ (z.B. „Hamburger = Einwohner/Hamburg“) für die erste, häufiger in den organischen Treffern von Google und BING vorkommende Bedeutung, und die Kategorie „2“ (z.B. „Hamburger = Gericht“) für die zweithäufigste Bedeutungsvariante. Zusätzlich gibt es eine Kategorie für Eigennamen/Firmennamen, eine als Sammelpunkt für weitere

Bedeutungsmöglichkeiten, sowie Wörterbucheinträge und das Vorkommen beider Hauptbedeutungen. Die weiteren Kategorien neben den Hauptbedeutungen „1“ und „2“ sind während des Pretests entwickelt worden, um alle Vorkommnisse in den Treffern abzudecken. Die Einordnung der Bedeutung erfolgte mithilfe der Trefferbeschreibungen. Bei Unklarheit wurde die vom RAT gespeicherte Website aufgerufen.

Das Bedeutungsverhältnis zwischen den beiden häufigsten Bedeutungen in den organischen Treffern wurde schließlich mit demjenigen aus den Klickdaten der T-Online-Suche korreliert. Diese Korrelation sollte einen Anhaltspunkt dafür liefern, inwiefern das Bedeutungsverhältnis in den organischen Treffern den tatsächlichen Nutzerinteressen entspricht. Klicks dürfen zwar nicht als absolutes Relevanzurteil, aber als klare Äußerung einer Präferenz der Nutzer verstanden werden (vgl. JOACHIMS 2007). Für die Ermittlung der organischen Klickdaten wurden alle Anfragen, die z.B. das Wort „Rock“ enthalten, von der T-Online-Suche über den Zeitraum von Januar 2012 - Dezember 2012 abgefragt (max. 2000 Datensätze) und den beiden Hauptbedeutungen zugeordnet. Dabei wurden die Anfragen, wenn möglich, zu einer sinnvollen Bedeutung zusammengefasst (z.B. Ballspiel, Pinball, Fußball zählen zu „Spielgerät“ gegenüber Tanzball, Ballkleider zu „Tanzveranstaltung“). Alle Verhältnisse wurden für die weitere Verwendung prozentual festgehalten.

Kann eine Korrelation festgestellt werden, bedeutet dies keinesfalls, dass es einen kausalen Zusammenhang zwischen den Nutzerpräferenzen und der Bedeutungsverteilung der organischen Treffer bei Google und/oder BING geben muss (vgl. BORTZ 2010). Eine Korrelation stellt lediglich fest, ob zwei Variablen positiv, negativ oder gar nicht in einem Zusammenhang stehen (vgl. BORTZ 2010). Sie können sich kausal bedingen, wahrscheinlicher ist aber meist der Einfluss weiterer Faktoren, die auf beide Variablen wirken (vgl. BORTZ 2010). Der Korrelationskoeffizient gibt ein Maß an, in dem die Variablen zusammenhängen. Bei einem Korrelationskoeffizienten nahe 1 besteht ein großer, positiver Zusammenhang, bei einem Wert nahe -1 ein großer, negativer Zusammenhang (vgl. BORTZ 2010).

Die Homographe wurden noch einmal manuell in beide Suchmaschinen eingetippt, um zu erheben, inwieweit Hinweise zur Disambiguierung gegeben werden und die integrierte Bildersuche mehrere Bedeutungen widerspiegelt.

Während also für Synonymie und Morphologie diverse Übereinstimmungsraten betrachtet werden, beruht die Beurteilung des Umgangs der Suchmaschinen mit Homographie auf einer Korrelation der Bedeutungsverhältnisse zwischen den organischen Treffern und den Klickpräferenzen, sowie auf dem Vorhandensein von Hinweisen zur Disambiguierung.

5.2 Testwortsammlung

Die Testwortsammlung wurde in Anlehnung an BAR-ILAN (vgl. GUGGENHEIM 2005, BAR-ILAN 2004) so konzipiert, dass die entsprechenden sprachlichen Probleme provoziert werden (vgl. Anhang a). Die Liste mit Homographen wurde von einem Homographie-Wörterbuch inspiriert, und durch Listen aus dem Internet ergänzt (vgl. FAMILIE 2013, DETHLOFF 2012), da das Wörterbuch sich kaum auf Substantive bezog (vgl. WEBER 1996). Die Synonyme entstammen dem DUDEN (vgl. DUDEN 2012). Es wurde darauf geachtet, möglichst eindeutige Synonyme zu verwenden (z.B. Abendessen und Abendbrot), um ähnliche Ergebnislisten voraussetzen zu können. Bei Verwendung sogenannter Quasi-Synonyme (z.B. Schleuder und Katapult) wäre von vorneherein zu erwarten, dass sich die Ergebnislisten unterscheiden, weil sie nicht exakt dieselbe Bedeutung haben.

Alle Synonyme wurden testweise einmal manuell in Google eingegeben, um Probleme durch Homographie auszuschließen. So ist „Orange“ beispielsweise ein ungeeignetes Testwort, weil es neben der Frucht auch eine Farbe bzw. einen Firmennamen darstellt – obwohl „Apfelsine“ ein echtes Synonym ist.

Die morphologischen Testworte sind durch eigenes Brainstorming entstanden und spiegeln etwa zur Hälfte Abfragen mit Singular/Plural-Unterschied, zur anderen Hälfte zusammengesetzte und auseinandergeschriebene Begriffe wider.

5.3 Pretest

Der Pretest diente der Überprüfung sowohl des Studiendesigns als auch des Umgangs mit dem Retrieval Assessment Tool (RAT), das Suchmaschinenergebnisse automatisch erfassen und speichern kann (vgl. HAW 2013). Für den Pretest wurden etwa ein Dutzend Begriffe automatisch über das RAT in Google und BING abgefragt, und die ersten 30 Ergebnisse begutachtet. Generell ergab der Pretest, dass sowohl die Datensammlung als auch die Auswertung über Excel unproblematisch sind.

Das ursprüngliche Konzept sah die Betrachtung von Begriffsrelationen vor, aber der Pretest zeigte, dass die Treffer kaum Rückschlüsse auf diese sprachliche Komponente ermöglichen und die Auswahl der Testworte nur subjektiv möglich ist. So sorgte allein der Unterschied, ob man dem Begriff „Brennmaterial“ „Steinkohle“ oder nur „Kohle“ zuordnete, für vollkommen andere Häufigkeitsverteilungen des Ober- bzw. Unterbegriffs in den Treffern. Eine weitere Änderung betraf die Einteilung der Testworte im Bereich Morphologie. Durch den Pretest konnte festgestellt werden, dass eine gleichwertige Aufteilung in den Bereich „Komposita“, bzw. „Singular/Plural“ bei der Auswertung zu spezifischeren Aussagen über den Umgang mit morphologischen Veränderungen führt. Im Bereich Homographie zeigte der Pretest die bereits festgestellte Schwierigkeit der Bewertung von Bedeutungsanteilen in den Ergebnissen, wodurch die Korrelation mit den T-Online-Daten umso wichtiger wurde. Auch die Abfrage der Klickdaten der T-Online Suche wurde in einem Pretest vor der eigentlichen Erhebung erprobt. Die Abfrage und der Export der Klickdaten erwiesen sich als unproblematisch.

5.4 Datensammlung

Für die Datensammlung bei Google und BING wurde das Retrieval Assessment Tool (RAT) der Hochschule für Angewandte Wissenschaften Hamburg verwendet (vgl. HAW 2013, LEWANDOWSKI 2012), das die Trefferdaten in Excel-Listen verfügbar machte und den zugehörigen HTML-Code auf einem lokalen Server speicherte. Die Auswertung erfolgte daher in Excel.

Das RAT fragte die Daten bei der lokalen „de-Freitextsuche“ von Google und BING ab. Es fanden keine weiteren Einschränkungen statt. Somit

wurden die Treffer nicht auf deutsche Websites oder die deutsche Sprache beschränkt, wie etwa z.B. bei GUGGENHEIM 2005. Für dieses Vorgehen gibt es zwei Gründe: Zum einen fand LEWANDOWSKI in einer Studie heraus, dass die Spracheinschränkung bei Google und MSN (heute Bing) unzuverlässig ist und eine Nutzung der Funktion daher nicht sinnvoll (vgl. LEWANDOWSKI 2008a), zum anderen ist anzunehmen, dass der normale User keine Einschränkungen in der Freitextsuche vornimmt und die Funktionen einer erweiterten Suche nicht nutzt (vgl. STOCK 2005, LEWANDOWSKI 2005 S.36ff).

Ausgewertet wurden insgesamt die Trefferpositionen 1-10 sowie 91-100 der organischen Treffer. Der Vergleich der TOP 10 gegenüber den weit hinten liegenden Trefferpositionen sollte Aufschluss darüber geben, ob sprachliche Besonderheiten bei den TOP 10 besser behandelt werden als bei den weiter hinten liegenden Treffern. Besser abschneidende Kennzahlen für die TOP 10 ließen dann vermuten, dass eine Berücksichtigung sprachlicher Besonderheiten im Rankingalgorithmus stattfindet.

6. Durchführung

Die Durchführung der Datensammlung erfolgte im April 2013 mithilfe des RAT Tools der HAW Hamburg. Aufgrund technischer Einschränkungen der automatischen Abfragen bei BING und Google fanden die Abfragen innerhalb einer Woche statt. Das RAT speicherte im Wesentlichen zuverlässig die Treffer-Snippets und die zugehörigen HTML-Codes der Treffer. Aufgrund eingeschränkten Zugangs zu manchen Websites oder anderen technischen Fehlern ergab sich eine Verlustquote bei der Datensammlung von 14 %. Fehler beinhalteten häufig fehlende Trefferbeschreibungen und URLs, eine nicht erfasste Rankingposition oder doppelt erfasste Positionen. Eine erneute Abfrage der Testworte mit den fehlerhaften Trefferdaten verringerte diese Verlustquote erfolgreich. Die fehlerhaften Daten wurden durch die beim zweiten Durchlauf meist korrekt erfassten ersetzt. Da für die Datensammlung aufgrund dieser vorher bekannten Problematik etwas mehr als die angesetzten 50 Testworte abgefragt wurden, ergab sich für die Ergebnisauswertung eine zufriedenstellende Menge an Daten.

Eine zusätzliche Fehlerquelle war die Wahl der Testworte selbst, wodurch die letztendliche Verlustquote wieder leicht höher ist als nach der zweiten Datensammlung und letzten Endes insgesamt 6,5 % beträgt. Trotz gründlicher Vorüberlegungen konnten erst bei der tatsächlichen Auswertung bei einigen Testworten Schwierigkeiten festgestellt werden, die dazu führten, dass diese Worte aus der Auswertung ausgeschlossen werden mussten. Einige Beispiele sind:

- Morphologie: brauner Bär/Braunbär → es stellte sich heraus, dass „brauner Bär“ ein Homograph ist, und ebenfalls eine Schmetterlingsart bezeichnet
- Homographie: keine ausreichenden Klickdaten zu seltenen Worten, bzw. nur unpassende Suchanfragen, weil ein Wort einfach zu oft Bestandteil gänzlich anderer Suchanfragen war.

Da für den Bereich Homographie mit den meisten Problemen gerechnet wurde, wurden hier auch die meisten Testworte abgefragt, nämlich insgesamt 70 statt der angesetzten 50. Für Synonymie und Morphologie ergaben sich nach der Auswertung Verlustquoten, die 5 % nicht überstiegen, lediglich für Homographie ist eine deutliche Verlustquote von 10 % vorhanden – doch aufgrund der hohen Testwortanzahl stand bei Homographie immer noch die größte Testwortanzahl für die Auswertung zur Verfügung. Für alle Bereiche wurden erfolgreich mehr als 50 Testworte bzw. Testpaare (Synonymie Beispiel für ein Testpaar: Abendbrot/Abendessen) ausgewertet. Folgende Tabelle fasst die Auswirkungen der genannten Fehlerquellen zusammen (vgl. Tabelle 2):

Tabelle 2: Datengrundlage nach der Durchführung (eig. Darst.)

Sprachliche Besonderheit	Anzahl Testworte / Testpaare	nicht verwendet	Anzahl der verwendeten Testworte/ Testpaare	Verlust in %
Synonymie	60	3	57	5,0 %
Morphologie	56	2	54	3,6 %
Homographie	70	7	63	10,0 %

Die Ermittlung der Klickpräferenzen der T-Online-Suche stellte wie im Pretest keinerlei Problem dar. Die Suchanfragen mussten jedoch vor einer manuellen Codevergabe für die Bedeutungszuordnung bereinigt werden. Das Homograph „Horn“ beispielsweise ist auch Bestandteil englischsprachiger Abfragen, die in irgendeiner Form das Wort „horny“ enthalten. Dergleichen kam überraschend oft vor. Für die Korrelation wurden am Ende nur die Codes für die erste und die zweite Hauptbedeutung verwendet. Daher wurden die Werte sowohl bei den Klickpräferenzen als auch bei den organischen Treffern normiert. Die Normierung der Werte stellte sicher, dass für die Korrelation der Einfluss der anderen vergebenen Kategorien herausgenommen wird, da sie die Werte in unterschiedlicher Weise und Größenordnung beeinflussen, aber für die Betrachtung der Korrelation keine Rolle spielen sollen. Die Erhebung weiterer Kategorien, wie in der Methode beschrieben (etwa „3“ = Eigennamen), war jedoch für die Einordnung der Datengrundlage sinnvoll.

Die Durchführung der Datensammlung, die Auswertung und die Erhebung der Kennzahlen, sowie die Erstellung entsprechender Diagramme erstreckten sich über einen Zeitraum von etwa zwei Monaten. Grundlegendes Arbeitsmittel war Excel (vgl. Anhang b). Trotz einiger Möglichkeiten, etwa URL-Vergleiche automatisch geschehen zu lassen, ergab sich durchaus ein recht hoher Aufwand an manueller Arbeit, allein schon durch den zusätzlichen Vergleich auf ähnliche Domains unabhängig von der Rankingposition. Für die ausgewertete Differenz wurden immer die nächstgelegenen gleichen Domains beachtet, wenn es mehr als eine Möglichkeit gab. Folgendes Schema veranschaulicht dieses Vorgehen, wobei der Pfeil für die ausgewertete Differenz derselben Domain (hier Domain A) steht und das Kreuz die nicht beachtete Differenz veranschaulicht (vgl. Tabelle 3):

Tabelle 3: Schema URL-Vergleich (eig. Darst.)

URL Testwort 1	URL Testwort 2
Domain A	Domain B
Domain C	Domain A
Domain D	Domain A

Die manuelle Bearbeitung war für einige Schritte aufgrund der notwendigen intellektuellen Bewertung erforderlich. Andere Schritte hätten eventuell mit genügend Zeit und Kenntnissen über eine Datenbank automatisiert ablaufen können. Für die Betrachtung der Ergebnisse muss festgehalten werden, dass ein Anteil menschlicher Fehler durch die manuelle Arbeit angenommen werden kann. Zudem ist eine intellektuelle Bewertung, etwa eine Zuordnung von Suchanfragen zu bestimmten Bedeutungen, immer auch subjektiv beeinflusst und diskutabel. Die meisten Zuordnungen, auch der organischen Treffer, waren jedoch nach eigenem Eindruck für diese Problematik unkritisch.

7. Ergebnisse

Die hier zusammengetragenen Ergebnisse zeigen deutlich, dass die Suchmaschinen Google und BING - wie der Stand der Forschung vermuten ließ - die Besonderheiten der deutschen Sprache nicht automatisch berücksichtigen, denn die wenigen Eindrücke, die so etwas vermuten lassen, können problemlos als zufällige Koinzidenz erklärt werden. Es wird im Einzelnen auf jeden der drei untersuchten Bereiche Synonymie, Morphologie und Homographie eingegangen. Zunächst mit einer Erläuterung der Bewertungsgrundlage, dann mit einer Veranschaulichung und Erläuterung der Kennzahlen, sowie einem abschließenden Fazit.

7. 1 Synonymie

Zwei Begriffe, die dasselbe bedeuten, sollten laut den Anforderungen an Retrievalsysteme auch zu demselben Suchergebnis führen. Oder zumindest für das Ranking der Suchmaschinen gleichermaßen einbezogen werden. Schon während der Durchführung fiel auf, dass es nur geringe Übereinstimmungen bei Eingabe zweier Synonyme gibt.

Bewertungsgrundlage für „Synonymie“

Zur Einordnung der Datengrundlage wurden die Treffer von Google und BING für dieselben Suchanfragen miteinander verglichen, weil damit die Wirkung eines Synonymie-Algorithmus ausgeschlossen werden konnte. Der Vergleich der Treffer ergab wie erwartet eine geringe Übereinstimmung. Mehr als 2/3 aller Treffer waren komplett unterschiedlich. Lediglich 5 % der Treffer stimmten sowohl in der URL, als auch in der

Rankingposition überein. Häufiger tauchte mit 13 % dieselbe URL auf einer anderen Position auf. Ein ähnliches Verhalten gilt für die Übereinstimmung von Domains (aber nicht URLs). Die Positionen stimmten selten überein. Somit basieren die nachfolgenden Ergebnisse nur zu 26 % auf ähnlichen Domains, bzw. URLs unabhängig von der Position bei Google und BING. Damit wird deutlich, dass unterschiedliche Ranking-Algorithmen bei den Suchmaschinen Anwendung finden, und diese zufälligen Übereinstimmungen wohl aufgrund der Inhalte der Treffer zustandekommen. Die genannten Übereinstimmungsraten geben ein Gefühl für die inhaltliche Kohärenz und sollen als Vergleichsmaßstab für die Übereinstimmungen bei Eingabe zweier Synonyme in eine Suchmaschine dienen. Wenn die Suchmaschinen Synonymie in ihren Algorithmen berücksichtigten, müssten die entsprechenden Werte für Übereinstimmungen zweier Synonyme höher liegen als diese zufälligen inhaltlichen Übereinstimmungsraten. Zur Überprüfung dieser Schlussfolgerung wurden auszugshaft für 10 Synonympaare die Übereinstimmungsraten zwischen Synonym 1 aus BING und Synonym 2 aus Google überprüft. Die Übereinstimmungsrate von Domain oder URL unabhängig vom Ranking liegt in diesem Fall bei 15 % und bestätigt die Vermutung, dass es unabhängig von irgendeinem Synonym-Algorithmus zu inhaltlichen Übereinstimmungen kommt.

Ergebnisse für „Synonymie“

Sofort fällt auf, dass beide Suchmaschinen mit Synonymen ähnlich umgehen – nämlich gar nicht. Vergleicht man für zwei Synonyme, wie zum Beispiel „Schlips binden“ und „Krawatte binden“, die URLs und Positionen, ergeben sich bei Google und BING lediglich 1 % bzw. 2 % Übereinstimmungen. Geringfügig größer wird der Grad der Übereinstimmung, wenn man zusätzlich die Domains und Positionen vergleicht, und steigt für BING auf 4 %, für Google auf 6 %.

Der Vergleich der TOP10 gegenüber den Positionen 91-100 zeigt bei beiden Suchmaschinen ein sehr ähnliches Bild. Wenn es Übereinstimmungen gibt, gleich ob es nur um die URL oder um URL oder Domain geht, tauchen die Übereinstimmungen eher in den TOP10 auf. Bei Google gilt dies jedoch nur für die Übereinstimmung der kompletten URL, deshalb

kann man hier nur von einer leichten Tendenz zugunsten der TOP10 sprechen. Der zusätzliche Einbezug derselben Domain auf demselben Rankingplatz begünstigt bei Google die TOP10 nicht. Folgende Grafik veranschaulicht den Vergleich der TOP10 mit den Plätzen 91-100 beispielhaft anhand von Google, basierend auf der Übereinstimmung von URL oder Domain. Dieselben Verhältnisse der TOP10 (links) gegenüber den Plätzen 91-100 (rechts) zeigen, dass Google die TOP10 in diesem Fall nicht begünstigt (vgl. Abbildung 1):

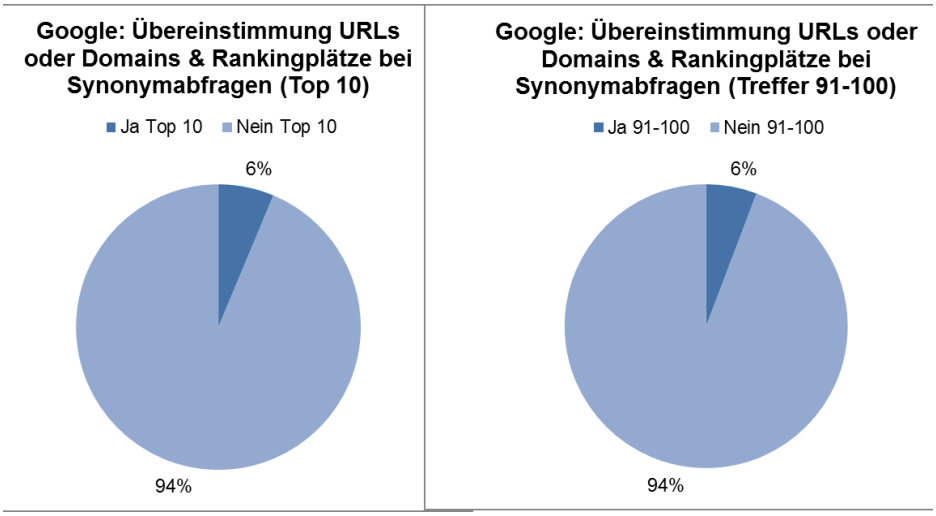


Abbildung 1: "Synonymie" - Google (eig. Darst.)

Der Vergleich der TOP10 mit den Treffern 91-100 ist allerdings aufgrund der allgemein geringen Übereinstimmungen kaum aussagekräftig.

Zusätzlich wurde betrachtet, inwieweit Domains bzw. URLs bei der Eingabe zweier Synonyme unabhängig vom Rankingplatz übereinstimmen. Bei dieser Betrachtung ergab sich zum ersten Mal ein wesentlicher Unterschied zwischen BING und Google. So zeigt Google eine Übereinstimmungsrate von 19 %, BING lediglich von 9 %. Ähnlich ist bei beiden Suchmaschinen wiederum der Anteil der tatsächlichen URL-Übereinstimmung unabhängig von der Rankingposition, sie liegt bei Google bei 3 %, bei BING bei 2 %. Die hohe Übereinstimmung von Domains unabhängig vom Ranking bei Google ist jedoch kritisch zu betrachten. Rechnet man starke und daher im Allgemeinen häufig vertretene Domains aus der Übereinstimmung heraus, liegt Googles Übereinstimmungsrate nur noch bei 9 %, BINGs noch bei 7

%. Für diese Betrachtung wurden diejenigen Domains nicht mit einberechnet, die mehr als zehnmal unabhängig vom Rankingplatz übereinstimmten. Bei BING betraf dies nur Wikipedia-Ergebnisse. Bei Google zusätzlich zu Wikipedia auch Wiktionary, Amazon und Duden. Das wirft die Frage auf, ob Google generell bestimmte Domains im Ranking bevorzugt im Vergleich zu BING, bzw. ob dieses unterschiedliche Verhalten durch eine entsprechende Weichenstellung im Algorithmus zustande kommt. Zumindest ist das häufige Vorkommen derselben Domains in den Suchergebnissen bei Google ein der Suchmaschine bekanntes Thema, dem im Algorithmus bereits entgegengewirkt wird (vgl. SCHWARTZ 2013).

Der Mittelwert der Ranking-Abweichung bei Übereinstimmung führt bei BING zu keiner weiteren Erkenntnis, da die Standardabweichung deutlich höher ist als der Mittelwert selbst. Bei Google sieht es für die Übereinstimmung der Domains ähnlich aus, für die Übereinstimmung der URLs ergibt sich allerdings ein Mittelwert der Rankingabweichung von 2 Plätzen und die entsprechende Standardabweichung beträgt lediglich 2 Plätze. Das heißt, wenn bei BING für ein Synonympaar Domain oder URL übereinstimmen und nicht von vorneherein derselbe Rankingplatz vorliegt, ergeben sich in der Regel gewaltige Rankingunterschiede. Stimmen bei Google hingegen die URLs überein, ist der Rankingunterschied generell gering, bei Domains ergibt sich jedoch eine ähnlich weite Streuung wie bei BING.

Insgesamt fällt bei einer Gegenüberstellung von Google und BING im Umgang mit Synonymen auf, dass in den meisten Fällen Google eine leicht höhere Übereinstimmungsrate zeigt. Allerdings sind die Unterschiede nicht signifikant, sondern betragen nur wenige Prozent.

Folgende Abbildung veranschaulicht, dass die Übereinstimmungsraten (y-Achse) generell etwas zunehmen, wenn man nach Übereinstimmungen unabhängig von der Rankingposition (x-Achse) sucht (vgl. Abbildung 2):

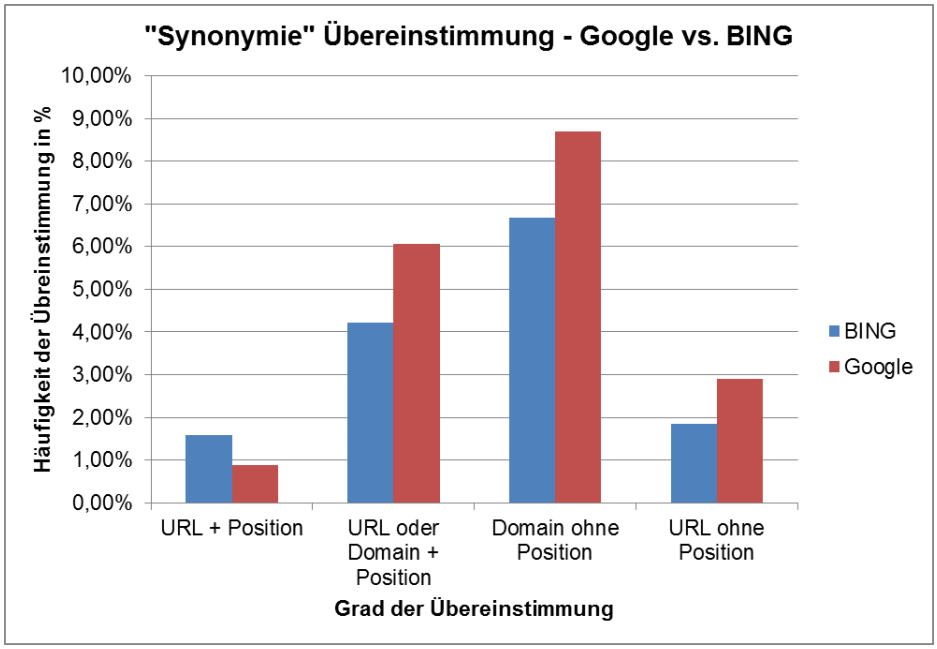


Abbildung 2: "Synonymie" - Google vs. BING (eig. Darst.)

Die geringen Übereinstimmungsraten sprechen nicht dafür, dass „Synonyme“ von Google und BING technisch berücksichtigt werden. Sie liegen unter den zufälligen Übereinstimmungsraten, die in der Bewertungsgrundlage (siehe S. 20-21) berechnet wurden, und zwar unabhängig davon, ob es nur um die URL oder auch die Domain geht. Übereinstimmungen aufgrund anderer Faktoren, zum Beispiel entsprechende Suchmaschinenoptimierung der Websites für beide Synonyme, sind folglich wahrscheinlicher.

7.2 Morphologie

Morphologische Veränderungen ein und desselben Wortes, etwa durch die Pluralbildung, sollten in Suchsystemen bei idealem Retrieval wenige Auswirkungen auf die Ergebnisse haben. Ob man Worte zusammensetzt oder nicht sollte ebenfalls für die Bedeutung des Gesuchten keine Rolle spielen. Die Ergebnisse für die Suchmaschinen BING und Google im Bereich „Morphologie“ zeigen generell große Ähnlichkeit zu denen im Bereich „Synonymie“.

Bewertungsgrundlage für „Morphologie“

Wie bei „Synonymie“ werden auch für „Morphologie“ zufällige, inhaltliche Übereinstimmungsraten anhand eines Vergleichs zwischen Google und BING erhoben. Die Auswertung von je 108 morphologischen Varianten ergab ebenfalls eine geringe Übereinstimmung zwischen den beiden Suchmaschinen. Wie bei „Synonymie“ sind 2/3 aller Treffer unterschiedlich. Die restliche prozentuale Verteilung der Übereinstimmung von URL und Position, nur URL, bzw. Domain und Position oder nur Domain ergibt ähnliche Werte wie die Untersuchung zur Synonymie. So stimmen in 4 % aller Fälle URL und Position überein, in 1 % Domain und Position, in 7 % die Domain unabhängig vom Ranking und in 12 % die URL unabhängig vom Rankingplatz. Wiederum gilt: Die Rankingposition ist seltener gleich, ähnliche Domains und URLs tauchen häufiger auf unterschiedlichen Positionen auf. Eine Unterscheidung bei der Prüfung der Übereinstimmung von Google und BING nach den beiden Testvarianten Singular/Plural bzw. Komposition/Zerlegung ergibt keine auffälligen Unterschiede im Vergleich zur Gesamtbetrachtung.

Ergebnisse für „Morphologie“

Wie bei Synonymie fällt insgesamt eine geringe Übereinstimmungsquote auf – jedoch gibt es im Allgemeinen mehr Übereinstimmungen für Singular/Plural-Abfragen, als für Komposition oder Synonyme. Auffällig ist, dass Abfragen im Bereich Komposition/Zerlegung (etwa Adventsgesteck, Gestecke Advent), fast auf das Prozent genau zu denselben Übereinstimmungsraten führen wie die Abfrage von Synonymen. Das gilt für die Gesamtbetrachtung und die Gegenüberstellung von TOP10 vs. Treffer 91-

100 – und für beide Suchmaschinen. Je unterschiedlicher die Suchbegriffe also sind, desto seltener kommt es zu Übereinstimmungen.

Generell ist die Übereinstimmung von URL und Position bei Komposition/Zerlegung also gering, mit 2 % bei Google und 1 % bei BING. Betrachtet man die Übereinstimmung von Domain und Rankingplatz, liegt Google mit 7 % Übereinstimmung deutlich vor BING mit 2 %. Wie auch bei Synonymen liegen Übereinstimmungen bei BING in den TOP10, wenn es sie gibt. Bei Google trifft dies auf die Übereinstimmung von URLs und Rankingplätzen zu, aber die zusätzliche Betrachtung von gleichen Domains gleicht die Anteile in TOP10 und Treffer 91-100 ziemlich aus.

Etwas höher als bei Synonymabfragen treten Übereinstimmungen bei Komposition/Zerlegung auf, wenn man diese losgelöst vom Rankingplatz betrachtet. So stimmen bei Google 25 % der Domains überein, wenn auch wie bei Synonymie mit gewaltigen Rankingunterschieden, die eine so große Streuung besitzen, dass der Mittelwert unerheblich ist. Immerhin 8 % dieser Übereinstimmungen gehen tatsächlich auf gleiche URLs zurück, für die wie bei Synonymie gilt, dass die Rankingunterschiede bei Google gering sind. Bei BING gibt es immerhin 13 % übereinstimmende Domains abgesehen vom Rankingplatz, von denen 4 % auf derselben URL basieren. Der deutliche Unterschied an Übereinstimmung zwischen Google und BING lässt sich wieder durch das häufigere Vorkommen bestimmter Domains bei Google erklären (hier Wikipedia, Amazon, Ebay). Entfernt man diese Anteile aus der Betrachtung, verringert sich der Unterschied zwischen Google und BING auf lediglich 4 % statt 12 %.

In den Daten fällt auf, dass bei Abfragen im Singular und Plural zwar keine gravierend großen Anteile von Treffern übereinstimmen, aber deutlich mehr als bei Komposition/Zerlegung oder Synonymie. Ob dies aber wirklich daran liegt, dass die Suchmaschinen mit Singular und Plural technisch umgehen können, oder einfach daran, dass auf relevanten Websites Singular und Plural-Formen gleichermaßen vorkommen, ist hier nicht zu unterscheiden. So stimmen bei Google 5 % (BING 3%) der URLs und Rankingpositionen überein, und bei weiteren 5 % Domains und Rankingpositionen (BING 2%). Wie bei Komposition/Zerlegung und Synonymen liegen bei

Google dieselben URLs in den TOP10, dieselben Domains beschränken sich nicht auf die TOP10. BING bevorzugt wie bei Synonymie die TOP10 bei den Übereinstimmungen.

Auch der Anteil gleicher Treffer unabhängig vom Rankingplatz ist bei Singular/Plural-Abfragen höher als bei den Ergebnissen zur Synonymie, aber vergleichbar mit der bei Komposition/Zerlegung. Die einzige Überraschung ist, dass bei Google hier der Anteil häufiger Domains generell geringer ist. Rechnet man die Domains mit mehr als 10 Übereinstimmungen unabhängig vom Ranking heraus (hier Amazon), ergibt sich statt 22 % eine Übereinstimmung von 17 %. Immerhin 10 % dieser Übereinstimmung basieren auch auf tatsächlich gleichen URLs. Bei BING immerhin 9 %.

Stellt man BING (blau) und Google (rot) einander gegenüber, ergibt sich bei Komposition/Zerlegung mit ca. 4 % Differenz in drei Übereinstimmungskategorien ein leichter Vorteil zugunsten Googles, so stimmt bei Google etwa die URL unabhängig von der Position 4 % häufiger überein als bei BING und nur bei der Übereinstimmung von URL und Position liegen BING und Google im Grunde gleichauf (vgl. Abbildung 3):

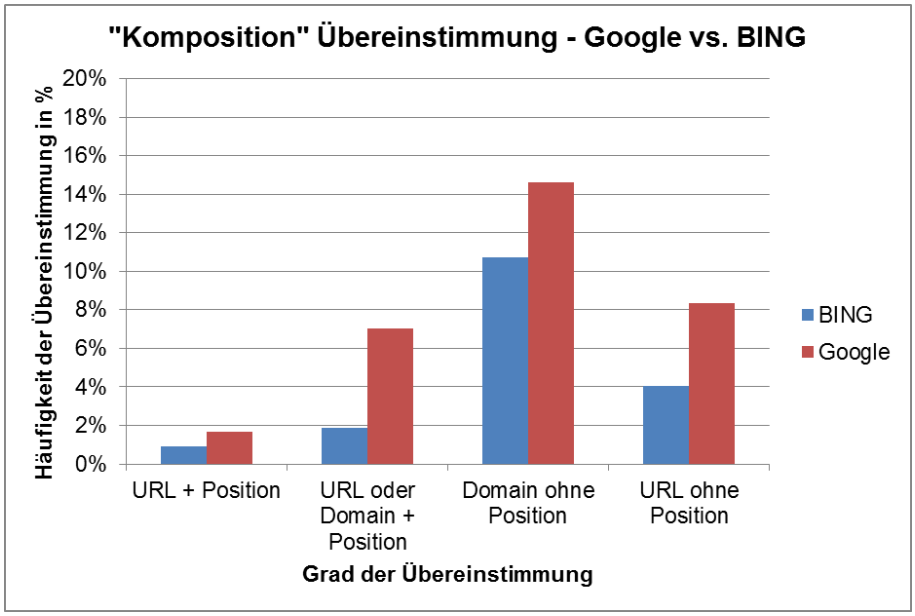


Abbildung 3: "Komposition" - Google vs. BING (eig. Darst.)

Bei einer Gegenüberstellung bei „Singular/Plural“ fällt ein ähnlicher Vorteil von Google (rot) gegenüber BING (blau) auf, allerdings vor allem im Bereich der „Domain-Übereinstimmung“, die in dieser Darstellung bereits von „häufigen Domains“ bereinigt wurde (vgl. Abbildung 4):

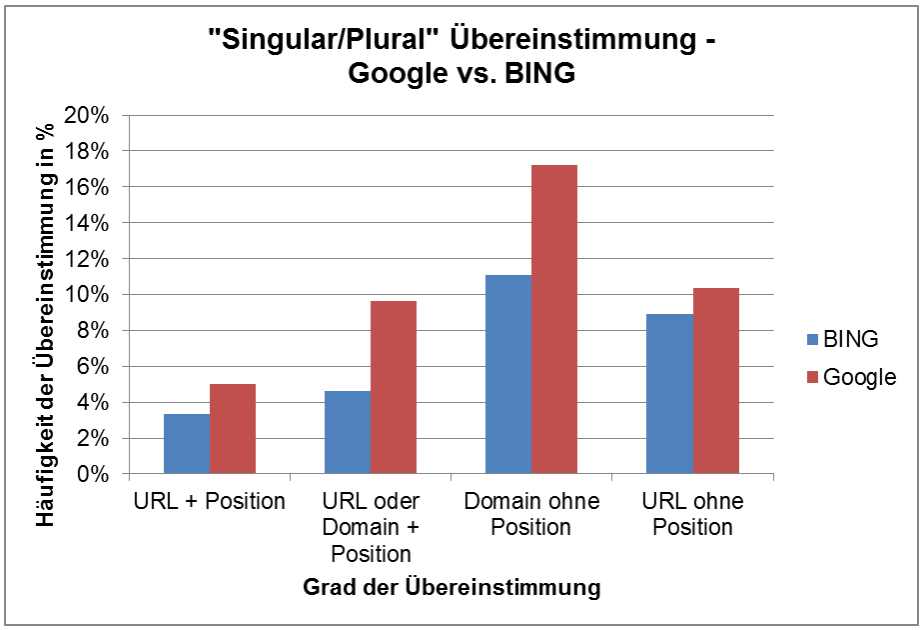


Abbildung 4: "Singular/Plural" - Google vs. BING (eig. Darst.)

Vergleicht man nun die Übereinstimmungsraten mit den zufälligen Quoten der Bewertungsgrundlage (siehe S. 25), fällt im Gegensatz zum Bereich „Synonymie“ auf, dass im Falle der Domains die Übereinstimmungsraten für Singular/Plural und Komposition/Zerlegung höher liegen als die zufälligen Quoten. So liegt die zufällige Übereinstimmung von Domains unabhängig vom Ranking bei 7 %, die entsprechende Rate für Singular/Plural bzw. Komposition/Zerlegung bei Google ist mehr als doppelt so hoch. BING liegt mit etwas mehr als 10 % Übereinstimmung näher an der zufälligen Quote. Da diese Beobachtung jedoch nur die Domains betrifft, bleibt fraglich, inwieweit dies mit dem Algorithmus von Google zusammenhängt. Gerade bei ähnlichen Wortbestandteilen wie im Bereich Morphologie können ein wenig höhere Werte für Übereinstimmungen auch durch die Optimierungsversuche (SEO) der Betreiber der betreffenden Domains verursacht werden. Insgesamt ist Google gegenüber BING durch höhere Übereinstimmungsraten im Bereich „Morphologie“ leicht im Vorteil.

7.3 Homographie

Homographie, also die gleiche Schreibweise eines Wortes mit mehreren Bedeutungen (z.B. Heide = Stadt, Landschaft, Name…), wird von den Suchmaschinen Google und BING im Allgemeinen ähnlich gehandhabt. Unterschiede ergeben sich bei einzelnen Begriffen hinsichtlich dessen, welche der untersuchten Bedeutungen am häufigsten in den Treffern vorkommt. Generell lässt sich aber von einer sehr ähnlichen Tendenz beider Suchmaschinen für eine Bedeutung sprechen. Diesen Eindruck bestätigt auch eine Auswertung der Korrelation der Suchergebnisse der beiden Suchmaschinen. Betrachtet man etwa die Korrelation der Häufigkeit von Bedeutung 1 (blau) bei Google (y-Achse) und BING (x-Achse) innerhalb der gesamten Treffer, ergibt sich ein klares Cluster im oberen rechten Quadranten, und ein Korrelationskoeffizient von über 40 %. Denselben Zusammenhang findet man für Bedeutung 2 (rot), nur dass diese sich größtenteils im Quadranten links unten sammelt (vgl. Abbildung 5):

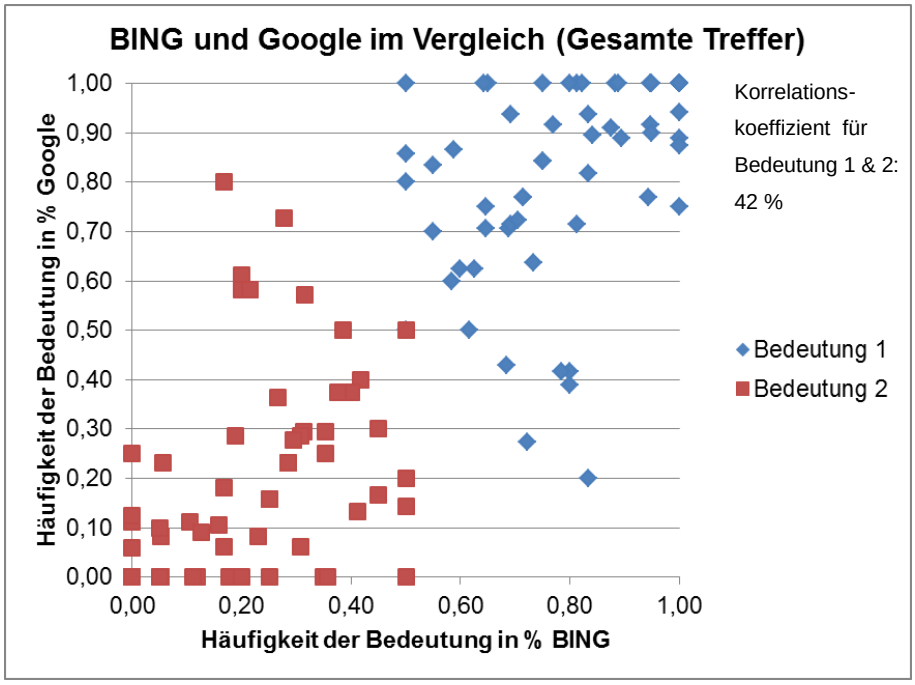


Abbildung 5: "Homographie" - Google vs. BING (eig. Darst.)

Eine Unterscheidung zwischen TOP10 und den Treffern 91-100 zeigt als einzige Überraschung eine deutliche stärkere Korrelation zwischen Google und BING in den Treffern 91-100 für Bedeutung 1.

Betrachtet man den Zusammenhang der organischen Treffer mithilfe des Korrelationskoeffizienten für je Google und BING, zeigt sich ebenfalls ein ähnliches Verhalten der beiden Suchmaschinen. BING und Google zeigen eine höhere Korrelation zwischen den TOP10 und den gesamten Treffern, als zwischen den Treffern 91-100 und den gesamten Treffern. Dieser Zusammenhang ist bei Google mit 84 % noch etwas stärker als bei BING mit 79 %. In den hinteren Treffern korreliert die Häufigkeit der Bedeutung 1 mit den gesamten Treffern deutlich geringer mit 49 % bei Google und 41 % bei BING. Die zweite Bedeutung zeigt hingegen durchaus noch einen ähnlich starken Zusammenhang bei den Treffern 91-100 wie bei den TOP10 mit 60 % bei BING und 71 % bei Google. Es lässt sich also nicht sagen, dass die Klickpräferenzen generell eher in den TOP10 zum Tragen kommen. Die Ursache der Korrelationsdifferenz im Ranking bei der ersten Bedeutung ist nicht klar und kann Zufall sein.

Bewertungsgrundlage für „Homographie“

Zur Einordnung und Bewertung der Ergebnisse soll kurz auf die Vergabe der anderen Kategorien eingegangen werden, falls Ergebnisse nicht als Begriff 1 oder 2 kategorisiert werden konnten. Unpassende oder Unklare Treffer waren selten, es gab diese in 83 % aller Fälle bei Google und bei BING in 89 % überhaupt nicht. Das heißt: Fast alle organischen Treffer konnten eindeutig zugeordnet werden. Die meist vergebene andere Kategorie war diejenige für Eigennamen. Das Vorkommen beider Hauptbedeutungen bei einem Treffer war selten festzustellen. Bei einigen Worten gab es mit über 10 % auffallend viele Wörterbucheinträge in den organischen Treffern (z.B. Blüte, Mangel…), die entsprechend die Datengrundlage für die Häufigkeitsverteilung der beiden Hauptbedeutungen verringerten. Wörterbucheinträge zeigt Google häufiger als BING und auch ein wenig häufiger in den TOP10 als auf den Plätzen 91-100. Die Kategorie „Andere Bedeutungen“ wurde insgesamt selten vergeben, beeinflusst aber einzelne Keywords in hohem Maß (5-25%). Über alle 63 abgefragten Homographe ist dieser Einfluss vernachlässigbar, lediglich der Einfluss der vorkommenden Eigennamen ist beachtlich. So lassen sich bei BING bis zu 20 % der Treffer bei etwas mehr als 40 % aller Suchanfragen in die Kategorie „Eigennamen“ einordnen. Bei Google sind es etwas mehr als 30 %

der Treffer. Dafür kommen bei Google Eigennamen mit etwas über 40 % bei etwa 20 % aller Suchanfragen vor, bei BING sind es lediglich circa 10 %. Mehr als die Hälfte der Ergebnisse werden nicht sehr häufig von Eigennamen bestimmt. Folgende Grafik veranschaulicht die Häufigkeit mit der Eigennamen bei BING und Google vorkommen, wobei auf der x-Achse Kategorien abgetragen sind. Die erste Kategorie zeigt an, dass BING beispielsweise keine Eigennamen in etwa 30 % aller Fälle in den organischen Treffern anzeigt. Die zweite Kategorie ist mit der nächsten Obergrenze beschriftet, hier gibt es bis zu 20 % Eigennamen in den organischen Treffern von BING in über 40 % aller Fälle (vgl. Abbildung 6):

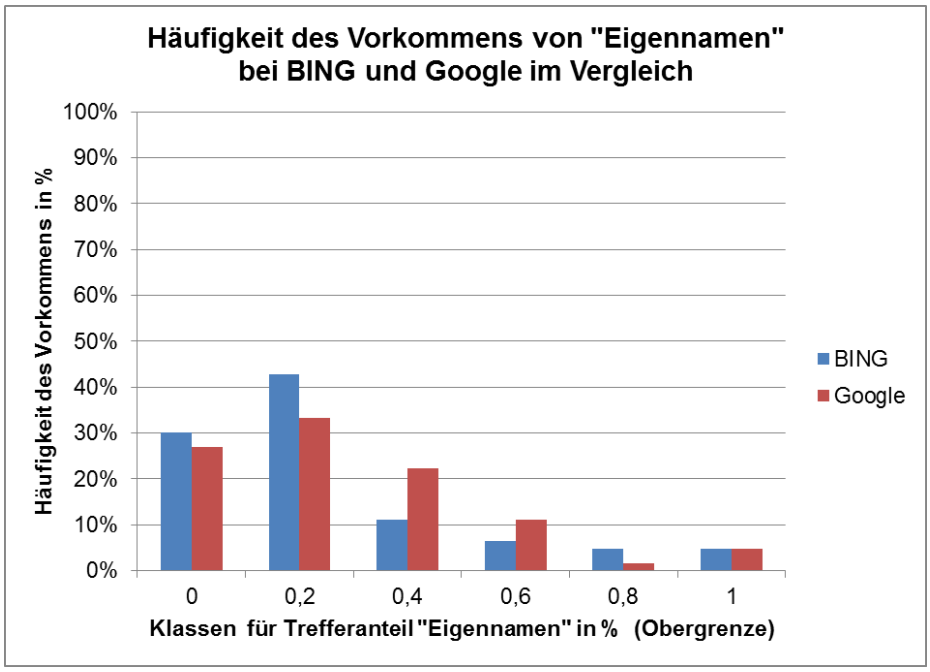


Abbildung 6: "Eigennamen" - Google vs. BING (eig. Darst.)

Bei einer Betrachtung der Korrelation muss somit berücksichtigt werden, dass unterschiedliche Datenmengen als Grundlage zum Tragen kommen, durch die Vergabe weiterer Kategorien und die daraus resultierende Normierung der Häufigkeiten.

Die Übereinstimmung zwischen den Suchmaschinen birgt keine Auffälligkeiten im Vergleich zum Bereich „Synonymie“ und „Morphologie“, und liegt wiederum bei etwa 26 %, das heißt auch hier kann von unabhängigen Algorithmen ausgegangen werden.

Ergebnisse für „Homographie“

Die Korrelation der Bedeutungsverteilung innerhalb der organischen Treffer (y-Achse) mit den entsprechenden Klickpräferenzen der T-Online-Klickdaten (x-Achse) ergibt ein sehr stimmiges Bild, wenn man Google und BING vergleicht. So zeigt sich für Bedeutung 1 klar, dass die Bedeutungen, die bei BING und Google häufig in den organischen Treffern auftreten, in der Regel auch die sind, die vermehrt geklickt werden (vgl. Abbildung 7 und 8 Cluster oben rechts):

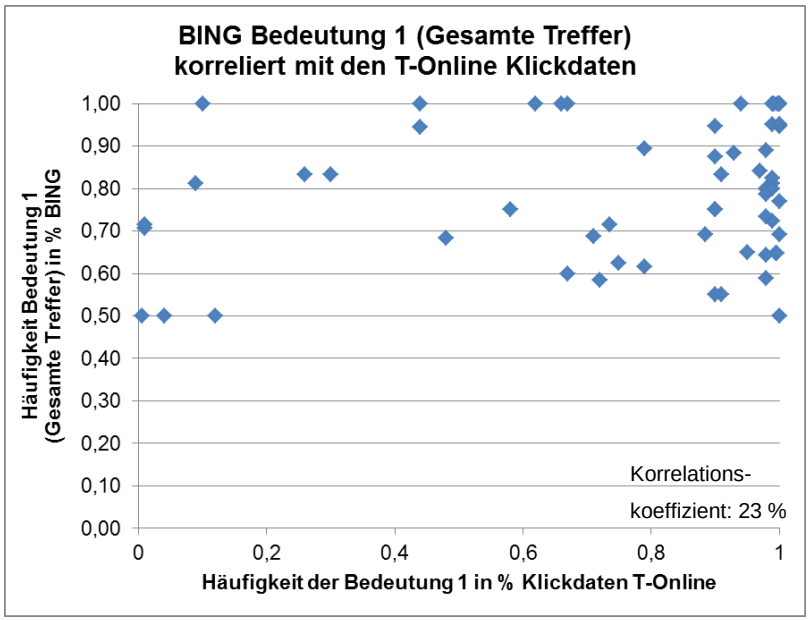


Abbildung 7: Korrelation „Bedeutung 1“ - BING (eig. Darst.)

Ob diese Korrelation zwischen der Häufigkeit der Bedeutung und der Klickpräferenz der T-Online-Daten aber auf einer tatsächlichen Berücksichtigung der meist geklickten Bedeutung basiert, ist fraglich. Möglich ist auch, dass durch die größere Zielgruppe präferierte Websites entsprechend mehr Suchmaschinenoptimierung betreiben, und sich daher ein gewisser Zusammenhang ergibt. Es gibt einige Ausnahmen, bei denen die Suchmaschinenergebnisse eine Bedeutung bevorzugen oder vernachlässigen, wie es nicht dem Klickverhalten entspricht (vgl. Abbildung 7 und Abbildung 8, z.B. Cluster oben links):

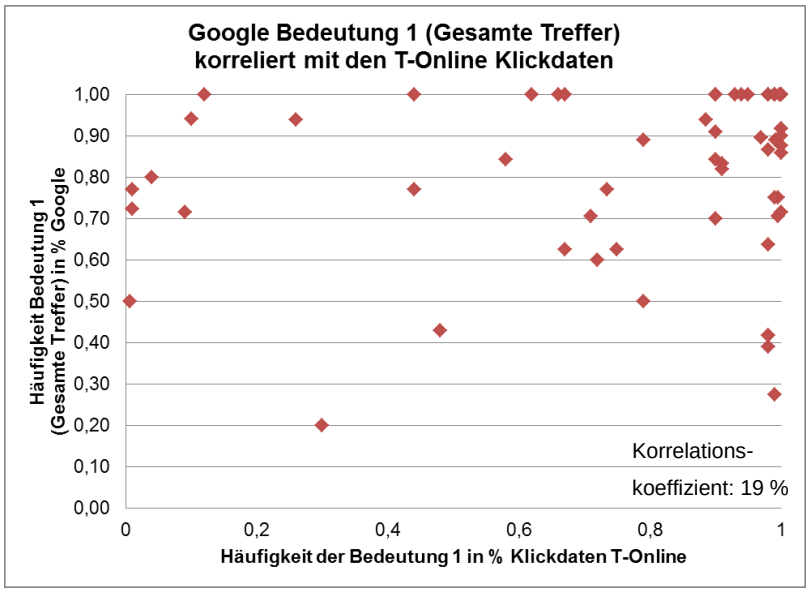


Abbildung 8: Korrelation „Bedeutung 1“ - Google (eig. Darst.)

Die zweite Bedeutung wird in der Regel weniger durch Klicks präferiert (x-Achse) und taucht entsprechend seltener in den organischen Suchergebnissen (y-Achse) auf (vgl. Abbildung 9 und 10, Cluster unten links). Aber auch hier gibt es einige Ausnahmen, die den Klicks entsprechend präferiert werden, aber in den organischen Ergebnissen seltener auftauchen (vgl. Abbildung 9 und 10, Cluster unten rechts):

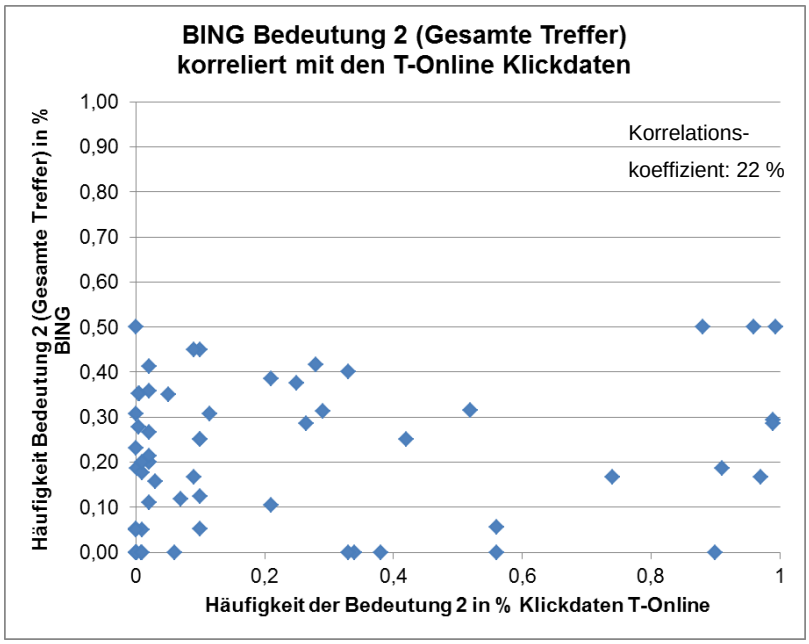


Abbildung 9: Korrelation „Bedeutung 2“ - BING (eig. Darst.)

Man erkennt auch ein kleines Mittelfeld, in dem sich Treffer sammeln, bei denen eine mittlere Klickpräferenz und auch eine mittlere Häufigkeit bei den organischen Treffern vorliegen (vgl. Abbildung 9 und 10, Mitte):

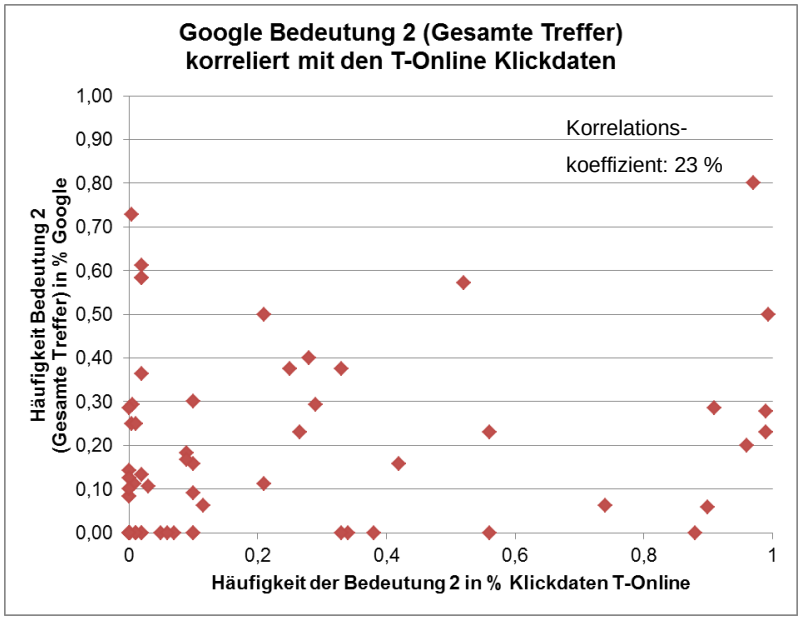


Abbildung 10: Korrelation „Bedeutung 2“ - Google (eig. Darst.)

Im Unterschied zu BING fällt bei der Korrelation der zweiten Bedeutung mit den Klickpräferenzen bei Google auf, dass es einige Ausreißer gibt, die mit über 50 % Häufigkeit in Googles organischen Treffern vorkommen, aber fast keinerlei Klicks erhalten und somit überhaupt nicht mit den Klickpräferenzen übereinstimmen (vgl. Abbildung 10).

Das etwas diffuse Bild der Korrelation wird durch die Korrelationskoeffizienten bestätigt. Für alle organischen Treffer beläuft sich der Zusammenhang bei Google und BING mit den Klickpräferenzen von T-Online gleichermaßen für beide Bedeutungen um die 20 %, mit geringen Unterschieden. Es kann also nur ein gewisser, positiver Zusammenhang bestätigt werden. Die Bedeutungen, die eher angeklickt werden, werden demnach durchaus ein wenig eher von den Suchmaschinen präferiert.

Eine Unterscheidung zwischen TOP10 und den Treffern 91-100 sieht so ähnlich aus in der Korrelationsgrafik, dass nur die Betrachtung des Korrelationskoeffizienten kleinere Unterschiede aufweist. So hängen bei BING in den TOP10 beide Bedeutungen mit knapp 10 % deutlich geringer mit den

Klickpräferenzen zusammen als in den Treffern 91-100 mit ca. 29 % Zusammenhang bei Bedeutung 1 und knapp 15 % Zusammenhang bei Bedeutung 2. Bei Google hingegen ergibt sich ein recht stimmiges Bild des Korrelationskoeffizienten für die organischen Treffer und die Klickpräferenzen, egal ob insgesamt oder aufgeteilt auf die TOP10 bzw. Treffer 91-100. Er schwankt um die 20 % für beide Bedeutungen und spricht somit für einen geringen, positiven Zusammenhang. Lediglich die zweite Bedeutung weist in den Treffern 91-100 einen deutlich geringeren Zusammenhang mit nur 13 % auf. Die Unterscheidung nach Rankingpositionen liefert also keinen Hinweis darauf, dass es eine Tendenz dazu gibt, die präferierte Bedeutung in den TOP10 stärker zu präsentieren als in den Treffern 91-100. Diesen Eindruck bestätigt auch eine generelle Betrachtung der Häufigkeit von Bedeutung 1 und 2 bei Google und BING unabhängig von den Klickpräferenzen. Hier gilt für beide Suchmaschinen, dass die Bedeutungen im Allgemeinen unabhängig vom Rankingplatz ähnlich häufig bzw. ähnlich selten vorkommen.

Betrachtet man die eindeutigen Ausreißer der Suchmaschinen, in denen die Tendenz der organischen Treffer für eine Bedeutung nicht mit den Klickpräferenzen übereinstimmen, kann man anhand der Datengrundlage keinen eindeutigen Grund für diese Schieflage feststellen. Die T-Online-Klickdaten für diese Ausreißer basieren z.B. nur teilweise auf einer geringen Anzahl von Suchanfragen. Die Daten der organischen Treffer sind bei den Ausreißern ebenfalls nur teilweise durch einen hohen Anteil von Eigennamen etc. verzerrt, aber nicht bei allen Ausreißern. Zudem kommt es vor, dass für einzelne Anfragen zwar Google und BING bei den organischen Treffern eine ähnliche Beeinträchtigung der Datengrundlage haben, BING aber mit der Klickpräferenz von T-Online besser übereinstimmt als Google. Diese Betrachtung legt den Schluss nahe, dass andere Faktoren bei den Ausreißern eine Rolle spielen. In Frage kommen zum Beispiel kommerzielle Interessen. Allerdings ist es bei den starken Ausreißern nicht so, dass die Suchmaschinen kommerzielle Interessen bevorzugen und die Klickpräferenzen eher für andere Bedeutungen sprechen – sondern genau umgekehrt. Es handelt sich jedoch nur um wenige Ausnahmefälle, daher wäre es übertrieben, von einer Tendenz zu sprechen.

Was die Verteilung der Bedeutungen in den integrierten Bildern der Ergebnisseite betrifft, zeigen sich zum ersten Mal deutliche Unterschiede zwischen Google und BING. Generell zeigt Google in 84 % aller Abfragen Bilder, BING nur in etwa der Hälfte aller Abfragen, wobei die integrierten „Maps“-Anzeigen nicht als Bild gezählt wurden (Ausnahme „Atlas“: Maps gilt als Bild für die Bedeutung „Karte“). Dafür stellt man bei BING eine deutliche Tendenz fest, die häufigere Bedeutung auch eher in den Bildern zu zeigen als die zweithäufigste Bedeutung. Bei Google ist das Bilderverhältnis für beide Hauptbedeutungen generell ausgeglichen.

Hinweise auf verschiedene Bedeutungen oder Erklärungen präsentiert nur Google gesondert und immerhin in 25 % der getesteten Fälle. Diese neu angebotenen Informationen stammen aus dem Google Knowledge Graph, der entwickelt wurde, um Nutzern zusammenhängendes Wissen zu präsentieren und auch sprachliche Probleme zu lösen (vgl. GOOGLE 2012). Die Bedeutungserklärungen stammen in den Beispielen dieser Studie entweder aus Wikipedia oder es gibt einen gesonderten Hinweis auf „Ergebnisse für…“, oben rechts direkt neben den ersten Treffern. Ein Hinweis für verschiedene Bedeutungsmöglichkeiten wurde für diese Studie nur gewertet, wenn dieser mehr als eine Definition einer Bedeutung zum Inhalt hatte, also auch einen entsprechenden Hinweis auf andere Ergebnisse. Beispielhaft sieht man diese Hilfestellung auf folgendem Screenshot (vgl. Abbildung 11 rechts):

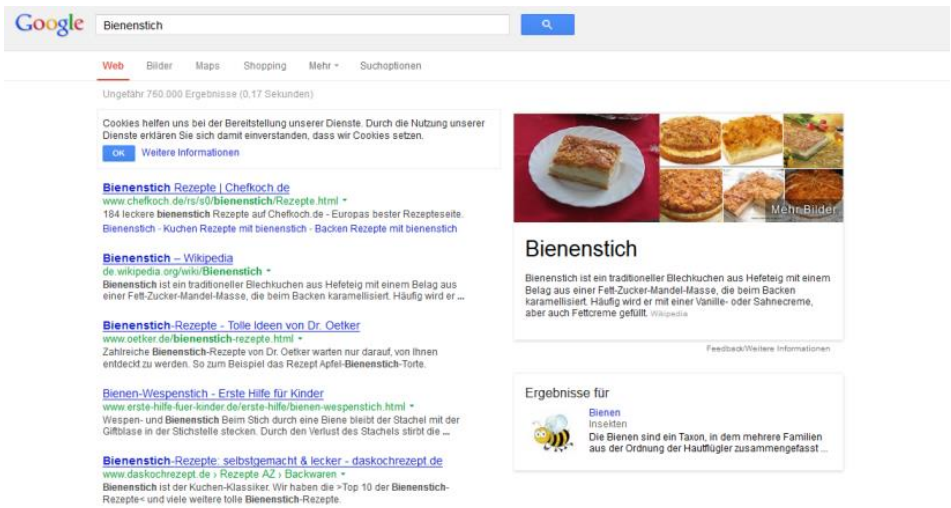


Abbildung 11: Google Disambiguierung „Bienenstich“ (Screenshot)

Am Beispiel „Hamburger“ sieht man deutlich die verspielte und visuell gut unterstützte Hilfestellung zur Bedeutungsklärung (vgl. Abbildung 12):

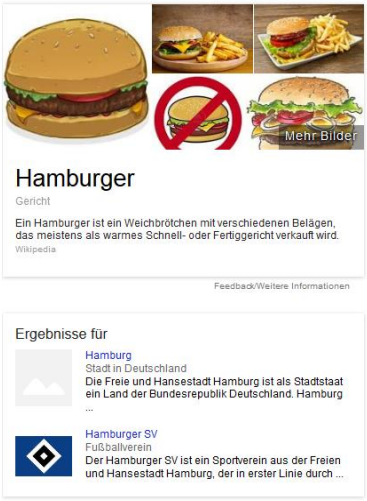


Abbildung 12: Google Disambiguierung “Hamburger” (Screenshot)

Solche Hinweise fehlen bei der Suchmaschine BING komplett. BING zeigt lediglich “Ähnliche Suchvorgänge” teilweise oben rechts an, und nicht nur unten wie Google. Diese auf Nutzerstatistiken beruhenden verwandten Suchanfragen zählen in dieser Studie nicht als Hilfestellung bei der Bedeutungsklärung. In dieser Hinsicht liegt der Nutzerservice von Google also vor BING, wobei eine Studie über die tatsächliche Beachtung und Annahme von Googles Hilfestellung durch echte Nutzer weiteren Aufschluss über die Sinnhaftigkeit geben könnte.

8. Diskussion der Ergebnisse

Vergleicht man die zu Beginn der Arbeit aufgestellten Hypothesen (vgl. S. 11) mit den im letzten Kapitel erläuterten Ergebnissen, erkennt man deutlich, dass beide Suchmaschinen nur im geringen Maß mit den untersuchten, sprachlichen Besonderheiten im Deutschen technisch umzugehen wissen. So zeigt sich:

- Bei Eingabe zweier Synonyme ergeben sich nur zu einem äußerst geringen Prozentsatz dieselben Treffer. Das spricht nicht für eine technische Berücksichtigung von „Synonymie“.

- Bei Eingabe zweier morphologischer Varianten zeigen beide Suchmaschinen ebenfalls nur geringe Übereinstimmungen zwischen den Trefferlisten. Eine Tendenz der Suchmaschinen, morphologische Varianten effektiv technisch zusammenzuführen, kann nicht bestätigt werden.
- Das Verhältnis der untersuchten Bedeutungen bei den organischen Treffern von BING und Google orientiert sich ansatzweise an den Nutzerpräferenzen. Direkte Hinweise auf verschiedene Bedeutungen zeigt Google nur in einem Viertel aller untersuchten Fälle, BING generell nicht.

Die Hypothesen konnten also mithilfe der durchgeführten Studie nicht bestätigt werden. Im Sinne klassischen Information Retrievals findet keine technische Berücksichtigung von Synonymie oder Morphologie der deutschen Sprache statt, wie sich aufgrund des Stands der Forschung bereits vermuten ließ (vgl. S. 4ff). Bei Homographie vermitteln die Ergebnisse den Eindruck, dass beide Suchmaschinen eine Tendenz hin zu den Nutzerpräferenzen zeigen. Allerdings in einem Maß, das aufgrund der unterschiedlichen Datengrundlage kaum als eindeutiger Hinweis auf eine technische Zusatzfunktion verstanden werden kann. Zusammengefasst lassen sich auf Basis der durchgeführten Studie folgende Kernaussagen treffen:

- Gibt man zwei Synonyme mit derselben Bedeutung bei BING oder Google ein, erhält man in der Regel fast komplett unterschiedliche Ergebnisse.
- Wenn man bei BING oder Google Singular und Plural eines Suchbegriffes eingibt, zeigt Google eine etwas höhere Übereinstimmungsrate als BING. Für beide Suchmaschinen gilt jedoch, dass der Großteil der Treffer unterschiedlich ist und man daher bei einer intensiven Informationsrecherche bessere Ergebnisse erzielt, wenn man beide Varianten abfragt.

- Ob man eine Suchanfrage zusammenschreibt oder zerlegt, führt zu sehr großen Unterschieden in den Ergebnissen und verschwindend geringen Übereinstimmungen.
- Im Allgemeinen stimmen bei Eingabe zweier Suchbegriffe mit einer ähnlichen Bedeutung eher die Domains und nicht die URLs überein. Dieses gilt praktisch unabhängig vom Ranking und spricht nicht unbedingt für eine technische Lösung, sondern für eine entsprechend breite Aufstellung der Worte im Inhalt der jeweiligen Domains zu einem spezifischen Thema.
- Die geringen Übereinstimmungsraten bei „Morphologie“ und „Synonymie“ lassen den Einfluss anderer Faktoren beim Ranking als eine sprachliche Bearbeitung vermuten, besonders im Vergleich zu den zufälligen Übereinstimmungsraten.
- Mehrdeutige Worte überfordern sowohl Google als auch BING, auch wenn Google mehr für die Bedeutungsaufklärung der Nutzer unternimmt als BING.
- Die organischen Ergebnisse zeigen bei mehrdeutigen Suchanfragen durchaus eine Tendenz zu einer Bedeutung, die mit der Klickpräferenz der Nutzer übereinstimmt. Allerdings variiert der Grad dieser Übereinstimmung signifikant und es gibt deutliche Ausreißer.
- Die heutigen Suchmaschinen Google und BING werden den bekannten sprachlichen Hürden des klassischen Information Retrievals nicht gerecht.
- Keine Suchmaschine erweist sich als eindeutiger Vorreiter auf den untersuchten sprachlichen Gebieten. Allerdings bietet Google in einigen Fällen eine durchaus nützliche Begriffsklärung für den Nutzer an und zeigt häufig leicht höhere Übereinstimmungsraten.

Die durchgeführte Studie legt nahe, dass die beiden meist genutzten Suchmaschinen Deutschlands Schwierigkeiten bei der technischen Umsetzung typischer deutscher Sprachhürden haben. Somit sollte jeder Nutzer, der sich umfassend informieren will, die Eingabe verschiedener Wortformen und Synonyme in Betracht ziehen. Bei mehrdeutigen Anfragen nehmen die

Suchsysteme den Nutzern weitgehend nicht die eigene intellektuelle Leistung ab. Suchmaschinenbetreiber könnten bei solchen Abfragen vermehrt die Klickpräferenzen der eigenen Nutzer verwenden, um eine Bedeutungsverteilung in den Ergebnissen im Interesse der meisten Nutzer zu gewährleisten.

9. Methodenkritik

Die Daten dieser Studie können allenfalls als Momentaufnahme des Status Quo verstanden werden, und werden aufgrund der agilen Suchmaschinenbranche innerhalb kurzer Zeit als veraltet gelten. Größer angelegte Studien sind natürlich immer aussagekräftiger, weil sie auf mehr Daten basieren, allerdings darf der entsprechende Aufwand nicht unterschätzt werden. Die Betrachtung der TOP10 vs. der Treffer 91-100 hat keinen so gravierenden Unterschied ergeben, als dass man in zukünftigen Studien den Kreis der untersuchten Treffer nicht einfacher wählen könnte. Insofern ist die Aussage, dass es im Grunde kaum einen Unterschied im Ranking gibt, eine lohnende Erkenntnis, aber es stellt sich zu Recht die Frage, wie einem deutliche Unterschiede bei der Interpretation so geringer genereller Übereinstimmungen bei „Synonymie“ und „Morphologie“ weitergeholfen hätten. Zukünftigen Studien in diesem Bereich wird daher empfohlen, Übereinstimmungen bzw. Korrelationen und Häufigkeiten von Bedeutungen für eine größere Trefferanzahl pro Suchbegriff zu überprüfen, mit der Frage im Hinterkopf, ob eine Betrachtung der TOP 50 gegenüber lediglich 20 Positionen die Kennzahlen beeinflussen. Gerade im Bereich Homographie liegt die Vermutung nahe, dass die Vergrößerung der statistischen Basis auch eine Annäherung an die Klickpräferenzen zur Folge haben dürfte. Zukünftige Arbeiten sollten ebenfalls von Anfang an mehr Testworte anlegen, als final als Mengengerüst angestrebt werden, da es immer Punkte geben wird, mit denen man nicht rechnet (z.B. mangelnde Klickdaten bei einigen Homographen, die die hohe Verlustquote von 10 % bei Homographie größtenteils verursachten).

Die Subjektivität der Bedeutungseinordnung bei den Homographen, sowohl bei den organischen Treffern als auch bei den Klickdaten, hätte durch den Einsatz weiterer beurteilender Personen reduziert werden können. Dies

war in dieser Masterarbeit aufgrund personeller Rahmenbedingungen nicht möglich. Zukünftige Studien sollten mehrere Personen zur Beurteilung einsetzen, um zu objektiveren Ergebnissen zu gelangen. Ähnlich kann bei der Erstellung der Testworte von einer gewissen Subjektivität ausgegangen werden und für zukünftige Studien wäre es interessant, tatsächliche Nutzeranfragen für die Eignung einer Studie wie dieser auszuwerten und zu benutzen, anstatt theoretisch entwickelter Anfragen.

Hinsichtlich der Datengrundlage müssen vor allem die Ergebnisse im Bereich „Homographie“ kritisch betrachtet werden. Problematisch ist allein schon die Verwendung von Klickdaten aus der T-Online-Suche, und nicht von Google und BING selbst. Es kann zwar grundsätzlich von einem ähnlichen Klickverhalten ausgegangen werden, aber es handelt sich um eine Übertragung. Wäre eine entsprechende Verfügbarkeit vorhanden, wäre es sinnvoller, die organischen Klickdaten von Google und BING für die verwendeten Homographe auszuwerten. Zudem beruhen die Verhältnisse der Klickpräferenzen für die Bedeutungen auf unterschiedlichsten Suchvolumen und Klickmengen. So gibt es Homographe, bei denen 2000 Suchanfragen für die Klickpräferenzen ausgewertet werden konnten, und andere, für die gerade einmal 100 Suchanfragen zur Verfügung standen. Die Bedeutungsverhältnisse, mit denen die organischen Treffer korreliert wurden, basieren somit auf unterschiedlich großen Datenmengen, und müssen daher auch als unterschiedlich zuverlässig angesehen werden. Dies alles führt dazu, dass die Ergebnisse der Korrelation allenfalls als Richtungshinweis, aber nicht als absolute Aussage zu verstehen sind. So problematisch die Zuverlässigkeit und Übertragungsfähigkeit der Klickdaten von T-Online auch zu betrachten sind – ohne sie hätte es keinerlei Anhaltspunkt gegeben, wie die Bedeutungsverhältnisse der organischen Treffer sich hinsichtlich der Nutzerpräferenzen verhalten.

Zusätzlich darf nicht vergessen werden, dass die zugrundeliegende Stichprobe in allen Fällen mit zwanzig Treffern pro Anfrage klein ist und ein zusätzlicher Wegfall weiterer Treffer für die Korrelation durch andere Kategorien wie „Eigennamen“ die Stichproben hier weiter verringert. Daher können alle Aussagen nur als Indizien gewertet werden.

10. Lessons Learned

Bei der Durchführung der Studie hat sich gezeigt, dass man den Aufwand für bestimmte Tätigkeiten erstaunlich gut unterschätzen kann. Die intellektuelle Leistung der Bedeutungszuordnung bei den Homographen hat bei der Datenbearbeitung den größten Anteil an Zeit eingenommen. Mit entsprechenden Möglichkeiten wäre zudem ein automatisierter Vergleich von URLs, Domains und Rankingpositionen eine nicht zu unterschätzende Arbeitserleichterung. Studien, die für eine größere Datenbasis angelegt sind, werden ohne entsprechende Programmierleistung nicht umzusetzen sein. Wenn man sich ansieht, welche einfachen Kernaussagen über das Verhalten heutiger Suchmaschinen mit sprachlichen Hürden und deren Einordnung in entsprechende Standards des Information Retrievals am Ende dieser Masterarbeit als zentrale Punkte zu nennen sind, ergibt sich die Frage, inwieweit sich der Aufwand gelohnt hat. Die wissenschaftliche Arbeit ist gerechtfertigt, weil diese Studie deutliche Unzulänglichkeiten der Suchmaschinentechnik beim Umgang mit der deutschen Sprache aufzeigt und neue Fragestellungen aufwirft.

11. Grenzen der Studie

Zu den bereits angesprochenen Einschränkungen der Aussagen, die sich hinsichtlich der Aktualität und der Datengrundlage ergeben, kommen die Einschränkungen, die sich aufgrund der nicht betrachteten Frage der Präzision stellen. Sucht man „Adventsgesteck“ oder „Gestecke Advent“ – bei welchem Suchbegriff ist die Präzision für die Nutzer höher? Eine entsprechend angelegte Studie könnte mithilfe von Juroren klären, inwieweit sich die Relevanz bei Synonymabfragen oder morphologischen Varianten unterscheidet und somit inwiefern eine technische Bearbeitung dieser sprachlichen Besonderheiten überhaupt im Web notwendig ist.

Einen weiteren Ansatzpunkt für Studien liefert die Frage nach einem internationalen Vergleich der hier untersuchten sprachlichen Hürden. Können Google und BING mit englischen Homographen besser umgehen als mit deutschen? Wie sieht es im Vergleich mit morphologischen Varianten und Synonymen aus? Ein Vergleich mit der deutschen T-Online-Suche kann

zudem aufzeigen, inwiefern die deutsche Suchmaschine ihrer eigenen Sprache im Vergleich zu internationalen Suchmaschinen gerecht wird.

12. Fazit

Die Analyse von Synonymie, Morphologie und Homographie in dieser Ausführlichkeit und mit dieser Datenbasis wurde zum ersten Mal für die deutsche Sprache durchgeführt. Die Wahl der Testworte hat sich durchaus an vorangegangenen Studien orientiert. Google und BING scheinen technisch gesehen in etwa gleichauf zu liegen, soweit es den Umgang mit „Synonymie“, „Morphologie“ und „Homographie“ betrifft. Google zeigt im Bereich „Homographie“ durch die Einbindung von Ergebnissen aus dem „Knowledge Graph“ eine etwas bessere Leistung bei der Bedeutungsklärung. Die Klickpräferenzen der Nutzer entsprechen von der Tendenz her häufig der tatsächlichen Bedeutungsverteilung der organischen Treffer, wobei es angeraten ist, bei der verwendeten Datengrundlage das Ergebnis nicht überzubewerten. Auf welchen Ursachen die festgestellte Korrelation beruht, etwa einer Berücksichtigung von Klickpräferenzen oder anderen Rankingfaktoren, konnte nicht ermittelt werden. Synonymie, Singular/Plural, oder die Komposition bzw. Zerlegung von Begriffen wirken sich stark auf die Ergebnislisten aus. Inwiefern sich dadurch eine schlechtere Relevanz ergibt, wurde in dieser Studie nicht untersucht. Die Suchmaschinen BING und Google sind den Standards des klassischen Information Retrievals nicht gewachsen, doch stellt sich zurecht die Frage, inwieweit dieses negativ bewertet werden muss. Es bleibt festzuhalten, dass die deutsche Sprache mit Synonymie, Morphologie und Homographie einige Fallstricke für die in Deutschland meist genutzten Suchmaschinen bietet, die bisher technisch nicht berücksichtigt werden. Insofern obliegt es den Nutzern, mit diesen Defiziten kompetent umzugehen. Ausschlaggebend sollte somit ihr Urteil sein, das von der Relevanz der organischen Treffer geleitet sein wird.

13. Glossar

Begriffsrelation	Beziehung zwischen Begriffen (z.B. Hierarchie: Ober- und Unterbegriff)
Crawler	Programm einer Suchmaschine, das URLs folgt und Websites indiziert
Disambiguierung	Bedeutungsklärung bei mehrdeutigen Begriffen
Domain	Hauptname einer Website, unter dem alle zugehörigen URLs zu finden sind (z.B. hamburg.de)
Homograph	Ein Wort, das mehrere Bedeutungen hat, z.B. Jaguar (Raubkatze / Automarke)
Homographie	Oberbegriff für das Vorkommen von gleich geschriebenen Worten mit unterschiedlichen Bedeutungen
Hypothese	Aussagen/Annahmen, die überprüfbar und in sich widerspruchsfrei sein müssen
Index	Verzeichnis einer Suchmaschine über alle ihr bekannten Websites
Indizieren	Inhaltlich & strukturell erfassen
Information Retrieval	Wissenschaft, die sich mit Retrievalsystemen und -techniken befasst, also dem technischen Prozess, wie Nutzer an Informationen gelangen
Informationsballast	Zu viele, auch nicht relevante Informationen werden angeboten
Informationsverlust	Wichtige, relevante Informationen werden nicht angeboten

Komposition	Zusammensetzung von Begriffen
Korrelation	Der mathematische Zusammenhang zwischen zwei Merkmalen, der jedoch keine Kausalität voraussetzt. Ein positiver Zusammenhang zeigt, dass Merkmal B im selben Maß zunimmt wie Merkmal A.
Korrelationskoeffizient	Prozentsatz, der anzeigt, wie stark zwei Merkmale korrelieren. Je höher die Prozentzahl, desto höher ist auch die Korrelation. Sie kann positiv oder negativ sein.
Lemmatisierung	Reduktion einer Wortform auf ihren Stamm anhand linguistischer Regeln
Morphologie	Oberbegriff für grammatikalische Veränderungen eines Wortes (z.B. Singular/Plural)
Organische Treffer	Suchmaschinenergebnisse, die keine Werbeanzeigen sind, also nicht „gekauft“ werden können
Phrase	Mehrere Worte, die sinntragend zusammenhängen (z.B. Deutsche Demokratische Republik)
Präzision	Relevanz der Suchmaschinenergebnisse (auch *Retrievaleffektivität*)
Relevanz	Bedeutung und Nutzen eines Suchmaschinentreffers für das Bedürfnis eines Nutzers
Retrievaleffektivität	siehe *Präzision*
Retrievalsystem	Ein System, das Informationen indiziert und auffindbar macht

Semantik	Wissenschaft, die sich mit der Bedeutung von Worten beschäftigt.
Stammformreduktion	siehe *Stemming*
Stemming	Reduktion eines Begriffs auf seine Stammform (z.B. gelaufen zu lauf), auch „Stammformreduktion“, basiert auf einfachen Basisregeln im Gegensatz zur *Lemmatisierung*
Stoppworte	Worte in einer Sprache, die häufig vorkommen, aber keinen Sinn tragen (z.B. Artikel, Präpositionen)
Suchmaschinen-optimierung	Sämtliche Maßnahmen, mit denen das organische Ranking von Suchmaschinen beeinflusst wird (z.B. Verlinkungen)
Synonym	Ein Wort ist dann ein Synonym, wenn ein anderes Wort existiert, dass dieselbe Bedeutung hat (z.B. Orange & Apfelsine)
Synonymie	Oberbegriff für das Vorkommen von Synonymen
Thesaurus	Nachschlagewerk für Begriffe und deren Beziehungen (Ober-, Unterbegriff, Synonyme etc.)
Treffer-Snippets	Trefferdaten der Suchergebnisse (Titel, Beschreibung, URL etc.)

14. Literaturverzeichnis

Aloufi 2010

ALOUFI, Khalid S. R.: Information Retrieval of Text with Diacritics. In: *International Journal of Computer Science and Network Security* 10 (2010), Nr. 8, S. 118-122. - URL http://paper.ijcsns.org/07_book/201008/20100818.pdf. - Abruf 2013-04-11

Alpkoçak 2012

ALPKOÇAK, Adil; CEYLAN, Meltem: Effects of diacritics on Turkish information retrieval. In: *Turkish Journal of Electrical Engineering & Computer Sciences* 20 (2012), Nr.5, S. 787-804. - ISSN 1303-6203. - URL http://journals.tubitak.gov.tr/elektrik/issues/elk-12-20-5/elk-20-5-9-1010-819.pdf. - Abruf 2013-03-26

Baeza-Yates 2011

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier: *Modern Information Retrieval : the concepts and technology behind search*. 2nd ed. Harlow [u.a.] : Addison Wesley, 2011. - ISBN 978-0-321-41691-9

Bar-Ilan 2004

BAR-ILAN, Judit; GUTMAN, Tatyana: How do search engines respond to some non-English queries? In: *Journal of Information Science* 31 (2005), Nr. 1, S. 13-28. - ISSN 2163-193X. - URL http://jis.sagepub.com/content/31/1/13.abstract. - Abruf 2013-02-05

BBC 2009

BRITISH BROADCASTING CORPORATION (BBC): *Microsoft and Yahoo seal web deal*. - URL http://news.bbc.co.uk/2/hi/business/8174763.stm. - Abruf 2013-04-29

Bortz 2010

BORTZ, Jürgen; SCHUSTER, Christof: *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin [u.a.] : Springer, 2010 (Springer Lehrbuch). - ISBN 978-3-642-12770-0. - ISSN 0937-7433.- S. 153-182

Büttcher 2010

BÜTTCHER, Stefan; CLARKE, Charles L.A.; CORMACK, Gordon V.: *Information Retrieval : Implementing and Evaluating Search Engines*. Cambridge, Mass. [u.a.] : MIT Press, 2010. - ISBN 978-0-262-02651-2

Choro 2005

CHORO, K.: Testing the effectiveness of retrieval to queries using polish words with diacritics. In: AWIC *Lecture Notes in Artificial Intelligence* 3528 (2005), S. 101-106

Chowdhury 2008

CHOWDHURY, G.G.: *Introduction to modern information retrieval*. 2. ed. London : facet, 2008. - ISBN 978-1-85604-480-6

Croft 2010

CROFT, W. Bruce; METZLER, Donald; STROHMAN, Trevor: *Search Engines : Information Retrieval in Practice*. Boston [u.a.] : Pearson, 2010. - ISBN 978-0-13-136489-9

Demirci 2007

DEMIRCI, R.G.; KISMIR, V.; BITIRIM, Y.: *An evaluation of popular search engines on finding Turkish Documents* (*Second International Conference on Internet and Web Applications and Services ICIW07*). Famagusta [Cyprus] : IEEE Computer Society, 2007. - ISBN 0-7695-2844-9. - URL http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4222963&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4222963. - Abruf 2013-08-16

Dethloff 2012

DETHLOFF, Jacqueline: *459 Teekesselchen*. - URL http://www.originell-betreut.org/300679_1553807.htm. - Abruf 2013-03-12

Dettweiler 2009

DETTWEILER, Marco; FRANKFURTER ALLGEMEINE ZEITUNG (Hrsg.): *Microsofts Suchmaschine : Die Maschine mit dem Bing*. Stand: 03.06.2009. - URL http://www.faz.net/aktuell/technik-motor/computer-internet/microsofts-suchmaschine-die-maschine-mit-dem-bing-1596438.html. - Abruf 2013-02-05

Duden 2012

DUDENREDAKTION (Hrsg.): *Duden - das Wörterbuch der Synonyme : rund 100.000 Stichwörter und Synonyme für den alltäglichen Schreibgebrauch*. Mannheim : Bibliographisches Institut, 2012. - ISBN 978-3-411-74482-4

Efthimiadis 2008

EFTHIMIADIS, E.N.; MALEVRIS, N.; KOUSARIDAS, A.; LEPENIOTOU, A.; LOUTAS, N.: An evaluation of how search engines respond to greek language queries. In: *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, *7-10 Jan. 2008*. S.136. - URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4438839&isnumber=4438696. - Abruf 2013-02-05

Eimeren 2012

VAN EIMEREN, Birgit; FREES, Beate: Ergebnisse der ARD/ZDF-Onlinestudie 2012 : 76 Prozent der Deutschen online – neue Nutzungssituationen durch mobile Endgeräte. In: *Media Perspektiven* (2012), Nr. 7–8, S. 362-379. - ISSN 0170-1754. - URL http://www.ard-zdf-onlinestudie.de/fileadmin/Online12/0708-2012_Eimeren_Frees.pdf. - Abruf 2013-02-05

EU 2011

EUROPÄISCHE KOMMISION (Hrsg.): *User language preferences online : survey conducted by The Gallup Organization, Hungary upon the request of Directorate-General Information Society and Media*. - URL http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf. - Abruf 2013-02-05

Familie 2013

FAMILIE-ONLINE.DE (Hrsg.): *Teekesselchen bei Familie-online*. - URL http://www.familie-online.de/tee.shtml. - Abruf 2013-03-12

Ferber 2003

FERBER, Reginald: *Information Retrieval : Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. 1. Aufl. Heidelberg : dpunkt, 2003. - ISBN 3-89864-213-5

Google 2012

GOOGLE (Hrsg.): *Introducing the Knowledge Graph : things, not strings*. Stand: 16.05.2012. - URL http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html. - Abruf 2013-06-28

Google 2013a

GOOGLE (Hrsg.): *Unternehmensprofil*. - URL http://www.google.de/about/company/. - Abruf 2013-02-05

Google 2013b

GOOGLE (Hrsg.): *Grundlegende Hilfe für die Suche*. - URL http://support.google.com/websearch/bin/answer.py?hl=de&answer=134479. - Abruf 2013-03-21

Google 2013c

GOOGLE (Hrsg.): *Spracheinstellungen ändern*. - URL http://support.google.com/toolbar/bin/answer.py?hl=de&answer=9279. - Abruf 2013-03-21

Google 2013d

GOOGLE (Hrsg.): *Die Option „Wortwörtlich“.* - URL http://support.google.com/websearch/bin/answer.py?hl=de&answer=1734130&topic=1221265&ctx=topic. - Abruf 2013-03-21

Google 2013e

GOOGLE (Hrsg.): *Suchoperatoren*. - URL http://support.google.com/websearch/answer/136861?hl=de&ref_topic=1221265. - Abruf 2013-04-14

Griesbaum 2002

GRIESBAUM, Joachim; BEKAVAC, Bernard; RITTBERGER, Marc: *Deutsche Suchmaschinen im Vergleich : AltaVista.de, Fireball.de, Google.de und Lycos.de*. Konstanz [u.a.]: Universität Konstanz/Heinrich-Heine-Universität Düsseldorf, 2002. - URL http://www.informatik.fh-kl.de/~amueller/vorlesungen/ir/griesbaum_rittberger_bekavac.pdf. - Abruf 2013-04-10

Griesbaum 2004

GRIESBAUM, Joachim: Evaluation of three German search engines : Altavista.de, Google.de and Lycos.de. In: *Information Research* 9 (2004), Nr. 4. - ISSN 1368-1613. - URL http://eprints.rclis.org/5746/1/paper189.html. - Abruf 2013-04-10

Griesbaum 2009

GRIESBAUM, Joachim; BEKAVAC, Bernard; RITTBERGER, Marc: *Typologie der Suchdienste im Internet*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen : Nutzerorientierung in Wissenschaft und Praxis.* Heidelberg : Akademische Verlagsgesellschaft, 2009. - ISBN 978-3-89838-607-4. - S. 18-52

Günther 2012

GÜNTHER, Markus: *Evaluierung von Suchmaschinen : Qualitätsvergleich von Google-und Bing-Suchergebnissen unter besonderer Berücksichtigung von Universal-Search-Resultaten*. - URL http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/SWiF2012/SWIF_Guenther.pdf. - Abruf 2013-02-23

Guggenheim 2005

GUGGENHEIM, Esther; BAR-ILAN, Judit: Tauglichkeit von Suchmaschinen für deutschsprachige Abfragen. - In: *Information Wissenschaft & Praxis* 56 (2005), Nr. 1, S. 35-40. - ISSN 1434-4653

Halim 2006

HALIM, Hananzita; KAUR, Kiran: Malaysian web search engines : a critical analysis. In: *Malaysian Journal of Library & Information Science* 11 (2006), Nr. 1, S. 103-122. - ISSN 1394-6234. - URL http://eprints.um.edu.my/631/1/web_search_engines_kiran.pdf. - Abruf 2013-02-05

Hammo 2009

HAMMO, Bassam H.: Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. In: *Information Retrieval* 12 (2009), Nr. 3, S. 300-323. - ISSN 1573-7659. - URL http://link.springer.com/content/pdf/10.1007/s10791-008-9081-9. - Abruf 2013-03-26

HAW 2013

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN (HAW) HAMBURG: *RAT : Relevance Assessment Tool*. - URL http://searchstudies.org/website/index.php/rat.de.html. - Abruf 2013-07-24

Joachims 2007

JOACHIMS, Thorsten; GRANKA, Laura; PAN, Bing; HEMBROOKE, Helene; RADLINSKI, Filip; GAY, Geri: Evaluation the accuracy of implicit feedback from clicks and query reformulations in web search. In: *ACM Transactions on Information Systems* 25 (2007), Nr. 2, Artikel 7. - URL http://sing.stanford.edu/cs303-sp11/papers/joachims_etal_07a.pdf. - Abruf 2013-06-21

Karanikolas 2009

KARANIKOLAS, Nikitas N.: *Bootstrapping the Albanian Information Retrieval* (*Fourth Balkan Conference in Informatics 2009*). S. 231-235. - URL http://www.informatik.uni-trier.de/~ley/pers/hd/k/Karanikolas:Nikitas_N=.html. - Abruf 2013-03-25

Kettunen 2012

KETTUNEN, Kimmo; ARVOLA, Paavo: *Generating variant keyword forms for a morphologically complex language leads to successful information retrieval with Finnish*. In: SALAMPASIS, Michail; LARSEN, Birger: *Multidisciplinary Information Retrieval (5th International Retrieval Facility Conference, IRFC 2012, Vienna, Austria, July 2-3, 2012)*. Berlin [u.a.] : Springer, 2012 (Lecture Notes in Computer Science 7356). - S. 113-126. - ISBN 978-3-642-31274-8. - URL http://link.springer.com/chapter/10.1007%2F978-3-642-31274-8_10?LI=true. - Abruf 2013-02-05

Lazarinis 2007a

LAZARINIS, Fortis: Web retrieval systems and the Greek language : do they have an understanding? In: *Journal of Information Science* 33 (2007), Nr. 3, S. 622–636. - ISSN 2163-193X

Lazarinis 2007b

LAZARINIS, Fortis (Hrsg.); VILARES, Jesus (Hrsg.); TAIT, John I. (Hrsg.): *ACM SIGIR 2007 workshop : improving non English web searching (iNEWS'07) held in conjunction with the 30th Annual International ACM SIGIR Conference*. - ISBN: 978-84-690-6978-3. - URL http://www.grupolys.org/biblioteca/LazVilTai2007b.pdf. - Abruf 2013-08-16

Lazarinis 2008

LAZARINIS, Fortis (Hrsg.); EFTHIMIADIS, Efthimis N.; VILARES, Jesus (Hrsg.); TAIT, John I. (Hrsg.): *Proceedings of the 2nd ACM workshop on improving non english web searching 2008*. - URL http://dl.acm.org/citation.cfm?id=1460027&picked=prox&CFID=303046425&CFTOKEN=25311402. - Abruf 2013-03-26

Lazarinis 2009

LAZARINIS, Fortis; VILARES, Jesús; TAIT, John; EFTHIMIADIS, Efthimis N.: Current research issues and trends in non-English Web searching. In: *Information Retrieval* (2009), Nr. 12, S. 230-250. - ISSN 1573-7659

Leturia 2012

LETURIA, Igor; GURRUTXAGA, Antton; ARETA, Nerea; ALEGRIA, Iñaki; EZEIZA, Aitzol: Morphological query expansion and language-filtering words for improving Basque web retrieval. In: *Language Resources and Evaluation* (2012). - ISSN 1574-0218. - URL http://link.springer.com/article/10.1007/s10579-012-9208-x#. - Abruf 2013-04-11

Levene 2010

LEVENE, Mark: *An introduction to search engines and web navigation*. - Hoboken [u.a.] : Wiley, 2010. - ISBN 978-0-470-52684-2. - URL http://edu.ercess.co.in/ebooks/internet/An-Introduction-to-Search-Engines-and-Web-Navigation.pdf. - Abruf 2013-04-10

Lewandowski 2005

LEWANDOWSKI, Dirk, OCKENFELD, Marlies (Hrsg.): *Web Information Retrieval : Technologien zur Informationssuche im Internet*. Frankfurt am Main : DGI, 2005 (Informationswissenschaft 7). - ISSN 0940-6662. - ISBN 3-925474-55-2. - URL http://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/Web_Information_Retrieval_Buch.pdf. - Abruf 2013-04-08

Lewandowski 2007

LEWANDOWSKI, Dirk: *Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen?* In: MACHILL, Marcel; BEILER, Markus (Hrsg.): *Die Macht der Suchmaschinen / The Power of Search Engines*. Köln : Herbert von Halem Verlag, 2007. - ISBN 978-3-938258-33-0. - URL http://eprints.rclis.org/16086/1/Kennzahlen_Qualitaet_von_Suchmaschinen.pdf. - Abruf 2013-04-10

Lewandowski 2008a

LEWANDOWSKI, Dirk: Problems with the use of Web search engines to find results in foreign languages. In: *Online Information Review* 32 (2008), Nr. 5, S. 668–672. - ISSN 1468-4527. - URL http://www.durchdenken.de/lewandowski/doc/OIR2008_Preprint.pdf. - Abruf 2013-04-11

Lewandowski 2008b

LEWANDOWSKI, Dirk: The retrieval effectiveness of web search engines : considering results descriptions. In: *Journal of Documentation* 64 (2008), Nr. 6, S. 915-937. - ISSN 0022-0418. - URLhttp://www.bui.haw-hamburg.de/fileadmin/user_upload/lewandowski/doc/retrieval_effectiveness_results_descriptions_JDoc2008.pdf. - Abruf 2013-04-11

Lewandowski 2009

LEWANDOWSKI, Dirk: *The retrieval effectiveness of search engines on navigational queries*. - URL http://eprints.rclis.org/17233/1/ASLIB2011_preprint.pdf. - Abruf 2013-04-10

Lewandowski 2011a

LEWANDOWSKI, Dirk: Evaluierung von Suchmaschinen. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2011. S. 203-228. - ISBN 978-3898386517

Lewandowski 2011b

LEWANDOWSKI, Dirk: The retrieval effectiveness of search engines on navigational queries. In: *ASLIB Proceedings 62* (2011), Nr. 4, S. 354-363. - URL http://eprints.rclis.org/17233/1/ASLIB2011_preprint.pdf. - Abruf 2013-04-11

Lewandowski 2011c

LEWANDOWSKI, Dirk: *Suchmaschinen-Update*. In: OCKENFELD, Marlies (Hrsg.): *Web 3.0 – wird es das Web der Informationsspezialisten*? (*Proceedings des 26. Oberhofer Kolloquiums*). Frankfurt am Main : Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis, 2011. - URL http://eprints.rclis.org/17230/1/Suchmaschinen-Update_2011_preprint.pdf. - Abruf 2013-04-16

Lewandowski 2012

LEWANDOWSKI, Dirk; SÜNKLER, Sebastian: *Relevance Assessment Tool : ein Werkzeug zum Design von Retrievaltests sowie zur weitgehend automatisierten Erfassung, Aufbereitung und Auswertung von Daten*. - In: OCKENFELD, Marlies; PETERS, Isabelle; WELLER, Katrin: *Proceedings der 2. DGI-Konferenz : Social Media und Web Science - das Web als Lebensraum*. Frankfurt am Main : DGI, 2012. - S. 237-249. - URL http://eprints.rclis.org/17261/1/RAT_DGI_Lewandowski_Suenkler_preprint.pdf. - Abruf 2013-04-11

Liu 2012

LIU, Yu-Chin; LIN, Chun-Wei: A new method to compose long unknown Chinese keywords. In: *Journal of Information Science* 38 (2012), Nr. 4, S. 366-382. - ISSN 2163-193X. - URL http://jis.sagepub.com/content/38/4/366. - Abruf 2013-02-05

Manning 2010

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich: *Introduction to Information Retrieval*. Cambridge : Cambridge University Press, 2010. - ISBN 978-0-521-86571-5

Microsoft 2013a

MICROSOFT (Hrsg.); YAHOO (Hrsg.): *Yahoo Bing network : about*. - URL http://yahoobingnetwork.com/en/about. - Abruf 2013-04-29

Microsoft 2013b

MICROSOFT (Hrsg.): *BING Hilfe*. - URL http://onlinehelp.microsoft.com/de-de/bing/help.aspx. - Abruf 2013-03-21

Microsoft 2013c

MICROSOFT (Hrsg.): *BING : Allgemeine Einstellungen*. - URL http://www.bing.com/settings.aspx. - Abruf 2013-03-21

Microsoft 2013d

MICROSOFT (Hrsg.): *BING Query Language*. - URL http://msdn.microsoft.com/en-us/library/ff795667.aspx. - Abruf 2013-04-15

Moukdad 2001

MOUKDAD, Haidar: Information retrieval from full-text arabic databases : can search engines designed for english do the job? In: *Libri* 51 (2001), S. 63-74. - ISSN 0024-2667. - URL http://www.librijournal.org/pdf/2001-2pp63-74.pdf. - Abruf 2013-02-05

Moukdad 2004

MOUKDAD, Haidar: Lost In Cyberspace : how do search engines handle arabic queries? In: *Proceedings of the 32nd Annual Conference of the Canadian Association for Information Science*. - URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.2087. - Abruf 2013-02-05

Moukdad 2005

MOUKDAD, Haidar; CUI, Hong: How do search engines handle chinese queries? In: *Webology* 2 (2005), Nr. 3, Article 17. - URL http://www.webology.org/2005/v2n3/a17.html. - Abruf 2013-02-05

Nohr 2003

NOHR, Holger: *Grundlagen der automatischen Indexierung : ein Lehrbuch*. Berlin : Logos, 2003. - ISBN 3-8325-0121-5

Rather 2008

RATHER, Rafiq Ahmad; LONE, Fayaz Ahmad; SHAH, Gulam Jeelani: Overlap in web search results : a study of five search engines. In: *Library Philosophy and Practice*, Artikel 226, 2008. - ISSN 1522-0222. - URL http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1228&context=libphilprac. - Abruf 2013-04-13

Savoy 2008

SAVOY, Jacques: *Searching strategies for the hungarian language*. In: *Information Processing & Management* 44 (2008), Nr. 1, S.310-324. - ISSN 0306-4573. - URL http://members.unine.ch/jacques.savoy/papers/huipm.pdf. - Abruf 2013-03-26

Schmetzke 1998

SCHMETZKE, Axel: The utility of German WWW search services for North American users. In: *Reference Services Review* 26 (1998), Nr. 1, S.43-50. - ISSN 0090-7324

Schmidt 2012

SCHMIDT, Holger: Suchmaschinen : Google bekommt neue Konkurrenz. In: *Focus Online*. Stand: 16.09.2012. - URL http://www.focus.de/digital/internet/netzoekonomie-blog/suchmaschinen-google-bekommt-neue-konkurrenz_aid_820575.html. - Abruf 2013-02-21

Schwartz 2013

SCHWARTZ, Barry: *Google's Matt Cutts : domain clustering to change again : fewer results from same domain*. Stand: 17.05.2013. - URL http://searchengineland.com/google-domain-clustering-change-159997. - Abruf 2013-07-01

SEW 2013

SEARCH ENGINE WATCH (SEW): *Articles on SEO*. - URL http://searchenginewatch.com/seo. - Abruf 2013-02-23

Shatnawi 2012

SHATNAWI, Mohammad Q.; YASSEIN, Muneer Bani; MAHAFZA, Reem: A framework for retrieving Arabic documents based on queries written in Arabic slang language. In: *Journal of Information Science* 38 (2012), Nr. 4, S. 350-365. - ISSN 1741-6485. - URL http://jis.sagepub.com/content/38/4/350. - Abruf 2013-03-26

Smith 2010

SMITH, Alastair G.: Internet search tactics. In: *Online Information Review* 36 (2010), Nr. 1, S. 7-20. - ISSN1468-4527. - URL http://www.emeraldinsight.com/journals.htm?issn=1468-4527&volume=36&issue=1&articleid=17019374&show=abstract. - Abruf 2013-04-14

Spiegel 2011

SPIEGEL, Sebastian Reiner: *Machine learning for the analysis of the morphologically complex languages*. University of Bristol, 2011. - URL http://www.cs.bris.ac.uk/Publications/Papers/2001371.pdf. - Abruf 2013-03-26

Spink 2006

SPINK, Amanda; JANSEN, Bernard J.; BLAKELY, Chris; KOSHMAN, Sherry: A study of results overlap and uniqueness among major web search engines. In: *Information Processing and Management* 42 (2006), Nr. 5, S. 1379-1391. - ISSN 0306-4573. - URL http://eprints.qut.edu.au/4755/1/4755.pdf. - Abruf 2013-04-10

Spree 2011

SPREE, Ulrike; FEIßT, Nadine; LÜHR, Anneke [u.a.]: *Semantic Search – State-of-the-Art-Überblick zu semantischen Suchlösungen im WWW*. In: LEWANDOWSKI, Dirk (Hrsg.): *Handbuch Internet-Suchmaschinen 2 : Neue Entwicklungen in der Web-Suche*. Heidelberg : Akademische Verlagsgesellschaft Aka, 2011, S. 203-228. - ISBN 978-3898386517

Sroka 2000

SROKA, Marek: Web search engines for polish information retrieval : questions of search capabilities and retrieval performance. In: *The International Information & Library Review* 32 (2000), Nr. 2, S. 87-98. - ISSN 1057-2317. - URL http://www.sciencedirect.com/science/article/pii/S1057231700901280. - Abruf 2013-08-23

Stacey 2004

STACEY, Alison; STACEY, Adrian: *Effective information retrieval from the internet : an advanced user's guide*. Oxford [u.a.] : Chandos, 2004. - ISBN 1-84334-077-1

Stock 2007

STOCK, Wolfgang G.: *Information Retrieval : Informationen suchen und finden*. München [u.a.] : Oldenbourg Verl., 2007. - ISBN 978-3-486-58172-0

Toth 2006

TOTH, Erzsébet: Exploring the capabilities of English and Hungarian search engines for various queries. In: *Libri* 56 (2006), S. 38-47. - ISSN 0024-2667. - URL http://www.librijournal.org/pdf/2006-1pp38-47.pdf. - Abruf 2013-02-05

Turney 2001

TURNEY, Peter: *Mining the Web for Synonyms : PMI-IR Versus LSA on TOEFL*. In: DE RAEDT, Luc (Hrsg.); FLACH, Peter A. (Hrsg.): *Machine Learning (EMCL 2001, 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001, Proceedings)*. Berlin [u.a.] : Springer, 2001 (Lecture Notes in Computer Science 2167). - ISBN 3-540-42536-5. - URL http://www.informatik.uni-trier.de/~ley/db/conf/ecml/ecml2001.html. - Abruf 2013-03-25

Wang 2007

WANG, Fu Lee; YANG, Christopher C.: Mining web data for Chinese segmentation. In: *Journal of the American Society for Information Science and Technology* 58 (2007), Nr. 12, S. 1820–1837. - ISSN 1532-2890. - URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20629/abstract. - Abruf 2013-03-26

Weber 1996

WEBER, Heinz-Josef: *Homographen-Wörterbuch der deutschen Sprache*. Berlin [u.a.] : de Gruyter, 1996. - ISBN 3-11-014641-X

Anhang

Anhang a - Testworte

Testworte „Synonymie“ (Nicht verwendete Testworte sind grau hinterlegt)

Birma	Myanmar
minderjährig	nicht volljährig
Abendessen	Abendbrot
Abfalleimer	Mülleimer
Toilette	WC
ewig	endlos
exakt	akkurat
Vorgesetzter	Chef
Selters	Sprudelwasser
Couch	Sofa
Forscher	Wissenschaftler
Banknoten	Geldscheine
Fantasie	Vorstellungskraft
Feind	Gegner
Gatte	Ehemann
Garnelen	Hummerkrabben
feuerfest	unbrennbar
Gehalt	Lohn
Internet	www
Karotte	Möhre
Keuschheit	Enthaltsamkeit
Lauch	Porree
Paniermehl	Semmelbrösel
Leberfleck	Muttermal
Maut	Straßenzoll
Vierundzwanzig Uhr	Mitternacht
Dessert	Nachtisch
Nilpferd	Flusspferd
Pacht	Miete
Geist	Gespenst
Pantoffel	Hausschuh
Feier	Fest
Papa werden	Vater werden
gesetzwidrig	rechtswidrig
Großmutter	Oma
PR	Öffentlichkeitsarbeit
Rabatt	Preisnachlass
haarlos	glatzköpfig
Reißzwecke	Heftzwecke
Hochzeit	Trauung
Jalousie	Rolladen
Rundfunk	Radio
Passwort	Kennwort

Sauna	Dampfbad
Radler	Alsterwasser
Schlachter	Fleischer
Satan	Teufel
Neid	Eifersucht
Schrebergarten	Kleingarten
Schwarzhandel	Schmuggel
Seherin	Weissagerin
Tatsachen	Fakten
Teigrolle	Nudelholz
Piano	Klavier
brandschatzen	in Brand setzen
Brautkleid	Hochzeitskleid
Eidotter	Eigelb
Freiheitsstrafe	Gefängnisstrafe
Realität	Wirklichkeit
Schlips binden	Krawatte binden

Testworte „Morphologie“ (Nicht verwendete Testworte sind grau hinterlegt)

Allergie	Allergien
Baum	Bäume
Bewerbungsmappe	Bewerbungsmappen
Buch	Bücher
Ei	Eier
Einhorn	Einhörner
Feind	Feinde
Fußballstadion	Fußballstadien
Gespenst	Gespenster
Heizung	Heizungen
Hochhaus	Hochhäuser
Information	Informationen
Kissenbezug	Kissenbezüge
Lamm	Lämmer
Licht	Lichter
Ohrring	Ohrringe
Pizza	Pizzen
Postkarte	Postkarten
Ragout	Ragouts
Regenschirm	Regenschirme
Schadstoff	Schadstoffe
Schneemann	Schneemänner
Sehenswürdigkeit	Sehenswürdigkeiten
Sonnendach	Sonnendächer
Strandkorb	Strandkörbe
Weinglas	Weingläser

Windlicht	Windlichter
Adventsgesteck	Gestecke Advent
Altersheim	Altenheim
Autoscheinwerfer	Schweinwerfer fürs Auto
Braunbär	brauner Bär
Druckerpapier	Papier Drucker
Fensterrahmen	Rahmen für Fenster
Flugzeugträger	Träger Flugzeuge
Glashandwerk	Handwerk Glas
Glaskanne	Kanne aus Glas
Heizung reparieren	Reparation Heizung
Heizungsmonteur	Monteuer für Heizung
Informationszeitalter	Zeitalter der Information
Kaffeebohnen	Bohnenkaffee
Kerzenhalter	Halter für Kerzen
Kinofilmstarts	Start Kinofilme
Krawattenhalter	Halter für Krawatten
Kristallgläser	Gläser aus Kristall
Lackschicht	Schicht aus Lack
Neuer Duschkopf	Duschkopf neu
Rasierapparat	Apparat zum Rasieren
Reiseversicherung	Versicherungen für Reisen
Restauranteröffnung	Restaurant eröffnen
Salatschüssel	Schüsseln für Salat
Satellitenfernsehen	Fernsehen mit Satellit
Schiffsreise	Reisen mit Schiff
Schmuckkästchen	Kästchen für Schmuck
Spiegelglas	Glas für Spiegel
Teeservice	Service für Tee
Tierfutter	Futter für Tiere

Testworte „Homographie“ (Nicht verwendete Testworte sind grau hinterlegt)

Kaschmir	Jaguar
Schloss	Schimmel
Fächer	Krebs
Wachtraum	Schuppen
Wachstube	Schwamm
Herde	Grillen
Heide	Flechten
August	Drosseln
Wolfram	Amerikaner
Jasmin	arm
Job	Atlas
Hamburger	Bank
modern	Bart

d

Bienenstich	Blüte
Bremse	Engländer
Fuchschwanz	Ente
Horn	Golf
Käfer	Mandeln
Knete	Uhu
Kuhfuß	Dietrich
Libelle	Fliege
Lippe	Fingerhut
Mars	Otter
Morgenstern	Tau
Pickel	Taube
Po	Fuge
Porto	Kater
Raupe	Speiche
Spiegel	Vitamin B
Star	Kiefer
Wanze	Krone
Zelle	Mangel
Zylinder	sieben
Watt	Ball
Becken	Laster

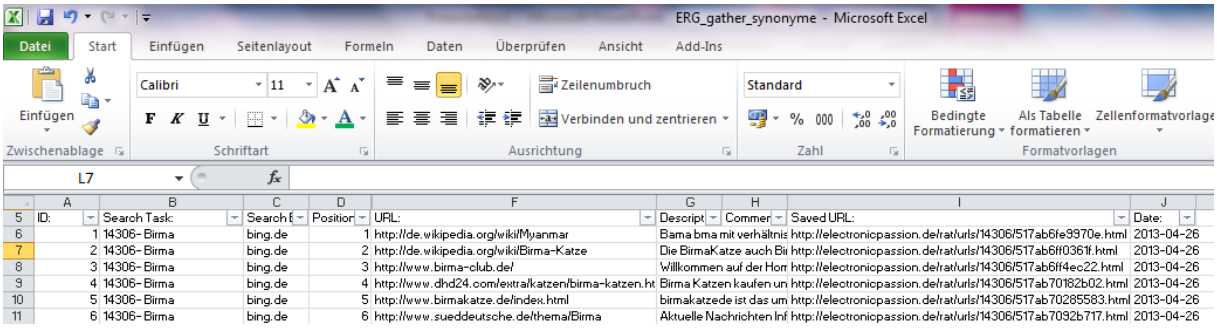
Anhang b – Verfahrensbeschreibung Auswertung

Verfahrensbeschreibung zur Auswertung im Bereich „Synonymie“

Da das Verfahren im Bereich „Synonymie“ weitgehend dem im Bereich „Morphologie“ entspricht, soll die angewendete Vorgehensweise bei der Synonymie exemplarisch beschrieben werden. Der einzige Unterschied der Auswertungsverfahren beläuft sich darauf, dass für „Morphologie“ die Berechnungen der Kennzahlen je für „Singular/Plural“ und für „Komposition/Zerlegung“ durchgeführt wurden.

Schritt 1: Datenkontrolle

Die vom RAT gelieferten Rohdaten wurden zunächst auf Fehler (fehlende Trefferdaten) hin untersucht. Hier ein Auszug der gelieferten Rohdaten:



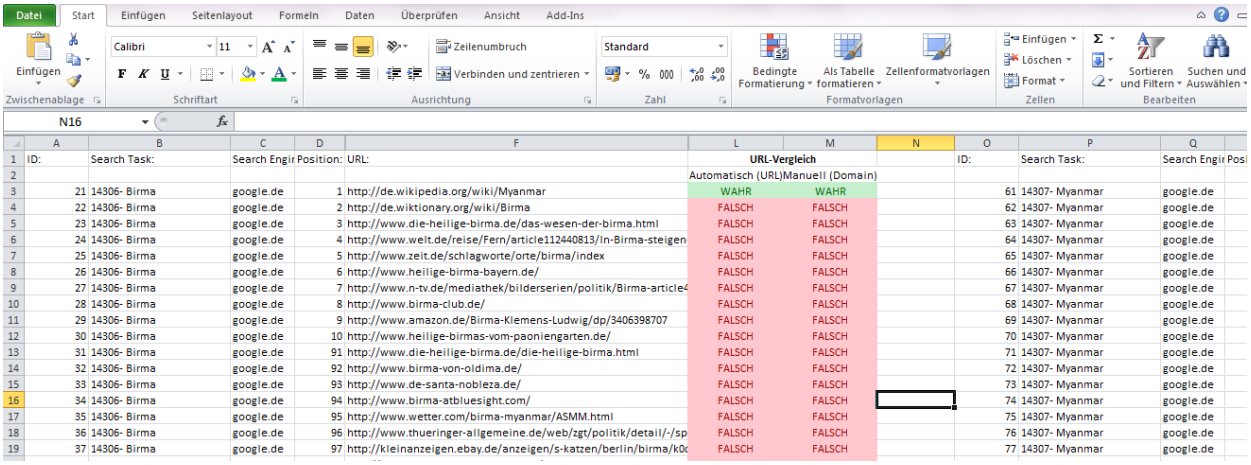
	A	B	C	D	F	G	H	I	J
5	ID:	Search Task:	Search E	Positior	URL:	Descript	Commer	Saved URL:	Date:
6	1	14306- Birma	bing.de	1	http://de.wikipedia.org/wiki/Myanmar	Bama bma mit verhältnis		http://electronicpassion.de/rat/urls/14306/517ab6fe9370e.html	2013-04-26
7	2	14306- Birma	bing.de	2	http://de.wikipedia.org/wiki/Birma-Katze	Die BirmaKatze auch Bi		http://electronicpassion.de/rat/urls/14306/517ab6ff0361f.html	2013-04-26
8	3	14306- Birma	bing.de	3	http://www.birma-club.de/	Willkommen auf der Hon		http://electronicpassion.de/rat/urls/14306/517ab6ff4ec22.html	2013-04-26
9	4	14306- Birma	bing.de	4	http://www.dhd24.com/extra/katzen/birma-katzen.ht	Birma Katzen kaufen un		http://electronicpassion.de/rat/urls/14306/517ab70182b02.html	2013-04-26
10	5	14306- Birma	bing.de	5	http://www.birmakatze.de/index.html	birmakatzede ist das um		http://electronicpassion.de/rat/urls/14306/517ab70285583.html	2013-04-26
11	6	14306- Birma	bing.de	6	http://www.sueddeutsche.de/thema/Birma	Aktuelle Nachrichten Inf		http://electronicpassion.de/rat/urls/14306/517ab7092b717.html	2013-04-26

Die Suchanfragen, bei denen fehlerhafte Daten auftauchten, wurden für eine zweite Datenabfrage verwendet, und die fehlerhaften Daten durch die nunmehr korrekt erfassten Daten ersetzt (und grün markiert). Die danach noch verbliebenen Fehler wurden grau markiert und endgültig aus der Bewertung entfernt.

Schritt 2: Ermittlung der Übereinstimmungsraten

In je einem neuen Datenblatt wurden die Daten für Google und für BING so zusammengestellt, dass sich die Treffer für „Synonym 1“ und „Synonym 2“ genau gegenüberstehen. In der Mitte wurden zwei Spalten dafür benutzt, die Übereinstimmungen der URLs zu überprüfen. In der linken Spalte fand eine automatische Überprüfung der kompletten URL mittels der Excel-Funktion „Identisch(X,Y)“ statt. In der rechten Spalte wurden manuell mit den Worten „WAHR“ (Übereinstimmung zutreffend) und „FALSCH“ (Übereinstimmung nicht zutreffend) Übereinstimmungen der Domains festgehalten.

Durch eine farbliche Hinterlegung der Worte „WAHR“ und „FALSCH“ wurde schnell ersichtlich, dass es nur wenige Übereinstimmungen gibt. Hier ein Auszug dieser Darstellung:



	A	B	C	D	F	L	M	N	O	P	Q
1	ID:	Search Task:	Search Engi	Position:	URL:	URL-Vergleich			ID:	Search Task:	Search Engi
2						Automatisch (URL)	Manuell (Domain)				
3	21	14306- Birma	google.de	1	http://de.wikipedia.org/wiki/Myanmar	WAHR	WAHR		61	14307- Myanmar	google.de
4	22	14306- Birma	google.de	2	http://de.wiktionary.org/wiki/Birma	FALSCH	FALSCH		62	14307- Myanmar	google.de
5	23	14306- Birma	google.de	3	http://www.die-heilige-birma.de/das-wesen-der-birma.html	FALSCH	FALSCH		63	14307- Myanmar	google.de
6	24	14306- Birma	google.de	4	http://www.welt.de/reise/Fern/article112440813/In-Birma-steigen	FALSCH	FALSCH		64	14307- Myanmar	google.de
7	25	14306- Birma	google.de	5	http://www.zeit.de/schlagworte/orte/birma/index	FALSCH	FALSCH		65	14307- Myanmar	google.de
8	26	14306- Birma	google.de	6	http://www.heilige-birma-bayern.de/	FALSCH	FALSCH		66	14307- Myanmar	google.de
9	27	14306- Birma	google.de	7	http://www.n-tv.de/mediathek/bilderserien/politik/Birma-article4	FALSCH	FALSCH		67	14307- Myanmar	google.de
10	28	14306- Birma	google.de	8	http://www.birma-club.de/	FALSCH	FALSCH		68	14307- Myanmar	google.de
11	29	14306- Birma	google.de	9	http://www.amazon.de/Birma-Klemens-Ludwig/dp/3406398707	FALSCH	FALSCH		69	14307- Myanmar	google.de
12	30	14306- Birma	google.de	10	http://www.heilige-birmas-vom-paoniengarten.de/	FALSCH	FALSCH		70	14307- Myanmar	google.de
13	31	14306- Birma	google.de	91	http://www.die-heilige-birma.de/die-heilige-birma.html	FALSCH	FALSCH		71	14307- Myanmar	google.de
14	32	14306- Birma	google.de	92	http://www.birma-von-oldima.de/	FALSCH	FALSCH		72	14307- Myanmar	google.de
15	33	14306- Birma	google.de	93	http://www.de-santa-nobleza.de/	FALSCH	FALSCH		73	14307- Myanmar	google.de
16	34	14306- Birma	google.de	94	http://www.birma-atbluesight.com/	FALSCH	FALSCH		74	14307- Myanmar	google.de
17	35	14306- Birma	google.de	95	http://www.wetter.com/birma-myanmar/ASMM.html	FALSCH	FALSCH		75	14307- Myanmar	google.de
18	36	14306- Birma	google.de	96	http://www.thueringer-allgemeine.de/web/zgt/politik/detail/-/sp	FALSCH	FALSCH		76	14307- Myanmar	google.de
19	37	14306- Birma	google.de	97	http://kleinanzeigen.ebay.de/anzeigen/s-katzen/berlin/birma/k0	FALSCH	FALSCH		77	14307- Myanmar	google.de

In einem neuen Datenblatt „Diagramme“ wurden dann je für Google und BING (diagramme 1 und 2) mit der Funktion „ZÄHLENWENN(Datenbereich, Kriterium)“ die Anzahlen von „WAHR“ und „FALSCH“ ermittelt und in absoluten Zahlen festgehalten. Schließlich wurden diese absoluten Zahlen durch die gesamte Trefferanzahl für die jeweilige Suchmaschine geteilt, um zu relativen Übereinstimmungsraten zu kommen. Diese Arbeitsschritte wurden einmal für alle Treffer von der jeweiligen Suchmaschine insgesamt und dann je für die TOP10 bzw. Treffer 91-100 durchgeführt. So ergaben sich beispielsweise folgende Übereinstimmungsraten (links Übereinstimmungsraten der automatischen Übereinstimmungen insgesamt und nach TOP10 bzw. 91-100, rechts manueller Abgleich der Domain entsprechend):

Übereinstimmung - Google			
	Ja	Nein	Gesamt
automatisch (URL) Gesamt	10	1130	1140
Gesamt	1%	99%	
	Ja Top 10	Nein Top 10	
1-10	10	560	570
in %	0,88%	49,12%	
	Ja 91-100	Nein 91-100	
91-100	-	570	570
in %	0,00%	50,00%	

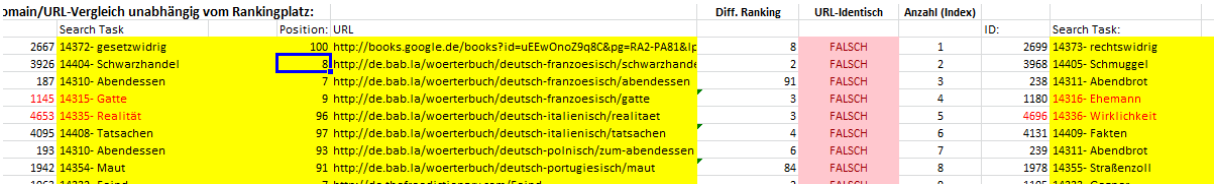
manuell (Domain) Gesamt	69	1071	1140
Gesamt	6%	94%	
	Ja Top 10	Nein Top 10	
1-10	36	534	570
in %	3,16%	46,84%	
	6,32%	93,68%	
	Ja 91-100	Nein 91-100	
91-100	33	537	570
in %	2,89%	47,11%	
	5,79%	94,21%	

Schritt 3: Domain- und URL-Vergleich unabhängig vom Ranking

Dieser vom Ranking unabhängige Vergleich wurde manuell durchgeführt. Dazu wurden die bisher ermittelten Daten aus Schritt 2 (Gesamte Treffer) kopiert und im entsprechenden Datenblatt der Suchmaschine weiter unten noch einmal eingefügt. Übereinstimmungen in der Domain wurden unabhängig vom Ranking für die entsprechend zusammengehörigen Synonyme gelb markiert, alle anderen aus der Liste gelöscht. So standen sich am Ende die zusammengehörigen Treffer unabhängig vom Ranking gegenüber.

Zusätzlich wurde die Übereinstimmung mit der bereits verwendeten Funktion „IDENTISCH()“ auf komplette URL-Übereinstimmungen überprüft, und entsprechende „WAHR“ und „FALSCH“-Werte erneut für die Ermittlung relativer Übereinstimmungsraten genutzt. Die Ranking-Differenz wurde per Excel-Funktion (Position 1 – Position2) berechnet. Diese Differenz wurde dann hinsichtlich ihres Mittelwerts und ihrer Standardabweichung überprüft,

für die gesamten Daten, sowie für die reinen Domain bzw. reinen URL-Übereinstimmungen. Hier ist ein Ausschnitt aus diesem Arbeitsschritt:



omain/URL-Vergleich unabhängig vom Rankingplatz:				Diff. Ranking	URL-Identisch	Anzahl (Index)		
	Search Task	Position:	URL				ID:	Search Task:
2667	14372- gesetzwidrig	100	http://books.google.de/books?id=uEEwOnoZ9q8C&pg=RA2-PA81&lp	8	FALSCH	1	2699	14373- rechtswidrig
3926	14404- Schwarzhandel	8	http://de.bab.la/woerterbuch/deutsch-franzoesisch/schwarzhande	2	FALSCH	2	3968	14405- Schmuggel
187	14310- Abendessen	7	http://de.bab.la/woerterbuch/deutsch-franzoesisch/abendessen	91	FALSCH	3	238	14311- Abendbrot
1145	14315- Gatte	9	http://de.bab.la/woerterbuch/deutsch-franzoesisch/gatte	3	FALSCH	4	1180	14316- Ehemann
4653	14335- Realität	96	http://de.bab.la/woerterbuch/deutsch-italienisch/realitaet	3	FALSCH	5	4696	14336- Wirklichkeit
4095	14408- Tatsachen	97	http://de.bab.la/woerterbuch/deutsch-italienisch/tatsachen	4	FALSCH	6	4131	14409- Fakten
193	14310- Abendessen	93	http://de.bab.la/woerterbuch/deutsch-polnisch/zum-abendessen	6	FALSCH	7	239	14311- Abendbrot
1942	14354- Maut	91	http://de.bab.la/woerterbuch/deutsch-portugiesisch/maut	84	FALSCH	8	1978	14355- Straßenzoll

Einige Domains tauchten auffällig häufig bei Google auf, wodurch sich hohe Unterschiede in der Übereinstimmung im Vergleich zu BING ergaben. Daher wurden diejenigen Domains aus der Berechnung entfernt (grau hinterlegt), die mehr als 10 Mal in der vom Ranking unabhängigen Auswertung auftauchten. Dadurch näherten sich die Werte für Google und BING an einander an.

Schritt 4: Erstellung der Diagramme

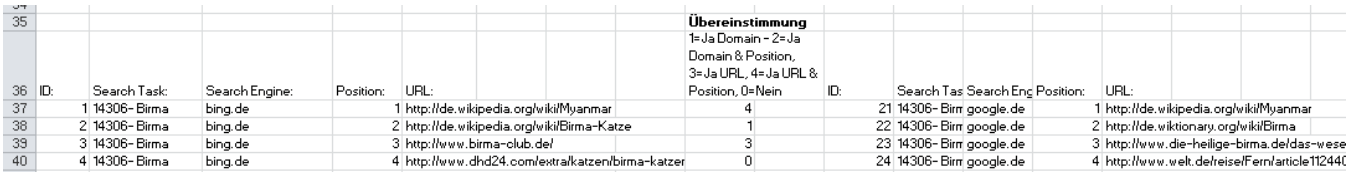
Für die jeweiligen Übereinstimmungsraten wurden entsprechende Tortendiagramme erstellt (Gesamte Treffer, TOP10 und Treffer 91-100). Zusätzlich wurden in einem weiteren Datenblatt „gesamt“ die verschiedenen Übereinstimmungskategorien bei Synonymie für Google und BING eingetragen, die sich durch die vom Ranking abhängigen und unabhängigen Vergleiche ergaben (URL+Position, Domain ohne Position etc.) und in einem Säulendiagramm veranschaulicht.

Schritt 5: Überprüfung der Datenbasis

Um die Ergebnisse hinsichtlich ihrer Signifikanz besser interpretieren zu können, wurde überprüft, inwieweit die Suchmaschinen unabhängig voneinander arbeiten. Dazu wurden im Datenblatt „Datenbasis“ die Treffer von Google und BING pro Suchanfrage gegenübergestellt und manuell verglichen. Folgendes Codesystem kam dabei zum Einsatz:

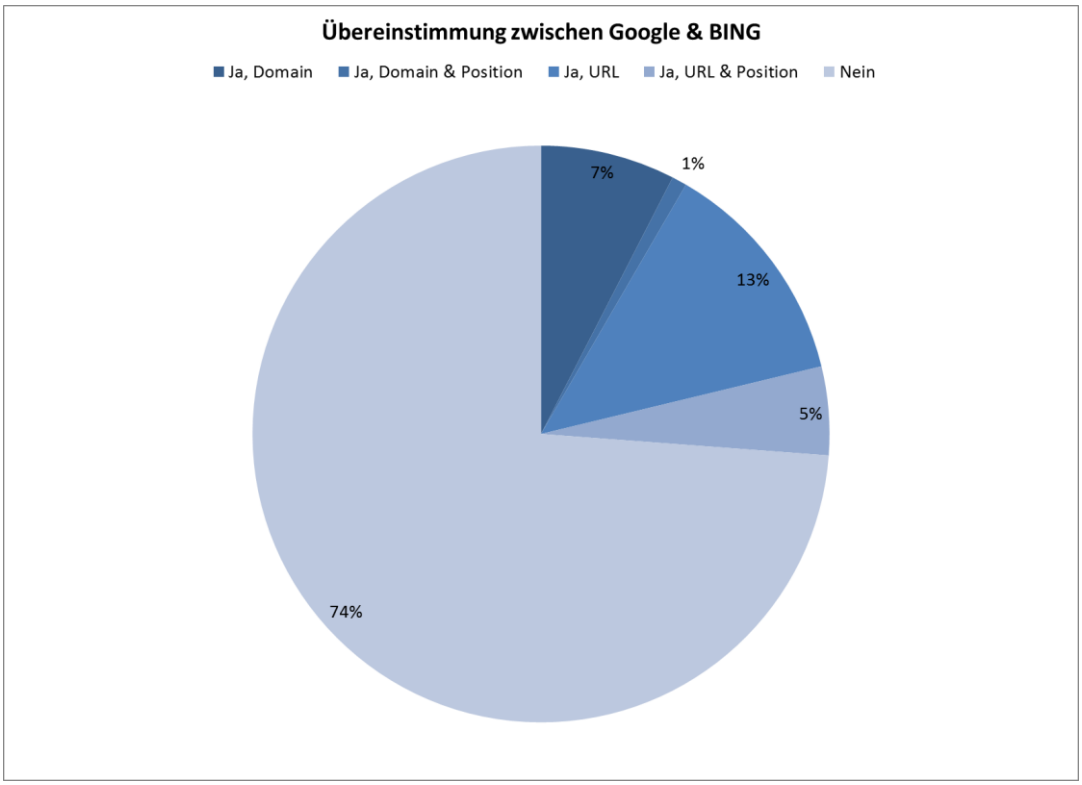
- 1 = Ja Domain
- 2 = Ja Domain & Position
- 3 = Ja URL
- 4 = Ja URL & Position
- 0 = Nein

Folgender Ausschnitt aus der Auswertungsdatei zeigt die Gegenüberstellung der Trefferdaten von BING (links) und Google (rechts), sowie die Vergabe der Übereinstimmungscodes in der Mitte:



35						Übereinstimmung						
36	ID:	Search Task:	Search Engine:	Position:	URL:	1=Ja Domain - 2=Ja Domain & Position, 3=Ja URL, 4=Ja URL & Position, 0=Nein	ID:	Search Tas	Search Eng	Position:	URL:	
37	1	14306- Birma	bing.de	1	http://de.wikipedia.org/wiki/Myanmar	4	21	14306- Birm	google.de	1	http://de.wikipedia.org/wiki/Myanmar	
38	2	14306- Birma	bing.de	2	http://de.wikipedia.org/wiki/Birma-Katze	1	22	14306- Birm	google.de	2	http://de.wiktionary.org/wiki/Birma	
39	3	14306- Birma	bing.de	3	http://www.birma-club.de/	3	23	14306- Birm	google.de	3	http://www.die-heilige-birma.de/das-wese	
40	4	14306- Birma	bing.de	4	http://www.dhd24.com/extra/katzen/birma-katzer	0	24	14306- Birm	google.de	4	http://www.welt.de/reise/Fern/article112440	

Mithilfe der Funktion „ZÄHLENWENN“ konnten somit hinterher alle Übereinstimmungen ausgewertet und durch die Teilung mit der Gesamttrefferanzahl zu relativen Häufigkeiten verarbeitet werden. Dadurch ergab sich folgendes Diagramm:



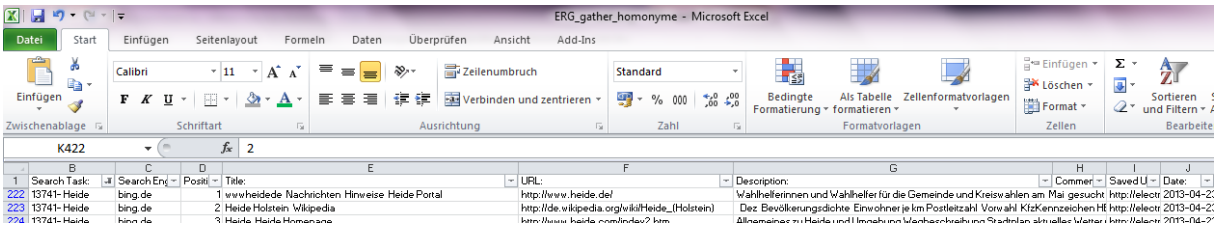
Zur Überprüfung einer Vermutung wurden auszugsweise 10 zufällig gewählte Synonyme hinsichtlich der Übereinstimmung zwischen „Synynom 1“ von Google und „Synonym 2“ von BING auf dieselbe Weise codiert und ausgewertet. Diese Raten zeigen eine rein inhaltliche Übereinstimmung, unabhängig eventueller spezifischer Algorithmen der Suchmaschinen.

Verfahrensbeschreibung zur Auswertung im Bereich „Homographie“

Die Auswertung der „Homographie“ unterscheidet sich grundlegend von der in den Bereichen „Synonymie“ und „Morphologische Varianten“, weil es hier nicht um einfache Übereinstimmungsraten ging, sondern um Bedeutungsanteile in den organischen Trefferlisten und eine Korrelation eben dieser mit den Klickdaten der deutschen Suchmaschine T-Online.

Schritt 1: Datenkontrolle

Die vom RAT gelieferten Rohdaten wurden zunächst auf Fehler (fehlende Trefferdaten) hin untersucht. Hier ein Auszug der gelieferten Rohdaten:

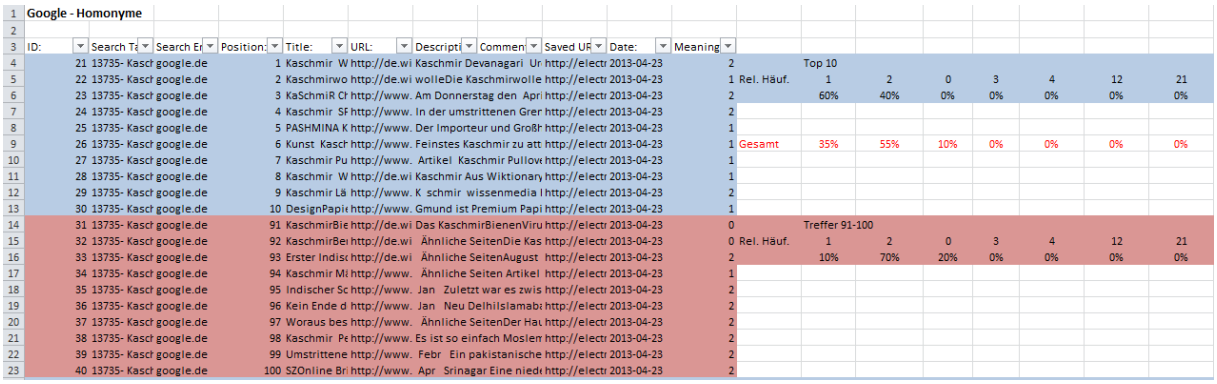


Die markierten Suchanfragen, die fehlerhafte Daten enthielten, wurden für eine zweite Datenabfrage verwendet, und die fehlerhaften Daten ggf. durch die nun korrekt erfassten ersetzt (grün markiert). Die danach noch verbliebenen Fehler wurden grau markiert und endgültig aus der Auswertung entfernt.

Schritt 2: Bedeutungszuordnung

Die kontrollierte Rohdatenliste wurde durchgesehen und mittels eines Codesystems wurden die verschiedenen Bedeutungen manuell den Treffern zugeordnet. Die Entscheidung, welches die erste Hauptbedeutung (Code: 1) und welches die zweite Hauptbedeutung (Code: 2) war, fiel anhand der Trefferliste. Die jeweils am häufigsten vorkommende Bedeutung bei Google und BING bekam den Code 1, die zweithäufigste den Code 2. Weitere Codes wurden vergeben (z.B. „3“ für Eigennamen, „0“ für unpassend/nicht zuzuordnen etc.). Danach wurde die gesamte Liste nach Google und BING aufgeteilt und in gesonderten Datenblättern gespeichert, um die relativen Häufigkeiten der jeweiligen Bedeutungen für jede Suchmaschine zu ermitteln. Letzteres geschah wieder mithilfe der Funktion „ZÄHLENWENN()“, die bestimmte Codes zählte. Diese absoluten Zahlen wurden durch eine entsprechende Division der zugehörigen Trefferzahl für den Suchbegriff (also 20 Treffer) geteilt und somit ergaben sich relative Häufigkeiten. Diese Häufigkeiten wurden für die gesamten 20 Treffer, sowie für die TOP10 und Treffer 91-100 ermittelt.

Die einzelne Übersicht pro Suchanfrage sah folgendermaßen aus, hier anhand des Beispiels „Kaschmir“ (TOP10: blau hinterlegt, Treffer 91-100: rot hinterlegt, rechts relative Häufigkeiten gemäß Codesystem, rote Schrift = relative Häufigkeiten für die gesamten 20 Treffer):



Google - Homonyme

ID:	Search Te	Search En	Position:	Title:	URL:	Descripti	Commen	Saved UR	Date:	Meaning								
21	13735-	Kasch	google.de	1	Kaschmir W	http://de.wi	Kaschmir	Devanagari	Ur	http://electr	2013-04-23	2		Top 10						
22	13735-	Kasch	google.de	2	Kaschmirwo	http://de.wi	wolleDie	Kaschmirwolle		http://electr	2013-04-23	1	Rel. Häuf.	1	2	0	3	4	12	21
23	13735-	Kasch	google.de	3	KaSchmiR Cl	http://www.	Am Donnerstag	den	Apri	http://electr	2013-04-23	2		60%	40%	0%	0%	0%	0%	0%
24	13735-	Kasch	google.de	4	Kaschmir SF	http://www.	In der umstrittenen	Grer		http://electr	2013-04-23	2								
25	13735-	Kasch	google.de	5	PASHMINA K	http://www.	Der Importeur und	Großh		http://electr	2013-04-23	1								
26	13735-	Kasch	google.de	6	Kunst Kasch	http://www.	Feinstes Kaschmir zu att			http://electr	2013-04-23	1	Gesamt	35%	55%	10%	0%	0%	0%	0%
27	13735-	Kasch	google.de	7	Kaschmir Pu	http://www.	Artikel Kaschmir Pullov			http://electr	2013-04-23	1								
28	13735-	Kasch	google.de	8	Kaschmir W	http://de.wi	Kaschmir Aus Wiktionary			http://electr	2013-04-23	1								
29	13735-	Kasch	google.de	9	Kaschmir Lä	http://www.	K schmir wissenmedia l			http://electr	2013-04-23	2								
30	13735-	Kasch	google.de	10	DesignPapi	http://www.	Gmund ist Premium Papi			http://electr	2013-04-23	1								
31	13735-	Kasch	google.de	91	KaschmirBie	http://de.wi	Das KaschmirBienenViru			http://electr	2013-04-23	0		Treffer 91-100						
32	13735-	Kasch	google.de	92	KaschmirBei	http://de.wi	Ähnliche SeitenDie Kas			http://electr	2013-04-23	0	Rel. Häuf.	1	2	0	3	4	12	21
33	13735-	Kasch	google.de	93	Erster Indis	http://de.wi	Ähnliche SeitenAugust			http://electr	2013-04-23	2		10%	70%	20%	0%	0%	0%	0%
34	13735-	Kasch	google.de	94	Kaschmir Mi	http://www.	Ähnliche Seiten Artikel			http://electr	2013-04-23	1								
35	13735-	Kasch	google.de	95	Indischer Sc	http://www.	Jan Zuletzt war es zwis			http://electr	2013-04-23	2								
36	13735-	Kasch	google.de	96	Kein Ende d	http://www.	Jan Neu Delhilslamabi			http://electr	2013-04-23	2								
37	13735-	Kasch	google.de	97	Woraus bes	http://www.	Ähnliche SeitenDer Hau			http://electr	2013-04-23	2								
38	13735-	Kasch	google.de	98	Kaschmir Pe	http://www.	Es ist so einfach Moslem			http://electr	2013-04-23	2								
39	13735-	Kasch	google.de	99	Umstrittene	http://www.	Febr Ein pakistanische			http://electr	2013-04-23	2								
40	13735-	Kasch	google.de	100	SZOnline Bri	http://www.	Apr Srinagar Eine niedr			http://electr	2013-04-23	2								

Schritt 3: Übersicht Bedeutungsverteilung

Die Verteilung der Häufigkeiten wurde in zwei neuen Datenblättern (diagramme 1 und 2) für jede Suchmaschine übersichtlich in eine Tabelle kopiert. Auf deren Basis erfolgten verschiedene Berechnungen und Darstellungen. Etwa Mittelwerte und Standardabweichungen, sowie eine Übersicht über die Häufigkeiten in Klassen, durch die z.B. ersichtlich war, in welchem prozentualen Grenzbereich die Hauptbedeutung am häufigsten vorkam. Durch diese Übersicht nach Klassen (wie häufig kommt Code 1 mit 20-40% in der Trefferliste vor?), konnte auch die Häufigkeit der vorkommenden Eigennamen betrachtet werden.

Schritt 4: Bedeutungsverteilung der T-Online-Daten

Über einen entsprechenden von der HAW bereitgestellten Zugang wurden die organischen T-Online-Klickdaten für die Homographe dieser Studie abgefragt. Mit einer Suche nach beispielsweise „Jaguar“ wurden alle Suchanfragen mit ihren organischen Klickdaten ausgegeben, die diese Buchstabenfolge aus dem Jahr 2012 enthielten. Nach dem Export ergab sich somit für jede der 70 angesetzten Suchanfragen eine Excel-Datei u.a. mit den organischen Klickdaten, maximal wurden 2000 zugehörige Anfragen ausgegeben. Jede der 70 Excel-Listen wurde daraufhin entsprechend dem Codesystem (Hauptbedeutung=1, Unpassend=0 etc.) genau wie die organischen Treffer codiert. Schematisch sah dies folgendermaßen aus (hier fiktive Anfragen und Daten):

Suchanfrage	Organ. Klickrate	Bedeutungszuordnung
Jaguar Raubkatze	1256	2
Jaguar Cabrio	5468	1
Jaguar kaufen	8892	1

Für die Berechnung der relativen Häufigkeiten (Anteile Klicks Bedeutung 1 / Anteile Klicks insgesamt) wurden nur die Klickraten der beiden Hauptbedeutungen verwendet und andere Codes vernachlässigt, da die weiteren Kategorien die Bedeutungsverteilung in unterschiedlichen Größenordnungen beeinflussen und dieser Einfluss für die Korrelation herausgenommen werden sollte (Normierung).

Die ermittelten und normierten relativen Häufigkeiten der organischen Klickdaten wurden in einem neuen Datenblatt „Korrelation“ für alle abgefragten Homographe eingetragen. Einige Homographe erwiesen sich nach diesem Arbeitsschritt als unpassend, da zu wenige Klickdaten verfügbar waren (unter 10 Anfragen, keine Anfragen) oder da das Homograph zu oft Bestandteil unpassender Suchanfragen war (z.B. die Buchstabenfolge „arm“).

Schritt 5: Korrelation

Vor der eigentlichen grafischen Korrelation und der Berechnung des Korrelationskoeffizienten (mittels der Excel-Funktion „KORREL()“), wurden auch die Daten der Häufigkeitsverteilung der organischen Treffer normiert, und somit der numerische Einfluss der weiteren Kategorien aus den Daten entfernt. Dies erfolgte indem die „Häufigkeit Code 1“ durch die „Häufigkeit von Code 1+2“ geteilt wurde.

Die Daten wurden für die gesamten Treffer, und die TOP10 und Treffer 91-100 normiert und in einer Tabelle eingetragen, wie in allen Auswertungen je für Google und BING. Daraus ergab sich folgende Übersicht (hier beispielsweise nur BING), wobei rot markiert diejenigen Begriffe sind, bei denen die Bedeutung überwiegt, die laut der T-Online-Klickdaten deutlich seltener erwünscht ist:

		Bedeutungen		BING (gesamt) - normiert		BING TOP 10 - normiert		BING Treffer 91-100 - normiert	
Index	Begriff	1	2	Bedeutung 1	Bedeutung 2	Bedeutung 1	Bedeutung 2	Bedeutung 1	Bedeutung 2
1	Kaschmir	Wolle	Region	0,80	0,20	0,70	0,30	0,90	0,10
2	Schloss	Gebäude	Türschloss, Zündschlos	0,95	0,05	0,90	0,10	1,00	0,00
3	Fächer	Luftfächer	Schulfächer	0,68	0,32	1,00	0,00	0,40	0,60
4	Herde	Tiergruppe	Elektrogerät	1,00	0,00	1,00	0,00	1,00	0,00

Auf Basis dieser Tabelle wurden die Korrelationsgrafiken erstellt, sowie die Korrelationskoeffizienten berechnet. Für die gesamten Treffer, die TOP10, die Treffer 91-100 etc.

Zudem wurde für die rot markierten Ausnahmefälle die Datenbasis kontrolliert, ob sie etwa auf den erhöhten Einfluss durch „Eigennamen“ oder eine geringe Klickdaten-Basis zurückzuführen sind.

Schritt 4: Überprüfung auf Disambiguierung und Bedeutungsverhältnisse der Bilder

Jedes in die Auswertung einfließende Homograph wurde einmal jeweils in BING und Google manuell eingegeben. Dann wurde die erste Trefferseite – und nur die erste Trefferseite – angeschaut, ob es direkte Hinweise auf Disambiguierung (Spalte H) gibt und welcher Art diese ggf. sind (Spalte I). Das Vorkommen von Bildern (Spalte E, 1 oder 0) und die Anzahl der Bilder pro Bedeutung (Spalte F und G) wurden zudem festgehalten, hier ein Beispiel:

	A	B	C	D	E	F	G	H	I
1	Index	Homonym	Bedeutung 1	Bedeutung 2	Bilder	Bilder für Bedeutung	Bilder für Bedeutung	Hinweis auf Disambiguierung	Art des Hinweises
2					*Ja (1), Nein (0)*	*X/Y*	*X/Y*	Ja(1), Nein (0)	
3	1	Kaschmir	Region	Wollgewebe	1	100%	0%	0	*Wikipedia-Def. Zu Kaschmir Region*
4	2	Schloss	Türschloss	Gebäude	1	0%	100%	0	*0*
5	3	Fächer	Luftzufächerer	Schulfächer	1	100%	0%	0	*0*
6	4	Herde	Geräte zum Kochen	Tiergruppen	1	0%	100%	0	*0*
7	5	Heide	Stadt	Landschaft	1	100%	0%	1	*Wikipedia-Hinweis zur Stadt Heide, Hinweis auf Ergebnisse für Heide Landschaft*

Eine entsprechende Auswertung ergab dann die relativen Häufigkeiten für das Vorkommen von Bildern gemäß der Hauptbedeutungen 1 und 2 bzw. des Vorkommen von Hinweisen auf Disambiguierung (wobei hier nur Hinweise auf weitere Ergebnisse gezählt wurden, nicht alle

festgehaltenen Hinweise wie etwa reine Wikipedia-Definitionen). Beispielsweise betragen die Werte für Google:

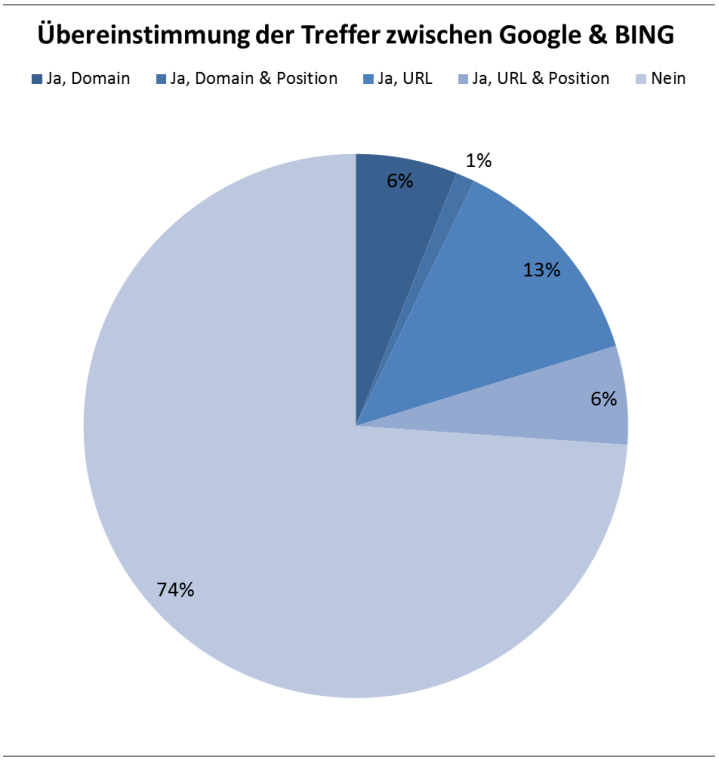
Bilder vorhanden	84%
Bild für Bedeutung 1:	35%
Bild für Bedeutung 2:	34%
Hinweis auf Bedeutungen:	25%

Schritt 5: Überprüfung der Datenbasis

Um die Ergebnisse besser hinsichtlich ihrer Signifikanz interpretieren zu können, wurde überprüft, inwieweit die Suchmaschinen unabhängig voneinander arbeiten. Dazu wurden im Datenblatt „Datenbasis“ die Treffer von Google und BING pro Suchanfrage gegenübergestellt und manuell verglichen, wie auch bei „Synonymie“ und „Morphologie“. Die Übersicht ist entsprechend (BING links, Google rechts, Codevergabe in der Mitte):

31														
32							Übereinstimmung							
33	ID:	Search Tas	Search Eng	Position:	Title:	URL:	1=Ja Domain - 2=Ja Domain & Position, 3=Ja URL, 4=Ja URL & Position, 0=Nein	ID:	Search Tas	Search Eng	Position:	Title:	URL:	
34	1	13735- Kas	bing.de	1	Kaschmir \	http://de.wikipedia.org/wiki/Kaschmir	4	21	13735- Kas	google.de	1	Kaschmir \	http://de.wikipedia.org/wiki/Kaschmir	
35	2	13735- Kas	bing.de	2	PASHMINA	http://kaschmir.net/	3	22	13735- Kas	google.de	2	Kaschmirw	http://de.wikipedia.org/wiki/Kaschmir	
36	3	13735- Kas	bing.de	3	Kaschmirw	http://de.wikipedia.org/wiki/Kaschmirwolle	3	23	13735- Kas	google.de	3	KaSchmiRI	http://www.ksrgym.de/	
37	4	13735- Kas	bing.de	4	Kaschmir b	http://www.bpb.de/themen/LKG7XK,0,0,Kaschr	0	24	13735- Kas	google.de	4	Kaschmir S	http://www.spiegel.de/thema/kaschmir	
38	5	13735- Kas	bing.de	5	Kaschmir S	http://www.kaschmir.net/Kaschmir-Schal-Herrer	0	25	13735- Kas	google.de	5	PASHMINA	http://www.kaschmir.net/	

Mithilfe der Funktion „ZÄHLENWENN“ konnten somit hinterher alle Übereinstimmungen ausgewertet und entsprechend durch die Teilung mit der Gesamttrefferanzahl zu relativen Häufigkeiten verarbeitet werden. Dabei kam ein Ergebnis heraus, dass dem im Bereich „Synonymie“ und „Morphologie“ sehr ähnlich ist:



Eidesstattliche Versicherung
„Ich versichere, die vorliegende Arbeit selbstständig ohne fremde Hilfe verfasst und keine anderen Quellen und Hilfsmittel als die angegebenen benutzt zu haben. Die aus anderen Werken wörtlich entnommenen Stellen oder dem Sinn nach entlehnten Passagen sind durch Quellenangabe kenntlich gemacht.“

_____	_____
Ort, Datum	Unterschrift