Trefwoord, tijdschrift voor lexicografie. Jaargang 2012. http://www.fryske-akademy.nl/trefwoord

Google Bücher aus dem Blickwinkel des Lexikographen

Dr. Dominik Brückner

Google Inc. und Google Bücher

1998 gründeten Sergey Brin und Larry Page die Firma Google Inc. Ihr Name ist abgeleitet von der Zahl 10^{100}, die der US-amerikanische Mathematiker Edward Kasner 1938 "Googol" genannt hatte[1]. Und dieser Name ist Programm: Mit ihren Produkten wie der Internetsuchmaschine "Google", "Google Maps", "Google Earth" oder "Street View" versucht die Firma, riesige Mengen an Daten zu sammeln, zu verwalten und zugänglich zu machen.

Seit 2005 wird mit Boogle Bücher (Google Books[2]) ein weiteres Online-Tool angeboten, dessen Zielvorgabe nichts weniger ist, als durch Digitalisierung das in Büchern gespeicherte Wissen der Welt für die Volltextsuche verfügbar zu machen. Dieses Tool findet in den letzten Jahren nicht zuletzt in den Textwissenschaften zunehmend Anwendung.

Solchen Zielvorgaben entsprechend ermöglicht es Boogle Bücher, in den Volltexten weltweit digitalisierter Bücher, Zeitschriften, Zeitungen, Comics etc. zu recherchieren. Um wie viele Einheiten es sich dabei handelt, gibt Google nicht bekannt und da es definitionsabhängig ist, was als eine Einheit gilt, dürfte es auch schwierig sein, diesbezüglich belastbare Zahlen vorzulegen. In Bezug auf die Gesamtzahl der zu digitalisierenden Texte sind Googles Ziele jedoch ebenso klar wie ehrgeizig: Geplant ist die vollständige Digitalisierung sämtlicher Bücher dieser Welt, in allen Sprachen und aus allen Zeiten.

Um dieses Zeil zu erreichen, arbeitet Google mit einer zunehmenden Zahl von Bibliotheken zusammen. Dazu gehören derzeit die Universitätsbibliotheken der University of Michigan, der Harvard University, der Stanford University, der University of Virginia, der University of Wisconsin-Madison, der Princeton University, der University of California und der University of Texas at Austin, dazu die New York Public Library, und in Europa die Bodleian Library der Oxford University, die Bibliothek der Universidad Complutense Madrid sowie die Biblioteca de Catalunya in Barcelona, die Bayerische Staatsbibliothek in München, die Bibliothèque Municipale de Lyon, die Österreichische Nationalbibliothek und seit 2007 auch die Universiteitsbibliotheek Gent.

Gleichzeitig lief ein Partnerprogramm mit Verlagen an. Die Verlage überlassen Google PDF-Dateien oder stellen dem Unternehmen, ähnlich wie die Bibliotheken, Bücher zur Verfügung. Die Bücher werden von

[1] Kasner, Edward/Newman, James: Mathematics and the Imagination. New York 1940.
[2] http://books.google.nl/

Google zunächst gescannt und dann mittels OCR (Optical Character Recognition, Optische Zeichenerkennung) bzw. ICR (Intelligent Character Recognition) als maschinenlesbare Texte in den Index aufgenommen. Die Details dieser Verfahren sind weitgehend geheim.

Ausrichtung

Die Unternehmung Google Bücher ist in den letzten Jahren aus den verschiedensten Gründen öffentlich heftig kritisiert worden. Man wirft dem Konzern Copyrightverletzungen, "Knebelverträge" für Partnerbibliotheken, Verlage und Autoren vor[3], zudem eine kulturelle "Hegemonie des Englischen" (J.-N. Jeanneney[4]) sowie eine Dominanz der Wirtschaft in einem Bereich, der von wirtschaftlichen Interessen ausgenommen sein sollte – was sich vor allem in der Konkurrenz Googles zu öffentlich finanzierten Digitalisierungsprojekten konkretisiert.

Diese Probleme und die darum ausgetragenen Diskussion sind weithin bekannt, daher richtet dieser Beitrag sein Augenmerk auf andere Aspekte des Tools. Die internen, technischen Probleme von Google Bücher, mit denen es der Nutzer tagtäglich zu tun bekommt, stehen jenen juristischen und kulturpolitischen nämlich in nichts nach, befinden sich aber in weit geringerem Ausmaß im Blickfeld der öffentlichen Kritik. Im Folgenden wird es deshalb darum gehen, ausgehend von der alltäglichen Praxis im Umgang mit Google Bücher die Inhalte und Suchmöglichkeiten des Online-Tools – einschließlich ihrer Grenzen und Probleme – kritisch zu beleuchten. Dabei orientiert sich die Diskussion an den Bedürfnissen der (germanistischen) Lexikologie und Lexikographie.[5] Es wird im Folgenden im Sinne einer Recherche mit Hilfe von Google Bücher vorgegangen: Zunächst werden die wichtigsten Suchoptionen der "erweiterten Buchsuche"[6] vorgestellt, danach werden die Ergebnislisten ins Blickfeld gerückt, vor allem im Hinblick auf ihre Zusammensetzung, die Anzahl der Ergebnisse und den Zugriff darauf. In der Folge werden ausgewählte Probleme diskutiert, die im Zusammenhang mit in "vollständiger Ansicht", als

[3] Zu einigen juristischen Aspekten aus deutscher Sicht s. z. B. Lewandowski, Dirk: Google Buchsuche. Bücher kostenlos zum Download. In: Password. 10/2006, S. 36, ders.: Wie verändert die Einigung mit Verlegern und Autoren die Buchwelt?. In: Password 12/2008, S. 13 sowie Weber, Klaus: Drei Jahre Freiheitsstrafe für alle Google-Mitarbeiter? Ein Beitrag zur Praxis des Urheberstrafrechts. In: Zeitschrift für Internationale Strafrechtsdogmatik 2010, S. 220 – 226.

[4] Jeanneney, Jean-Noël: Googles Herausforderung. Für eine europäische Bibliothek. Mit einem neuen Vorwort des Autors zur dt. Ausg. Nachwort Klaus-Dieter Lehmann. Übers. Sonja Finck, Nathalie Mälzer-Semlinger. Stiftung Preußischer Kulturbesitz Berlin. Wagenbach-Verlag, Berlin-Hamburg 2006.

[5] Die folgenden Beobachtungen ergaben sich aus der täglichen Arbeit am Deutschen Fremdwörterbuch (DFWB), einem lexikographischen Projekt, das den Kernbereich der geläufigen, in die deutsche Standardsprache der Gegenwart fest integrierten Fremdwörter und Fremdwortfamilien in ihrer historischen Entwicklung beschreibt und dokumentiert: Deutsches Fremdwörterbuch. Begonnen von Hans Schulz, fortgeführt von Otto Basler. 2. Auflage, völlig neu erarbeitet im Institut für Deutsche Sprache. Berlin/New York 1995ff.
Vgl. zum Folgenden auch Brückner, Dominik: Die Google-Buchsuche als Hilfsmittel für die Lexikographie. In: Sprachreport 3/2009. S. 26-31. Mannheim: Institut für Deutsche Sprache, 2009.

[6] http://books.google.nl/advanced_book_search (August 2012).

"eingeschränkte Leseprobe" und in "Auszügen" vorhandenen Büchern[7] auftreten können. Zum Schluss wird versucht werden, den Nutzen von Google Bücher für Lexikographie und Lexikologie im Lichte dieser Diskussion zusammenfassend auf den Punkt zu bringen.

Suchoptionen

Google Bücher bietet dem Nutzer unter der Rubrik "erweiterte Buchsuche" eine ganze Reihe von frei wählbaren Suchoptionen. Die für den Lexikologen/Lexikographen interessantesten darunter sind:
"Mit allen Wörtern" (dabei findet Google diejenigen Texte, in denen sämtliche Suchausdrücke vorkommen, unabhängig von der Reihenfolge ihrer Eingabe), "mit der genauen Wortgruppe" (dabei findet Google diejenigen Texte, in denen alle Suchausdrücke in der Reihenfolge ihrer Eingabe vorkommen), "mit irgendeinem der Wörter" (dabei findet Google verschiedene Suchausdrücke) und "ohne die Wörter" (dabei werden alle diejenigen Texte vollständig aus dem Suchergebnis ausgeschlossen, in denen der eingetragene Ausdruck mindestens einmal vorkommt).
Hinzu treten eine Reihe optionaler Sucheinschränkungen. Google bietet die Möglichkeit an, nach Sprachen zu filtern, sowie Titel, Autornamen, Verlag sowie ISBN/ISSN in die Suchmaske einzugeben. Zudem kann die Suche mithilfe der Eingrenzung des Veröffentlichungszeitraums eingeschränkt werden.
Diese Suchoptionen sind aus verschiedenen Gründen problematisch, was insbesondere dann ärgerlich ist, wenn diese Probleme keine nachvollziehbaren technischen Hintergründe haben. So ist etwa von Google keine Trunkierung vorgesehen, wie sie doch bei vergleichbaren Programmen heutzutage ansonsten weithin Standard ist. Daher muss jede einzelne Flexionsform oder Schreibvariante im Feld "mit irgendeinem der Wörter" per Hand eingegeben werden. Dies kann sich bekanntlich sehr schnell zu einer erheblichen Aufgabe auswachsen, insbesondere dann, wenn man es mit historischen Schreibvarianten zu tun bekommt. Man denke etwa an die <k>/<c>-Varianz, an die <I>/<J>-Varianz oder an die <u>/<v>-Varianz, von denen mehrere zusammen auftreten können, noch dazu in verschiedenen Flexionsformen. Dabei ist zu beachten, dass Google Bücher die Eingabe ohne Not auf 32 Formen beschränkt hat – was zunächst zwar durchaus komfortabel erscheinen mag, sich aber sehr schnell als zu limitiert erweisen kann. Problematischer ist es jedoch, dass Google dem Nutzer dies nicht mitteilt, sondern einfach die überzähligen Formen stillschweigend unberücksichtigt lässt. Dazu kommt, dass Google, ohne dass klar

[7] Im folgenden wird versucht, den Ausdruck "Buch" soweit möglich in derselben Weise zu verwenden, wie Google das tut, um den Bezug zu dem, was Google ein "Buch" nennt, zu wahren. Damit sei aber keinesfalls zu Ausdruck gebracht, dass Google eine eindeutige, brauchbare oder auch nur auffindbare Definition von "Buch" anbietet. Entsprechendes gilt im gleichen Fall für andere Bezeichnungen und Formulierungen, die Google Bücher verwendet, etwa "Wort", "Wortgruppe", "Ergebnis" oder "Sortierung".

würde, auf welche Weise und warum dies geschieht, für jede Eingabe neue, und zwar verschiedene Ergebnislisten zusammenstellt. Dadurch kann es geschehen, dass man bei einer Eingabe von nur wenigen Wortformen Textstellen findet, die bei der Eingabe von mehr Formen nicht mehr gefunden werden, auch wenn die ursprünglichen Formen in diesen enthalten sind:

> Eine Suche nach *indizieren indiziren indiciren indicieren* im Zeitraum zwischen 1750 und 1760 ergab am 15. Juli 2012 eine Liste von 10 Ergebnissen, eine Suche nach *indizieren indiziren indiciren indicieren indiziere indizire indicire indiciere* im gleichen Zeitraum ergab 12 Ergebnisse. Unten diesen fehlten drei der Ergebnisse, die die erste Suche zutage gefördert hatte.
> Als die Suche nach *indizieren indiziren indiciren indicieren* eine Stunde später wiederholt wurde, fanden sich nur noch 6 Ergebnisse auf der Liste. Google Bücher gibt also für zwei identische Suchanfragen nicht auch zwei identische Suchergebnisse aus, vielmehr verändern sich die Ergebnislisten – nach nicht nachvollziehbaren Kriterien. Dabei wäre eine Zunahme der Ergebnisanzahl durch die stetige Zunahme der Digitalisate zu erklären, eine Abnahme muss allerdings rätselhaft bleiben. Möchte man einen verschwundenen Beleg wiederfinden, so besteht die Möglichkeit, den Rechner zu wechseln, auf anderen Rechnern werden vermisste Ergebnisse oft wiedergefunden. Auch eine Suche einige Tage später fördert Verschwundenes oftmals wieder zutage. Eine Garantie dafür gibt es allerdings nicht.

Andererseits gibt Google Bücher auch Ergebnisse aus, die überhaupt nicht gesucht werden sollten:

> Eine Suche nach *Indizien* fördert seit einiger Zeit auch zusätzlich Ergebnisse für *indizieren* und *Indizierung* zutage und eine Suche nach *Imperatorin* erbrachte Ergebnisse für *Imperator, in* – ohne dass der Benutzer die Möglichkeit hatte, diese Funktion abzustellen.

Auch die anderen Felder funktionieren nicht befriedigend. Das Feld "Sprache Antwortseiten, geschrieben auf" mag als Sprachfilter angelegt worden sein – er funktioniert jedoch nicht. Diese Feststellung mag reichlich ungenau klingen, exaktere Schlüsse lassen die Resultate, die mithilfe dieser Option erzielt werden können, aber nicht zu: Ergebnisse aus anderen Sprachen werden nämlich weiterhin mit ausgegeben, teilweise sogar dieselben wie ohne Nutzung dieser Option. Rückschlüsse auf die Funktionsweise des Sprachfilters, die man für einen händischen Workaround gut gebrauchen könnte, lassen sich daraus nicht ziehen.
Das Feld "Autor Bücher von folgendem Autor" ist unbrauchbar, solange Google nicht zwischen einem Autor und einem Herausgeber unterscheiden kann. Und auch Suchen nach Titel, Verlag und ISBN sind

unsicher, in allererster Linie weil diese Angaben in Folge ihrer Kürze in besonderem Maße von der Qualität der OCR/ICR abhängig sind.

Einige dieser Suchoptionen mögen für die Arbeit des Lexikologen/Lexikographen nicht von großer Bedeutung sein, eine ist es aber ganz sicher: Die Suchoption "Veröffentlichungsdatum Büchern [sic] mit Veröffentlichungsdatum zwischen", mittels derer (durch die freie Eingabe zweier Jahreszahlen) eine zeitliche Einschränkung der Suche möglich ist. Aus lexikographischer Sicht dürfte diese Option in erster Linie genutzt werden, um:

- so genannte (Früh- und) Erstbelege zu suchen
- gezielt Beleglücken zu füllen
- Wörterbuch-Buchungen durch "echte" Belege zu ersetzen
- zu eng gefasste Belegschnitte zu erweitern

Ergebnisliste

Eine solche Suche führt zur Ausgabe einer Liste von Ergebnissen, die sich (bedingt durch copyrightbedingte Einschränkungen) aus heterogenem Material zusammensetzt:

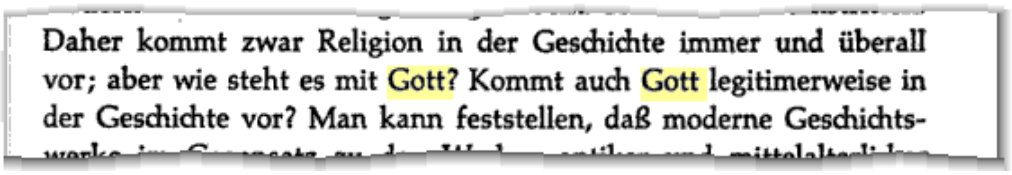
"vollständige Ansicht"
Texte, die von Google so gekennzeichnet wurden, sind für sämtliche Nutzer vollständig einsehbar. Diese Texte können zudem in Form von pdf-Dateien heruntergeladen, gespeichert und ausgedruckt werden.

"Eingeschränkte Leseprobe"
Aus diesen Texten werden nur ausgewählte Seiten angezeigt. Dies kann zum einen dadurch bedingt sein, dass der herausgebende Verlag Google bestimmte Nutzungsbeschränkungen auferlegt, es kann aber auch mit Googles Umsetzung des Urheberrechts zusammenhängen: Im Rahmen des Projekts werden zwar auch aktuellere Publikationen digitalisiert, Bücher, die nach 1864 erschienen sind, sind aber (von einigen wenigen Ausnahmen abgesehen) für Nutzer außerhalb der USA nicht in vollständiger Ansicht verfügbar.[8]

"Snippet-Ansicht"

[8] Die Grenze, die das Jahr 1864 darstellt, wird allerdings nicht immer genau eingehalten. Erfahrungsgemäß sind auch viele Texte aus der zweiten Hälfte der 1860er Jahre noch in vollständiger Ansicht verfügbar.

Von solchen Ergebnissen gibt Google Bücher gibt nur Minimalkontexte aus, die in den meisten Fällen nicht einmal ganze Sätze beinhalten:


Daher kommt zwar Religion in der Geschichte immer und überall vor; aber wie steht es mit Gott? Kommt auch Gott legitimerweise in der Geschichte vor? Man kann feststellen, daß moderne Geschichts-

"Keine Vorschau verfügbar"

Von derartig gekennzeichneten Texten sind nicht einmal Minimalkontexte verfügbar, Google hält allerdings bibliographische Daten vor. Leider sind diese Daten OCR-/ICR-abhängig und damit alles andere als verlässlich (mehr dazu unten).

Die Suche kann in verschiedener Weise angepasst werden, so dass eine Einschränkung der Ergebnisliste möglich ist: Der Nutzer kann somit alle Bücher durchsuchen, oder sich auf eingeschränkte Leseproben und in vollständiger Ansicht vorhandene Bücher beschränken, er kann zudem ausschließlich in vollständiger Ansicht vorhandene Bücher durchsuchen und sich seit einiger Zeit auch auf Google eBooks beschränken.

Sortierung der Ergebnisliste

Google Bücher gibt die Ergebnisliste per Default "nach Relevanz sortiert" aus. Wieder bleibt unklar, was Google damit meint, nach welchen Kriterien das System also Relevanz bestimmt. Der Nutzer bekommt zwar den Eindruck, dass eine nach Relevanz sortierte Ergebnisliste eine größere Menge der jeweils landessprachigen Ergebnisse "nach oben" sortiert, nachweisbar ist dies allerdings nicht.

Eine zweite Sortierungsmöglichkeit besteht in der Option "nach Datum sortiert", dem Pendant der Suchoption "Veröffentlichungsdatum" auf der Ebene der Ergebnislisten. Diese bietet die Möglichkeit, eine bereits erzeugte Ergebnisliste chronologisch ordnen zu lassen. So wird es möglich, noch einfacher und gezielter auf Textstellen aus bestimmten Zeiträumen zuzugreifen. Allerdings wird durch die Benutzung dieses Tools die Ergebnisliste nicht nur umsortiert, auch in ihre Zusammensetzung wird eingegriffen: Google fügt nämlich einer bereits generierten Ergebnisliste im Zuge ihrer chronologischen Sortierung weitere Einträge hinzu oder verkürzt die Liste, bisweilen gar auf Null. Die Gründe hierfür sind

erwartungsgemäß nicht zu eruieren. Erweitert sich die Ergebnisliste allerdings, kann das Phänomen sogar nützlich sein, auch wenn eine chronologische Sortierung eigentlich gar nicht benötigt wird: Denn auf diese Weise hat man Zugriff auf eine größere Zahl von Textstellen:

> Eine Suche nach *indizieren indiziren indiciren indicieren indiziere indizire indicire indiciere* im Suchfeld "mit irgendeinem der Wörter" mit der Vorgabe "zwischen 1750 und 1760" ergab am 17. Juli 2012 eine Ergebnisliste, die 7 Ergebnisse aufwies. Durch die chronologische Sortierung erhöhte Google Bücher diese Zahl auf 8.

Diese Liste war immerhin korrekt chronologisch sortiert – was allerdings nicht immer der Fall ist:

> Der Sucheintrag *Imperialismus Imperialismen* im Feld "mit **irgendeinem** der Wörter" mit der Vorgabe *Frantz* im Feld "**Autor**" hatte am 1. April 2011 nach chronologischer Sortierung eine Ergebnisliste erzeugt, die folgendermaßen sortiert war:
>
> 1873, 1882, 1902, 1921, 1924, 1925, 1967, 1859, 1862, 1970, 1859, 2005, 2009
>
> Die Liste ließ sich auch durch mehrmaliges Umsortieren nicht in eine chronologische Reihenfolge bringen.
>
> Eine Wiederholung dieser Suche am 10. August 2012 erbrachte folgende Reihung:
>
> 2009, 2005, 1862, 1859, 1859, 1859, 1859, 1925, 1924, 1924, 1966, 1882, 1968
>
> Erneut ließ sich die Liste durch mehrmaliges Umsortieren nicht in eine chronologische Reihenfolge bringen.

Unter welchen Bedingungen die chronologische Sortierung funktioniert und unter welchen nicht, konnte nicht festgestellt werden, ebensowenig, ob sich ihre Funktionalität mit der Zeit verbessert. Worin das Problem dabei besteht, ist ebenfalls kaum nachzuvollziehen, da eine chronologische Sortierung nun wirklich kein unlösbares technisches Problem darstellt. Bei der überschaubaren Menge von 7 oder 8 Ergebnissen der Suche nach *indizieren* mag das Problem nicht gravierend sein, sieht man sich aber einer mehrere hundert Ergebnisse umfassenden Liste gegenüber, kann man die Konsistenz der chronologischen Sortierung nicht mehr ohne Weiteres prüfen. Die Funktion ist also nicht zuverlässig. Das wird umso deutlicher, je intensiver man sie nutzt. Um zu dokumentieren, wie gravierend die damit verbundenen Probleme sein können, sei nun ein Beispiel aus Brückner 2012[9] zitiert:

[9] Brückner, Dominik: Noch einmal: Die Google-Buchsuche. In: Sprachreport 2/2012. S. 16-20. Mannheim: Institut für Deutsche Sprache, 2012, S. 17f.

Der Sucheintrag *Imperativismus Imperativismen* im Feld "mit irgendeinem der Wörter" mit der Vorgabe "vor 1980" erbrachte am 15. 11. 2010 eine Ergebnisliste, die 13 Einträge umfasste. Eine chronologische Sortierung der Liste mit Hilfe des Tools "nach Datum sortiert" generierte weitere fünf, also insgesamt 18 Ergebnisse, was wenig überraschend war, da dieser Effekt zu diesem Zeitpunkt bereits bekannt war. Als frühestes Ergebnis wurde in dieser chronologisch sortierten Liste eine Textstelle aus dem Jahr 1970 angegeben.

Nun wurde die gleiche Suche erneut vorgenommen, allerdings mit dem Unterschied, dass jetzt nur Textstellen vor 1900 ausgegeben werden sollten. Aufgrund der Erfahrungen aus der ersten Suche war allerdings zu erwarten, dass diese Suche kein einziges Ergebnis generieren würde. Tatsächlich gab Google Bücher aber fünf neue Ergebnisse aus, der früheste Beleg stammte diesmal aus dem Jahr 1876.

Welche Textstellen zwischen 1900 und 1970 hatte das System unterdrückt? Und warum?

Um dies zu überprüfen, wurde nun in einem nächsten Schritt erneut im Feld "mit irgendeinem der Wörter" nach *Imperativismus Imperativismen* gesucht, diesmal willkürlich zwischen 1890 und 1950. Diese Suche erbrachte (bereits chronologisch sortiert) 62 Ergebnisse, zwei Tage später (bereits chronologisch sortiert) 60 Ergebnisse und noch einmal eine ¾ Stunde später (bereits chronologisch sortiert) 59 Ergebnisse.[10]

In einem letzten Schritt wurde nun *Imperativismus Imperativismen* im gesamten Zeitraum zwischen 1891 und 1969 gesucht. Dies ergab zunächst 47, nach erfolgter chronologischer Sortierung 76 Ergebnisse. Das Resultat: Google unterschlug bei der ersten Suche (vor 1980) insgesamt mindestens 81 Ergebnisse, einschließlich des Erstbelegs (Suchen vor 1876 ergaben keine weiteren Ergebnisse – zumindest nicht an diesem Tag).

Umfang der Ergebnislisten

All diese Beobachtungen werfen nun die Frage auf, wie viele Ergebnisse es nun eigentlich wirklich gibt. Dies hängt zuallererst davon ab, was man unter "Ergebnis" eigentlich versteht. Dass Google Bücher dies nirgendwo klar definiert, dürfte an dieser Stelle keinen Leser mehr überraschen.

Zwar gibt Google Bücher über jeder Ergebnisliste eine ungefähre Ergebnisanzahl an ("Ungefähr […] Ergebnisse"), solange aber unklar ist, was der Nutzer unter "Ergebnis" zu verstehen hat, ist diese Angabe

[10] Dass die Ergebniszahlen durch neu eingescannte Texte zunehmen, ist nachvollziehbar, dass sie aber abnehmen, muss dem Benutzer unverständlich bleiben.

unbrauchbar. Zudem lässt sich bei Ergebnislisten von geringerem Umfang leicht nachzählen, dass die von Google unter "Ungefähr […] Ergebnisse" angegebene Zahl nicht mit der tatsächlich Zahl der gelisteten Ergebnisse nicht übereinstimmt.

Dies wirft die Frage auf, wie diese ungefähre Ergebniszahl zustande kommt. Auf dem Auszählen der dem Nutzer tatsächlich ausgegebenen Belege kann sie nicht beruhen. Dies wiederum führt uns zu unserer Ausgangsfrage zurück: Was ist ein Ergebnis?

Die Zusammensetzung der Ergebnislisten lässt darauf schließen, dass Google als Ergebnis ein Buch zählt, in dem das Suchwort mindestens ein Mal bzw. mindestens eines der Suchwörter vorkommt. Das bedeutet, dass ein Buch, in dem sich ein Suchwort mehrfach findet, genauso als ein Ergebnis zählt wie ein Buch, in dem sich das Suchwort nur ein einziges Mal findet.[11]

Die Ursachen für diese offenbar nur vage Bestimmung sind nur zum Teil Google anzulasten. Häufig wird das gleiche Buch, also derselbe Text in der gleichen Ausgabe, in mehreren Bibliotheken eingescannt. In solchen Fällen ist die Frage nach der Natur des Ergebnisses unproblematisch. Wie aber soll man verfahren, wenn es sich zwar um den gleichen Text, aber um verschiedene Ausgaben handelt? Oder um eine Textstelle, die in unterschiedlichen Versionen eines Texts (etwa in verschiedenen Fassungen eines Gedichts) vorkommt, für sich genommen in beiden Versionen aber identisch ist? Vergleichbares gilt auch für Zitate: Soll man das Zitat einer Textstelle als eigenständiges Ergebnis zählen? Und wie soll Google Bücher ein Zitat als solches erkennen?

Vernünftigerweise zählt Google in diesen letztgenannten Fällen jedes Vorkommen des Suchausdrucks als ein eigenständiges Ergebnis. Die Fragen, die zu beantworten wären, wollte man diese Probleme anders lösen, würden auch weit über das hinausgehen, wozu ein Programm wie Google Bücher gedacht ist. Was in solchen Fällen als Ergebnis zu zählen hat, muss zudem dem Benutzer anheim gestellt bleiben, der dies in Abhängigkeit von Erkenntnisinteresse und methodischem Ansatz entscheiden muss.

Ein ernstes Problem werfen die Homographen auf. Diese zu unterscheiden, dürfte Google Bücher und vergleichbaren Systemen noch auf lange Zeit hinaus Schwierigkeiten bereiten. In Google Bücher als einem einzelsprachübergreifenden System verschärft sich das Problem dadurch, dass homographe Elemente aus verschiedenen Sprachen auftreten können (man denke etwa an <the> im Englischen, Französischen und Italienischen oder an <thee> im Niederländischen und Englischen). Dies könnte durch einen funktionierenden Sprachfilter entschärft werden, der jedoch, wie oben beschrieben, nicht zur

[11] Ähnliches gilt auch für die vollständige Ansicht der Bücher, hier in durchaus sinnvoller Weise: Erscheint ein Wort z. B. in den Kopfzeilen mehrerer Seiten, etwa als lebender Kolumnentitel, so wird dieses Vorkommen zwar gelb unterlegt, aber nicht durch eine Markierung am rechten Rand als Treffer markiert.

Verfügung steht. Das Problem wird im Gegenteil täglich größer, da die über Google recherchierbare Textmenge stetig zunimmt.

Über die Ursachen anderer Probleme lässt sich nur spekulieren: Bereits angesprochen wurden die Abhängigkeit der Ergebniszahlen von der Sortierungsweise der Ergebnisliste, das Phänomen, dass Google bereits gefundene Textstellen bei einer späteren Suche nicht wiederfindet, die Suchfunktion, die in Form von Ableitungen und Zusammensetzungen Ergebnisse mit ausgibt, die gar nicht gesucht worden waren, sowie die Tatsache, dass Google bei zeitlichen Eingrenzungen im System vorhandene Textstellen nicht anzeigt.

Besonders schwerwiegend ist ein Problem, dass dadurch verursacht wird, dass die Ergebnislisten maximal nur etwa sechshundert Einträge anzeigen, und zwar unabhängig davon, wie viele dem Suchausdruck entsprechende Textstellen im System tatsächlich vorhanden sind. Da bei den meisten Suchen nicht solche großen Ergebniszahlen benötigt werden, ist dieses Phänomen nicht sehr bekannt. Auf die von Google angegebene Ergebnismenge ("Ungefähr […] Ergebnisse") hat der Nutzer also gar keinen Zugriff, die entsprechende Zahl stimmt daher vielleicht sogar, ist so aber auch nicht verifizierbar. Diese "Deckelung" der Ergebnislisten fällt nur dann auf, wenn man Ergebnismengen, die mehr als 600 Belege umfassen, mit Hilfe des Tools "nach Datum sortiert" chronologisch sortieren lässt: Dabei bemerkt man, dass Textstellen aus ganzen Zeiträumen fehlen und man sich auch nicht durch die Ergebnislisten zu ihnen hindurchklicken kann. Diese Ergebnisse sind nur durch eine erneute Suche zu erreichen, bei der der Suchzeitraum entsprechend eingeschränkt wird, nämlich so, dass die Suche nicht mehr als 600 Ergebnisse generiert. Bei hochfrequenten Suchwörtern führt das schnell zu einer immensen Zahl an einzelnen Suchaufträgen, was folglich äußerst mühsam sein kann.

Auf all diese Schwierigkeiten hat zudem die Qualität der OCR/ICR großen Einfluss: Es kann frustrierend sein, sich durch hunderte von Belegen für *iudiciren* zu arbeiten, wenn man eigentlich Belege für *indiciren* sucht. Auch wenn dies zumindest teilweise mit der Druckqualität der Vorlagen zusammenhängt und man als Nutzer andererseits den Eindruck hat, dass die Qualität der OCR/ICR über die Jahre zunimmt, so können derartige Verlesungen bisweilen doch starken Einfluss auf die Ergebniszahlen haben.

Vor diesem Hintergrund erscheinen Aussagen über die Häufigkeit von Phänomenen, die auf Daten von Google Bücher beruhen, in erheblichem Maße unsicher. Dasselbe gilt vermutlich auch für andere Google-Produkte, etwa die Internet-Suchmaschine. Wenn darüber hinaus Ergebnisse nur zufallsabhängig (wieder) auffindbar sind, muss selbst ihr bloßes (Nicht-)Vorhandensein in Frage gestellt werden. Solange Google für zwei identische Suchanfragen nicht auch zwei identische Suchergebnisse produziert, sind

wissenschaftliche Aussagen, die auf Google-Daten beruhen, zudem nicht nachprüfbar und damit letztlich wertlos.

Bibliographische Angaben

Die bibliographischen Angaben der Google-Buchsuche stellen sich häufig als unzuverlässig heraus, sind unvollständig oder fehlen ganz. Dies hängt in erster Linie damit zusammen, dass diese äußerst wichtigen Daten auf sehr geringen Textmengen beruhen, also für Lesefehler der OCR/ICR in besonderem Maße anfällig sind. Ein Fraktur-S ähnelt einem Fraktur-G oft allzu sehr, als dass das Programm sie auseinanderhalten könnte. So verzeichnet das System z. B. den Autor Isidor Sutter sowohl unter seinem richtigen Namen als auch als Isidor Gutter.

Auch die von Google angegebenen Erscheinungsjahre sind alles andere als verlässlich. Die Goethe-Ausgabe von 1659 ist schon fast sprichwörtlich. Befindet sich mehr als eine Jahreszahl auf einer Titelseite, so weiß Google nicht, bei welcher Angabe es sich um das Erscheinungsjahr handelt. Es sind aber vor allem Zeitschriften, bei denen die Datierung häufig fehlerhaft ist. Google datiert diese nämlich meist nach dem Erscheinungsjahr der ersten Nummer. Im schlimmsten Fall muss man sich dann sowohl den angegebenen Jahrgang als auch die angegebene Nummer der Zeitschrift besorgen – und selbst dann ist nicht garantiert, dass man die gesuchte Textstelle auch in einem der beiden Bände findet.

Denn für die Seitenzahlen gilt ähnliches: Google Bücher gibt in einem eigenen Feld die Seitenzahl der jeweils aufgeschlagenen Seite an (und auch in der Ergebnisliste ist eine Seitenzahl zu lesen). Bei Büchern, die in vollständiger Ansicht zu sehen sind, fällt jedoch bisweilen auf, dass die vom System angegebenen Seitenzahlen und die auf den Scans zu lesenden nicht übereinstimmen. Die Seitenzahlangaben müssen daher immer anhand des Scans rückgeprüft werden. Ist eine vollständige Ansicht nicht verfügbar (z. B. bei den "eingeschränkten Leseproben"), ist es unumgänglich, die Seitenzahl am gedruckten Buch zu verifizieren, bevor eine Textstelle übernommen werden kann.

Dabei erlebt man häufig Überraschungen wie diese:

> Bei einer Google-Recherche fiel auf, dass die Seitenzählung mitten in einem Buch wieder von vorn begann. Eine genauere Prüfung ergab, dass der erste Band der von Guhrauer herausgegebenen

> Sammlung "Leibnitz's Deutsche Schriften" (1838 erschienen) zusammen mit dem zweiten Band (1840 erschienen) als ein einziger Titel registriert worden war.[12]

Dies kann verschiedene Gründe haben:

> 1. Bekanntlich binden Bibliotheken aus unterschiedlichen Gründen physisch separat erschienene Bücher zusammen, wenn eine unmittelbar einleuchtende Zusammengehörigkeit besteht (etwa bei Zeitschriftenjahrgängen oder mehrbändigen Werken) oder andere Gründe dafür sprechen. Werden solche Bände von Google eingescannt, so wird dies entsprechend ins System übernommen, d. h. der Scan wird im Sinne Googles als ein Buch in der Datenbank abgelegt.
>
> 2. Bisweilen werden durch den Scanvorgang aber auch Bücher bloß digital zusammengefügt. Diese können, müssen aber nicht unbedingt, in einem bestimmten Zusammenhang zueinander stehen. Zum Teil findet man Bücher aus verschiedenen Jahrhunderten und zu vollkommen verschiedenen Themen, die zu einem einzigen Google-Buch verschmolzen wurden.

In unserem Beispiel sind beide Erklärungen denkbar. In Bezug auf die Schwierigkeiten, die dieses Phänomen nach sich ziehen kann, ist es aber besonders illustrativ: Bei der Übernahme von Daten aus dem zweiten Teil eines solchen Digitalisats kann es nämlich passieren, dass versehentlich die bibliographischen Angaben von der Titelseite des ersten Teils übernommen werden. Diese Gefahr ist, wie an unserem Beispiel schön zu sehen ist, umso größer, je mehr sich die zusammengebundenen oder zusammengescannten Bände im Layout ähneln. Zwar sind solche Fälle selten, sie kommen aber deutlich häufiger vor, als man erwarten würde – und allein die Tatsache, dass sie überhaupt vorkommen, sollte den Nutzer vorsichtig machen.

Auch in diesem Fall gilt besondere Vorsicht bei in "eingeschränkter Vorschau" oder in "Auszügen" vorhandenen Büchern: Eine Überprüfung anhand der von Google bereitgestellten bibliographischen Daten ist nämlich nicht möglich, da das Digitalisat nicht in Gänze durchsuchbar ist. Der Griff zum Buch ist unumgänglich.

[12] Guhrauer, G. E. (Hrsg.): Leibnitz's Deutsche Schriften. Erster Band, Berlin 1838, und zweiter Band, Berlin 1840. http://books.google.de/books?id=Vk5xa7wmty4C&printsec=frontcover&hl=de&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Fazit

Trotz aller beschriebener Probleme findet Google Bücher in den letzten Jahren zunehmend auch wissenschaftliche Anwender.[13] Und tatsächlich kann Google Bücher ein wertvolles Hilfsmittel für den Lexikologen/Lexikographen sein: Allein der Umfang der Textmenge, die Google bereitstellt, macht es schwer, dieses Tool zu ignorieren. So ermöglicht es etwa das Auffinden von Frühbelegen, die zum Teil zu deutlichen Früherdatierungen von Wortverwendungen führen[14], den Nachweis, dass es sich bei einem selten belegten Wort eben doch nicht um einen Okkasionalismus handelt[15] oder das womöglich sogar ausführliche Belegen bislang wenig bekannter historischer Bedeutungen[16].

Solche Nachweise gelingen allerdings nur, wenn der Nutzer dafür sorgt, dass die vielen Probleme, von denen hier nur einige der gravierendsten beschrieben wurden, die Ergebnisse nicht verfälschen. Welche Probleme, die dem Nutzer gar nicht bewusst werden, darüber hinaus Einfluss auf die Suchergebnisse haben könnten, muss ohnehin dahingestellt bleiben. Dass die entsprechenden Maßnahmen vom Nutzer getroffen werden müssen und nicht vom System vorgesehen sind, ist als ein schwerwiegender Mangel des Tools anzusehen.

Dem steht wiederum die große und stetig wachsenden Menge an Texten gegenüber, die Google verfügbar macht. So muss man eine Einschätzung seiner Funktionalität aus wissenschaftlicher Sicht wohl folgendermaßen umreißen:

Google Bücher darf keinesfalls für ein elektronisches Korpus gehalten werden, auch nicht für ein offenes Korpus, da Größe und Zusammensetzung des Inhalt zu keinem Zeitpunkt festgestellt werden können. Nach korpuslinguistischen Kriterien zusammengestellte Korpora kann Google Bücher allenfalls punktuell ergänzen, nicht aber ersetzen. Als alleinige Basis für (die meisten) lexikologisch-lexikographischen Fragestellungen ist es daher ungeeignet. Stattdessen muss man Google Bücher als eine Möglichkeit sehen, an bestimmte Texte und Textstellen leichter oder überhaupt erst heranzukommen. Aus Google Bücher gewonnene Ergebnisse können aber keinesfalls als wissenschaftlich tragfähig erachtet werden und sollten mit google-externen Hilfsmitteln auf ihre Verlässlichkeit überprüft werden.

---

[13] Jedenfalls gewinnt man diesen Eindruck – dazu, wer Google Bücher wozu wie nutzt, äußert sich der Konzern nicht.

[14] Vgl. etwa den derzeit ältesten bekannten Beleg für "Hobby" im Deutschen, in einem Brief Mendelssohns an Lessing aus dem Jahr 1763, also etwa 150 Jahre vor dem bislang vermuteten Entlehnungszeitraum (wie einige andere der Frühbelege noch in der Form *Hobby-Horse*), s. DFWB, Bd. VII, habilitieren – hysterisch, 2010, S. 328, oder einen Beleg für den Gebrauch des Wortes *Homöopathie* bereits im 18. Jahrhundert, also deutlich vor der Einführung des von Samuel Hahnemann eingeführten Heilverfahrens, mit der die Bezeichnung bisher immer in Verbindung gebracht wurde, ebd., S. 347f.

[15] Vgl. etwa die Ableitungen *Humbug(g)er*, *humbug(g)en*, *humbugisch* zu Humbug, ebd., S. 476ff.

[16] So konnte im DFWB die Bedeutung 'Rangordnung der Engel', spezifiziert zu 'eine der drei mal drei Gliederungsebenen in der Rangordnung der Engel' von *Hierarchie* erstmals ausführlich dokumentiert werden, zudem konnten die entsprechenden Verwendungen der Ableitungen *hierarchisch* und *Hierarch* (also 'Engel') nachgewiesen werden. Vgl. ebd., S. 246ff.

Literatur

Brückner 2009: Brückner, Dominik: Die Google-Buchsuche als Hilfsmittel für die Lexikographie. In: Sprachreport 3/2009. S. 26-31. Mannheim: Institut für Deutsche Sprache, 2009.

Brückner 2012: Brückner, Dominik: Noch einmal: Die Google-Buchsuche. In: Sprachreport 2/2012. S. 16-20. Mannheim: Institut für Deutsche Sprache, 2012.

Jeanneney 2006: Jeanneney, Jean-Noël: Googles Herausforderung. Für eine europäische Bibliothek. Mit einem neuen Vorwort des Autors zur dt. Ausg. Nachwort Klaus-Dieter Lehmann. Übers. Sonja Finck, Nathalie Mälzer-Semlinger. Stiftung Preuß. Kulturbesitz Berlin. Wagenbach-Verlag, Berlin-Hamburg 2006.

Kasner/Newman 1940 Kasner, Edward/Newman, James: Mathematics and the Imagination. New York 1940.

Lewandowski 2006: Lewandowski, Dirk: Google Buchsuche. Bücher kostenlos zum Download. In: Password. 10/2006, S. 36.

Lewandowski 2010: Lewandowski, Dirk: Wie verändert die Einigung mit Verlegern und Autoren die Buchwelt?. In: Password 12/2008, S. 13.

Weber 2010: Weber, Klaus: Drei Jahre Freiheitsstrafe für alle Google-Mitarbeiter? Ein Beitrag zur Praxis des Urheberstrafrechts. In: Zeitschrift für Internationale Strafrechtsdogmatik 2010, S. 220 – 226.

DFWB: Deutsches Fremdwörterbuch. Begonnen von Hans Schulz, fortgeführt von Otto Basler. 2. Auflage, völlig neu erarbeitet im Institut für Deutsche Sprache. Berlin/New York 1995ff.