Heinrich C. Mayr, Martin Pinzger (Hrsg.): INFORMATIK 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 227

Sinnsuche noch günstiger: wie Google an Inbegriffen scheitert

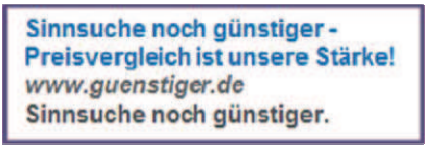
Irene Teich[1], Peter Schnupp[2]

Abstract: Die semantischen Erkennungseigenschaften auch hochentwickelter Suchmaschinen sind noch immer unbefriedigend. Der Aufsatz spürt den Gründen nach und findet sie in der mangelnden Berücksichtigung des heutigen „Wissens über Wissen“: der Konstruktiven Epistemologie, Wittgensteins „Sprachspielen“ und den daraus abgeleiteten „Inbegriffen“ als Wortsymbole, die abhängig von dem jeweiligen Kontext („Wirklichkeit“ des Benutzers) gänzlich unterschiedliche Bedeutungen haben können. Es wird gezeigt, dass in diesen Fällen die derzeitige Chomskysche Linguistik als Grundlage des Analyse- und Suchvorgangs zwar nicht theoretisch aber praktisch nicht verwendbar ist. Eine brauchbare Alternative bietet eine assoziative sprachliche „Mustererkennung“ auf der Basis der *Wittgensteinschen* Linguistik. Deren Implementierung und die ersten marktreifen Produkte werden kurz beschrieben.

Keywords: Semantische Suche, Konstruktivismus, Wirklichkeit, Chomsky, Wittgenstein, Sprachspiele, Linguistik.

1 Einleitung

Eine Google-Recherche nach „Sinnsuche“ überraschte kürzlich neben einigen hunderttausend anderen Ergebnissen mit dem Resultat



Dass es sich hier nicht um ein Standard-Suchergebnis sondern wohl um eine danebengegangene „semantische Recherche“ für eine *Adword*-Anwendung handelt, macht den Unsinn nicht sinniger sondern stellt ganz im Gegenteil eine grundlegenden Mangel jeder suchwort-orientierten Recherche heraus.

Dieser Beitrag erklärt die wissenstheoretisch sehr tiefen Grundlagen dieser generellen Schwäche sowie ein bereits implementiertes und auf dem Markt eingeführtes, völlig anders arbeitendes Konzept, das sich an der seit fast einem halben Jahrhundert dahinschwindenden, assoziativ-wissensbasierten Literaturrecherche umfassend gebildeter Bibliothekare und Dokumentare orientiert. Es beruht auf der Konstruktiven

[1] AIQ.go Enterprises GmbH & Co. KG, Zentrale, Im Mediapark 5, 50670 Köln, teich@aiq-go.de
[2] AIQ.go Enterprises GmbH & Co. KG, Entwicklung, Senftenberger Straße 19, 02977 Hoyerswerda, schnupp@aiq-go.de

Epistemologie von *Watzlawick* und *Rupert Lay*. Zur Implementierung der Grundsoftware sowie der ersten, bereits auf dem Markt eingeführten Applikationen verwendet es anstelle der bisher in der Informatik praktisch ausschließlich verwendeten Linguistik von *Chomsky* diejenige von Wittgenstein.

2 Konstruktive Epistemologie, Wirklichkeiten, Sprachspiele, Inbegriffe

Um diese Implementierung zu verstehen, ist es notwendig, kurz einige Begriffe und Ergebnisse von Nachbarwissenshaften (Epistemologie, Konstruktivismus, *Wittgensteins* „Philosophischer Linguistik", und der von Rupert Lay als Folge unterschiedlicher Wirklichkeiten eingeführten „Inbegriffe") zu referieren, die bis jetzt in der Informatik kaum bekannt sind und noch weniger verwendet werden.

Beginnen wir also mit dem Konstruktivismus. Nach der in den letzten Jahrzehnten vor allem von *Watzlawick* und *Rupert Lay* [Wa2005, La2015a, La2015b] entwickelten *Konstruktiven Epistemologie* ist die von der (ersten) Aufklärung propagierte Auffassung, dass alle „vernünftigen" Menschen die gleiche, sich an der *Realität* orientierende *Wirklichkeit* besitzen oder zumindest anstreben, schlicht falsch – eine Einsicht, die übrigens schon Sokrates in den platonischen Dialogen immer wieder demonstrierte.

Stattdessen *konstruiert* sich jeder Mensch *Wirklichkeiten* aus seinen Erfahrungen, Wünschen, Überzeugungen sowie, vor allem, seinen sozialen Umgebungen und deren Sprachgemeinschaften. *Wittgenstein* nennt sie *Sprachspiele*. Aus diesen stammen die *Begriffe* sowie die sie bezeichnenden *Wortsymbole.* Und da ein Mensch bestenfalls anstreben kann, dass seine Wirklichkeiten der unbekannten Realität entsprechen, sind auch durch dasselbe Wortsymbol benannte Begriffe semantisch sehr unterschiedlich – eine Erkenntnis, die schon durch *Thomas von Aquin*[3] formuliert aber in den Jahrhunderten dazwischen wieder vergessen wurde: Alles Erkannte wird im Erkennenden nach der Weise des Erkennenden erkannt.

Wem all dies zu abstrakt oder zu spekulativ erscheint, können zwei inzwischen leider ebenfalls vergriffene Bücher empfohlen werden. Beide beschreiben sehr unterschiedliche Wirklichkeiten und die in ihnen entwickelten Sprachspiele – sofern dieser harmlose Namen als wissenschaftlicher Term auch in dem dort beschriebenen Kontext gestattet ist: in beiden Büchern ist dieser das zum Glück nur 12 Jahre währende aber auf Tausend Jahre geplante Dritte Reich, dessen damals „Weltanschauung" genannte Wirklichkeit und die spezielle Sprache, die *Victor Klemperer* [Kl1947] im Buchtitel als Lingua Tertii Imperii bezeichnet. Als jüdischer Philologe, der diese zwölf Jahre vor allem auch durch die Loyalität seiner nicht-jüdischen Ehefrau in Dresden überlebte und seine sprachliche Umwelt protokollierte und analysierte, liefert sein Buch

[3] Cognitum autem est in cognoscente secundum modum cognoscentis (Thomas von Aquin, Summa theologiae, I, q.12, a. 4)

vor allem sehr gute Beispiele für die Umwertung von Wortbedeutungen. Diese vor inzwischen siebzig Jahren gewonnene Einsicht ist das Beste den Verfassern bekannte Beispiel für ein erst vor kurzem von *Rupert Lay* (wieder-) entdecktes Konzept des heutigen Konstruktivismus, die „Inbegriffe".

Auch das zweite dieser Bücher entstand unmittelbar nach dem Ende des Dritten Reichs. Der Wiesbadener Philologe *Heinrich Reichert* [Re1948] beschäftigt sich in ihm mit zwei sehr unterschiedlichen Wirklichkeiten und den daraus entstandenen Sprachspielen: dem des römischen Rechts-, Geschäfts- und Alltagslebens („Human und Urban", wie der Titel der zweiten Auflage seines Buchs lautet) im Gegensatz zur Weltanschauung des gerade vergangenen und unmittelbar erlebten Dritten Reichs. Während *Klemperer* seine Analysen weitgehend auf einzelne Worte und damit auf die Herausarbeitung von „Inbegriffen" in der Sprache des Dritten Reichs konzentriert, arbeitet *Reichert* die fundamentalen Unterschiede der römischen Wirklichkeit und der nationalsozialistischen Weltanschauung aus den „Lateinischen Sentenzen" (dem ursprünglichen Titel seines Buchs) heraus, genau der semantische Ansatz, den Wittgenstein in seiner Linguistik verfolgt und fordert.

Die gerade eingeführten *Inbegriffe* spielen als Bezeichnung für viele, oft sehr unterschiedlich konstruierte Begriffe unterschiedlicher Wirklichkeiten in der konstruktiven Erkenntnislehre eine zentrale Rolle. Die *Sinnsuche* aus unserem Titelbeispiel gehört ebenso dazu wie die Mehrzahl der für jeden intellektuellen Diskurs besonders wichtigen Worte wie *Gerechtigkeit*, *Freiheit*, *Bildung*, ... Eine „naive" wortorientierte Suche in großen Wissensbasen oder gar im gesamten Internet muss daran zwangsläufig scheitern.

Wesentlich ist, dass die Vielfalt der in einem Inbegriff enthaltenen und verstandenen Begriffe dem suchenden Menschen noch nicht einmal bewusst ist. Damit ist er denn auch der Interpretation eines solchen Inbegriffs durch eine Suchmaschine hilflos ausgeliefert, es sei denn, die Ergebnisse sind so offensichtlich unsinnig wie die obige „Sinnsuche".

Die Lösung liegt in einem anderen Begriffs- und Sprachmodell für die Computer-Linguistik. Ein entsprechendes, wissenstechnologisches Objektmodell und eine darauf beruhende „konstruktivistischen Recherche-Software" sind inzwischen implementiert.

3 Inbegriffe und Wirklichkeiten

Die Implementierung einer Ontologie und einer befriedigenden Semantischen Recherche über einer großen, vielfältigen Textbasis wird umso schwieriger und schließlich hoffnungslos, je unterschiedlicher die Wirklichkeiten und die (In-)Begriffsdefinitionen der Autoren einerseits und der Recherchierenden andererseits sind. Wenn ein Inbegriff schon je nach dem ihm verwendenden Sprecher sehr unterschiedliche Inhalte haben kann: was hindert mich dann, ihn einfach einmal zu verwenden, wenn mir gerade keine

passende Bezeichnung einfällt? Etwa die „Gerechtigkeit“ – da sie leider ein Inbegriff ist, schützt sie niemand davor, dass sich ein Politiker, ein Wirtschaftler, ein Philosoph auf sie beruft, wenn ihm die richtigen Worte fehlen? „*Gewöhnlich glaubt der Mensch, wenn er nur Worte hört, es müsse sich dabei doch auch was denken lassen*“ sagt Mephisto in der Hexenküche von Faust I. So dass sein Schöpfer, *Goethe*, durchaus auch den Anspruch auf die Erfindung der Inbegriffe erheben könnte, wenn er es denn nötig hätte.

Nun aber zu Alltagsanfragen, wie sie täglich zu hunderten an Google oder andere Suchmaschinen gestellt werden. Und vor allem auch zu den auch heute schon von ihnen verlangten und auch geleisteten semantischen Interpretationen von Anfragen und dem „Sinn“ der in ihnen enthaltenen Suchworte. Zuerst ein einfaches Beispiel:

Mercedeshändler Köln.

Natürlich geht der Anfragende davon aus, dass die Suchmaschine bei der Aufbereitung und Verarbeitung der Anfrage einige einfache semantische Umformungen und Ergänzungen vornimmt. Sie sollte also das erste Wort in seine beiden Bestandteile zerlegen. Und sie sollte vermutlich den (In)Begriff „Händler“ auch durch einige andere Inhalte wie „Vertretung“, „Vertreter“, „Niederlassung“ ergänzen. Ob sich in „seiner Wirklichkeit“ ein bestimmter Autohändler lieber als Niederlassung sieht, und dass dies für ihn aus rechtlichen, geschäftlichen oder sonstigen Gründen vielleicht sogar sehr wichtig ist, interessiert den prospektiven Autokäufer nicht. Damit sind diese beiden Wirklichkeiten auch problemlos mit einander verträglich. Und es ist auch völlig unerheblich, ob diese Unterschiede jemandem bewusst werden oder nicht.

Nicht ganz so einfach ist es mit dem ersten Teilwort, „Mercedes“. Will der Anfragende tatsächlich unbedingt einen Mercedes, oder ist das für ihn nur ein Sammelbegriff für ein Premium-Auto, so dass eine semantische Suche ihm auch Audi-, BMW- oder Jaguar-Händler nachweisen sollte? Und wenn nicht – wie ist es mit einer „Daimler-Vertretung“? Hier lohnt es sich schon eher, vor der Konstruktion und Implementierung eines semantischen Recherchesystems, über Unterschiede in Wirklichkeiten und Inbegriffen nachzudenken. Aber „falsche“ Entscheidungen sind nicht sehr folgenreich. Werden nach „Mercedes“, vielleicht mit einer etwas geringeren Bewertung durch den Suchalgorithmus, auch andere Premium-Marken gesucht und als Ergebnisse ausgegeben, wird dies den Anfragenden nicht allzu sehr stören, selbst wenn in seiner Wirklichkeit ein Mercedes „das einzige ordentliche Auto“ wäre.

Und umgekehrt – beschränkt sich die Suchmaschine ausschließlich auf Mercedes, kann der Recherchierende ja einfach alle ihn interessierenden Marken in seiner Anfrage nennen und sich darauf verlassen, dass diese bei der Verarbeitung als Alternativen interpretiert werden. Oder er kann einfach für jede ihm sonst noch in den Sinn kommende Automarke eine entsprechende Anfrage nachreichen. Wobei es aber wichtig ist, wie schnell er die Antwort auf die Recherche erhält. Sind es Sekunden, so ist das überhaupt kein Problem. Dehnen sich aber die Antwortzeiten auf Stunden, Tage, Wochen aus, wie es in vor-digitalen Jahrhunderten immer und heute noch bei anspruchsvollen Recherchen etwa für Gerichtsgutachter oder in der Wissenschaft die

Regel ist, gilt das nicht mehr. Fehlinterpretationen, oder in unserem Modell Fehlanpassungen von Wirklichkeiten und Inbegriffen, können dann nicht mehr „praktisch umsonst“ durch erneute Anfragen korrigiert werden. Wir werden darauf noch zurückkommen.

Zuerst aber eine kleine Veränderung der Suchanfrage:

Gotteshäuser Köln.

Jetzt ist es plötzlich nicht mehr uninteressant, was in den Wirklichkeiten des Anfragenden und der Autoren der Dokumente in der Wissensbasis ein „Gotteshaus“ ist. Der Kölner Dom wird vermutlich von allen als solcher verstanden werden, genauso wie die Pfarrkirche einer christlichen Gemeinde in einer Kölner Vorstadt. Aber wie ist es mit einer Synagoge oder einer Moschee? Mit einer Meditations-Stätte einer Sekte oder in einem Krankenhaus? Mit einer kleinen Waldkapelle irgendwo im Grünen oder einem Versammlungs- und Verwaltungshaus von Scientologen? Oder der Ruine einer Klosterkirche aus dem Mittelalter?

Schon wird es bedeutsam, wie Vertreter unterschiedlicher Wirklichkeiten diesen Inbegriff interpretieren. Und es sind nicht nur emotionale Verärgerungen, wenn etwa jemand mit einer mohammedanischen Wirklichkeit seine Moschee nicht als Gotteshaus anerkannt sieht. Auch „sachlich“ kann für ihn, ebenso wie die Vertreter anderer Wirklichkeiten, die Klärung dieser semantischen Frage sehr wichtig und relevant sein: etwa wenn eine Gemeinde die Baugenehmigung für eine Moschee verweigert und es darum geht, ob dies gegen im Grundgesetz garantierte Freiheiten verstößt.

Man könnte nun versuchen, das Problem einfach auf die Ontologen abzuwälzen, welche die semantischen Definitionen für die Suchmaschine erstellen: sollen die doch bitte „Gotteshaus“ so definieren, dass alle vernünftigen Wirklichkeiten sich in dieser Definition wiederfinden. Diese Aufgabe ist unlösbar, aus verschiedenen Gründen.

Aber, wie man in einem anderen, ebenfalls empfehlenswerten Buch von Rupert Lay über den Konstruktivismus [La2015b], nachlesen kann, sind wertende Attribute wie „vernünftig“ oder „richtig“ auf Wirklichkeiten nicht anwendbar. Das ergibt sich schon daraus, dass das einzige Maß dafür die Realitätsnähe wäre. Und dass ein Mensch die Realität nicht erkennen kann, weiß man seit mehr als zweitausend Jahren. Die einzige, leidlich „vernünftig“ einer Wirklichkeit zuschreibbare Eigenschaft ist „nützlich“ im ethischen Sinn. Aber die stellt sich leider, wenn überhaupt, erst im Nachhinein an den Folgen heraus: dass der Nationalsozialismus und seine Wirklichkeit nützlich waren, wird angesichts des von ihm angerichteten Elends wohl kaum noch ein „vernünftiger“ Mensch behaupten.

Aber setzen wir uns einmal über diese ganze Wirklichkeits-Theorie hinweg und nehmen als Informatiker einfach einmal an, wir müssten den Begriff „Gotteshaus“ in das Objektmodell einer Ontologie einbauen, vielleicht mit einer Parametrisierung der zugrundeliegenden Wirklichkeit als „christlich“ „mohammedanisch“, „agnostisch“ …

Wir werden vermutlich damit beginnen, dass wir ihn, ähnlich wie den „Mercedeshändler“ im ersten Beispiel, in die Begriffe „Gott“ und „Haus“ zerlegen. Und wenn wir das „Haus“ nicht unmittelbar als „Wohnhaus“ definieren sondern ein Gebäude, das irgendwie „Gott“ gewidmet ist oder ihm verbundenen Funktionen oder Handlungen dient, werden uns vermutlich Vertreter aller drei betroffenen Wirklichkeiten folgen.

Aber was meint „Gott“? Ist das ein Inbegriff? Vermutlich nicht – was der derzeitige Konstruktivismus dazu meint, werden wir am Ende dieses Abschnitt noch kurz ansprechen. Vorerst einmal: wie „konstruieren“ wir einen Gottesbegriff. Schon da gibt es die ersten Probleme: einige religiöse Wirklichkeiten, die mohammedanische ebenso wie die jüdische, verbieten dies sogar ausdrücklich. Also müssen wir einen anderen Weg suchen, zum Beispiel die Übernahme des abrahamitischen Gottesbegriffs, den zumindest Christen, Juden und Mohammedaner gleichermaßen anerkennen. Auch Agnostiker finden ihn meist akzeptabel: wenn es „in Wirklichkeit“ keinen Gott gibt, ist es auch gleichgültig, wie andere Wirklichkeiten ein solches Abstraktum konstruieren. Und das gilt auch für viele anderen religionsartigen Gemeinschaften – auch für sie gibt es entweder keinen Gott im abrahamitischen Sinn, oder es gibt viele Götter, und da kommt es, auf die Eigenschaften von einem von ihnen nicht mehr besonders an.

Doch wir stoßen auf andere Schwierigkeiten. Zum einen schließt die Beschränkung auf den abrahamitischen Gott einige der obigen Prätendenten auf die Bezeichnung Gotteshaus aus. Das wäre noch zu ertragen. Aber für einen Juden oder einen Mohammedaner ist zwar Gott Vater abrahamitisch, aber nicht Jesus. Im Verständnis eines Mohammedaners ist er allenfalls der zweitgrößte Prophet nach Mohammed. Und für einen Juden ist er wohl noch nicht einmal das. Noch schlimmer steht es um den Heiligen Geist. Sogar aus Sicht mancher Christen ist er nur ein Konstrukt, das Kaiser Konstantin 325 während des Konzils zu Nikäa einführte, um Angesichts der Anerkennung von Jesus als „Sohn Gottes“ den Anspruch des Christentums als monotheistischer Religion durch eine typische Kompromiss-Lösung aufrechterhalten zu können: man erfindet einen dritten Begriff und fasst sie als „Dreieinigkeit“ zu verschiedenen Erscheinungen des gleichen Gotts zusammen.

Wenn auch dies einem Agnostiker vielleicht noch gleichgültig ist, kann man es einem in einer mohammedanischen oder jüdischen Wirklichkeit lebenden Menschen nicht übelnehmen, wenn er dies bei aller Toleranz nicht nachvollziehen kann. Und wenn kaum bestreitbar ist, dass im Kölner Dom neben Gott Vater auch Jesus und der Heilige Geist verehrt werden – disqualifiziert es ihn nicht vielleicht als Gotteshaus?

Aus einem halben Jahrhundert Erfahrung in der Software-Entwicklung und in der Künstlichen Intelligenz: so etwas kann nur scheitern, schon in der Konzeption unter dem herkömmlichen Linguistik- und Objektmodell. Wir können nicht gegen die Einsichten des Konstruktivismus und die Inbegriffe angehen, sondern wir müssen unmittelbar auf beidem aufbauen.

Ach ja: ist „Gott“ eigentlich ein normaler Begriff, wie „Auto“, „Haus“, oder auch „Mercedes“? Sicher nicht – dazu sind die Bedeutungen in den verschiedenen

Wirklichkeiten viel zu unterschiedlich. Ist es also ein Inbegriff? Auch das ist er nicht: viele, vor allem religiöse Wirklichkeiten akzeptieren weder viele verschiedene, durch das Wort bezeichnete aber unterschiedliche Begriffe. Das ist viel zu nahe an Vielgötterei oder zumindest Konstrukten wie der christlichen Dreieinigkeit. Und einen Gottes-Inbegriff zu konstruieren verbieten, wie bereits gesagt, viele Religionen mit gutem Grund. Dass es nicht funktionieren kann, haben wir ja gerade gesehen.

Der moderne Konstruktivismus [La2015a] führt deshalb eine weitere Begriffsklasse ein, welche er als „Urbegriffe“ bezeichnet. Er nimmt an, dass sie in allen oder zumindest den meisten Menschen entweder von Geburt an oder in seiner frühen Sozialisation angelegt und dann in seine Wirklichkeiten eingebaut werden. Neben „Gott“ gehören dazu vor allem emotional besetzte Begriffe wie „Mutter“, „Liebe“, „Gewissen“. Ob er sie dann im Laufe seines Lebens irgendwie rationalisiert oder immer „aus dem Bauch heraus“ erfühlt, hängt von seiner Persönlichkeit, seiner Bildung und seiner Einbettungen in Kommunikations- und Lebensgemeinschaften ab. Auch hier sind, wie bei allen Wirklichkeiten „richtig“, „falsch“ oder „wahr“ schlicht keine anwendbaren Attribute.

Auch auf diese Urbegriffe und den Umgang mit ihnen könnte wieder Goethe die Urheberrechte anmelden. In Herrmann und Dorothea sagt er von den „Kindern“, auch einem solchen Urbegriff: „So wie Gott sie uns gab, so muss man sie haben und lieben.“ Und wir Nachfolger sollten keine Ontologie, keine Semantik und keine semantische Recherche akzeptieren, die nicht mit Gott, Kindern, Müttern und ähnlichen Urbegriffen umgehen kann …

4 Konstruktive Wissensrecherche

Repetieren wir kurz die wesentlichen Probleme einer Recherche. Sie soll sich nicht auf ein oder wenige Suchworte beschränken. Statt dessen soll wenigstens insoweit „semantisch“ sein, dass sie, wie in unserem obigen Beispiel, etwa zusammengesetzte oder flektierte Worte in die Stammformen der Bestandteile zerlegen und leidlich mit mehreren Interpretationen eines solchen Textelements umgehen kann. Eine Aufgabe ist, mit verschiedenen, aus einem Inbegriff wirklichkeitsabhängig zu extrahierenden Bedeutungen umzugehen. Entsprechend hat sie die einfachere, aber ähnliche Aufgabe der Unterscheidung von Synonymen, etwa für das von *Wittgenstein* als Beispiel verwendete Wort „König“, das einmal die Schachfigur und ein andermal das Staatsoberhaupt einer Monarchie zu erfüllen. Wie *Wittgenstein* bemerkt, ist aus dem Einzelwort nicht zu ersehen, was gemeint ist. Findet man es in einem längeren Text, so ist meist sofort klar, was gemeint ist. In aller Regel weiß der Leser, ob er gerade einen Text über Schach, über Politik oder in einem Lifestyle-Blatt liest. Aus dem jeweiligen Kontext, dem Sprachspiel der besprochenen Wirklichkeit, weiß er dann auch, welcher König gemeint ist.

Natürlich kann auch Google nicht erkennen, was für einen König es suchen soll, wenn man bloß das einzelne Wort eingibt. Vielleicht kann eine Semantik-Software durch

Annahme aus Bezügen zu anderen eingegebenen Worten eine mehr oder weniger sichere Klassifizierung vornehmen. Ist das zweite Suchwort „Turm“ oder „Dame“, meint der Recherchierende vermutlich die Schachfigur, und ist es „Rochade“, so ist diese Interpretation so gut wie sicher. Aber kann man vom Benutzer die Eingabe der „richtigen“ Zweit- oder Dritt-Suchwörter dem Benutzer überlassen?

Ist es nicht besser, zu den alten, bewährten Rechercheverfahren zurückzukehren. Handelte es sich um komplexere Rechercheaufgaben in kaum oder gar nicht bekannten „Wirklichkeiten“, ging man zu seinem Bibliothekar oder Dokumentar. Und natürlich warf man dem nicht einfach ein oder zwei Worte an den Kopf, etwa „Gotteshäuser Köln“ oder die „Sinnsuche“ aus dem Titel. Sondern man gab ihm ein oder mehrere Texte und bat ihn um ähnliches, verwandtes Material. Das konnte eine Kundenanfrage sein, Zeitungsartikel, Gerichtsurteile – was immer die gerade anliegende oder vermutete Wirklichkeit so hergab. Nach ein paar Tagen, oder in schwierigen Fällen ein paar Wochen legte ein guter Dokumentar seinem Auftraggeber die Ergebnisse seiner „semantischen Recherche“ vor: Dokumente als Kopien oder Sonderdrucke, Zettel mit den wichtigsten Begriffen, ihrer Bedeutung im aktuellen Kontext und Referenzen auf die Fundstellen, „Teaser“, also kurze, relevante Textauszüge – ähnlich unseren obigen Goethe-Zitaten.

Wie das unser „Digitaler Dokumentar 2.0“ auf dem Bildschirm als Ergebnis einer Recherche nach Eingabe eines Suchttexts tut, zeigt Abbildung 1.

Abb. 1 Das Rechercheergebnis des digitalen Dokumentars

Natürlich braucht er dazu nicht Tagen oder Wochen sondern einige Sekunden, wie wir es

von unseren Computern und dem Internet gewohnt sind. Die Fundobjekte sind kleine viereckige Knöpfe, die dann das jeweilige Fundobjekt bei Darüberfahren mit der Maus oder nach einem Mausklick anzeigen oder aufblättern.

Schon haben wir den digitalen Dokumentar erfunden – wir müssen ihn nur noch implementieren. Aber da stoßen wir auf praktische Probleme, die wir bereits oben theoretisch abgehandelt haben. Listen wir sie auf, bevor im nächsten Abschnitt kurz ihre Lösung beschreiben.

- Die Software muss „irgendwie" aus den Eingabetexten neben den Inhalten und Aussagen des Verfassers auch die „Wirklichkeit" erkennen und entnehmen, in und aus der Autor seinen Text verfasst hat.
- Sie soll aus der Dokumenten- oder Wissensbasis die Dokumente finden, „assoziieren", die nicht nur den Inhalten sondern auch der Wirklichkeit des Suchtexts am besten entsprechen.
- Dabei muss sie diejenigen Begriffe – vor allem auch als Begriffsinhalte der Inbegriffe – extrahieren, welche diese Assoziation nahelegen und verursachen.
- Sie sollte dem Auftraggeber nicht nur diese Begriffe sondern auch Referenzen auf ihre Herkunft und ergänzende Informationen liefern, welche dem Benutzer helfen, diese Begriffe sowohl im Kontext der Wirklichkeiten, sowohl die der Dokumentenautoren als auch seiner eigenen zu verstehen und zu bewerten.
- Und schließlich sollte sie dem Benutzer nicht nur Begriffe sondern auch kurze Textausschnitte vorlegen, welche beim Verständnis der gefundenen Dokumente und den Auffassungen und Informationen ihrer Verfasser am besten helfen.

Wenn wir behaupten würden, wir hätten den Digitalen Dokumentar entweder top-down von der gezeigten Benutzeroberfläche oder bottom-up aus einem entsprechenden Objektmodell nach den Regeln der Softwaretechnologie entworfen und realisiert – der Leser würde es uns doch nicht glauben. Deshalb beschränken wir uns lieber darauf, auf welche Weise wir die einzelnen Komponenten fanden und zusammenstellten.

5 Der Digitale Dokumentar

Beim Versuch, diese Teilleistungen des Digitalen Dokumentars durch ein Objektmodell und seine Methoden softwaremäßig zu implementieren, stößt man auf mancherlei ungelöste Probleme. Die meisten entstehen durch den Konstruktivismus als Modell für die Wissensgewinnung, Wissensanalyse, Wissensverdichtung und Darstellung des dem Auftraggeber vorgelegten Rechercheergebnisses, das wir als „Wissensobjekt" bezeichnen und dessen Visualisierung Abb. 1 zeigte. Die meisten dieser Schwierigkeiten brachte die Akzeptanz gleichberechtigter aber oft sehr unterschiedlicher (aber als solche noch nicht einmal erkannter) Wirklichkeiten und entsprechend der sehr unterschiedlichen individuellen Begriffsinhalte der von der Sprachgemeinschaft zu ihrer

Konstruktion und Beschreibung verwendeten Wortsymbole, der Inbegriffe. Manche konnten mit bereits existierenden Modellen und Techniken, allenfalls nach leichten Anpassungen, behandelt und gelöst werden. Andere benötigten, zumindest nach unserer Einschätzung, das völlige Verlassen der eingeführten Grundlagen (der „Paradigmen" nach *Thomas Kuhn* [Ku1991] und Übergang zu einem ganz neuen Konzept. Dies gilt vor allem für das verwendete Linguistik-Modell und, eng damit verbunden, der Ontologie zur Beschreibung der wesentlichen Begriffe. Hier adoptierten wir anstelle der in die Informatik allgemein eingeführten *Chomsky*-Linguistik diejenige von *Wittgenstein*.

5.1 Erkennung und Analyse von Kontexten und „Wirklichkeiten"

Zuerst einmal: die uns interessierenden Wissensbasen sind nicht „Big Data", also keine großen, verteilten und im Wesentlichen relational organisierten Datenbanken. Sie sind eher „Big Wisdom", unformatierte Texte in zahlreichen Nationalsprachen und – wie wir inzwischen von Wittgenstein gelernt haben – noch weit zahlreicheren „Sprachspielen". Die Grundlage dafür ist deshalb die Erkennung und Analyse von Satz- und Wortstrukturen und ihrer Zusammenhänge. *Chomsky* verwendet dazu eine im Wesentlichen lineare Syntaxanalyse, wobei die Klassifizierung in Wortarten und deren (mutmaßliche) Bedeutungen durch Nah-Kontexte erfolgt.

Wittgenstein behauptet und belegt statt dessen, dass dies zumindest dann nicht ausreicht, wenn man Wortsymbole im Kontext von Sprachspielen und der dadurch induzierten Vielfalt der Inbegriffe (in heutiger Sprechweise) semantisch verstehen und verarbeiten will. Dazu braucht man längere Phrasen und Sätze als primäre Elemente des Analyseprozesses, die dann einerseits der ersten Erkennung und Klassifizierung der in ihnen verwendeten Wortsymbole dienen und andererseits durch das über den ganzen Text aufgespannte Netz dieser Teilelemente, der „Sätze", die genaue semantische Diskriminierung der Wortsymbole ermöglichen, also im obigen Beispiel erkennen lassen, von was von einem „König" die Rede ist.

Auf den ersten Blick ist nicht zu sehen, wieso eine Chomsky-Linguistik nach entsprechender Erweiterung und Ausbau nicht auch diese, von Wittgenstein gestellte Erkennungsaufgabe leisten kann. Auch wir haben daran nicht gezweifelt und wurden darin sogar bestärkt, als wir auf ein Buch namens „Grammar of English Grammars" stießen, eine von *Goold Brown* [Br1851] verfasste „Metagrammatik" des Englischen, die wir deshalb auch in [TS2016] ausführlicher referieren. Es ist eine ausführliche, verbale aber trotzdem in den für uns wichtigsten Abschnitten sehr formale Anleitung, wie man eine englische Grammatik so schreiben soll, dass nach ihr entweder ein korrekter Text in dieser Sprache erstellt oder ein vorhandener analysiert („parsed") werden kann, und zwar nicht nur syntaktisch sondern, soweit möglich, auch semantisch korrekt.

Das Buch ist in mehrerer Hinsicht unglaublich. Zum einen könnte nach ihm ein leidlich geschickter Softwareentwickler, zwar mit beträchtlichem Zeitaufwand aber ohne große Probleme einen Parser für englische Texte in einem bestimmten Sprachspiel schreiben – „könnte", weil selbst in der Heimatstadt von *Goold Brown*, New York, Computer und

Software-Technologen erst über einhundert Jahre später Einzug hielten: das Buch wurde 1851 veröffentlicht. Und zeitaufwendig wäre es, selbst bei der exzellenten Vorbereitung, auch – das Buch enthält weit über 2000 Seiten, sämtlich mühelos und zeitweise sogar direkt kurzweilig zu lesen, aber trotzdem eine recht umfangreiche „Spezifikation". Das hätte uns noch nicht abgeschreckt. Aber das immer wieder, wenigstens für die wichtigsten natürlichen Sprachen und für jede Wirklichkeit und die von ihr induzierten Sprachspiele – also letztlich mehrfach für jeden betroffenen Menschen … nichts gegen *Chomsky* als theoretische Alternative, aber praktisch kam er für uns nicht mehr in Frage, nachdem wir das realisiert hatten.

5.2 Assoziative Recherche mit konventionellen Suchmaschinen

Was aber dann? Alles, von der Basissoftware bis zur Benutzeroberfläche neu zu konzipieren und zu implementieren kam nicht in Frage, selbst wenn uns ein großes Unternehmen oder ein großer Wagnisfinanzierer etliche Millionen zur Verfügung gestellt hätte. 50 Jahre Programmiererfahrung lehrten den älteren der beiden Autoren, dass die ideale Größe eines Entwicklerteams eine Person ist, oder allenfalls zwei, falls einer von ihnen an einen Baum fährt oder etwas ähnlich Unsinniges tut.

Also mussten wir uns die einzelnen Komponenten zusammensuchen, möglichst eingeführte Industriesoftware oder bewährte Standard-Techniken, und die wenigen verfügbaren Mannjahre für den programmtechnischen Kitt dazwischen aufsparen. Das meiste machte auch keine Probleme: die Einbettung in Microsoft-Office und vor allem *Word*, sowie Ben Shneiderman's „TreeMaps" [Sh1998]. Bloß – wo gab es eine für unsere Zwecke geeignete Suchmaschine, für das Internet sowie proprietäre Wissensbasen. Eine, die sich nicht nur „semantisch" nannte, sondern auch wenigstens einige der obigen Grundfunktionalitäten wie die Aufspaltung von zusammengesetzten Worten in die Komponenten, möglichst mit Erkennung von deren Grundformen und die Angabe von Synonymen, wie etwa „Verkauf" oder „Niederlassung" neben „Handlung" in unserem Mercedes-Beispiel? Und das mit einem „Verstehen" von Wirklichkeiten, Sprachspielen und Inbegriffen bei der Auswahl, der Assoziation von Dokumenten in der Datenbasis zu einem eingegebenen Suchtext?

Es gab sie nicht. Und ein entsprechend semantisches Google in vielleicht zwei Jahren neu zu schreiben – das trauten sich die beiden Programmierer unseres Entwicklerteams auch bei aller Selbstüberschätzung nun doch nicht zu. Aber sie waren erfahren genug, Programmierung einfach durch Software-Architektur zu ersetzen:

- einen Präprozessor, der aus dem vorgelegten Text einen wortorientierten Suchausdruck extrahierte, welchen eine verfügbare Suchmaschine – in unserem Fall Microsoft Bing – gerade noch verarbeiten konnte,
- einen Compiler, welcher die zurückgelieferten Fundtexten (ähnlich denen von Google) parste und so gut wie möglich in die Struktur aus Dokumenten, Begriffsbeschreibungen und „Semantische Splitter" übersetzt, die wir schon oben

als „Wissensobjekt“ bezeichneten, und

- eine proprietäre Implementierung von Ben Shneiderman’s TreeMaps, neben dem Compiler für die Bing-Ausgabe die einzige wirklich neu konzipierte und geschriebene Komponente des Gesamtsystems.

Erfreulicherweise reichen die etwa 30 Suchworte die Bing in einer Anfrage verarbeiten kann zusammen mit den gebotenen Möglichkeiten ihrer Verknüpfung durch logische Operatoren (OR, AND, NOT) sowie der eingebauten, wenn auch beschränkten semantischen Unterstützung aus, durch eine entsprechende Aufbereitung der Eingabe und Nachbearbeitung der Ergebnis-Textstruktur aus, eine sicher nicht vollkommene aber in vielen Fällen durchaus brauchbare „Wittgenstein“-Assoziation der Dokumente in der Wissensbasis zu erreichen.

5.3 Begriffsextraktion und -analyse

Der Mechanismus beruht im Wesentlichen auf dem schon oben beschriebenen, mustererkennungs-ähnlichen Phänomen, dass in vielen Fällen einfach das gleichzeitige Vorkommen bestimmter Worte in einem Text ausreicht, ihn und die diesen Worten zugeordneten Begriffe einem Sprachspiel zuzuordnen: in *Wittgensteins* Beispiel waren dies „König“, „Turm“, „Bauer“, „Rochade“, „matt“ oder ähnliche, um das Thema als „Schach“ und den „König“ als Schachfigur zu identifizieren.

Wenn vor allem in sehr großen Dokumentenbasen und durch die Beschränkung auf etwa dreißig Suchworte die vollautomatische Erstellung der Suchanfrage und die Möglichkeiten der Nachbearbeitung durch den Compiler nicht ausreichen, gibt es zwei Wege, die Ergebnisse nachhaltig zu verbessern. Der Benutzer kann durch die Eingabe zusätzlicher Wort-Bezeichner für das „Thema“, seine „Absicht“ oder „unerwünschte Dokumente“ dem System Hinweise für die Suche und deren Aufbereitung geben. Der Präprozessor baut sie in den hauptsächlich aus Worten des vorgegebenen Recherchetexts bestehenden Suchausdruck mit geeigneten logischen Operatoren ein.

Zum anderen erwies es sich als erstaunlich produktiv, einfach die Recherche in wiederholten Schritten zu verfeinern: entweder indem der Benutzer die ihm besonders zusagenden Objekte (Dokumente, Begriffe, Semantische Splitter) in eigene Wissensobjekte sammelt, die dann persistent in seiner persönlichen Wissensbasis gesammelt werden. Oder er kann diese besonders guten Treffer, seien es Dokumente, Begriffe oder Textsplitter, einfach als Ausgangstexte für eine erneute Recherche verwenden. Auf diese Weise reichern sich die semantischen Kriterien zur Auswahl eines bestimmten Sprachspiels immer mehr an, und die nächsten, durch die Suche und den Compiler erzeugten Wissensobjekte entsprechen immer mehr dem Sprachspiel und der aktuellen Wirklichkeit des Benutzers.

Wie schon oben erwähnt, ist dies nur praktikabel, weil die Software für eine gründliche, semantische Recherche eben nicht Tage oder Wochen sondern nur noch Sekunden

braucht. Dies ist der eine, wesentliche Fortschritt der neuen Recherche-Software.

Den zweiten entdeckten wir rein zufällig: weil die linguistische Analyse hier nicht auf einer Syntax-Analyse aufbaut sondern diese nach *Wittgenstein* durch eine „linguistische Mustererkennung“ ersetzt, funktioniert sie unabhängig von der natürlichen Sprache der Such- und Zieldokumente. Wir erprobten dies bereits, ohne jede Änderung der Software, mit deutschen, englischen, spanischen und russischen Eingabe-Texten. Und mit chinesischen, aber dort konnten wir die Qualität der gefundenen Ergebnisse auf Grund unserer ungenügenden sprachlichen Kompetenz nicht nachprüfen …

5.4 Der WisARD als integriertes Software-Hardware-Produkt

Inzwischen gibt es diesen Digitalen Dokumentar nicht nur als Basissoftware zur Implementierung von Anwendungen auf PCs, in lokalen Netzen oder in der Cloud.

Abb. 2 WisARD: der fertig konfigurierte Digitale Dokumentar

Es gibt ihn auch unter dem Namen WisARD (Wissens-Akquisition, Recherche und Darstellung) auf dem Laptop AIQ.Research, fertig installiert vor allem für Studenten, Wissenschaftler und Wissensarbeiter in der Politik oder Wirtschaft.

Und es gibt weiter ausgebaute Systeme, zur konsensorientierten Entscheidungsfindung in Teams oder Diskussionsgruppen von Personen auch mit unterschiedlichen Wirklichkeiten. Oder zur Entscheidungsorientierten Projektplanung und Steuerung nach Hubbard [Hu2014]. Oder zur Einrichtung, der laufenden Fortschreibung und optimalen Nutzung der Qualitätsdokumentation nach der ISO 9001:2015. Aber über diese Applikationen und deren Möglichkeiten werden andere Aufsätze berichten.

Literaturverzeichnis

[Br1851] Brown, G.; Grammar of English Grammars, 1851, Kindle Edition, Amazon Media EU S.a.r.l., 2011.

[Kl1947] Klemperer, V.: LTI – Notizbuch eines Philologen. Reclam Verlag, Leipzig, 1947.

[Re1948] Reichert, H. G.: Lateinische Sentenzen, Essays. Dieterich'sche Verlagsbuchhandlung, Wiesbaden, 1948.

[Sh1998] Shneiderman, B.: Treemaps for space-constrained visualization of hierarchies, 1998/2014, http://www.cs.umd.edu/hcil/treemap-history/.

[Hu2014] Hubbard, Douglas W.: How to Measure Anything, 3rd Edition, Wiley, Hoboken, N.J. 2014.

[TS2016] Teich, I.; Schnupp, P.: Semantical Grammar. AIQ.go, Köln, Kindle Edition 2016.

[La2015a] Lay, R.: Konstruktivismus, Grundlagen der Wissenstechnologie, Band 3, Amazon Kindle, 2015.

[La2015b] Lay, R.: Die zweite Aufklärung – Einführung in den Konstruktivismus, Verlagshaus Monsenstein und Vannerat, Münster 2015.

[Ku1991] Kuhn, Thomas: Die Struktur wissenschaftlicher Revolutionen. Suhrkamp, Frankfurt am Main, 1991.

[Wa2005] Watzlawick, Paul: Wie wirklich ist die Wirklichkeit - Wahn, Täuschung, Verstehen; Piper, München, 2005. Original 1976.