Großer Beleg

zum Thema

Automatische Fließtexterstellung aus Entitätsfakten in einer Wissensbasis

Technische Universität Dresden

Fakultät Informatik

Institut für Systemarchitektur

Lehrstuhl Rechnernetze

Bearbeitet von:
Christian Hensel

Betreut von:
Dipl.-Medien-Inf. David Urbansky

Eingereicht am 14. 10. 2011

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen wörtlich oder sinngemäß übernommenen Gedanken sind als solche kenntlich gemacht. Ich erkläre ferner, dass ich die vorliegende Arbeit an keiner anderen Stelle als Prüfungsarbeit eingereicht habe oder einreichen werde.

Christian Hensel Dresden, den 08. Oktober 2011

Inhalt

1. Einleitung

1.1 Problemstellung

Natürlich klingenden Fließtext aufbauend auf Stichpunkten und Schlüsselwörtern zu schreiben, stellt für ein menschliches Lebewesen meist eine einfach zu bewältigende Aufgabe dar. Ein Mensch ist in der Lage, aus einzelnen gegebenen Wortbausteinen zusammenhängende Sätze zu kreieren, welche sowohl im Hinblick auf Ausdruck als auch grammatikalisch gewissen Ansprüchen genügen, darüber hinaus einen inhaltlichen Zusammenhang aufweisen und sich dem Leser verständlich präsentieren.

All diese Anforderung stellen für Computer eine weitaus größere Hürde dar. Es gibt auf diesem Feld noch große Schwierigkeiten, ähnlich gute Ergebnisse zu liefern, wie sie von Menschenhand erstellt werden können. Der Wunsch ist hier, den Computer mit einzelnen Informationen aus bspw. einer Datenbank zu versorgen und dieser generiert automatisch auf dieser Grundlage gut lesbaren, möglichst einfach zu erfassenden, natürlichen Text. Warum sind heutige Computersysteme dazu nicht ausreichend in der Lage? Haben Programmierer bis heute noch keine angemessenen Algorithmen zur Problembewältigung gefunden oder ist es möglich, dass ein Computer nie dazu in der Lage sein wird?

Textverständnis ergibt sich nicht nur aus den aneinandergereihten einzelnen Wörtern, sondern insbesondere auch aus deren Kontext zueinander. So ist der Satz „Die Suppenschüssel kämpft gegen den Musiker." zwar grammatikalisch korrekt, macht aber semantisch keinen Sinn. Einem Computer derartige Beziehungen klarzumachen, ist eine nicht triviale Problemstellung.

1.2 Motivation

Aus der Praxis kommt ein starker Bedarf nach derartiger Software. Ein konkretes Beispiel sei ein Service zur "Pollenflug Vorhersage in Schottland".[1] Das System nimmt als Eingabe sechs Zahlen, welche den voraussichtlichen Pollenflug klassifizieren und gibt als Ausgabe eine kurzen Bericht zu den angegeben Pollenwerten:

„Grass pollen levels for Friday have increased from the moderate to high levels of yesterday with values of around 6 to 7 across most parts of the country. However, in Northern areas, pollen levels will be moderate with values of 4."

Möchte man auf Basis von einzelnen Wetterdaten automatisch den täglichen Wetterbericht erstellen; möchte man aus einzelnen Zahlen und Statistiken das Wirtschaftswachstum des vergangenen Jahres in Worten ausdrücken oder möchte man die neusten Fußballergebnisse bekanntgeben; es findet sich ein breites Spektrum vieler Anwendungsmöglichkeiten, bei denen automatische Textgenerierung zum Einsatz kommt bzw. kommen könnte. Die genannten Einsatzszenarien haben gemeinsam, dass Daten in knappen Stichpunkten oder in

[1] **http://www.csd.abdn.ac.uk/research/nlg/cgi_bin/pollen.html**

Form von Zahlen oder Fakten vorliegen. Diese Informationen in zusammenhängenden Volltext zu überführen, ist eine gute Möglichkeit sie in übersichtlicher, besser zu erfassender Weise zu präsentieren.

1.3 Zielstellung dieser Belegarbeit

Möchte man sich im Internet konkrete Informationen über eine Person oder eine Sache beschaffen, so stellt man meistens eine Anfrage an eine Suchmaschine. Zurück bekommt man eine Fülle von Informationen unter denen man sich oft mühselig das gewünschte Wissen heraussuchen muss. Zudem kann es vorkommen, dass an mehreren Stellen im Internet veraltete oder sich widersprechende Informationen zu dem jeweiligen Thema stehen können.

Das an der TU-Dresden entwickelte Projekt *WebKnox*[2] nimmt sich derartiger Probleme an, indem es gezielt Informationen über verschiedene Themen aus dem Web sammelt und diese in einer Wissensbasis archiviert. In dieser Wissensbasis finden sich dann konkrete Aussagen wie bspw. „Mount Everest, Höhe 8848m“ oder „Erstbesteigung:1953“.

Die Zielstellung dieser Belegarbeit ist es, eine Anwendung zu entwickeln, die derartige Aussagen und Stichpunkte als Eingabe bekommt und auf diesen Grundlagen automatisch natürlich klingenden Fließtext in englischer Sprache generieren kann. Alle Aussagen über ein Thema sollen in verständlicher, leicht zu erfassender Form präsentiert werden. Der Text sollte inhaltlich nach Absätzen gruppiert werden, er sollte grammatikalisch korrekt sein, einen adäquaten Ausdrucksstil aufweisen und alle enthalten Informationen sollten redundanzfrei sein. Schwerpunkt liegt dabei auf der Extraktion von Textphrasen aus dem Internet, die als Basis der Textgenerierung verwendet werden. Es werden Mittel und Methoden untersucht geeignete Phrasen zu finden und diese inhaltlich nachvollziehbar im Text anzuordnen.[3]

1.4 Forschungsfragen

Aus einzelnen Fakten natürlich klingenden Fließtext zu generieren, ist ein Prozess, der mit einer Vielzahl von Problemstellungen konfrontiert ist.

Wie plant man die Struktur und den Inhalt eines Textes auf Basis des Konzepts und der Fakten einer Entität und deren semantischen Zusammenhangs? Hierbei müssen zwei Fragen geklärt werden. Zum einen wie die vollständigen Satzbausteine gewonnen werden sollen und zum anderen wie man diese anordnet. Ein besonderer Schwerpunkt dieser Arbeit liegt dabei in der Extraktion von Textphrasen aus dem Internet. Aus der Praxis kommen sehr viele Beiträge zu diesem Thema. Es werden daher Ideen und Ansätze miteinander verglichen,

[2] **http://www.inf.tu-dresden.de/index.php?node_id=578&ID=117**
[3] vgl. Aufgabenstellung :**http://www.inf.tu-dresden.de/index.php?node_id=580&arbeit=932&ln=de**

diese Problemstellung anzugehen. Des Weiteren wird ein Konzept zum strukturierten Aufbau des zu erzeugenden Textes vorgestellt.

Wie geht man mit extrahierten Textphrasen weiter vor um den Text aufzubauen? Es müssen Mittel und Methoden untersucht werden, wie man bspw. einzelne Textphrasen syntaktisch miteinander verknüpfen könnte, um längere Satzstrukturen zu erzeugen. Es kann darüber hinaus der Einsatz von Synonymen und Redewendungen in Betracht gezogen werden, um den rhetorischen Stil des Textes zu erhöhen. Des Weiteren müssen grammatikalische Fragestellungen gelöst werden. Darunter richtige Punkt- und Kommasetzung; die Frage, ob Aktiv- oder Passivsatz; oder wo Haupt – und Nebensätze angebracht sind. Darüber hinaus muss gewährleistet werden, dass Fakten im Text nicht mehrfach beschrieben werden.

Wie lassen sich die extrahierten Textphrasen bzw. die erzeugten Texte auf Brauchbarkeit und Güte (grammatikalisch, inhaltlich, rhetorisch…) überprüfen? Hier muss man menschliche und automatische Komponenten einbeziehen und Kriterien anführen nach denen Phrasen und die Gesamttexte bewertet werden können. Daraus sollten Schlussfolgerungen und Verbesserungsvorschläge gewonnen werden.

1.5. Thesen

Zu den meisten Sachverhalten, wie z.B. dem Geburtstag eines Schauspielers, finden sich in Menschenhand gemachten Texten meist gleiche Formulierungen, die diesen Sachverhalt ausdrücken. So ist es denkbar, dass man auf Formulierungen stößt, wie: „*Jim Carrey is a comedian and he was born on 17th January 1962*.“ und “*He was born on 17th January 1962 to Kathleen and Percy Carrey*“ In beiden Sätzen findet man die Wortkette „*he was born on 17th January 1962*“, in der der Fakt “Geburtstag” enthalten ist. Man kann nun in Betracht ziehen diese Formulierung immer dann zu nutzen, wenn man den Geburtstag eines Schauspielers ausdrücken möchte. Man muss nur an der entsprechende Stelle das eigene Datum einfügt.

Es soll daher die Vermutung angestellt werden, dass zu jedem Fakt derartige Formulierungen existieren, und man diese auch finden kann, insofern man genügend viele Texte untereinander abgleicht. Dadurch findet man genügend Textmaterial, mit dem man verallgemeinert einen Fließtext zu bestimmten Klassen von Entitäten (z.B. Schauspielern) aufbauen kann.

Um die Anordnung des Textinhaltes zu planen, kann man mit ähnlicher Strategie vorgehen. Da sich viele Fakten einer Entität in einer semantischen Beziehung befinden (z.B. Geburtstag, Geburtsort, Alter), erscheinen diese in einem natürlichen Text fast immer zusammenhängend. Darüber hinaus findet man oft wiederkehrende globale Textaufbaustrukturen. So sind bspw. Wikipedia Einträge über Schauspieler oft gleichartig aufgebaut. Anfangs findet man meist die Biografie, gefolgt von der Filmografie und abschließend werden Auszeichnungen und Leistungen aufgelistet. Vergleicht man den Aufbau dieses Wikipedia Artikels mit Beiträgen andere Webressourcen, so findet man oft die gleiche oder eine sehr ähnliche Aufteilung.

Daher soll die These aufgestellt werden, dass man einen Textaufbauplan erstellen kann, indem man auch hier genügend menschliche Texte auf ihren Aufbau untersucht. Aus den gewonnenen Erkenntnissen kann man einen allgemeinen Aufbau für ein Konzept generieren und darüber hinaus Rückschlüsse ziehen, welche Fakten in semantischer Beziehung stehen.

Zur Gewährleistung syntaktischer Feinheiten (Punkte, Kommas…) und der Grammatik kann man auf eine der originalen Textphrasen zurückgreifen, aus denen das Template erstellt wurde, in dem man diesem eine Referenz zu Originalquellen beifügt. Eine weitere Möglichkeit ist das Implementieren eines Lexikons, wie es in der Praxis oft zum Einsatz kommt. Dadurch könnte jedem Wort eines Templates die entsprechende Wortart (Substantiv, Verb, Adjektiv…) zugeordnet werden. Analog kann man den Einsatz einer Grammatikbibliothek in Betracht ziehen.

Nach dieser letzten Phase soll der gewünschte natürlich klingende Volltext vorliegen.

2. Grundlagen

Da diese Arbeit für das Projekt *WebKnox* entworfen wird und auf diesem aufbaut, sollen im Vorfeld wichtige Begriffe und Zusammenhänge erläutert werden, die im Laufe der Belegarbeit wiederholt genutzt werden.

2.1. Projekt Webknox

Möchte man sich im Internet über einen bestimmten Sachverhalt informieren (z.B. über berühmte Schauspieler, Länder, Berge etc.), dann sind erste Anlaufpunkte im World Wide Web meist Suchmaschinen oder Online Enzyklopädien. Diese Webangebote liefern umfangreiche Informationen über den gesuchten Sachverhalt; dies allerdings in teils erschlagender Form. Meistens ist man gezwungen eine Fülle Informationen zu durchforsten, bis man das gewünschte Wissen findet, nach dem man gesucht hat. Möchte man sich beispielsweise über die „Export von Kaffee“ informieren, so kann man z.B. den Wikipedia Artikel über „Kaffee“ als erstes Ziel ansteuern. Dort erhält man allerdings erst einmal eine Flut an Informationen über die Kaffeepflanze, über die Geschichte und Entdeckung des Kaffees und über seine Zubereitung, bis man auf die gesuchten Informationen trifft. Alternativ kann man auch eine Suchanfrage an Google stellen, wie „export kaffee“. Die ersten Suchergebnisse (darunter wieder der Wikipedia Artikel) beziehen sich wieder auf Kaffee allgemein und die gewünschten Informationen lassen sich nur mit erhöhtem Aufwand finden. Darüber hinaus wird man mit vielen Werbeangeboten zum Thema Kaffee konfrontiert, was die Suche erschwert.

An dieser Stelle kommt das Projekt Webknox ins Spiel. WebKnox ist die Kurzform für „Web Knowledge eXtraction“. Es handelt sich um ein Projekt, das gezielt Informationen über verschiedene Sachverhalte (z.B. berühmte Schauspieler) aus dem Internet extrahiert und diese in strukturierter Form präsentiert. Der Extraktionsprozess ist dabei zweigeteilt. Zum einen werden Suchanfragen zu einem Thema an unterschiedliche Suchmaschinen (Google, Bing…) gestellt. Die dadurch gewonnenen Ergebnisse werden nach Fakten zu diesem Thema untersucht (z.B. Exportzahlen von Kaffee). Nach diesen Fakten werden erneut Suchanfragen gestellt, um sie zum einen zu erweitern und zu vervollständigen und zum anderen auf Korrektheit zu überprüfen. Erst dann werden sie in die Wissensdatenbank eingetragen. Die dadurch gewonnenen Informationen zu einem Thema sind also nicht einfach nur „Suchergebnisse“ sonder „extrahiertes Wissen“.

2.2. Wichtige Befriffe

Die Personen, Objekte oder Themen nach denen man im Internet sucht (z.B. „Kaffee“, „Jim Carrey“ oder „Mount Everest“) bezeichnet Webknox als **Entitäten**. Eine Entität verfügt über eine Liste **Fakten**. Dies ist die Sammlung extrahierten Wissens über die Entität. Beispiele für Fakten sind „Geburtstag“ oder „Alter“ der Entität „Jim Carrey“. Fakten kapseln mindestens

eine oder mehrere FactValues (Faktwerte). Dies sind konkrete Daten des Faktes. Bspw. sind „17.01.1962“ und „17th January 1962“ FactValues, die dem Fakt „Geburtstag“ zugeordnet sind. Eine Entität hält überdies eine Referenz zu dem, ihr zugeordneten **Konzept**.
Ein Konzept beschreibt mit Hilfe einer Liste von **Attributen**, welche Fakten die von ihr abgeleiteten Entitäten enthalten. Des Weiteren werden Informationen, wie das Datum der letzten Suche oder Synonyme verwaltet. Jede Entität ist genau einem Konzept zugeordnet; aber ein Konzept kann mehrere Entitäten beschreiben. So gehört bspw. das Konzept „Schauspieler“ zu der Entität „Jim Carrey“. Neben „Jim Carrey“ kann es noch weitere Schauspieler Entitäten geben („Bruce Willis“, „Jackie Chan“, „Meg Ryan“ ...).
Die Attribute eines Konzepts bestimmen, welche Fakten die Entität kapselt. Im Konzept Schauspieler enthalten sind z.B. die Attribute „Alter“, „Auszeichnungen“, „Filme“ usw. Dementsprechend findet man Fakten der zugeordneten Entitäten. Die Strategie der Entitäten und Konzepte erinnert an Klassen und Objektinstanzen bei objektorientierter Programmierung.

3. Stand der Technik

3.1. Einordnung in Fachgebiet

Das automatische Generieren von natürlich klingenden Texten wird in der Fachwelt als „Natural Language Generation (NLG)“ bezeichnet. Zusammen mit dem „Natural Language Extraction (NLE), bildet es die Disziplin des „Natural Language Processing(NLP). Dies ist per Definition das Gebiet „der Analyse und Repräsentation von natürlich klingendem Text auf Basis mehrerer linguistischer Anforderung, mit dem Zweck von Menschenhand geschaffene Sprache maschinell zu verarbeiten und zu generieren“. NLP wird in das Gebiet der „Künstlichen Intelligenz“ eingeordnet. [Lid01]

NLE bezieht sich auf die Analyse von Text mit dem Ziel eine geeignete Repräsentation zu finden. Beispielsweise wird NLE eingesetzt, um einen Text in Stichpunkten zusammenzufassen oder um relevante Daten und Fakten zu extrahieren. Die Rolle eines NLE Systems kann als das Lesen und Einordnen / Strukturieren von Text verstanden werden [Lid05]. NLG Systeme sind dementsprechend für Textausgabe / zur Textgenerierung vorgesehen. Beide Ansätzen teilen sich gleiche Methoden und Strategien um ihre Aufgaben zu lösen; NLG Systeme benötigen darüber hinaus meist eine Art Textaufbauplan, mit dessen Hilfe spezifiziert werden kann, welche Informationen im Text erscheinen sollen und an welcher Stelle.[Hov 98]

3.2. Historischer Überblick

Die ersten Forschungen auf dem Gebiet des Natural Language Processing können bis in die 40iger Jahre zurückverfolgt werden. Im zweiten Weltkrieg gab es erste Versuche Geheimcodes und die Sprache der gegnerischen Parteien maschinell übersetzen zu lassen. Ein Pionier auf dem Gebiet der maschinellen Übersetzung war Warren Weaver (1894 – 1978), ein US-Amerikanischer Mathematiker, der mit seinen Memoiren das Interesse der Öffentlichkeit an automatisieren Übersetzen anregte. Seine Ideen waren oft Inspiration für viele spätere Projekte. Er schlug vor, Ansätze aus der Kryptographie und der Informationstheorie miteinander zu kombinieren. Innerhalb weniger Jahre begannen erste Forschungen in den Vereinigten Staaten.

In den späten 50iger Jahren glaubten die Menschen, es werde bald möglich sein, voll automatisch hoch qualitative Übersetzungen für jede Sprache per Computer generieren zu können, die nicht mehr von menschlichen Übersetzungen zu unterscheiden sind. Aufgrund der damaligen Kenntnisse in Linguistik und der noch nicht soweit entwickelten Computersysteme war dieser Wunsch allerdings noch sehr unrealistisch. [Lid 01]

In den 70iger Jahren widmete man sich Problemen wie semantischer Bedeutung eines Textes, Kommunikationsziel und –ablauffluss. Viele Erkenntnisse aus anderen Bereichen wie etwa der Sprachforschung trugen zur Weiterentwicklung auf diesem Gebiet bei. Neben theoretischen Fortschritten wurden in dieser Zeit auch eine Reihe Prototyp-Systeme

entwickelt. Winograd's „SHRDLU“ [Win80] war ein Roboter, der Bauklötze auf einem Tisch erkennen und stapeln konnte. Befehle erhielt dieser Roboter über Spracheingabe; diese Eingabe musste erkannte und analysiert werden, was für damalige Verhältnisse dem Roboter sehr gut gelang. Weizenbaum's „ELIZA“ [ELI66] war ein Software Projekt, das aus den Gesprächen von Psychologen mit ihren Patienten aller wichtigen Fakten extrahieren konnte, und daraus das Krankheitsbild des Patienten ermittelte.

In den 80iger Jahren, als Computer immer leistungsfähiger wurden, kam es zu einem zweiten Aufschwung. Die Öffentlichkeit interessierte sich in zunehmendem Maße für Rechensysteme und der Wunsch nach NLP Software in der Praxis wurde größer. Forscher griffen auf frühere Ansätze zurück und bauten diese weiter aus.

Ab den 90iger Jahren bis heute ist das Gebiet weiter stark vorangeschritten. Dafür verantwortlich ist zum einen die bereits angesprochene Verfügbarkeit immer leistungsfähigerer Rechnersysteme und zum zweiten das Aufkommen des Internets und der damit verbundenen immer größere Menge an digitalen Daten. Viele jüngere Ansätze zeichnen sich durch statistisch bessere Ergebnisse aus und sind in der Lage mit vielen generischen Problemstellungen aus der Praxis umzugehen. [Lid 01]

3.3. Heutiger Stand

Forscher arbeiten an einer Vielzahl neuer oder verbesserter Methoden, um den verschiedensten Problemstellungen zu begegnen. Derartige Methoden sind z.B. „Lexikalische Analyse“, [Lyn99] womit man eine Eingabe in eine Folge logischer zusammengehöriger Einheiten, so genannte Tokens zerlegen kann. Ein weiterer interessanter Aspekt mit dem sich Forscher beschäftigen, ist das automatische „Dazulernen“ [Dan01] von neuen Satzbau- und Ausdrucksmöglichkeiten. Des Weiteren macht man sich Gedankten über „Semantische Analyse und Wortbedeutung“, [Wil03] wodurch Aufgabenstellungen angegangen werden können wie: Synonyme, Frage-Antwort Systeme oder Sprachübersetzung. Interessant ist auch das Gebiet der „Wissensrepräsentation“, um Wissen in Datenbanksystemen formal abzubilden. Die genannten Ansätze sind ein Schritt in Richtung „semantisches Web“. [Lid 01]

Es ist schnell klar geworden, dass ein intelligentes NLG Systeme meist mehrere Stufen der Planung erfordert. Typisch sind u.a.:
Strukturierung: Allgemeines Ordnen aller Information, inhaltliche Zusammenhänge z.B. in Paragraphen zusammenfassen.
Aggregation: Zusammenfassen von ähnlich klingenden Textauszügen, um Lesbarkeit und natürlich Wirken zu verbessern.
Lexikalische Änderungen: Das Verwenden von Synonymen und Redewendungen. [Wik]

Es haben sich eine Vielzahl an Interessengruppen gebildet, die sich intensiv mit derartigen Thematiken beschäftigen und bereits vieles zu diesen Themen beigetragen haben.[4]

So gut viele Ansätze bereits sein mögen; die Qualität eines von Menschenhand geschaffenen Textes können sie leider immer noch nicht zu 100% erreichen. [Lid 05] Dies erklärt sich hauptsächlich, da die menschliche Sprache sehr komplexe Ausmaß annimmt: es existieren Doppeldeutigkeiten, spezielle Redewendungen, regelbrechende Sonderfälle u.v.m.

Alles in allem ist das Fachgebiet des „Natural language Processing“ heute an einem Punkt, an dem seine Nutzen für den Menschen bereits spürbar ist. Entwicklungen auf diesem Gebiet schreiten weiter voran und bisherige Ergebnisse sind ermutigend. [Lid 01]

3.4. Verwandte Arbeiten

Im Laufe der Zeit wurden sehr viele Beiträge und wissenschaftliche Arbeiten zu diesem Thema verfasst. An dieser Stelle sollen jene Arbeiten vorgestellt werden, die wichtige Denkanstöße zur Anfertigung dieser Belegarbeit beigetragen haben. Im abschließenden Abschnitt sollen die daraus gewonnenen Erkenntnisse erläutert werden.

Ein hilfreicher Artikel zum Thema „Natural Language Generation“ stammt von Eduard Hovy [Hov 98]. Darin wird geschildert, in welche Teilphasen das Problem der natürlichen Textgenerierung in Praxis sehr häufig aufgeteilt wird:

Ausgehend von einem bestimmten Kommunikationsziel ist es Aufgabe der ersten Phase, der *Inhalts* – oder auch *Textplanerphase,* zum einen für den Aufbau des Textes benötigtes Textmaterial auszuwählen und dieses zweitens inhaltlich anzuordnen (z.B. nach der Strategie „Einleitung, Hauptteil, Schluss“) . Ein typisches Kommunikationsziel wäre „Schildere Urlaubserlebnis“ oder „Beschreibe Haus“. Davon ausgehend ließe sich bspw. folgender Textplan erstellen: „[Beschreibe Haustür mit Textphrasen 12,14] und [Beschreibe Dach und Schornstein mit Textphrasen 15,13, 9][Beschreibe Fenster mit Textphrasen 1,4,5]“. In diesem Beispiel wird in geordneter Struktur festgelegt, welcher Teil des Hauses wann beschrieben werden soll und mit welchen Textphrasen (Textmaterial). Des Weiteren wird gekennzeichnet, wie einzelne Abschnitte miteinander verbunden werden sollen (z.B. durch ein simples „und“). Um die Textplaner Phase zu realisieren, werden in der Praxis oft [Kat84] Schemata verwendet; eine Art Text „Schablone“, die meist dann eingesetzt werden, falls der Text einem festen Muster folgt, wie etwa bei einem Enzyklopädie Artikel und man im Vorfeld bereits weiß, welche Inhalte im Text erscheinen werden. Schemata legen fest, in welcher Reihenfolge Inhaltsmaterial präsentiert werden sollen (Im oben erwähnten Beispiel etwa: „Erst Haustür, dann Dach, dann Fenster“). Die entsprechenden Stellen müssen dann nur noch mit konkretem Inhalt gefüllt werden. Das Resultat der Textplaner Phase liegt meist in Form einer Baum- oder geordneten Listenstruktur vor.

[4] siehe z.B. **http://www.siggen.org/**

Die zweite Phase ist die sogenannte *Sentence Planer Phase*. [Hov 98] Zu den Aufgaben dieser Phase zählen: Festlegen von Satzaufbau und Verknüpfen mehrerer Sätze (z.B. mit „und“); Austauschen von Wörtern mit Synonymen für besseren rhetorischen Stil; Festlegen von Passiv- oder Aktivsätzen und anderer grammatikalischer Spezifikationen. Das Resultat dieser Phase ist meist eine Liste von Textphrasen, angereichert mit Zusatzinformationen, um z.B. zwei Phrasen miteinander in Beziehung zu bringen. In dem Beispiel zur Beschreibung eines Hauses wäre als Output denkbar: „1: Das Haus mit Hausnummer 12a hat eine weiße Wohnungstür – 8 | 2: das Flachdach des Hauses ziert ein roter Schornstein und – 3 | 3: es hat zwei runde Fenster- EOF“). Am Ende jeder Phrase befindet sich ein Trennsymbol und eine Referenz zur nächsten Textstelle. Syntaktische Feinheiten, wie etwa Groß- und Kleinschreibung oder Kommas fehlen an dieser Stelle noch. Dies ist meist Aufgabe der letzten Phase – der *Sentence Realisation*.

In diesem abschließenden Teil wird dem Text der „letzte Schliff“ gegeben. Wie geschildert, werden Groß – und Kleinschreibung realisiert; Punkte und Kommas werden gesetzt; grammatikalische Korrektheit wird geprüft u.v.m. Dabei bedient man sich oft einer Wissensdatenbank, wie etwa einem Lexikon oder einer Grammatikbibliothek. Am Ende dieser Phase liegt der Text in gewünschter Form vor.

Optional kann man darüber hinaus von einer „Stilistic Control Einheit“ Gebrauch machen. Diese Einheit stellt Mittel und Methoden zur Verfügung während des Generierungsprozesses einen bestimmten rhetorischen Stil aufrechtzuhalten. So kann etwa der Formalitätsgrad des Textes spezifiziert werden. Folgende Abbildung gibt einen Überblick über alle geschilderten Phasen der natürlichen Textgenerierung.

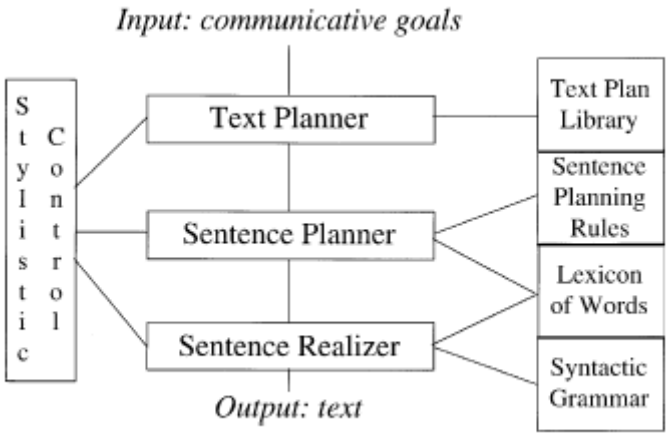


Abb 3.1 – Phasen der natürlichen Textgenerierung, Quelle: [Hov98]

Ein weiterer Artikel [Ehu95], der einen guten Überblick über die Materie schafft, Vor- und Nachteile verschiedener NLG Techniken

Auch hier wird das Problem in die drei geschilderten Phasen aufgeteilt. Zur Analyse und Textplanungsphase bedient man sich einer von Menschenhand angefertigten Vorlage, in Praxis oft „Textkorpus“ genannt. Ein Text ist nicht nur eine zufällig Aneinanderreihung von

Informationen; die Daten befinden sich meist in einer strukturierten Ordnung. Ein einfaches Beispiel ist eine Erzählung oder ein Märchen. Diese bestehen aus Einleitung, Hauptteil, Schluss. Viele Texte sind darüber hinaus sehr viel komplexer strukturiert. Man erhält Des Weiteren Auskunft über den semantischen Zusammenhang von Fakten und Daten. Im Beispiel „Mount Everest" wird man zum Datum der Erstbesteigung oft den Namen des Erstbesteigers finden. Also kann man zwischen diesem Datum und Namen eine semantische Beziehung erkennen und dieses auf andere Entitäten des Konzeptes Berg übertragen.

Es wird darüber hinaus vorgestellt, wie nutzbares Textmaterial gefunden werden kann. Man sucht dabei nach dem Vorkommen eines bestimmten Wortes innerhalb des Textes (z.B. nach dem Wort „Höhe" oder „Meter", wenn man sich für die Höhe eines Berges interessiert). Es haben sich die Begriffe „Entität", „Konzept" und „Relation" ausgeprägt. Zur Veranschaulichung dieser Begriffe ein Beispiel: Der Mount Everest und der Mount Blanc sind konkrete Entitäten. Beide gehören zu dem Konzept „Berg". Diesem Konzept zugeordnet sind Relationen, wie „Höhe" oder „Erstbesteigung", zu denen es je einen konkreten Fakt bei Mount Everest und Mount Blanc gibt und die eine Beziehung ausdrücken zwischen z.B. einem Datum und einem Berg. Diese Strategie erinnert an die Klassen im Projekt WebKnox.

Im weiteren Verlauf des Artikels werden Strategien der Sentence Planer Phase vorgestellt. In diesem Abschnitt werden Methoden aufgelistet, die einzelnen Textphrasen im Text anzuordnen. Dieser als „Discourse Planning" bezeichnete Teilabschnitt realisiert z.B. Textphrasen miteinander zu kombinieren, um komplexere Textabschnitte zu gewinnen.

In der letzten Phasen, der linguistischen Realisation (in der vorherigen Arbeit als Satzrealisation bezeichnet) werden ähnliche Methoden vorgestellt, wie sie im vorherigen Artikel bereits geschildert wurden. Man greift auch hier meist auf Bibliotheken und Datenbanken zurück; untersucht (falls nicht schon in der vorherigen Phase geschehen) einzelne Wörter nach Typ (Substantiv, Verb, Objekt…) und kann meist ein Template anlegen, um die Syntax des Textes zu vervollständigen.

In Artikel [Kat84] wird das Software System „*TEXT*" vorgestellt; das Paragraphen-ähnliche Antworten generiert zu Fragen die sich auf Fakten aus einer Datenbank beziehen(Vgl. Abb 1.). Befinden sich bspw. in der Datenbank Informationen über den Mount Everest (z.B. Lage, Höhe, Datum der Erstbesteigung…), so kann daraus ein Paragraph erstellt werden, wie: "*Der Mount Everest ist mit einer Höhe von 8848 Meter der höchste Berge der Erde. Gelegen im Himalaya war seine Erstbesteigung am 29. Mai 1953…* ". Um den Inhalt der Paragraphen festzulegen, kommt dabei ein Modell zum Einsatz, das von Menschenhand verfasste Texte analysiert und daraus gewonnene Textphrasen zur Erzeugung der eigenen Paragraphen verwendet. Dieses „Text Generation Model" geht dabei in ähnlicher Weise, wie schon in den vorherigen Artikeln geschildert, vor: zum einen die „Strategie Phase", in der Struktur des Paragraphen, Anordnung der Fakten und extrahierte Textphrasen festgelegt werden und zum anderen die „Taktische Phase", die die von der Strategie Phase bereitgestellten Informationen in englische Sätze überführt auf Basis von Grammatik- und Wörterbuch Datenbanken. Phase 1 ist wiederum unterteilt in das Extrahieren von relevanten Informationen für die Antwort.

Dabei werden alle zur Verfügung stehenden Texte aus der Datenbank nach Auftreten der Fakten untersucht und - wenn ja - in einen „Pool von relevantem Wissen“ zusammengefasst. Im zweiten Unterabschnitt wird eine sogenannte rhetorische Strategie ausgewählt, die u.a. ausgehend von Verwendungszweck des Paragraphen und Grad an Formalität, festlegt, welche Aussagen und Satzteile aus dem „relevantem Pool“ geeignet sind. Ein „Focus Mechanismus“ überwacht dabei die ausgewählten Satzteile und zugehörigen Fakten und gibt in Rücksprache mit der rhetorischen Strategie Auskunft über aufeinanderfolgende Fakten, um letztendlich den Aufbau des Paragraphen zu steuern. Nachdem alle Fakten und Satzteile verarbeitet wurden, wird das Resultat an die taktische Komponente weitergegeben. Dort wird jede einzelne Satzgruppe Wort für Wort untersucht und mit Hilfe eines Lexikons werden Substantive, Verben, Adjektive etc. ermittelt. Durch eine Grammatik Datenbank werden die Satzgruppen dann in korrekte englische Sätze überführt.

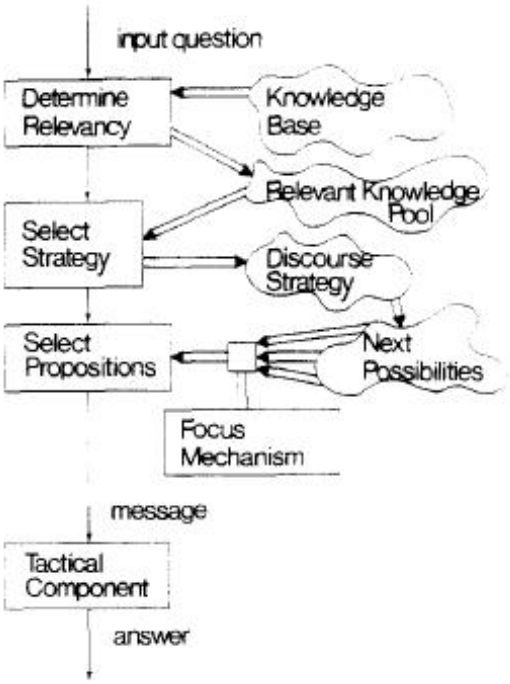
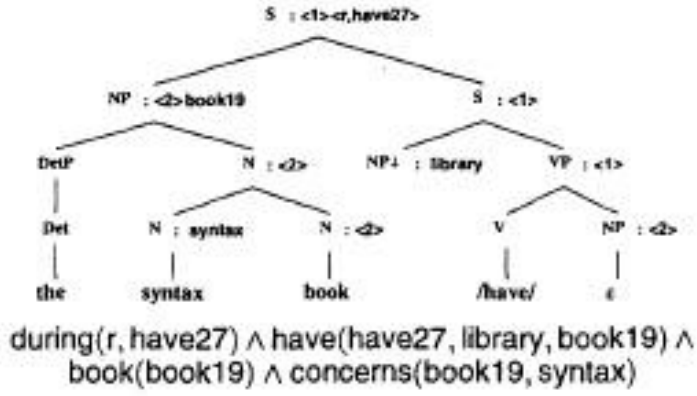


Abb. 3.2 – System „TEXT“ Überblick, Quelle: [Kat84]

Eine vierte Arbeit [Mat98] ist „Sentence Planning as Description Using Tree Adjoining Grammar” von Matthew Stone und Christine Doran (*Referenz).* Hier wird ein Algorithmus vorgestellt, der sowohl Syntax als auch Semantik eines Satzes simultan entwirft und dabei von einer sogenannten „Lexicalized Tree Adjoining Grammar (LTAG)“ Gebrauch macht. Diese Grammatik bietet abstrakte Informationen, wie Wortgruppen syntaktisch kombiniert werden können. Diese Idee wird in dieser Arbeit weiter ausgebaut, indem der LTAG Syntax beschreibenden Metainformationen beigefügt werden, um semantische Zusammenhänge auszudrücken. Der dadurch entstehende sogenannte „Sentence Planner using Descriptions (SPUD)“ wendet nun beschreibende Spezifikationen an, um festzulegen, welche Fakten beschrieben werden sollen und mit Hilfe welcher Informationen.

Jeder einzelne Fakt erhält dabei eine Beschreibung, in der u.a. festgelegt wird, welchen Sinn dieser Fakt verfolgt; bspw. könnte der Fakt „29. Mai 1953“ ein Geburtsdatum sein, oder der Tag der Erstbesteigung des Mount Everest. In der Beschreibung dieses Faktes gibt es überdies sogenannte Operatoren, die einzelne Wörter und Wortgruppen auswählen können, um den Sinn des Faktes auszudrücken. Um Beziehungen zwischen mehreren Fakten auszudrücken, werden Relationen festgelegt (z.B.: „owns(library,book)“ – eine Relation zwischen Bibliothek und Buch). Darüber hinaus wird das Konzept eines „LTAG – Baumes“ genutzt, um komplette Sätze zu kreieren.

Derartige Bäume können nach folgendem Schema erstellt werden: a.) Wähle eine zufällige Relation und erstelle zu jedem dazugehörigen Fakt einen Ast. b.) Finde zu jedem neu entstandenem Blatt weitere Relationen und erstelle weitere Äste. c.) Wiederhole Vorgang bis alle Relation im Baum vorhanden. Diese Baumstrukturen werden darüber hinaus, mit semantischen Informationen angereichert, um inhaltliche Zusammenhänge auszudrücken. Möchte man bspw. den Satz „*Does the Library own the Book for Syntax and Pragmatics?*“ beschreiben, so ergibt sich folgender LTAG-Baum:



during(r, have27) ∧ have(have27, library, book19) ∧
book(book19) ∧ concerns(book19, syntax)

Abb. 3.3 - Beispiel LTAG – Baum, Quelle: [Mat98]

Einzelne Äste können mit logischen Operationen(z.B. und, oder) aneinandergereiht werden. Auf diese Weise kann rekursiv und inkrementell der gesamte Text erstellt werden.

Zum Abschluss soll noch eine Arbeit [Mir03] genannt werden, die mit stochastischen Methoden in der Textplanung experimentiert hat und interessante Ideen und Strategien vorstellt, aus einzelnen Fakten einen kompletten Text zu generieren. Aufbauend auf rhetorischen Strukturbäumen, wie sie in der vorherigen Arbeit bereits zum Einsatz kommen, wird in diesem Beitrag aus einer Sammlung von Fakten und Relationen mit Hilfe von Zufallsalgorithmen eine Liste von Texten erstellt und unter diversen Kriterien als verwerf- oder brauchbar eingestuft. Abbildung 3 zeigt ein Beispiel, in welcher Form Fakten und Relationen vorliegen:

```
fact('this item','is','a figurative jewel',f6).
fact(bleufort,'was','a french designer',f3).
fact(shiltredge,'was','a british designer',f7).
fact('this item','was made by',bleufort,f8).
fact(titanium,'is','a refractory metal',f4).

rel(contrast,f7,f3,[]).
rel(elab,F1,F2,[]) :-
   mentions(F1,0),
   mentions(F2,0),
   \+ F1=F2.
```

Abb.3.4 – Fakten und Relationen Beispiel, Quelle: [Mir03]

Zusammenhängende Wortgruppen innerhalb eines Faktes werden in Hochkommas gestellt und ihre Reihenfolge ist gleichbedeutend mit ihrer Reihenfolge im Satz. Jeder Fakt bekommt eine Bezeichnung (z.B. „f6). Eine Relation erhält einen Namen (z.B. „contrast“), die zwei in Beziehung stehenden Fakten und eine Liste von weiteren Fakten (in diesem Beispiel leer), die im Vorfeld verarbeitet werden mussten, bevor diese Relation benutzt werden kann. Eine Relation kann als einzelner Baustein existieren (Relation „contrast“) oder als Zusammensetzung anderer Relationen definiert werden (Relation „elab“).

Wie im vorherigen Artikel bereits geschildert, werden auf dieser Basis rhetorische Strukturbäume generiert, mit denen dann nach folgendem Ablauf verfahren wird: 1.) Konstruiere eine Sammlung zufälliger rhetorischer Bäume bis ein Zeitlimit erreicht ist. 2.) Klassifiziere Bäume nach gewissen Qualitätskriterien. 3.) Wähle die besten Kandidaten aus. 4.) Nutze diese, um weitere zufällige Kandidaten zu erzeugen. 5.) Füge wiederum die besten Kandidaten hinzu und verdränge schlechtere Varianten. Der Algorithmus kann jederzeit gestoppt werden und es wird der Beste Vertreter ausgegeben, den man bis dahin ermittelt hat. Es kann davon ausgegangen werden; je größer die genehmigte Rechenzeit, desto größer die Güte des besten Resultats. Es muss an dieser Stelle noch geklärt werden, wie genau die Güte eines Resultats bestimmt wird. Ein Resultat liegt in Form eines rhetorischen Baumes vor, aus dem der Text erstellt werden kann. Diesem Baum wird eine Zahl („Score“) zur Klassifikation der Güte zugeordnet, die sich als Summe der Scores einzelner Gütekriterien zusammensetzt. Beispiele für Kriterien sind: 1.)Textinhalt – Es wird ermittelt, ob alle gewünschten Fakten und Relationen im Text vorhanden sind und ob der Aufbau dem eines von Menschenhand geschriebenen Textes ähnelt. Es werden hier Scores vergeben zwischen -30 bis +20 Punkte. 2.) Textumfang: Je größer die Anzahl an Wörtern, um eine Relation und zugehörige Fakten zu beschreiben, desto schlechter ist das Verständnis des Textes. Die Scores liegen zwischen -5 bis +5 Punkte. 3.) Fakteninhalt – Sollte ein Fakt eine Entität beschreiben, die bereits in vorhergehenden Fakten genannt wurde, dann werden -3 Punkte abgezogen bzw. es werden +3 Punkte vergeben, falls eine neue Entität auftaucht. Neben diesen Kriterien können noch weitere Anforderung untersucht werden.

Wie bereits erwähnt, nutzt man die besten Vertreter, um neue Kandidaten zu gewinnen. Dies geschieht, indem Teilbäume eines guten Kandidaten ausgetauscht werden und der dadurch erzeugte Baum erneut auf Gütekriterien untersucht wird. Dieses Tauschen von Teilbäumen geschieht, indem alle Fakten und Relationen innerhalb eines Astes des Teilbaumes als Sequenz niedergeschrieben werden (siehe Abb.4). Es wird dann ein Fakt aus dieser Sequenz gewählt und an beliebiger Stelle neu eingeführt. Dadurch ergibt sich ein neuer Ast und die

Summe aller neuen Äste einen neuen Teilbaum. Darüber hinaus besteht die Möglichkeit, Fakten zweier Äste miteinander zu tauschen.

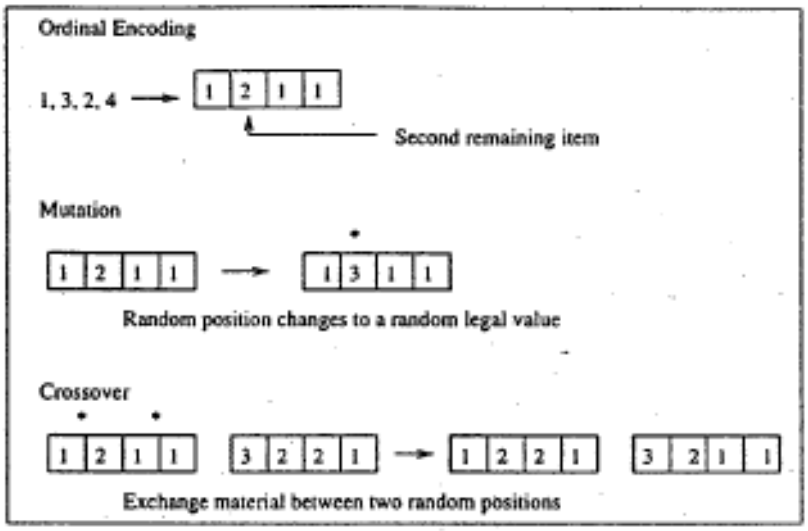


Abb. 3.5 - Fakten Sequenz und darauf angewendete Operationen,
Quelle: [Mir03]

3.5. Erkenntnisse aus vorgestellten Arbeiten

Stellt man die genannten Arbeiten gegenüber, lassen sich sehr viele Gemeinsamkeiten, aber auch teils große Unterschiede feststellen. Festhalten kann man in jedem Falle, dass jedes System zur automatischen Textgenerierung den Zweck hat ein bestimmtes Kommunikationsziel in natürlich klingender Form auszudrücken. Als Eingabe liegt dabei meist ein Bestand an Daten (Fakten) vor, der in Textform gebracht werden (z.B. möchte das Projekt WebKnox einem Leser die Fakten einer bestimmten Entität in intuitiv verständlicher Form präsentieren).

Eine Gemeinsamkeit, die sich nicht nur in den geschilderten Arbeiten, sondern in vielen anderen Textgenerierungssystemen wiederfinden lässt, ist die Unterteilung des Problems in einfachere Teilabschnitte. In Hinblick auf Textgenerierung werden Schritt für Schritt zunächst Grobstrukturen (Textaufbauplan), und dann immer feinere Detailstufen angewendet (erst Satzbau bestimmen, am Schluss letzte Details einfügen).

Man beginnt üblicherweise in der ersten Phase mit dem Planen des Textinhaltes. Man ermittelt dabei zum einen welche Inhalte in den Text eingefügt werden sollen und zum anderen in welcher Reihenfolge, bzw. an welcher Stelle. Dieses Problem wird sowohl häufig in der Praxis als auch in den vorgestellten Arbeiten mit Hilfe von vorgefertigten Wissensstrukturen gelöst. Der Aufbau des Textes ist mit Hilfe eines Plans festgelegt, der entweder von Hand oder durch das Analysieren inhaltlich ähnlicher Texte erstellt wurde. Die gleiche Strategie kommt auch bei dem Aufbau der Textphrasen zum Einsatz. Entweder diese Phrasen liegen bereits in einer Datenbank vor oder befinden sich in anderen Strukturen (z.B. Relationen) gekapselt. Man erkennt also, dass im Vorfeld bereits ein Wissensbestand angelegt

wurde (per Hand oder maschinell) und dass in jedem Falle immer eine menschliche Komponente zur Erzeugung dieses Wissens im Spiel ist.

Die nächste Phase erhält als Eingabe eine geordnete Struktur (z.B. Sequenz oder Baum), auf deren Basis sie mit dem Aufbau einzelner Abschnitte (Sätze, Phrasen) beginnt. An dieser Stelle geben Ansätze aus der Praxis viele teils sehr unterschiedliche Methoden an, wie sie auf ihre Weise die Satzplanungsstufe bewältigen (Relationen verknüpfen, LTAG Bäume manipulieren, Templates nutzen usw.). Wie bereits dargestellt, haben alle Ansätze ihre Vor- und Nachteile, bieten aber je nach individuellem Szenario einen Lösungsweg zu diesem Abschnitt der Textgenerierung. Man kann also die Schlussfolgerung ziehen, dass sich die Strategie, die man für das eigene Projekt in dieser Phase entwirft, stark nach der eigenen Ausgangslage, dem Kommunikationsziel und der erwünschten Endform des Textes richtet (z.B.: WebKnox will keinen komplexen und langen Text zu einer Entität generiert haben, sondern einen übersichtlichen intuitiv verständlichen Text, in dem alle Fakten geschildert werden -> „So viel wie nötig, so wenig wie möglich.“). Am Schluss dieser Phase hat man dann entweder bereits den fertigen Text vorliegen oder es müssen nur noch letze syntaktische Änderungen (Kommas, Punkte…) gemacht werden.

Der Einsatz der dritten Stufe hat vor allen den Vorteil, dass man im Abschluss erneut maschinell prüfen kann, ob der Text in gewünschter Form vorliegt und ob bspw. die Grammatik eingehalten wurde.

Die geschilderten Arbeiten bieten einen guten Überblick über Methoden zur Erzeugung natürlich klingender Texte. Wie im nächsten Kapitel noch näher ausgeführt, wird sich diese Belegarbeit ebenfalls an der Strategie orientieren, das Problem in die drei genannten Teilphasen zu zerlegen, wobei der Schwerpunkt auf der Textplaner Phase und speziell der Gewinnung von Textmaterial liegen wird. Des Weiteren wird geklärt, welche geschilderten Methoden zur Durchführung der jeweiligen Phase in Hinblick auf das Projekt WebKnox geeignet sind. Im Rahmen einer Evaluierung werden die gewonnenen Resultate untersucht und im abschließenden Fazit ein Ausblick zur Verbesserung der geschilderten Strategien und offenen Problemen gegeben.

4. Konzept

4.1. Vorüberlegungen

Wie bereits in den verwandten Arbeiten ersichtlich, ist der Prozess der natürlich klingenden Fließtexterstellung ein komplexer Vorgang, der sich in mehrere Teilprobleme aufgliedert. Um adäquate Ergebnisse zu erzielen, ist daher im Vorfeld ein gewisses Maß an Vorüberlegungen und Planung anzustellen. Es gibt für jedes Teilproblem diverse Lösungsansätze, die in Hinblick auf die eigenen Anforderungen auf ihre Vor- und Nachteile untersucht werden sollen.

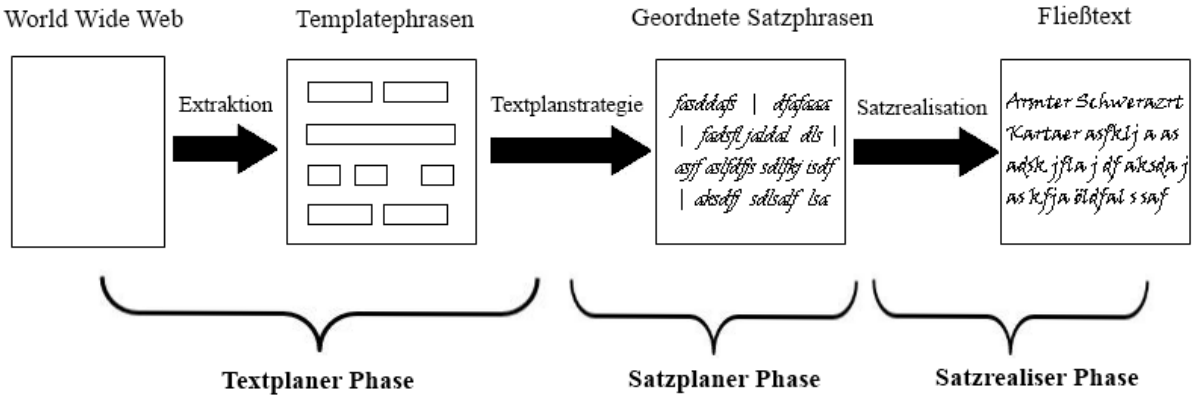


Abb. 4.1 - Projektplan

4.1.1 Bereitstellen von Textmaterial

An einem gewissen Punkt der Textgenerierung müssen die vorhanden Fakten in komplette Sätze oder zumindest Satzphrasen eingesetzt werden. Das Problem ist die Auswahl dieser Textphrasen und insbesondere deren Wortinhalt. Zum einen muss der Sinn ausgedrückt werden, den man mit den Fakten darstellen will (Beispiel Geburtsort) und zum anderen sollten die Sätze einen natürlich klingenden Stil aufweisen.

Wie bereits angesprochen, soll das Konzept der *Templates* zum Einsatz kommen. Dies sind einfache Satzschablonen, in die dann an die passende Stelle jeweils der Fakt eingefügt werden kann. (Beispiele: *„[Name] is [Age] years old.*“ oder „[*Name] celebrates on [Birthdate] his [Age]th Birthday.*“) Dieser Bestand kann entweder von Hand angefertigt werden, was bei größeren Texten schnell sehr aufwendig werden kann oder man extrahiert Textphrasen aus anderen Texten. Dabei untersucht man im einfachsten Falle einzelne Sätze nach dem Auftreten bestimmter Schlüsselwörter (bspw. einem Fakt) und nimmt bei Erfolg einen Satz in den Bestand der eigenen Textphrasen auf. Templates können einem oder mehreren Fakten zugeordnet werden.

Deren Vorteil ist vor allen Dingen die Simplizität in Hinblick auf Implementierung und da diese Textphrasen ursprünglich von einem Menschen formuliert wurden, klingen diese natürlich. Außerdem kann man jederzeit entweder per Hand oder mittels automatischen

Analysierens von vorhandenen Texten neue Templates für eine Gruppe von Fakten finden. Dadurch stehen mehr Formulierungen zur Verfügung und generierte Texte sind abwechslungsreicher. Außerdem lassen sich jedem Konstrukt diverse Eigenschaften zuordnen, wie etwa Grad der Formalität.

Nachteilig an diesem Verfahren ist, dass man nicht immer gewährleisten kann, dass eine gefundene Textphrase genau den semantischen Zusammenhang ausdrückt, den man wiedergeben will. Möchte man bspw. eine Angabe zum Geburtsort eines Schauspielers machen, so durchsucht man relevante Texte nach dem Vorkommen dieses Ortsnamens. In den gefundenen Phrasen befindet sich zwar immer der Name des Ortes, allerdings kann man keine Aussage treffen, ob dieser Name gerade als Geburtsort genannt wird oder in einem gänzlich anderen Zusammenhang. Um derartige Probleme zu umgehen und relevante von nicht relevanten Textphrasen zu trennen, muss erhöhter Aufwand betrieben werden. Eine Möglichkeit wäre, eine große Anzahl Texte zu durchsuchen und gefundene Textphrasen untereinander abzugleichen. Sollten bestimmte Wortgruppen oder Formulierungen mehrmals im Zusammenhang mit einem oder bessere noch mehreren Schlüsselwörtern auftreten, so ist die Chance erhöht, dass diese Wortgruppen genau den gewünschten Sinn ausdrücken. Diese Strategie setzt allerding voraus, dass ein genügend großer Bestand an Texten zur Verfügung steht, die man analysieren kann.

Eine Alternative zur Textphrasenextraktion wäre ein voll automatisches Generierungsverfahren. Der Aufbau eines Satzes wird von der Grammatik der jeweiligen Sprache bestimmt. Eine Grammatik legt fest, an welcher Stelle welche Wortart steht bzw. stehen kann. So beginnt ein Satz bspw. oft mit einem Artikel, gefolgt von Substantiv, Verb, Adjektiv. Ein Generierungsalgorithmus könnte so um Fakten herum einen Satz aufbauen. Dies könnte realisiert werden, indem jedes Wort eines Faktes mit bspw. einer Lexikondatenbank auf die jeweilige Wortart überprüft wird. Dadurch erhält man überdies Auskunft über mögliche folgende Wortarten (bspw. kommt hinter einem Artikel immer ein Substantiv).

Man ist allerdings mit Problemen konfrontiert, wie etwa der Wahl, ob männlicher, weiblicher oder sächlicher Artikel und weiterer grammatikalischer Probleme. Eine zusätzliche Grammatikbibliothek könnte hier zum Einsatz kommen. Nachdem die entsprechende Wortart an einer bestimmten Stelle eines Satzes festgelegt wurde, ist aber immer noch nicht eindeutig die Wahl eines konkreten Wortes an dieser Stelle bestimmt. Möchte man mit dieser Methode einen Satz erstellen, der wie im vorherigen Beispiel ausdrückt, an welchem Ort ein Schauspieler geboren ist; so wird der Algorithmus den Ortsnamen als Substantiv klassifizieren. Es wird außerdem ermittelt, dass auf den Ortsnamen ein Verb folgen soll. Hier kommt das geschilderte Problem zum Tragen. Woher soll der Algorithmus wissen, welches Verb genau an dieser Stelle passend ist? Man könnte allgemeine Verben verwenden wie „hat“ oder „ist“ („Der Geburtsort von Jim Carrey *ist* Montreal“). Unter diesen Verben muss man dann aber immer noch das Passende ermitteln. Wenn man außerdem nur eine derartig begrenzte Menge Wörter zu einer Wortart verwendet, leidet der rhetorische Stil des Textes und er klingt insbesondere nicht mehr natürlich.

Eine andere Alternative wäre, dass man selbstständig spezifiziert, zu welchem Fakt, welche Wörter passen und der Algorithmus muss nur noch auf die richtige Grammatik achten (z.B. Gibt man an, dass die Wörter „geboren sein“ zu „Geburtsort“ passen). Aber auch das kann sehr aufwendig werden. Daher ist der große Nachteil dieses Verfahrens, dass die generierten Texte entweder sehr eintönig klingen und meist nicht den natürlichen Charakter haben, wie es mit Textphrasenextraktion möglich ist; oder dass man diesen Umstand mit eigenem Aufwand ausgleichen muss; was man ja allerdings durch das komplett automatische Generieren der Sätze vermeiden wollte.

4.1.2 Ansätze zur Textplanung

Um den Aufbau des Textes festlegen nutzt man, wie bereits erläutert, oft vorgefertigte Vorlagen. Dabei verwendet man entweder selbst erstellte Textaufbaupläne (Texttemplates) oder man analysiert andere von Menschenhand angefertigte Texte (Textkorpora). Gemeinsamkeit beider Verfahren ist die menschliche Komponente, an der sich der Textaufbau orientiert. Dies kann zu natürlicher klingenden Resultaten beitragen und eignet sich insbesondere bei solchen Texten, deren Aufbau stets nach gleichem Muster erfolgt (z.B. Enzyklopädie Einträge). Als Nachteil kann man anführen, dass das Erstellen des Textes nicht vollautomatisch durchgeführt wird, sondern man immer noch von menschlichen Komponenten abhängig ist. Außerdem lässt sich diese Strategie nur dann anwenden, wenn im Vorfeld bekannt ist, aus welchen Fakten/ welchem Inhalt der Text bestehen wird.

Eine weitere Möglichkeit sind *Relationen*. Diese vermitteln Beziehungen zwischen Wörtern, Sätzen und weiteren Relationen. Eine Relation definiert zum einen abstrakt welche Aussage mit ihr ausgedrückt werden soll (z.B. „Beschreibe Haustür“), zum anderen welche Fakten und/oder Textphrasen benutzt werden. Des Weiteren lässt sich beschreiben, in welchem Kontext diese Relation zu anderen steht; was etwa nachfolgende Relationen sein können. Sie lassen sich mit logischen Operationen verknüpfen und z.B. als Sequenz- oder Baumstruktur anordnen (Beispiel LTAG-Baum). Jede einzelne Relation ist ein kleiner Baustein, die in der Summe den Aufbau des Textes definieren.

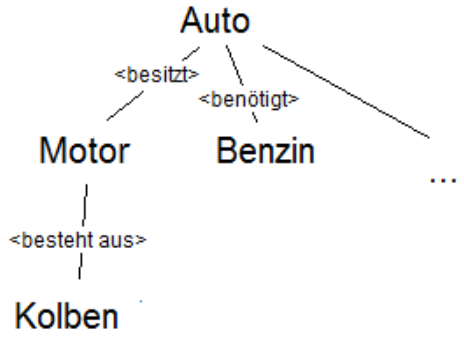


Abb. 4.2 – Entitätsbaum mit Relationen

Durch das Austauschen, Hinzufügen oder Verwerfen einzelner Bausteine lässt sich ein Text sehr viel dynamischer erzeugen, auch in Hinblick auf einen unbekannten Bestand an

Eingangsdaten oder -fakten. Außerdem lässt sich z.B. mittels Zufallsverfahren bestimmen, welcher Teilast eines Relationenbaums zuerst durchlaufen werden soll oder es lassen sich einzelne Relationengruppen (entspricht z.B. Teiläste im Baum) austauschen. Dadurch erzeugte Texte folgen einem inhaltlichen Sinn, der durch die Relationen gewährleistet wird. Sie sind darüber hinaus in ihrem Aufbau abwechslungsreicher, was z.B. der Lesbarkeit zu Gute kommt. Als Nachteil könnte man mit erhöhtem Verwaltungs- und Implementierungsaufwand argumentieren.

Man hat nun die Möglichkeit entweder jedem einzelnen Konzept, auf dem eine Entität beruht, derartige Relationen mit ein zu implementieren oder aber man fertigt unabhängig von dem jeweiligen Konzept eine Menge von Relationen an, aus denen sich jedes Konzept lediglich die passenden Relationen herausgreift, da gewisse Relationen, wie etwa „besitzt“ öfters Verwendung finden.

Großer Vorteil dieses Verfahren ist der Umstand, dass bei dem Projekt Webknox eine Entität und ihre Fakten bereits in Baumstruktur vorlegen und Relationen dann nur noch ergänzt werden müssten. Dies wäre einfach zu implementieren und Des Weiteren sind auch hier Möglichkeiten der Erweiterbarkeit einfach zu realisieren. Einer Relation kann man jeder Zeit passende Textphrasen hinzufügen, die diese Relation ausdrücken. Beispielsweise kann man der Relation „besitzt“ neben bereits bestehenden Templates wie „[Entität] besitzt [Fakt]“ noch weitere Bausteine wie „[Entität] verfügt über [Fakt]“. Als Eingabe erhält man also mit diesem Verfahren einen Webknox Entitätenbaum und als Ausgabe einen Entitätenbaum mit zugefügten Relationen zwischen den einzelnen Knoten. Nachteil dieses Verfahren ist allerdings immer noch ein fehlender Zusammenhang zwischen mehreren Relationen untereinander, um semantische Beziehungen im Gesamttext darzustellen, da eine Relation immer nur eine Beziehung zwischen zwei Fakten bzw. einer Entität und einem Fakt herstellt. Man könnte dieses Problem angehen, in dem zu einer Relation ein Verweis inklusiver passender Textphrasen zur nächsten Relation geknüpft wird, was allerdings schnell sehr aufwendig und unübersichtlich bei größeren Entität Bäumen werden kann.

Ein zweiter Ansatz ist das *Attributemapping*. In diesem Verfahren werden semantisch zueinander gehörige Attribute (Bspw. Geburtstag, Geburtsort, Alter) in Gruppen zusammengefasst, wobei jede Gruppe genau einem Index, mindestens ein Attribut, und optional einen oder mehrere Verweise auf die jeweils folgenden Attributgruppen kapselt. Das festlegen derartiger Verweise geschieht entweder manuell oder durch Analyse menschlicher Texte mit gleichem Inhalt.

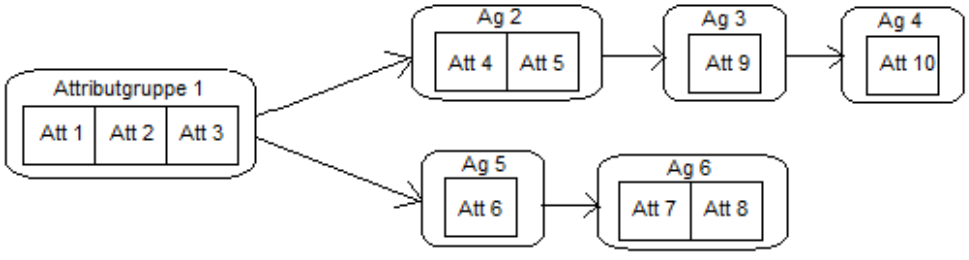


Abb. 4.3 - Beispiel Attributemapping

Es ergeben sich Pfade entlang der einzelnen Gruppen, wodurch zum einen rhetorische und semantische Beziehungen zwischen diesen Gruppen ausgedrückt werden kann und zum anderen der Gesamtaufbau des Textes definiert wird.

Vorteile dieses Verfahrens sind, dass es nun zum einen möglich ist für derartig zusammenhängende Attribute komplexere Textphrasen angeben zu können, die in semantischer Beziehung stehen und könnte man für die Verweise auf weitere Gruppen ebenfalls einen Satzbaustein verwenden. Falls eine Attributegruppe über mehrere Verweise zu nächst möglichen Gruppen verfügt, kann dann z.B. mittels Zufallsverfahren bestimmt werden, welcher Gruppenpfad zuerst verfolgt werden soll, wodurch die Textgenerierung dynamischer wird. Ein eventueller Nachteil dieses Verfahrens ist die Erweiterbarkeit von Textphrasen. Da die Satzbausteine durch mehrere Fakten bestimmt werden und damit komplexer werden können, ist es schwerer automatisch durch Lernverfahren weitere derartige Bausteine hinzuzufügen.

4.1.3 Ansätze für Satzplanung

Wenn einmal ein Aufbauplan und Textphrasen vorliegen, in denen die gewünschten Fakten adäquat ausgedrückt werden, kann man nun den Text Phrase für Phrase aufbauen. An dieser Stelle muss man sich über Probleme wie bspw. die lexikalische Anordnung der Phrasen Gedanken machen. Eine einfache Lösung wäre, jede Phrase für sich als einen Satz auszurichten. Die dadurch gewonnen Resultate klingen aber meist sehr abgehackt und nicht natürlich (Beispiel: „Das Haus hat eine Tür. Die Tür ist weiß. Die Tür hat einen Briefschlitz.“).

Zur Erhöhung des rhetorischen Stils kann man auf die Textphrasen eine Menge bestimmter Operationen anwenden, um natürlich klingendere Resultate zu erzeugen.

In der Praxis wird dieses Teilgebiet als „Discourse Planning“ bezeichnet [Mic98] und es werden sehr viele Vorschläge gemacht, die einzelnen Textphrasen zu komplexeren Textabschnitten zusammenzufügen. Zum einen Satzaggregation; die Methode um zwei Sätze oder Satzteile miteinander zu verbinden. Es werden verschiedene Möglichkeiten zur Durchführung skizziert; über das einfache Verknüpfen zweier Sätze mit einem „und“, zu komplexeren Verfahren zur Erzeugung von Haupt- und Nebensätzen (Beispiel: „die ersten auf dem Gipfel waren Edmund Hillary und Tenzing Norgay“. „Edmund Hillary ist ein neuseeländischer Bergsteiger.“ Diese zwei Sätze können zusammengefasst werden zu: „Die ersten auf dem Gipfel waren Edmund Hillary, ein neuseeländischer Bergsteiger, und Tenzing Norgay.“).

Eine weitere interessante Strategie ist Lexikalisation. Diese Methode befasst sich mit Problemen der Synonymfindung oder dem Einsatz von Redewendungen. Meist greift man hier auf eine Datenbank von bereits vorhandenen Ausdrucksmöglichkeiten zurück und sucht nach Wörtern oder Wortgruppen im Satz, mit denen diese Ausdrücke ersetzt werden können. Dies dient vor allem der rhetorischen Qualität des Textes.

Eine dritte Strategie wird als Satzbeziehungsplanung bezeichnet. Dabei wird analysiert, welcher Inhalt als Nachfolge für einen bereits vorhandenen Satz geeignet ist. Beispielsweise bietet es sich an, die Fauna des Mount Everest zu beschreiben, nachdem die Flora beschrieben wurde. Dies steht eng im Zusammenhang mit der Textstrukturplanung, bietet darüber hinaus allerdings auch Vorteile, grammatikalische Beziehungen über mehrere Sätze auszudrücken (Beispiel: „*Moosgräser* wachsen vor allen Dingen an den Hängen der Berge. *Sie* sind Nahrung für die Bergziegen der Region.“).

4.1.4 Ansätze für Satzrealisierung

Wie geschildert, werden Groß – und Kleinschreibung, Punkte, Kommas und weitere grammatikalische Korrektheiten überprüft. Extrahierte Textphrasen haben an dieser Stelle den Vorteil, dass derartige Syntax meist schon vorhanden ist und lediglich überprüft werden muss. Dabei kommt ebenfalls oft eine Wissensdatenbank zum Einsatz, wie etwa ein Lexikon oder einer Grammatikbibliothek, um z.B. zu ermitteln an welcher Stelle ein Komma gesetzt werden muss. Nachdem diese Phase abgeschlossen wurde, liegt der Text in gewünschter Form vor.

4.2 Geplanter Ablauf

Der Aufbau der Arbeit wird sich stark an den drei vorgestellten Phasen orientieren (Textplanung, Satzplanung, Satzrealisation), wie sie in der Praxis zum Einsatz kommen, wobei der Schwerpunkt auf der Gewinnung von Textmaterial beruhen wird.
Die Ausgangssituation ist ein großer Bestand Entitäten samt Fakten, die einem Konzept zugeordnet sind. Das Resultat ist ein automatisch generierter Fließtext, der zum einen alle Fakten enthält und deren semantischen Sinn wiedergeben kann und zum anderen so wirkt, als wäre er von Menschenhand erstellt worden. Abbildung 4.4 zeigt den schlussendlichen Gesamtvorgang der Texterstellung. Im weiteren Verlauf sollen die einzelnen Teilabschnitte näher erläutert werden.

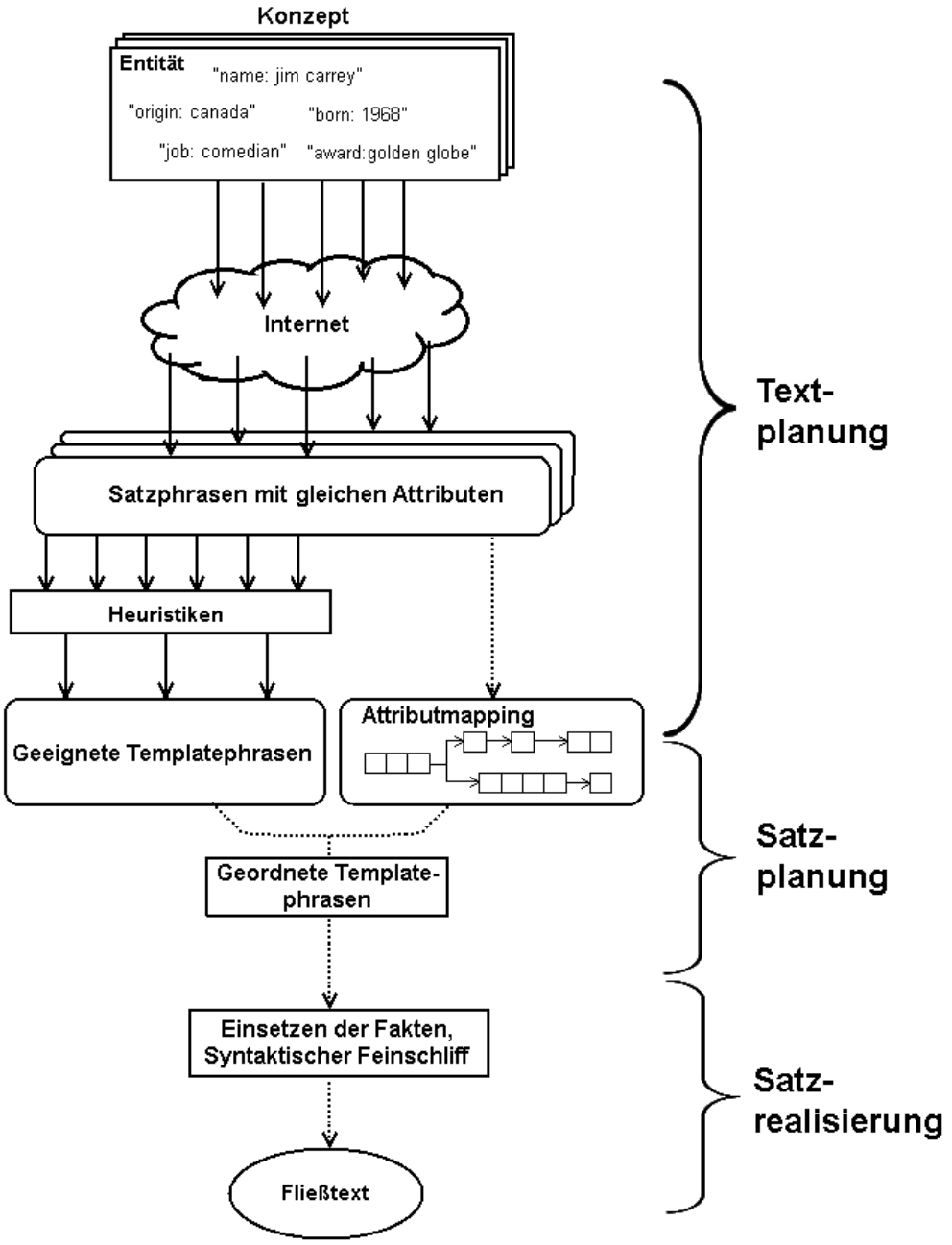


Abb 4.4. - Projektplan

4.2.1 Template Extraktion

Der erste wichtige Schritt zur Textgenerierung ist das Bereitstellen von natürlich klingenden Textphrasen. Da das Projekt WebKnox darauf ausgelegt ist, Wissen aus dem Internet zu extrahieren, liegt es sehr nahe ähnliche Methoden bei der Suche nach Textmustern zu verwenden.

Es soll die Strategie zum Einsatz kommen, passende Textphrasen aus dem Internet zu extrahieren. „Passende Textphrasen" bedeutet hierbei konzeptbezogene Satzmuster, in die an den richtigen Stellen ein konkreter Fakt einer Entität eingefügt werden kann. Folgendes Beispiel soll dies näher erläutern: *„[FullName] was born on [Birthday] in a town called [Birthplace]*". Dieses Beispiel kann als Satzmuster für das Konzept „Schauspieler" interpretiert werden, wobei mit dieser Textphrase die drei Attribute „FullName", „Birthday" und „Birthplace" ausgedrückt werden. Eine Entität vom Typ Schauspieler kann seine konkreten Fakten an diesen Stellen einfügen, um einen vollwertigen Satz zu erzeugen.

Zuerst müssen natürlich derartige Satzmuster gewonnen werden. Wie man dem Plandiagramm entnehmen kann, geschieht dies in mehreren Teilphasen. Als Ausgangspunkt dient eine Menge Entitäten samt Fakten, die dem gleichen Konzept und seiner Attribute zugeordnet sind. Mit Hilfe dieses Grundwissens stellt man unterstützt durch WebKnox pro Entität mehrere Suchanfragen mit dem Namen der Entität an das Internet. Man erhält eine Anzahl Webseiten, die auf die jeweiligen Entitäten bezogen sind. Aus diesen Inhalten extrahiert man sämtliche Sätze, in denen einen Fakt der Entität vorkommt. Da jedem Fakt ein Attribut des Konzepts zugeordnet ist, kann man das entsprechende Vorkommen des Faktes mit dem Attributnamen ersetzen, um so bereist erste Verallgemeinerungen zu erzielen. Beispiele für extrahierte Sätze zum Konzept Schauspieler sind:

- *[name] has won [award] awards and has also been nominated for numerous other awards.*
- *fox and [name] ([award] for " the truman-show ")*
- *[Birthplace] is a town in the heart of kansas*
- *He is a [origin]-american [job], [job], screenwriter, musician, winemaker and ufologist.*
- *introducing funny-man [name].*
- *and simply a fantastic web resource of all related to [name].*
- *links to all media articles covering [name] including all interviews and award coverage.*
- *[name] news and life.*
- *weekly polls on [name] the website has been recently launched and many of the features mentionned above will be gradually released online*

Dieses Beispiel verdeutlicht, dass gewonnene Textphrasen zum einen noch sehr entitäts-spezifisch sind und sich nicht konzeptübergreifend einsetzen lassen. Des Weiteren findet man sehr viele Textphrasen, die nicht den gewünschten semantischen Sinn ausdrücken, den man

dem jeweiligen Fakt zuordnen möchte (z.B. *„[Birthplace] is a town in the heart of kansas*“ drückt nicht aus, dass der Schauspieler in diesem Ort geboren wurde).

Es sind weitere Schritte notwendig, um eine Textphrase für ein Konzept nutzbar zu machen. An dieser Stelle macht man sich zu Nutze, dass mehrere Entitäten vorliegen und diese alle dem gleichen Konzept angehören. Also hat jede Entität Fakten zu je einem Attribut. Durch die extrahierten Textphrasen wird immer mindestens ein Attribut oder oft auch mehrere Attribute ausgedrückt. Man kann nun Textphrasen einer Entität, die eine bestimmte Kombination von Attributen enthalten, mit all den Textphrasen anderer Entitäten vergleichen, die die gleiche Kombination Attribute enthalten. Vergleichen bedeutet, man sucht nach dem Vorkommen gleicher Wortketten in zwei Textphrasen. Wichtig ist; man sollte Textphrasen mit gleichen Attributen nicht miteinander vergleichen, wenn sie bei der gleiche Entität extrahiert wurden. Denn dadurch geht man das Risiko ein, in den Templatephrasen auf Formulierungen zu stoßen, die immer noch zu entitätsspezifisch sind. Nur wenn ein Vorkommen einer Zeichenkette bei zwei Textphrasen auftritt, die von unterschiedlichen Entitäten stammen, kann diese Textphrase als Templatekandidat in Frage kommen.

Dies kann in verschiedener Weise realisiert werden. Zum einen könnte man nach dem längsten gemeinsamen Teilstring zweier Textphrasen suchen. Zur Veranschaulichung sei das Konzept Schauspieler gegeben und dazu zwei Entitäten „Jim Carrey“ und „Adam Sandler“. Aus dem Internet hat man für die Attribute „FullName“, „Birthday“, „Birthplace“ und „Job“ folgende Textphrasen extrahiert:

Jim Carrey:

- *[fullname] was born in [birthplace] on [birthday] and is a well known [job] and [job]*
- *[Job] [fullname] was born on [birthday] in ontario, [birthplace].*
- *born:[birthday] in [birthplace], biography:arguably the top screen [job] of the 1990s*
- *[fullname] is a [job] who was born in [birthplace] on [birthday]*

Adam Sandler:

- [*fullname] was born in [birthplace] on [birthday] and is a jewish-american [job], [job], and [job]*
- *born in [birthplace] on [birthday] to judy sandler [fullname] lived in new york*
- *[fullname] [birthday] (age 44)[birthplace] , new york, u.s. is a [job]*
- *[fullname] is a [job] who was born on [birthday]*

Vergleicht man mit “Teilstring-Methode“ jede Textphrase von Jim Carrey mit jeder Textphrase von Adam Sandler, erhält man folgende Ergebnisse:

- *[fullname] was born in [birthplace] on [birthday] and is*
- *born in [birthplace] on [birthday]*
- *[fullname]*
- *on [birthday]*
- *[fullname] was born*
- *was born*
- *in [birthplace]*
- *[birthplace]*
- *was born in [birthplace] on [birthday]*
- *born in [birthplace] on [birthday]*
- *[fullname] is a [job] who was born*

Diese Ergebnisse sind noch nicht zufriedenstellend. Die erste und die achte Ergebnissphrase ließe sich als Template Kandidat verwenden, die üblichen Ergebnisse sind nicht verwendbar. Außerdem findet sich in keinem Ergebniss alle Attribute wieder. Insofern genügend Textphrasen vorhanden sind, ließen sich mit dieser Methode einige Templatekandidaten finden; betrachtet man allerdings die Ausgangstextphrasen, könnte man bessere Ergebnisse aus ihnen gewinnen.

Eine zweite Möglichkeit ist zwei Textphrasen auf gleiche Aneinanderreihungen von Wörtern zu untersuchen. Das Vorgehen ist dabei wie folgt: Man vergleicht je ein Wort beider Textphrasen. Bei Übereinstimmung vergleicht man das jeweils nächste Wort beider Phrasen. Bei Nichtübereinstimmung der n-ten Wortpaare, wird Wort n aus Textphrase 1 mit Wort n+1 aus Textphrase 2 verglichen. Im Allgemeinen wird bei Übereinstimmung der Zeiger auf das jeweils zu untersuchende Wort bei beiden Textphrasen um eins erhöht; bei Nichtübereinstimmung nur der Zeiger von Textphrase zwei. Sollten sich mindestens zwei aufeinanderfolgende Worte gleichen, werden sie dem Ergebnisstring angefügt. Man erreicht dadurch, dass entitäts-spezifische Wörter bzw. Formulierungen einer Textphrase entfernt werden und nur allgemeinere Wortaneinanderreihungen übrig bleiben. Diese Methode auf das obige Beispiel angewendet, ergibt folgende Ergebnisse:

- *[fullname] was born in [birthplace] on [birthday]) and is a [job], and [job]*
- *born in [birthplace] on [birthday]*
- *is a*
- *was born on [birthday]*
- *[fullname] was born on [birthday]*
- *[birthplace] on [birthday]*
- *[fullname] is a [job]*
- *[fullname] was born*

- *on [birthday]*
- *[fullname] is a [job] who was born on [birthday]*

Unter diesen Ergebnissen findet sich besseres Material. Es ist aber auch ersichtlich, dass weitere Filterarbeiten nötig sind. Zum einen kann man sich zu Nutze machen, dass manche Templatephrasen in anderen Kandidaten vorkommen (Beispiel Phrase 9 „[fullname] was born“ findet sich in Phrase 1 und 5). Deshalb können die kürzeren Phrasen entfallen, da ihr Inhalt in den längeren Kandidaten mit ausgedrückt wird. Des Weiteren wird das Verhältnis von Attributwörtern zu anderen Wortvorkommen in einer Phrase vergleichen. Eine Textphrase, die aus zu viele Attributen im Verhältnis zu anderen Worten besteht, eignet sich nicht, um damit einen Satz zu erzeugen (z.B. „[birthplace] on [birthday]“); eine Phrase mit zu vielen Worten auf der anderen Seite drückt mit erhöhter Wahrscheinlichkeit nicht den Sinnzusammenhang der in der Phrase vorkommenden Attribute aus. Des Weiteren finden sich Phrasen, in denen keine Verben vorkommen (z.B. „[Name], son of [Father] and [Mother].“). Dieses Problem lässt sich beheben, indem man einen Bestand der gängigsten Verben der englischen Sprache verwendet und überprüft, ob in der extrahierten Phrasen wenigstens eines vorhanden ist. Die Verben könnte man in einer Textdatei abspeichern und zur Laufzeit einlesen. Für höhere Qualität kann man den Bestand um weitere Verben erweitern.

Prinzipiell sollten die Filterkriterien sehr streng sein; es wird nach der Philosophie verfahren: „Lieber einen guten Templatekandidaten zu viel herausgefiltert, als eine schlechte Phrase zu viel drin gelassen.“ Gewonnene Templatephrasen können z.B. in XML-Notation abgespeichert und zur späteren Textgenerierung für jede Entität des gleichen Konzepts genutzt werden.

4.2.2 Attributemapping

Das Planen des Textaufbaus soll mit Hilfe des Attributemappings erfolgen. Der Hauptgrund besteht darin, dass im Gegensatz zu den angesprochenen „Relationen“, hier die Beziehungen zwischen mehreren Attributen ausgedrückt werden können. Dies ist vor allen Dingen sehr nützlich, da auch in den gewonnenen Templatephrasen immer eine Kombination von Attributen ausgedrückt wird. Daher kann man einer Attributgruppe eine bestimmte Menge Templates zuordnen, in denen diese Attribute vorkommen (dazu Näheres in Abschnitt 4.2.3).

Es muss im Vorfeld erst einmal geklärt werden, an welcher Stelle im Text, welches Attribut und damit welcher Fakt einer Entität vorkommen soll und welche Attribute inhaltlich zusammengehören. Bspw. können „FullName“, „Birthday“ und „Birthplace“ als eine Attributgruppe angesehen werden, die meist am Anfang eines Textes stehen.

Woher weiß man, welche Attribute sich inhaltlich zu Gruppen zusammenfassen lassen? Hier kann man sich zu Nutze machen, dass in den aus dem Internet gewonnen Textphrasen oftmals mehrere Attribute in einem Satz vorkommen. Sollte dies in mehreren Sätzen der Fall sein, kann man die Vermutung anstellen, dass diese jeweiligen Attribute zusammengehören.

Die nächste Frage ist, in welcher Reihenfolge die Attributgruppen und damit auch die Attribute im Text angeordnet werden sollen und welche Attributgruppe auf die jeweils Nächsten verweist. Dazu kann man während der Extraktion der entitätsbezogenen Satzphrasen aufpassen, in welcher Reihenfolge die Fakten der Entität auftauchen. Insofern genügend Texte durchsucht wurden, lassen sich Rückschlüsse ziehen, an welche Stelle eine Attributgruppe eingeordnet werden kann und welche Gruppen folgen können.

Definiert wird ein Attribut Mapping in Xml-Notation. Ein Attribut Mapping besteht aus mindestens einer bis unendlich vielen Attributgruppen. Eine Attributgruppe verfügt über genau eine ID und kapselt zum einen das Element „NextGroup“, welches eine Liste von IDs der folgenden Attributgruppen führt und zum anderen mindestens ein Element vom Namen Attribut. Ein Attribut beschreibt in diesem Sinne ein Attribute des Koncepts.

4.2.3 Templates und Attributgruppen zuordnen

Ausgangspunkt an dieser Stelle sind eine Menge von Templates, die eine bestimmte Kombination von Attributen kapseln und eine Attributmap, die die Anordnung der Attribute festlegt. Jede Attributmap besteht aus einer Menge Attributgruppen, in denen wiederum eine Menge Attribute enthalten sind, die inhaltlich zusammengehören.

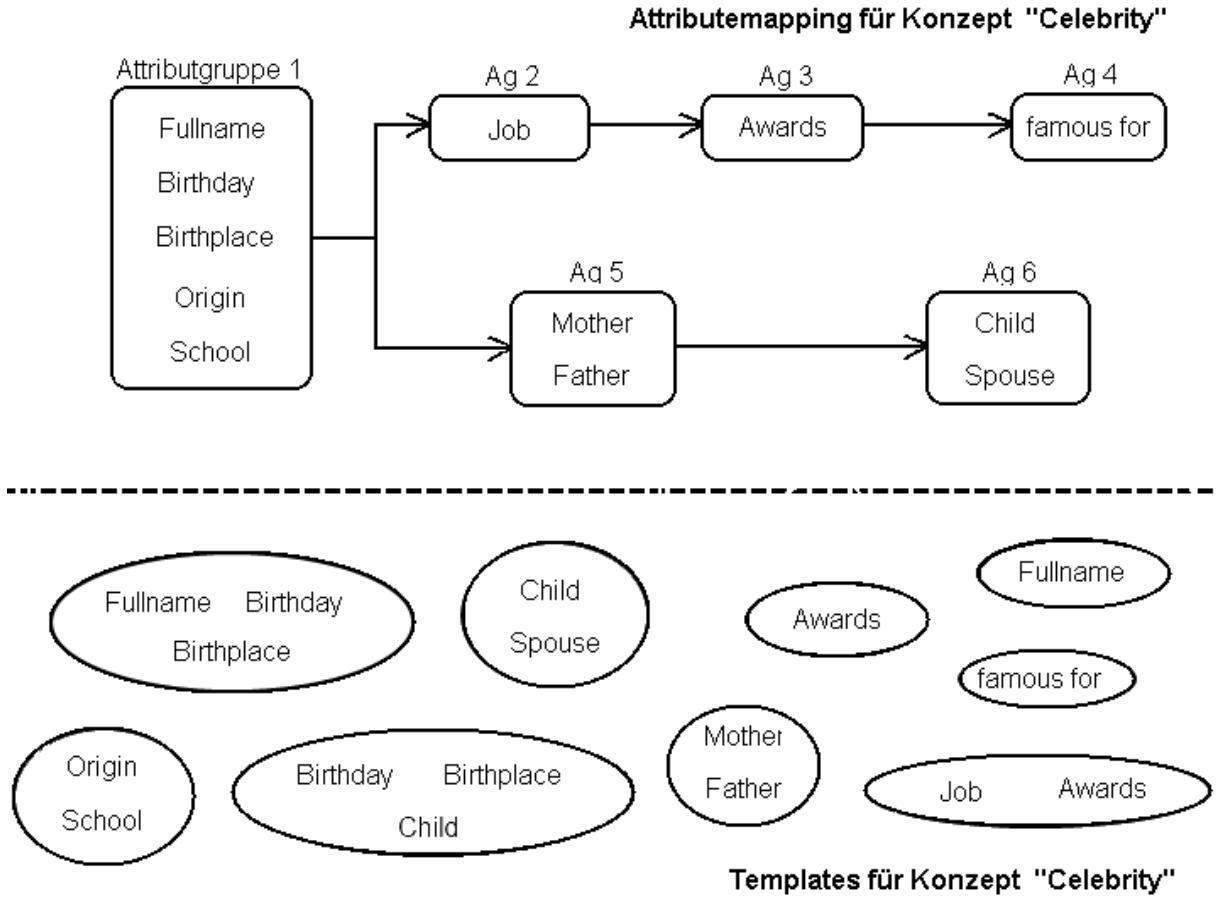


Abb 4.5. - Zuordnung zwischen Templates und Attributmapping

Es ist nun Aufgabe aus der Menge der Templates geeignete Kandidaten auszuwählen, damit zum einen alle Attribute der Attributmap ausgedrückt werden, zum zweiten durch Attributgruppen zusammengehörende Attribute möglichst durch Templates mit der gleichen Attributkombination ausgedrückt werden und zum dritten kein Attribut doppelt vorhanden sein darf, um später im Text keine Fakten doppelt auszudrücken.

Die obige Abbildung stellt die Beziehung zwischen der Attributmappe des Konzepts „Celebrity“ dar und die dafür gewonnenen Templates. Es ist ersichtlich, dass Templates mit unterschiedlichen Attribut-Kombinationen vorliegen, unter denen ausgewählte Vertreter einer Attributgruppe zugeordnet werden. Dabei kann sowohl einer Gruppe mehrere Templates zugesprochen werden (Attrgr. 1), als auch ein Template mehrere Attributgruppen vereinen(Attrgr. 2 und Attrgr. 3). Leider ist Abbildung 4.5. ein Idealfall, in der es für jede Attributgruppe geeignete Templatekombinationen gibt und kein Attribut ausgelassen wird oder doppelt vorkommt. In der Praxis muss dies nicht immer gegeben sein. Denkbar wäre bspw. ein Szenario, in dem das Attribut „Job“ nur in einem einzelnen Template zusammen mit dem Attribute „Birthday“ vorkommt. Diese Kombination lässt sich mit obiger Attributmappe nicht ohne weiteres einplanen. Da das Attribut „Birthday“ durch dieses Template bereits ausgedrückt wird, müssen zum einen Templatekombinationen für die übrigen Attribute aus Gruppe 1 gefunden werden. Des Weiteren muss man von der Anordnung der Attributgruppen in der Attributmappe abweichen. Eine zufriedenstellende Lösung für diese Probleme zu finden ist keine triviale Aufgabe.

Der gewählte Lösungsansatz ist, aus der Menge aller Templates zunächst alle Kombinationen zu ermitteln, in denen jedes Attribut des Konzepts genau einmal vertreten ist und dies erst einmal unabhängig von der Attributmappe. Aus der Gruppe dieser Kombinationen wählt man sich dann einen Kandidaten aus und versucht dessen Templates in die Attributmappe einzuplanen. Sollte dies nicht gelingen, greift man auf den nächsten Templatekandidaten zurück. Dieses Verfahren hat zwei große Vorteile. Zum einen kann man unter der Menge der passenden Kandidaten bei jedem Textgenerierungsvorgang stets einen anderen auswählen (zufällig oder nach bestimmten Kriterien), was dem Abwechslungsreichtum der erstellten Texte zu Gute kommt. Zweitens, indem zur Not alle möglichen Kombinationen ausprobiert werden, lässt sich stets eine einplanbare Kombination finden, insofern eine vorhanden ist. Sollte dies nicht der Fall sein, konnte man zeigen, dass mit dem momentanen Stand an Templates das Einplanen nicht möglich ist. Man könnte an dieser Stelle bspw. die Extraktion weitere Templatephrasen in die Wege leiten.

4.2.4 Templatephrasen verknüpfen

Nach Template Einplanung wählt man aus jedem angeordneten Template eine zufällige Textphrase aus und reiht diese aneinander, um den Text zu erzeugen. Am Beispiel des Konzepts „Celebrity“ soll aufgezeigt werden, in welcher Form der resultierende Text an dieser Stelle vorliegt:

[Fullname], who was born [Birthday] in [Birthplace] is a [job] || he is most famous for [famous for] || he won a [award] for [film] || he attended [school] in [origin] || he and his wife [Spouse] have child [Child]

Einzelne Textphrasen sind noch mit einem Trennzeichen voneinander abgegrenzt. Diese Stellen kann man mit Verknüpfungsphrasen zusammenfügen, um so den endgültigen Fließtext zu generieren. Beispiele für Verknüpfungsphrasen sind „ and“, „furthermore“ oder „in addition“. Alternativ kann man einige Templatephrasen auch einfach mit einem Punkt abschließen. Darüber hinaus erhöhen sich durch zufälliges Auswählen derartiger Verknüpfungen der Abwechslungsreichtum und der natürlich klingende Stil des resultierenden Textes. Nachdem dieser Vorgang abgeschlossen wurde, kann man den resultierenden Text noch „den letzten Schliff verpassen“, indem u.a. korrekte Zeichensetzung und Groß- und Kleinschreibung ergänzt werden. In den resultierende Text müssen nun nur noch die verallgemeinerten Attributstellen mit den konkreten Fakten der Entität ersetzt werden und das Resultat ist der gewünschte Fließtext:

„James Jim Eugene Carrey, who was born January 17th 1962 in Newmarket Ontario, is a canadian-american comedian. He is famous for his role in Ace Ventura and won a Golden Globe for The Truman show. He attended Blessed Trinity Catholic School in Newmarket. In Addition he and his wife Melissa Womer have a child, Jane Carrey.”

5. Implementierung

5.1. Spezifikationen

Wie man dem Klassendiagramm 5.2. entnehmen kann, wird der gesamte Textgenerierungsprozess durch drei große Klassen verwaltet, wobei jede auf die zentrale Klasse Template zurückgreift.

Ein Template ist in diesem Sinne keine einzelne Textphrase, sondern eine Klasse. Gekapselt wird eine Kombination von Attributnamen und alle Templatephrasen, in denen diese Attribute vorhanden sind. Jedes Template erhält eine ID, für spätere Auswahlstrategien der Texterstellung.

Aufgabe eines TemplateManagers ist es, eine Menge derartiger Templates zu verwalten. Sollte man der Menge ein neues Template hinzufügen, werden bereits vorhandene Templates auf ihre Attributkombinationen überprüft. Sollte ein Template die gleiche Kombination, wie die des Einzufügenden besitzen, werden dem bereits bestehenden Template lediglich die neuen Satzphrasen übergeben. Dadurch gewährleistet man, dass jede Attributkombination nur einmal vorhanden ist. Der Templatemanager hat darüber hinaus die Aufgabe, mit Hilfe seiner Templates die Attributmappe zu generieren. Des Weiteren können alle Templatekombinationen ermittelt werden, in denen jedes Attribut genau einmal vertreten ist, um später inhaltliche Überschneidungen zu vermeiden. Der TemplateManager speichert sowohl die Attributmappe, als die Templates und deren Kombinationen in Xml-Dateien und kann diese zu einem späteren Zeitpunkt auch wieder einlesen.

Der Templateextraktor hat zum einen die Aufgabe das Internet nach möglichen Kandidatenphrasen zu durchsuchen. Er erhält hierbei bei Initialisierung eine Referenz auf ein Konzept und eine Menge dazugehörige Entitäten. Unterstützt durch Palladian wird pro Entität eine einstellbare Anzahl Suchanfragen an eine SearchEngine gestellt und die empfangenen Textresultate auf das Vorkommen der FactValues der Entitäten untersucht. Pro Entität wird eine Xml-Datei erstellt, mit einer Liste Textphrasen, geordnet nach der Kombination an Factvalues, die in ihnen vorkommen. Die zweite Funktion des Extraktors ist die Gewinnung geeigneter konzeptübergreifender Templatephrasen. Wie im vorherigen Kapitel beschrieben, werden Textphrasen mit gleichen Attributvorkommen untereinander auf gleiche Wortfolgen untersucht und die dadurch gewonnenen Templatekandidaten durch strenge Heuristiken gefiltert, um so geeignete Templates zu finden.

Ein Textgenerator kapselt ein Konzept, sämtliche dazugehörige Templates, die ermittelten Templatekombinationen und die entsprechende Attributmap. Nun kann für jede Entität des Konzepts der Text generiert werden, indem eine Templatekombinationen gewählt wird und die Templates entsprechend der Attributmap angeordnet werden. Danach wird aus jedem Template eine zufällige Textphrase entnommen und die jeweiligen Attributvorkommen mit den dazugehörigen Fakten der Entität ersetzt. Die Textphrasen werden aneinandergereiht und ergeben den Gesamttext. Dieser wird zum Schluss syntaktisch aufgebessert (Kommas, Groß-, Kleinschreibung) und kann schließlich ausgegeben werden.

5.2. Benutzung

Möchte man für ein Konzept passende Templates finden, erstellt man sich eine Instanz der TemplateExtraktor Klasse. Dieser übergibt man im Konstuktor oder später eine Liste von Entitäten. Danach kann man die „extractCandidates“ Methode aufrufen, wobei für jeden Faktwerten jeder Entität eine - als Parameter mitgegebene - Anzahl Suchanfragen ans Internet gestellt wird und gefundene Textphrasen in XML-Dateien abgespeichert werden. Danach kann man die „findTemplates“ Methode nutzen, um aus diesen Textphrasen passende Templates und die Attributemappe zu generieren.

Sobald diese Datenstrukturen einmalig generiert sind, kann zu einem beliebigen Zeitpunkt ein Objekt vom Typ „Textgenerator“ genutzt werden, um für eine Entität und dessen Fakten den Fließtext zu erstellen. Dazu muss man lediglich die Methode „createText“ mit einer Referenz zur Entität aufrufen. Zusätzlich gibt man einem zweiten Parameter an, ob die Entität männlich, weiblich oder neutral ist. Man kann hierbei auf statische Membervariablen der Textgeneratorklasse zurückgreifen. Der Rückgabewert der „createText“-Methode ist der gewünschte Gesamttext als String.

```
ArrayList[Entity] mountains = createMountainEntities();

TemplateExtractor tempEx = new TemplateExtractor(mountainConcept);
tempEx.addAllEntities(mountains);


tempEx.extractCandidates(50);
tempEx.findTemplates();

TextGenerator tg = new TextGenerator(mountainConcept);

String myText = fm2.createText(mount_EeverestEntity, TextGenerator.NEUTRAL);

System.out.println(myText);
```

Abb 5.1.- Codebeispiel

5.3. Klassendiagramm

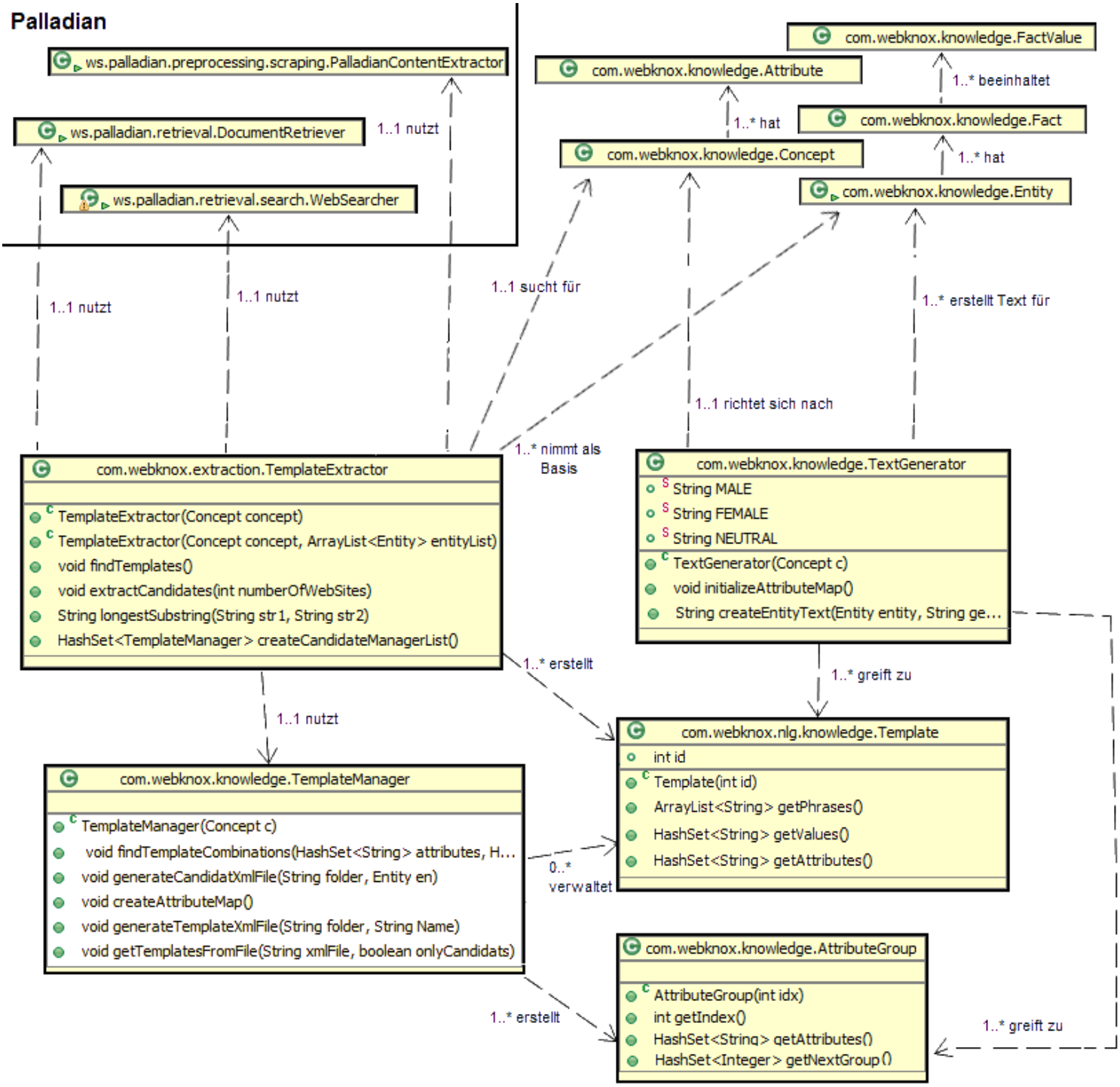


Abb 5.2.- Klassendiagramm

6. Evaluierung

Die Evaluierung unterteilt sich in drei große Abschnitte. Der Schwerpunkt der Arbeit lag auf der Extraktion von Konzeptphrasen. Diese Phrasen werden durch das untereinander Abgleichen der Satzphrasen mehrerer entsprechender Entitäten gewonnen. Deshalb soll im ersten Teil ein grobes Maß bestimmt werden, welche Mindestanzahl an Entitäten genutzt werden sollte, sodass für jedes Attribut wenigstens eine Templatephrase gefunden werden kann.

Im zweiten Teil werden die - für mehrere Beispielkonzepte gewonnenen - Templatephrasen auf ihre Güte analysiert. Konkret werden diese zum einen nach ihrem syntaktischen Aufbau eingeschätzt; insbesondere ob die Grammatik korrekt ist. Des Weiteren werden sie auf ihren semantischen Inhalt überprüft. Es wird ermittelt, ob diese den gewünschten Sinn ausdrücken für die in ihnen enthaltenen Kombinationen von Attributen. Die Phrasen, die sowohl syntaktisch als auch semantisch keine Mängel aufweisen, können dann noch auf ihre rhetorische Güte untersucht werden.

Der dritte Evaluierungsabschnitt bezieht sich auf die Auswertung der Gesamttexte. Dazu bekommen Probanden im Rahmen einer online Webauswertung einige generierte Texte vorgezeigt und diese müssen mit Hilfe eines Punktesystems nach bestimmten Kriterien eingeschätzt werden. Unter den angebotenen Texten befinden sich einige Ausschnitte, die von Wikipedia stammen, um einen Vergleich mit menschlichen Texten anstellen zu können.

6.1. Abschätzung der benötigten Entitäten -Mindestanzahl

Die Vorgehensweise richtet sich danach, für ein Beispielkonzept eine wachsende Anzahl Entitäten anzulegen und dadurch zum einen die gewonnene Anzahl Templatephrasen zu bestimmen und zum anderen die Anzahl Attribute für die keine Templatephrase gefunden wurde. Für das Konzept „Celebrity“ wurden 12 Attribute angelegt. Entsprechenden Entitäten wurden pro Attribut ein Fakt zugewiesen.[5]

[5] **Siehe Anhang A, Seite 53**

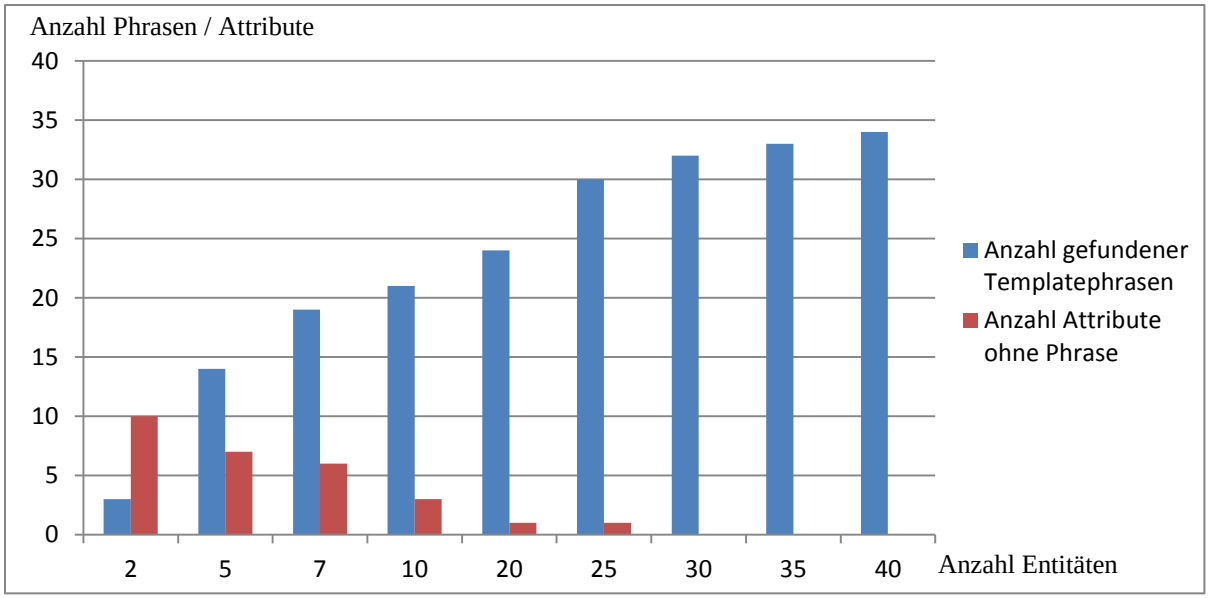


Abb 6.1 – Zusammenhang gefundener Templatephrasen und Entitätenzahl

Abbildung 6.1 kann man entnehmen, dass die Anzahl gefundener Templatephrasen zunächst beständig mit der Anzahl der Entitäten wächst, während die Anzahl an Attributen, für die keine Templatephrase gefunden werden konnte, beständig abfällt. Ab 25 Entitäten werden nur vereinzelt neue Phrasen entdeckt: bei einer Erhöhung der Entitätenzahl um 15 wurden nur noch 3 neue Phrasen gefunden. Ab 30 Entitäten ist für jedes Attribut mindestens ein Template vorhanden.

Dieser Verlauf kann erklärt werden, indem man sich die Vorgehensweise des Algorithmus vor Augen führt: es werden je Satzphrasen mit gleichen Attributvorkommen untereinander abgeglichen, um so immer wiederkehrende Formulierungen zu finden. Nach dem Erreichen einer bestimmten Anzahl Satzphrasen hat man die gängigsten Formulierungen für eine bestimmte Attributkombination gefunden und stößt nur noch selten auf neue Vertreter.

Ein Vorschlag ist daher die Anzahl an Entitäten auf den Wert festzulegen, bei dem für jedes Attribut wenigstens eine Templatephrase gefunden wurde.

6.2. Evaluierung der Templategüte

6.2.1. Vorgehensweise

Als Ausgangspunkt dienen 6 Beispielkonzepte (*Celebrities, Mountains, Mobiles, Athletes, Cities* und *Universities*) mit je 6 Attributen.[6] Zu diesen wurden je 20 Entitäten erstellt mit einem Fakt pro Attribut. Für alle Entitäten und jedem Fakt wurden bis zu 50 Webseiten

[6] **Siehe Anhang A**

extrahiert aus denen alle Satzphrasen gewonnen wurden, in denen diese Fakten vorkommen. Es wurden für alle 120 Entitäten insgesamt 103431 Satzphrasen mit unterschiedlichen Attributkombinationen gefunden. Mit Hilfe der Templateextraktorklasse wurden für alle Konzepte schlussendlich insgesamt 80 Templatephrasen gewonnen[7]. Wie bereits erwähnt werden Phrasen auf syntaktische, semantische und geeignete Phrasen auch auf ihre rhetorische Güte überprüft.

Phrasen werden syntaktisch danach eingeschätzt ob sie entweder:

- grammatikalisch korrekt sind
- einzelne oder leichte Mängel aufweisen (z.B. vergessene Artikel, falsche Deklination)
- komplett ungeeignet sind (z.B. fehlende Verben, Substantive)

Zweitens werden Phrasen danach unterteilt, ob sie den gewünschten Sinn der in ihnen enthalten Attribute ausdrücken. Hier wird kategorisiert ob sie entweder:

- den Sinnzusammenhang ihrer Attribute komplett richtig wieder geben
- leichte Mängel aufweisen (z.B. einzelne entitätsbezogene Wörter vorhanden sind)
- komplett falsch liegen

Grammatikalisch korrekte Phrasen können drittens noch auf ihre rhetorische Güte bewertet werden. Man unterteilt in:

- simple Phrasen: Sätze nach dem Muster "Substantiv Verb Objekt" (z.B."[name] lies in [country]." oder "[name] is a mountain.")
- bessere Phrasen: Passivsätze oder Sätze in denen mehr als ein Objekt oder Verb vorkommt (z.B. „the mountain [name] is located in [country].")
- hochwertige Phrasen: Sätze in denen mehr als zwei Attribute ausgedrückt werden oder Sätze in denen mehr als zwei Verben und mindestens ein Adjektiv oder Adverb vorkommen (z.B. „[name] is a well known mountain located in [country]" oder „[name], located in [country], is a mountain of the [gebirge] region.")

An folgenden extrahierten Templatephrasen für das Konzept „Mountain" soll beispielhaft aufgezeigt werden, wie deren Güte eingeschätzt wird:

"[name] is located in southcentral [county]"

Diese Phrase weist keine grammatikalischen Mängel auf. Semantisch muss ein Abstrich gemacht werden, da in ihr das entitätsbezogene Wort „southcentral" erhalten geblieben ist. Die rhetorische Güte kann als „besser" eingeschätzt werden, da Passivsatz.

[7] **Die Textphrasen im Detail in Anhang A**

“[name]'s height is [height]”

Diese Phrase ist sowohl syntaktisch als auch semantisch fehlerlos. Die rhetorische Güte ist simpel.

„the coordinates of [coordinates]“

Ein Beispiel für eine grammatikalisch ungeeignete Phrase.

“it is the [first_ascent]”

Diese Phrase ist grammatikalisch korrekt, drückt aber nicht den gewünschten Sinn aus. („Wann war die Erstbesteigung?“)

„[name], located in the [gebirge], has a height of [height]”

Ein Beispiel für eine tadellose Phrase mit hoher rhetorischer Güte.

6.2.2. Template Gesamteinschätzung

Die Gesamteinschätzung kann den folgenden Tabellen entnommen werden. Jede Zeile entspricht einer gefundenen Phrase. Eine Liste der ausformulierten Phrasen befindet sich in Anhang A, ab Seite 64. Zeilen die durchgehend grau sind, entsprechen automatisch generierten Sätzen, für deren enthaltene Attribute keine Phrase gefunden werden konnte.

Einschätzung der Templatephrasen des Konzepts „Mountain“

Phrase	Attribute	Syntax	Semantik	Rhetorik
M1	County			
M2	Name, Gebirge			
M3	Name, County			
M4	Name, Gebirge, Höhe			
M5	First Ascent			
M6	Height			
M7	Coordinates			

Tab.6.2 - Mountain Template Evaluierung

keine Mängel, bzw. hochwertiger Ausdruck
einzelne Abstriche, bzw. mittlerer Ausdruck
simpler Ausdruck
ungeeignet
nicht eingeschätzt

Einschätzung der Templatephrasen des Konzepts „Athlete“

Phrase	Attribute	Syntax	Semantik	Rhetorik
A1	Sport			
A2	Sport			
A3	Name, Sport			
A4	Name, Sport			
A5	Name, Sport			
A6	Name, Team			
A7	Achievement			
A8	Achievement			
A9	Birthday			
A10	Spouse			
A11	Hobby			

Tab.6.3 - Athlete Template Evaluierung

Einschätzung der Templatephrasen des Konzepts „City“

Phrase	Attribute	Syntax	Semantik	Rhetorik
Ci1	Name, Country			
Ci2	Name, Country			
Ci3	Name, Country			
Ci4	Name, Country			
Ci5	Name, Country			
Ci6	Name, Country			
Ci7	Name, Country			
Ci8	Name			
Ci9	Sight			

Ci10	University, Country			
Ci11	University, Country			
Ci12	Population			
Ci13	Transportation			
Ci14	Mayor			

Tab.6.4. - City Template Evaluierung

Einschätzung der Templatephrasen des Konzepts „University“

Phrase	Attribute	Syntax	Semantik	Rhetorik
U1	City, Country			
U2	City, Country			
U3	City, Country			
U4	City, Country			
U5	Type			
U6	City, Type			
U7	Establ.			
U8	Name, Type			
U9	President			
U10	Motto			

Tab.6.5. - University Template Evaluierung

Einschätzung der Templatephrasen des Konzepts „Mobile“

Phrase	Attribute	Syntax	Semantik	Rhetorik
Mo1	Manufacturer			
Mo2	Manufacturer			
Mo3	Manufacturer			
Mo4	Manufacturer			
Mo5	Manufacturer			
Mo6	OS			
Mo7	OS			
Mo8	OS			
Mo9	OS			
Mo10	OS			
Mo11	OS			
Mo12	OS			
Mo13	Resolution			
Mo12	Resolution			
Mo10	Memory			
Mo11	Memory			
Mo12	Weight			
Mo13	Sar-Wert			

Tab.6.6. - Mobile Template Evaluierung

Einschätzung der Templatephrasen des Konzepts „Celebrity“

Phrase	Attribute	Syntax	Semantik	Rhetorik
Ce1	Name, Job			
Ce2	Name, Job			

Ce3	Job			
Ce4	Job			
Ce5	Birthday			
Ce6	Name, Birthplace			
Ce7	Name, Parents			
Ce8	Name, Job, Fullname			
Ce9	Origin			
Ce10	Origin			
Ce11	Origin			
Ce12	Fullname, Birthplace			
Ce13	Name, Birthday			
Ce14	Leistung			
Ce15	Spouse			
Ce16	Spouse			
Ce17	Spouse			
Ce18	Award			
Ce19	Award			
Ce20	Award			
Ce21	Award			
Ce22	Award			
Ce23	Child			
Ce24	School			

Tab.6.7. - Celebrity Template Evaluierung

Zusammenfassend lässt sich entnehmen, dass unter den 72 untersuchten Templatephrasen 4 (5,5%) als grammatikalisch ungeeignet klassifiziert wurden. Die übrigen wiesen fast durchgehend eine fehlerfreie Grammatik auf (93%). Unter semantischen Gesichtspunkt mussten 7 Phrasen (9%) ausgemustert werden, die nicht den gewünschten Sinn ihrer Attribute ausdrücken konnten.

In 22 Phrasen (35%) kamen entitätsbezogene Formulierungen vor, die den Inhalt des Gesamttextes mit inkorrekten Aussagen anreichern können (Bsp: „[name] is the highest mountain in [gebirge].“). Die Anzahl der Templatephrasen die sowohl syntaktisch als auch semantisch absolut fehlerfrei sind, beträgt 40 (55,5%). Die überwiegende Mehrzahl der Phrasen ist auf einem simplen bis besseren rhetorischen Level (92%). Vereinzelt finden sich aber auch Formulierungen mit hochwertigem Ausdruck.

Alles in allem kann man mit den Ergebnissen zufrieden sein, da die überwiegende Mehrzahl der gewonnenen Phrasen keine grammatikalischen Fehler aufweisen und demzufolge die größte Mehrzahl generierter Gesamttexte syntaktisch korrekt sein werden. Darüber hinaus haben die meisten Phrasen einen adäquaten Ausdrucksstil, sodass die meisten Endresultate vermutlich eine gute natürlich klingende Form haben werden. Kritisch betrachten muss man die Anzahl Phrasen in denen inhaltliche Fehler vorkommen. Diese werden in den Fließtext mit übertragen. Leider könnte dies algorithmisch nur sehr schwer und nur durch erhöhten Aufwand korrigiert werden.

6.3. Evaluierung der Gesamttexte

Im Zuge der Gesamtauswertung wurde eine kleine Webanwendung angelegt, mit deren Hilfe Probanden die Güte der generierten Fließtexte auf Basis verschiedener Kriterien einschätzen können.

Beispiel:

“Samsung Wave Y is the latest Samsung. Its sar is 0,318 and its weight is 103 g. It comes with 320 x 480 pixels and has 150 MB memory. It is running on Bada.”

Die Anwendung ist ein Webservice[8] zur Durchführung und Auswertung von Umfragen. Dabei wurden für insgesamt 10 Enitäten der bereits genutzten Beispielkonzepte Fließtexte generiert. Des Weiteren wurden 3 Texte erstellt, die aus Auszügen von Wikipedia Artikeln bestehen und die als Referenz zu von menschenhand geschriebenen Texten dienen. Dabei wurden streng all jene Sätze des Artikels verwendet, in denen ein Attribut vorkommt, das auch in den Beispielkonzepten enthalten ist (siehe Anhang B, Seite 65).

[8] **http://www.equestionnaire.de/deu/**

Beispiel:

"Aconcagua is the highest mountain in the Americas at 6,962 m. It is located in the Andes mountain range, in the Argentine province of Mendoza. […] The coordinates are 32°39'12.35"S 70°00'39.9"W.[…] The first attempt on Aconcagua by a European was made in 1883 by a party led by the German geologist and explorer Paul Güssfeldt."

Probanden werden zu jedem Text eine Frage pro Kriterium gestellt. Die Kriterien beziehen sich ähnlich der Templateevaluation auf syntaktische und semantische Faktoren. Insgesamt wurden 8 Kriterien untersucht:

- Grammatik – Wieviele Grammatikfehler kann man im Text finden?
- Rechtschreibfehler – Wieviel Rechtschreibfehler?
- Umfang – Ist der Text zu kurz bzw. zu lang in Hinblick auf die enthaltenen Fakten.
- Reihenfolge – Ist die Anordnung der Fakten nachvollziehbar.
- Ausdruck – Wie gelungen ist der Ausdrucksstil?
- Verständnis – Sind die Inhalt des Textes einfach zu erfassen?
- Abwechslungsreichtum - Klingen die Sätze abwechslungsreich oder monoton?
- Natürlichkeit – Ist der Text von Mensch oder Maschine?

Jedes Kriterium kann mit einer von 5 möglichen Antworten bewertet werden (Detaillierte Liste aller Fragen und Antwortmöglichkeiten in Anhang C, ab Seite 68).

	Gut gelungen.	Ganz gut	Befriedigend	Teils unglücklich formuliert	Unleserlich
Wie gefällt ihnen der Ausdrucksstil des Textes?	○	○	○	○	○

Abb. 6.8. - Antwortmöglichkeiten für Kriterium „Ausdruck"

Je nach Kriterium müssen Antwortmöglichkeiten nicht immer aufsteigend von gut nach schlecht geordnet sein.

Die Auswertung wurde mit 3 Probanden durchgeführt. Alle 13 Texte wurden in zufälliger Reihenfolge präsentiert und jeweils nach den 8 Kriterien eingeschätzt. Bildet man den Mittelwert aller Einschätzungen für das jeweilige Kriterium erhält man folgende Ergebnisse:

Kriterium	Art der Punktevergabe	Durchschnitt für generierten Texte	Durchschnitt für Wikipedia Auszüge
Grammatik	1(Fehlerlos) - 5(mehr als 7 Fehler)	1,831	1,553
Rechtschreibfehler	1(Fehlerlos) - 5(mehr als 7 Fehler)	1,465	1,444
Umfang	1(Zu lang) – 3(Angemessen) – 5(Zu	2,696	3,333
Reihenfolge	1(Sehr gut strukuriert) – 5(Keine Strukur	2,498	2,000
Verständnis	1(sehr gut) – 5 (unverständlich)	1,933	1,996
Ausdruck	1(Sehr gut) – 5(unleserlich)	2,833	2,111
Abwechslungsreichtum	1(sehr abwechslungreich) –	2,433	1,996
Natürlichkeit	1(menschenhand) – 5 (computer generiert)	3,431	2,444

Tab.6.9 - Gesamtauswertung der Gütekriterien

Die Ergebnisse sind recht zufriedenstellend. In den meisten Texten gab es einzelne syntaktische Fehler, deswegen wurden Grammatik und Rechtschreibung meist mit Note 2 bewertet. Interessant ist, dass auch die Wikipedia Auszüge nicht mit perfekten Noten versehen wurden. Probanden gaben an, auch hier Grammatik- und Rechtschreibfehler gefunden zu haben.

Der Umfang der Texte in Hinblick auf die Anzahl der darin enthaltenen Fakten war den Probanden tendenziell eher zu kurz und knapp ausgelegt, hier schneiden die Wikipedia Auszüge besser ab. Erklären lässt sich die knappe Form der generierten Texte, da die meisten Templates eher von simpler Rhetorik sind (Substantiv, Verb, Objektiv), was auch Einfluss auf Ausdruck und Abwechslungsreichtum der Texte hatte. Die Reihenfolge der auftretenden Fakten war in den meisten Fällen angemessen; vereinzelte Fakten wurden ab und zu an unangebrachten Stellen eingeordnet, wodurch Reihenfolge und Textverständnis tendenziell schlechter eingeschätzt wurden. Alles in allem wurde der Textinhalt allerdings meist als „Gut verständlich“ bewertet. Interessanterweise haben auch die Wikipedia Auszüge keine perfekten Benotungen erhalten.Auf Grund der einzelnen Fehler in den generierten Texten konnten die Probanden in fast allen Fällen die computergenerierten Texte von den menschlichen Texten unterscheiden (Die detaillierten Umfrageergebnisse in Anhang C).

7. Zusammenfassung und Ausblick

7.1 Zusammenfassung

Im Zuge der Evaluierung war zu erkennen, dass unter den generierten Fließtexten oft kleinere Mängel von syntaktischer und semantischer Natur vorkamen und man die computererzeugten Texte in den meisten Fällen von den Wikipedia Auszügen unterscheiden konnte. Wie aufgezeigt wurde, ist ein Großteil der extrahierten Templatephrasen korrekt; es aber gibt einen gewissen Prozentsatz, der Fehler aufwies. Je länger der Gesamttext, desto größer die Chance eine solche Phrase zu verwenden. Die generierten Texte sind aber alles in allem von zufriedenstellender Qualität. Sie weisen zu einem überwiegenden Teil eine korrekte Grammatik aus; haben natürlich klingenden Charakter und angemessene Rhetorik. Mit der Qualität eines menschlichen Textes können sie sich an vielen Stellen messen, erreichen diese aber auf Grund der genannten vereinzelten Fehler noch nicht in vollem Umfang.

Aus der Arbeit kristallisiert sich die Vermutung heraus, dass es bei der Erstellung von natürlich klingendem Fließtext wohl immer eine menschliche Komponente geben sollte. Ein vollautomatischer Algorithmus, der z.B. mit Hilfe von Grammatikbibliotheken und Lexika arbeitet, weist meist einen geringeren natürlichen Charakter auf. Gründe dafür sind die Vielzahl an Abschweifungen, Sonderregeln, Redewendungen etc. die in der menschlichen Sprache vorkommen. In dieser Arbeit durch Internetphrasen realisiert, in anderen Arbeiten durch weitere Möglichkeiten umgesetzt, wird in den meisten Fällen auf einen Datenbestand menschlicher Originaltexte bzw. –sätze zurückgegriffen, um adäquate Ergebnisse zu erzielen. Ein hochwertiger Textgenerierungsalgorithmus wird sich wohl daran messen, wie gut er diese Komponente verarbeiten kann.

Desweitern lässt sich sagen: je höher das Vorwissen, desto bessere die Qualität der resultierenden Texte. Speziell in dieser Arbeit wurden qualitativ hochwertigere Ergebnisse dann erzielt, wenn ein größerer Bestand Attribute, Fakten, Entitäten etc. vorhanden war, auf dem man aufbauen konnte. Dadurch wurden komplexere Templatephrasen gewonnen, die zu einer höheren rhetorischen Güte beitragen. Des Weiteren wurde der Zusammenhang der Attribute untereinander ersichtlicher, wodurch intelligentere Attributmappings erstellt werden konnten.

7.2 Offene Problemstellungen

Es ist erkenntlich geworden, dass durch weiteren Aufwand die Qualität der Texte noch näher an die Güte eines menschlichen Textes heran gebracht werden kann. Bspw. lässt sich der Einsatz von Haupt- und Nebensatz Strukturen umsetzen, indem zwei Satzphrasen miteinander verknüpft werden, in denen das Attribut „Name“ enthalten ist. Z.B. kann aus „Jim Carrey is a comedian“ und „Jim Carrey was born in Canada.“ die Phrase “Jim Carrey, who was born in Canada, is a Comedian” generiert werden. Diese Strategie kann nicht nur auf die konkreten Entitätensätze angewendet werden, sonder auch auf konzeptbezogene Templatephrasen, um so noch komplexere Phrasen zu erzeugen. Das Umwandeln einiger Aktiv- in Passivsätze lässt

sich in analoger Weise umsetzen.
Des Weiteren kann man bei der Suche nach Templatephrasen mit besseren Extraktionsalgorithmen arbeiten. Anstatt lediglich nach dem Vorkommen bestimmter Fakten in einem Satz zu suchen, kann man darüber hinaus nach dem Vorkommen ganzer Wortgruppen und Formulierungen suchen. Dadurch wird es möglich komplexere Formulierungen, wie Redewendungen zu entdecken und zweitens kann dadurch die Korrektheit bereits vorhandener Phrasen bestätigt werden, wenn die gleiche Formulierung in anderen Texten erneut gefunden wird.
Darüber hinaus muss der bereits angesprochene Umstand verbessert werden, dass 35% der Phrasen semantische Mängel aufwiesen, d.h. noch entitätsbezogene Wörter enthalten haben. Auch hier kann man durch vermehrtes Abgleichen mit Sätzen aus dem Internet und der Phrasen untereinander die inhaltliche Korrektheit der Phrasen für ein Konzept verallgemeinert weiter verbessern.

7.3 Fazit

Die Forschungsfragen dieser Belegarbeit waren zum einen, ob sich genügend auf ein Konzept anwendbares Textmaterial finden lässt, indem man nur genügend menschliche Texte untereinander abgleicht. Zweitens, ob sich aus den gewonnenen Erkenntnissen ein Textaufbauplan erstellen lässt, um Fakten in nachvollziehbarer Weise inhaltlich anzuordnen und drittens, ob man dieses Wissen verbinden kann, um natürlich wirkenden Text zu erstellen. Die ersten beiden Vermutungen konnten untermauert werden; es konnte für jede gewählte Entität ein Fließtext erstellt werden, in dem alle Fakten enthalten waren, die ausgedrückt werden sollten. Auf Grund der angesprochenen Fehler waren die erstellten Texte aber immer noch von denen eines Menschen unterscheidbar. Alles in allem aber sind die generierten Texte an vielen Stellen bereits mit denen eines menschlichen Textes vergleichbar und weisen auf Grund der extrahierten Templatephrasen eine zufriedenstellende rhetorische Güte auf.

Wie sich in der Arbeit gezeigt hat, ist der Vorgang der automatischen Fließtexterstellung ein Prozess in den sehr viele teils unterschiedliche Ideen und Ansätze einfließen können, um adäquate Ergebnisse zu erzielen. Keine der verschiedenen Varianten ist *die* Lösung, um das Problem der natürlichen Fließtexterstellung zu bewältigen. Vielmehr ist jede Variante an die eigene konkrete Aufgabenstellung angepasst, um diese zu lösen. Das Resultat ist eine Vielzahl von Projekten, die mit ihren Ansätzen einen kleinen Teil zur Weiterentwicklung des Gebiets beitragen.

Alles in allem ist dieses Gebiet ein Teilbereich, das zwar immer noch um sehr viele neue Erkenntnisse bereichert werden kann; das aber auf Grund der bestehenden Vielzahl an Ideen und Überlegungen, auf Grund der Praxistauglichkeit und insbesondere in Hinblick auf die stetig steigende Qualität der generierten Texte in nicht allzu ferner Zukunft gemeistert werden kann.

Anhang A

Entitäten und Fakten

Konzept „Mountain“

Entität	Height	Gebirge	County	First Ascent	Coordinates
Mount Everest	8848	Himalaya	Tibet	hillary	27°59’17”N 86°55’31”E
Mont Blanc	4810	Alps	france	balmat	45°50’01”N 006°51’54”E
Mount Elbrus	5642	caucasus	russia	grove	43°21’18”N 42°26’21”E
Mount McKinley	6194	north america	alaska	stuck	63°04’10”N 151°00’27”W
Mount Aconcagua	6962	andes	argentine	zurbriggen	32°39’20”S 70°00’57”W
Brocken	1141	harz	germany	1572	51°48’02”N 10°37’02”E
K2	8611	Karakoram	pakistan	Achille Compagnoni	35°52’57”N 76°30’48”E
Kangchenjunga	8586	Himalaya	nepal	Joe Brown	27°42’09”N 88°08’54”E
Lhotse	8516	Mahalangur Himal	china	Fritz Luchsinger	27°57’42”N 86°56’00”E
Makalu	8481	Himalaya	china	Lionel Terray	27°53’21”N 87°05’19”E
Cho Oyu	8201	Mahalangur	china	Herbert Tichy	28°05’39”N 86°39’39”E
Dhaulagiri	8167	Dhaulagiri Himal	nepal	first ascent	28°05’39”N 86°39’39”E
Manaslu	8156	Mansiri Himal	nepal	first ascent	28°33’0”N 84°33’35”E
Nanga Parbat	8126	baltistan	pakistan	Hermann Buhl	35°14’15”N 74°35’21”E
Annapurna	8091	Himalaya	nepal	Maurice Herzog	28°35’46”N 83°49’13”E
Shishapangma	8013	himalaya	tibet	first ascent	28°21’8”N 85°46’47”E
Broad Peak	8051	Karakoram	pakistan	first ascent	35°48’39”N 76°34’06”E
Monte San Valentin	4058	Andes	chile	first ascent	46°35’42”S 73°20’45”W
Vinson Massif	4896	Sentinel Range	Antarctica	Nicholas Clinch	78°31’31.74”S 85°37’1.73”W
Mount Kilimanjaro	5895	Sentinel Range	Tanzania	Hans Meyer	3°4’33”S 37°21’12”E

Konzept „Mobile Phone“

Entität	Manufacturer	Resolution	OS	Memory	Weight	Sar
Nokia 2323	nokia	128 x 160	Series 40 Version 5.1	32.00 MB	90 g	0.94

Samsung B2100	samsung	128 x 160	Samsung OS	10 MB	103 g	0,716
Nokia N8	nokia	360 x 640	Symbian 3	16 GB	135 g	1.020
Motorola Gleam	motorola	240 x 320	proprietary	32 MB	105 g	0,405
Motorola Wilder	motorola	240 x 320	Brew MP	5 MB	92 g	1,
Nokia C2-06	nokia	240 x 320	Nokia Series 40	10 MB	115 g	sar
Nokia C5	nokia	240 x 320	Symbian	50 MB	90 g	sar
Nokia X6	nokia	360 x 640	S60	16 GB	122 g	sar
Motorola Defy	motorola	480 x 854	Android	2 GB	118 g	sar
Samsung Galaxy Ace	samsung	320 x 480	Android	158 MB	113 g	sar
HTC Rhyme	htc	480 x 800	Android	4 GB	130 g	sar
Samsung Wave Y	Samsung	320 x 480	Bada	150 MB	103 g	0,318
Nokia 500	nokia	360 x 640	Serie 40	2 GB	93 g	0,82
Nokia 6300	nokia	240 x 320	Serie 40	8 MB	91 g	0,57
LG Optimus Pro	LG Handy	240 x 320	android	150 MB	129 g	0,55
Sony Ericsson XPERIA Active	sony	320 x 480	Windows Mobile	320 MB	111 g	1,03
HTC Sensation	htc	540 x 960	Android	1 GB	148 g	0,36
Google Nexus S	google	480 x 800	Android	16 GB	129 g	0,82
HTC EVO 3D	htc	560 x 960	Android	1 GB	170 g	0,52
Blackberry Torch 9860	blackberry	480 x 800	Blackberry	4 GB,	135 g	0,82

Konzept „Athlete“

Entität	Sport	Achievment	Birthday	Spouse	Team	Hobby
Michael Schumacher	racing driver	medal	January 3, 1969	Corinna Betsch	Mercedes GP team	Motorcycle
Steffi Graf	tennis player	gold medal	Juni 14, 1969	Andre Agassi	World Team Tennis	bicycle
Ronaldinho	football	Champions League	March 21, 1980	Flamengo	music	Ronaldinho
Dirk Nowitzki	basketball player	NBA Championship	June 19, 1978	Jessica Olsson	the Dallas Mavericks	saxophone
Timo Boll	table tennis	World Cup	March 8, 1981	Rodelia Jacobi	with Borussia Düsseldorf	electronic
Miroslav Klose	football	FIFA World Cup	June 9, 1978	Sylwia Klose	Lazio	fishing
Michael Ballack	football	Bundesliga	July 1, 1976	Simone Lambe	Bundesliga club Bayer Leverkusen	Golf

Tiger Woods	golf	PGA Tour	December 30, 1975	Elin Nordegren	team usa	reading
Kevin Garnett	basketball	NBA Champion	May 19, 1976	brandi Garnett	the Boston Celtics	pole vaulting
Carlos Delgado	baseball	all-time home run	June 25, 1972	Betzaida Garcia	Toronto Blue Jays	kino
Oscar De La Hoya	boxer	Gold Medal	February 4, 1973	Shanna Moakler	team usa	golf
Serena Williams	tennis	Grand Slam	September 26, 1981	Common	U.S. tennis team	Fashion
Michael Strahan	football	Player of the Year	November 21, 1971	Wanda Hutchins	New York Giants	cooking
Scottie Pippen	basketball	NBA Champion	September 25, 1965	Larsa Pippen	Chicago Bulls	fishing
Venus Williams	tennis	gold medal	June 17, 1980	Rodelia Jacobi	U.S. tennis team	electronic
Mo Vaughn	baseball	Silver Slugger Award winner	December 15, 1967	Gail Turkovich	New York Mets	Horse
Juwan Howard	basketball	NBA All-Star	February 7, 1973	Jenine Wardally	the Miami Heat	Golf
Manny Ramirez	baseball	All-Star	May 30, 1972	Juliana Ramirez	Tampa Bay Rays	Golf
Alex Rodriguez	baseball	All-Star	July 27, 1975	Cynthia Rodriguez;Violet Chang	the New York Yankees	fishing
Lennox Lewis	boxing	WBC Champion	September 2, 1965	Cynthia Rodriguez;Violet Chang	-	kino

Konzept „City“

Entität	Country	Population	Sight	Mayor	Transport	University
New York	usa	8,175,133	statue of liberty	Michael Bloomberg	mass transit	New York University
Tokyo	japan	13,185,502	Tokyo Tower	Shintaro Ishihara	rail	University of Tokyo
Madrid	spain	3,273,049	Plaza Mayor	Alberto Ruiz	metro	Complutense University of Madrid
Amsterdam	netherlands	780,152	Red light district	Eberhard van der Laan	Cycling	University of Amsterdam
Hong Kong	China	7,061,200	The Peak	Donald Tsang	tram	University of Hongkong
London	United Kingdom	7,825,200	tower bridge	Boris Johnson	bus	University of London
Moscow	russia	11,514,330	Kremlin	Sergey Sobyanin	Metro	Lomonosov Moscow State University
Budapest	Hungary	1,733,685	Parliament Building	István Tarlós	rail	Budapest University
Berlin	germany	3,471,756	alexander	Klaus Wowereit	metro	Humboldt University
Sydney	australia	4,575,532	Opera House	Lord Mayor	Metroads	University of Sydney
Kiev	Ukraine	2,611,300	World War II Museum	Leonid Chernovetskyi	rail	Kiev National Taras Shevchenko University
Rio De Janeiro	brazil	6,323,037	Christ the Redeemer	Eduardo Paes	metro	Rio de Janeiro State University

Chicago	usa	2,695,598	Water Tower	Rahm Emanuel	transit	University of Chicago
San Francisco	usa	7,468,390	golden gate bridge	Edwin M. Lee	San Francisco Municipal Railway	University of California
New Orleans	usa	343,829	French Quarter	Mitch Landrieu	streetcars	Tulane University
Minsk	Belarus	1,836,808	Cathedral of the Holy Spirit	Mikalai Ladutska	public transport system	Belarusian State University
Rome	italy	2,761,477	Colosseum	Gianni Alemanno	radial network of roads	John Felice Rome Center
Paris	France	10,247,794	Eiffel Tower	Bertrand Delanoë	head of barge	University of Paris
Oslo	norway	605,005	parliament of norway	Fabian Stang	Norway's most extensive public transport system,	University of Oslo
Stockholm	sweden	1,372,565	Ericsson Globe	Sten Nordin	extensive public transport system	University of stockholm

Konzept „University“

Entität	Established	Type	President	Country	City	Motto
TU Dresden	1828	public	Hans Müller-Steinhagen	germany	dresden	Wissen schafft Brücken
Harvard	Harvard	private	Drew Gilpin Faust	usa	Cambridge	Veritas
Princeton	1746	private	Tilghman	usa	princeton	Dei sub numine viget
Yale	1701	private	levin	usa	new haven	Lux et veritas
California Institute of Technology	1891	private	chameau	usa	pasadena	The truth shall make you free
Humboldt University of Berlin	1810	Public	Jan-Hendrik Olbertz	germany	berlin	Veritas, Iustitia, Libertas
University of Sydney	1850	Public	Marie Bashir	australia	sydney	Sidere mens eadem mutat
University of London	1836	Public	The Princess Royal	united kingdom	london	Esse quam videri
Moscow State University	1755	Public	Viktor Sadovnichiy	russia	moscow	Science is clear learning of truth and enlightenment of the mind
University of Oxford	1167	Public	Lord Patten of Barnes	united kingdom	oxford	Dominus Illuminatio Mea
Stockholm University	1878	public	Filip Solsjö	sweden	stockholm	Ad utrumque
University of Oslo	1811	public	Ole Petter Ottersen	norway	oslo	Universitas Osloensis
Belarusian State University	1921	public	Siarhey Ablameyka	belarus	minsk	The university that works for you
Tulane University	1834	private	Scott Cowen	usa	new orleans	Non Sibi Sed Suis
University of California	1868	public	Mark Yudof	usa	Oakland	Fiat lux

University of Chicago	1890	Private	Robert Zimmer	usa	chicago	Crescat scientia; vita excolatur
University of Georgia	1785	Public	Michael F. Adams	usa	Athens	Et docere et rerum exquirere causas
University of Leipzig	1409	Public	Beate Schücking	germany	leipzig	Crossing Boundaries out of Tradition
University of Munich	1472	Public	Bernd Huber	germany	munich	Science is clear learning of truth and enlightenment of the mind
Complutense University of Madrid	1293	Public	José Carrillo	spain	madrid	Libertas Perfundet Omnia Luce

Konzept „Celebrity“

	FullName	Birthday	Birthplace	Origin	Job
Jim Carrey	James "Jim\" Eugene Carrey; James Eugene Redmond Carrey; James Eugene Carrey	January 17th 1962; 17th January 1962; january 1962; 17.01.1962; January 17; 1962	Newmarket,Ontario; Ontario Newmarket; Newmarket,Ontario; Ontario,Newmarket; Newmarket-Ontario; Ontario-Newmarket; Newmarket Ontario; Ontario, Newmarket	Canadian merican; Canadia,;Canada; American canadian; Canadian, american; American, Canadian; Canadian-American; Canadian-American	Comedian Actor
Lady Gaga	Stefani Joanne Angelina Germanotta; Stefani Germanotta;	28.03.1986; 03.28.1986; 28th march 1986; march 1986; march 28th 1986; march 28, 1986;	New York	am erican	Singer; Composer; Writer
Tom Cruise	Thomas Cruise Mapother	07.03.1986; 03.07.1986; 3rd july 1986; july 1986; july 3rd 1986; july 3, 1986	New York	american	producer; actor
Keanu Reeves	Keanu Charles Reeves; Keanu Reeves	02.09.1964; 09.02.1964; 2nd september 1964;september 1964;september 2nd 1964; september 2, 1964	Beirut	canadian	actor; musician; bassist
Jackie Chan	Chan Kong-sang;Chan Kong sang	07.04.1954;04.07.1954;7th april 1954;april 1954;april 7th 1954;april 7, 1954	Hong Kong	chinese	actor;choreographer;filmmaker
Barack Obama	Barack Hussein Obama	08.04.1961;04.08.1961;8th august 1961;august 1961;august 8th 1961;august 8, 1961	Honolulu	hawaii	president;senator;politician
Adam Sandler	Adam Richard Sandler	09.09.1966;9.9.1966;9th september 1966;september 1966;september 9th 1966;september 9, 1966	Brooklyn	USA	comedian;actor
Eddie Murphy	Edward "Eddie" Regan	03.04.1961;04.03.1961;3rd april 1961;april	Brooklyn	USA	comedian;actor

	Murphy	1961;april 3rd 1961;april 3, 1961			
Mike Meyers	Michael John "Mike" Myers;Michael John Myers	25.05.1963;05.25.1963;25th may 1963;may 1963;may 25th 1963;may 25, 1963	Ontario	canadian	comedian;actor;producer
Leonardo DiCaprio	Leonardo Wilhelm DiCaprio	11.11.1974; november 11 1974; 11th november 1974; november 1974; november 11th 1974; november 11, 1974	Los Angeles	american	producer;actor
Bruce Willies	Walter Bruce Willis	19.03.1955; 03.19.1955; 19th march 1955; march 1955; march 19th 1955; march 19, 1955	Idar-Oberstein	american	producer;actor;musician
Ben Stiller	Benjamin "Ben" Edward Stiller;Benjamin Edward Stiller	30.11.1965;11.30.1965;30th november 1965;november 1965;november 30th 1965;november 30, 1965	New York	american	producer;actor;comedian
Nicolas Cage	Nicholas Kim Coppola	07.01.1964;01.07.1964;7th january 1964;january 1964;january 7th 1964;january 7, 1964	Long Beach	american	producer;actor;director
Will Ferrell	John William "Will" Ferrell;John William Ferrell	16.07.1967;7.16.1967;7th july 1967;july 1967;july 7th 1967;july 7, 1967	Irvine	american	impressionist;actor;comedian
Johnny Depp	John Christopher "Johnny" Depp II; John Christopher "Johnny" Depp ;John Christopher Depp II	9.06.1963;6.9.1963;9th june 1963;june 1963;june 9th 1963;june 9, 1963	Owensboro	american	musician;actor
Angelina Jolie	Angelina Jolie Voight	4.06.1975;6.4.1975;4th june 1975;june 1975;june 4th 1975;june 4, 1975	Los Angeles	american	model;actress
Madonna	Madonna Louise Ciccone	16.08.1958;8.16.1958;16th august 1958;august 1958;august 16th 1958;august 16, 1958	Bay City	american	singer;entrepreneur;recording
Meg Ryan	Margaret Mary Emily Anne Hyra	19.11.1961	Bethel	american	actress;producer
Cameron Diaz	Cameron Michelle Diaz	30.8.1972;8.30.1972;30th august 1972;august 1972;august 30th 1972;august 30, 1972	San Diego	american	actress;model
Alyssa Milano	Alyssa Jayne Milano	19.12.1972;12.19.1972;19th december 1972;december 1972;december 19th 1972;december 19, 1972	Brooklyn	american	actress;singer;producer
Matt Damon	Matthew Paige "Matt" Damon	October 8th 1970	Cambridge, Massachusetts	American;Canadian	Comedian;screenwriter;philanthropist
Angela Merkel	Angela Dorothea Merkel	July 17th 1954;07.17.1954;17th July 1954;July 1954;17.07.1954;July 17, 1954	Hamburg	german;germany	politician;Chancellor
Charlie Sheen	Carlos Irwin	03.03.1965	New York	american	actor;comedian

	Estevez;Carlos Estevez				
Bill O'Reilly	William James "Bill" O'Reilly, Jr.;William James O'Reilly, Jr.;William James O'Reilly	September 10th 1949;09.10.1949;10th September 1949;September 1949;10.09.1949;September 10, 1949	Manhattan, New York;New York Manhattan;Manhattan,New York;New York,Manhattan;Manhattan-New York;New York-Manhattan;Manhattan New York;New York, Manhattan	American;America	commentator;television host
Ellen DeGeneres	Ellen Lee DeGeneres	January 26th 1958;01.26.1958;26th January 1958;january 1958;26.01.1958;January 26, 1958	Metairie, Louisiana;Louisiana Metairie;Louisiana,Metairie;Metairie-Louisiana;Louisiana-Metairie;Metairie Louisiana;Louisiana, Metairie	American;america	Comedian;Actor
Oprah Winfrey	Orpah Gail Winfrey;William James O'Reilly, Jr.;William James O'Reilly	January 29th 1954;01.29.1954;29th January 1954;january 1954;29.01.1954;January 29, 1954	Kosciusko, Mississippi;Mississippi Kosciusko;Kosciusko,Mississippi;Mississippi,Kosciusko;Kosciusko-Mississippi;Mississippi-Kosciusko;Kosciusko Mississippi;Mississippi, Kosciusko	American;America	media proprietor;businesswoman;talk show host
Chuck Norris	Carlos Ray "Chuck" Norris;Carlos Ray Norris;Carlos Norris	March 10th 1940;03.10.1940;10th March 1940;March 1940;10.03.1940;March 10, 1940	Ryan, Oklahoma;Ryan oklahoma;oklahoma,Ryan;Ryan,oklahoma;oklahoma-Ryan;Ryan-oklahoma;oklahoma Ryan;Ryan, Newmarket	American	martial artist;Actor
David Letterman	David Michael Letterman	April 12th 1947;04.12.1947;12th April 1947;April 1947;12.04.1947;April 12, 1947	Indianapolis, Indiana;Indianapolis Indiana;Indiana,Indianapolis;Indianapolis,Indiana;Indiana-Indianapolis;Indianapolis-Indiana;Indiana Indianapolis	American;america	television host;comedian
Seth Green	Seth Benjamin Gesshel-Green;Seth Gesshel-Green;Seth Benjamin Green	February 8th 1974;02.8.1974;8th February 1974;February 1974;8.02.1974;February 8, 1974	Overbrook Park, west philadelphia;west philadelphia Overbrook Park;Overbrook Park,west philadelphia;Overbrook Park-west philadelphia;west philadelphia-Overbrook Park;Overbrook Park west philadelphia;west philadelphia, Overbrook Park	American;America	Comedian;Actor;television producer
Britney Spears	Britney Jean Spears	December 2th 1981;12.2.1981;2th December 1981;December 1981;2.12.1981;December 2, 1981	Kentwood, Louisiana;kentwood louisiana;louisiana,kentwood;kentwood,louisiana;louisiana-kentwood;kentwood-louisiana;louisiana kentwood;kentwood, Newmarket	america	recording artist;entertainer
Robbie Williams	Robert Peter "Robbie" Williams;Robert Peter Williams;Robert Williams	February 13th 1974;01.13.1974;13th February 1974;February 1974;13.01.1974;February 13, 1974	Stoke-on-Trent, Staffordshire;Staffordshire Stoke-on-Trent;Stoke-on-Trent,Staffordshire;Staffordshire,Stoke-on-Trent;Stoke-on-Trent-Staffordshire;Staffordshire-Stoke-on-Trent;Stoke-on-Trent Staffordshire;Staffordshire, Stoke-on-Trent	british	singer;Actor
Brendan Fraser	Brendan James Fraser	December 3th 1968;12.3.1968;3th December 1968;December 1968;3.12.1968;December 3, 1968	Indianapolis, Indiana;Indianapolis Indiana;Indiana,Indianapolis;Indianapolis,Indiana;Indiana-Indianapolis;Indianapolis-Indiana;Indiana Indianapolis	Canadian American;Canadian;Canada;American Canadian;Canadian, American;American, Canadian;Canadian-American	Actor
Michael Caine	Maurice Joseph Micklewhite;Maurice Micklewhite;Sir Michael Caine	March 14th 1933;01.14.1933;14th March 1933;March 1933;14.01.1933;March 14, 1933	Rotherhithe, southwark;southwark Rotherhithe;Rotherhithe,Southwark;southwark,Rotherhithe;Rotherhithe-southwark;southwark-Rotherhithe;Rotherhithe southwark;southwark, Rotherhithe	british	Actor
James Hetfield	James Alan Hetfield	August 3th 1963;17.3.1963;3th August 1963;August 1963;3.17.1963;August 3, 1963	Downey, California;downey california;california,downey;downey,california;california-downey;downey-california;california downey;downey, Newmarket	america	lead vocalis;main songwriter;rhythm guitarist

Hugh Laurie	James Hugh Calum Laurie;James Hugh Laurie;James Calum Laurie	June 11th 1959;06.11.1959;11th June 1959;June 1959;11.06.1959;June 17, 1959	england, oxford;oxford england;england,oxford;oxford,england;england-oxford;oxford-england;england oxford;oxford, england	british	comedian;Actor;voice artist
Rowan Atkinson	Rowan Sebastian Atkinson	January 6th 1955;01.6.1955;6th January 1955;january 1955;6.01.1955;January 6, 1955	county durham, consett;consett county durham;Consett, County Durham;consett,county durham;county durham-consett;consett-county durham;county durham consett	british	Comedian;Actor
Phil Collins	Philip David Charles "Phil" Collins;Philip David Charles Collins;Philip David Collins	January 30th 1951;01.30.1951;30th January 1951;january 1951;30.01.1951;January 30, 1951	chiswick, london;london chiswick;chiswick,london;london,chiswick;chiswick-london;london-chiswick;chiswick london;london, chiswick	british	singer;songwriter;vocalist
John Lennon	John Winston Ono Lennon;John Winston Lennon;John Ono Lennon	October 8th 1940;10.8.1940;8th October 1940;October 1940;8.10.1940;October 8, 1940	Liverpool, england;england Liverpool;Liverpool,england;england,Liverpool;Liverpool-england;england-Liverpool;Liverpool england;england, Liverpool	british	singer;musician
Bob Marley	Robert Nesta "Bob" Marley;Robert Nesta Marley;Robert Marley	February 6th 1945;02.6.1945;6th February 1945;February 1945;6.02.1945;February 8, 1945	Saint Ann Parish, nine mile;nine mile Saint Ann Parish;Nine Mile,Saint Ann Parish;Saint Ann Parish-nine mile;nine mile-Saint Ann Parish;Saint Ann Parish nine mile;nine mile, Saint Ann Parish	Jamaica	singer;musician

	Leistung	award	Child	Father	Mother	School	Spouse
Jim Carrey	The Truman Show, Man on the Moon, Ace Ventura	Golden Globe	Jane Carrey	Perry	Kathleen	Blessed Trinity Catholic School	Melissa Womer; Lauren Holly
Lady Gaga	Poker Face; Just Dance; Paparazzi	Grammy Award	-	Joseph Germanotta	Cynthia	Convent of the Sacred Heart	-
Tom Cruise	Mission Impossible; Top Gun; The Last Samurai	Golden Globe	Isabella Jane; Connor Antony; Suri	Thomas	Mary Lee	Robert Hopkins Public School	Mimi Rogers; Nicole Kidman; Katie Holmes
Keanu Reeves	speed; the matrix; Bill and Ted's Excellent Adventure	Taurus Stunt Award	Ava Archer	Samuel Nowlin	Patricia Bond	Etobicoke School of the Arts	Jennifer Syme
Jackie Chan	battle creek brawl;rush hour;drunken master	award	Jaycee Chan	Charles	Lee-Lee	Nah-Hwa Primary School	Lin Feng-jiao
Barack Obama	juris doctor;Senate campaign;presidential campaign	Nobel Peace Prize	Malia;Sasha	Barack Obama, Sr	Stanley Ann Dunham	Besuki Public School	Michelle Robinson
Adam Sandler	deeds;larry;waterboy	golden globe	Sadie Madison; Sunny Madeline	Stanley	Judy	Manchester Central High School	Jacqueline Samantha Titone
Eddie Murphy	pluto nash;Beverly Hills Cop;Trading Places	golden globe	Eric Murphy;Christian Murphy;Miles Mitchell	Charles Edward	Lillian	Roosevelt Junior-Senior High School	Nicole Mitchell;Tracey Edmonds;Melanie Brown
Mike Myers	austin powers;wayne;shrek	golden globe	-	Eric Myers	Alice E. Hind	Bishopbriggs Academy	Robin Ruzan;Kelly Tisdale
Leonardo DiCaprio	titanic;aviator; departed	oscar	-	George DiCaprio	Irmelin	John Marshall High School	Bar Refaeli;Gisele Bündchen;Emma Miller

Bruce Willis	die hard;armageddon; the fifth element	oscar	Rumer;Scout; Tallulah	David Willis	Marlene	Penns Grove High School	Demi Moore;Brooke Burns;Emma Heming
Ben Stiller	zooloander;Something About Mary;Dodgeball	oscar	Quinlin Dempsey;Ella Olivia	Jerry Stiller	Anne Meara	Calhoun School	Christine Taylor
Nicolas Cage	the rock;ghost rider;Adaptation	golden globe	Kal-El	August Coppola	Joy Vogelsang	Beverly Hills High School	Patricia Arquette;Lisa Marie Presley;Alice Kim
Will Ferrell	old school;The Other Guys;elf	golden globe	Magnus Paulin;Mattias Paulin;Axel Paulin	Roy Lee Ferrell	Betty Kay	Rancho San Joaquin Middle School	Viveca Paulin
Johnny Depp	Edward Scissorhands;Pirates of the Caribbean;Charlie and the Chocolate Factory	golden globe	Lily-Rose Melody;John Christopher	John Christopher Depp	Betty Sue Wells	-	Winona Ryder;Kate Moss;Vanessa Paradis
Angelina Jolie	Lara Croft;Mr. and Mrs. Smith;Changeling	golden globe	Shiloh Nouvel;Knox; Vivienne Marcheline	Jon Voight	Marcheline Bertrand	Beverly Hills High School	Jonny Lee Miller;Billy Bob Thornton;Brad Pitt
Madonna	Like a Virgin;vogue;hung up	golden globe	Lourdes Maria Ciccone Leon;Rocco Ritchie;David Banda Mwale Ciccone Richie	Silvio Anthony Ciccone	Madonna Louise	Rochester Adams High School	Sean Penn;Guy Ritchie
Meg Ryan	Sleepless in Seattle;City of Angels;When a Man Loves a Woman	golden globe	Jack Henry;Daisy True	Harry Hyra	Susan Jordan	Bethel High School	Dennis Quaid;John Mellencamp
Cameron Diaz	The Mask;My Best Friend's Wedding;Something About Mary	golden globe	-	Emilio Diaz	Billie	Long Beach Polytechnic High School	Matt Dillon;Justin Timberlake;Alex Rodriguez
Alyssa Milano	who is the boss;charmed;Melrose Place	Young Artist Award	-	Thomas M. Milano	Lin	Buckley School	Cinjun Tate;David Bugliari ;Alex Rodriguez
Matt Damon	Good Will Hunting;Saving Private Ryan;The Talented Mr. Ripley	Golden Globe	Isabella; Gia Zavala; Stella Zavala	Kent Telfer Damon	Nancy Carlsson-Paige	Cambridge Alternative School	Luciana Bozán Barroso;Minnie Driver;Alex Rodriguez
	Leistung	award	Child	Father	Mother	School	Spouse
Jim Carrey	The Truman Show, Man on the Moon, Ace Ventura	Golden Globe	Jane Carrey	Perry	Kathleen	Blessed Trinity Catholic School	Melissa Womer; Lauren Holly
Lady Gaga	Poker Face; Just Dance; Paparazzi	Grammy Award	-	Joseph Germanotta	Cynthia	Convent of the Sacred Heart	-
Tom Cruise	Mission Impossible; Top Gun; The Last Samurai	Golden Globe	Isabella Jane; Connor Antony; Suri	Thomas	Mary Lee	Robert Hopkins Public School	Mimi Rogers; Nicole Kidman; Katie Holmes
Keanu Reeves	speed; the matrix; Bill and Ted's Excellent Adventure	Taurus Stunt Award	Ava Archer	Samuel Nowlin	Patricia Bond	Etobicoke School of the Arts	Jennifer Syme
Jackie Chan	battle creek brawl;rush hour;drunken master	award	Jaycee Chan	Charles	Lee-Lee	Nah-Hwa Primary School	Lin Feng-jiao
Barack Obama	juris doctor;Senate campaign;presidential campaign	Nobel Peace Prize	Malia;Sasha	Barack Obama, Sr	Stanley Ann Dunham	Besuki Public School	Michelle Robinson
Adam Sandler	deeds;larry;waterboy	golden globe	Sadie Madison; Sunny	Stanley	Judy	Manchester Central High School	Jacqueline Samantha Titone

			Madeline				
Eddie Murphy	pluto nash;Beverly Hills Cop;Trading Places	golden globe	Eric Murphy;Christian Murphy;Miles Mitchell	Charles Edward	Lillian	Roosevelt Junior-Senior High School	Nicole Mitchell;Tracey Edmonds;Melanie Brown
Mike Myers	austin powers;wayne;shrek	golden globe	-	Eric Myers	Alice E. Hind	Bishopbriggs Academy	Robin Ruzan;Kelly Tisdale
Leonardo DiCaprio	titanic;aviator; departed	oscar	-	George DiCaprio	Irmelin	John Marshall High School	Bar Refaeli;Gisele Bündchen;Emma Miller
Bruce Willis	die hard;armageddon; the fifth element	oscar	Rumer;Scout; Tallulah	David Willis	Marlene	Penns Grove High School	Demi Moore;Brooke Burns;Emma Heming
Ben Stiller	zooloander;Something About Mary;Dodgeball	oscar	Quinlin Dempsey;Ella Olivia	Jerry Stiller	Anne Meara	Calhoun School	Christine Taylor
Nicolas Cage	the rock;ghost rider;Adaptation	golden globe	Kal-El	August Coppola	Joy Vogelsang	Beverly Hills High School	Patricia Arquette;Lisa Marie Presley;Alice Kim
Will Ferrell	old school;The Other Guys;elf	golden globe	Magnus Paulin;Matti as Paulin;Axel Paulin	Roy Lee Ferrell	Betty Kay	Rancho San Joaquin Middle School	Viveca Paulin
Johnny Depp	Edward Scissorhands;Pirates of the Caribbean;Charlie and the Chocolate Factory	golden globe	Lily-Rose Melody;John Christopher	John Christopher Depp	Betty Sue Wells	-	Winona Ryder;Kate Moss;Vanessa Paradis
Angelina Jolie	Lara Croft;Mr. and Mrs. Smith;Changeling	golden globe	Shiloh Nouvel;Knox; Vivienne Marcheline	Jon Voight	Marcheline Bertrand	Beverly Hills High School	Jonny Lee Miller;Billy Bob Thornton;Brad Pitt
Madonna	Like a Virgin;vogue;hung up	golden globe	Lourdes Maria Ciccone Leon;Rocco Ritchie;David Banda Mwale Ciccone Richie	Silvio Anthony Ciccone	Madonna Louise	Rochester Adams High School	Sean Penn;Guy Ritchie
Meg Ryan	Sleepless in Seattle;City of Angels;When a Man Loves a Woman	golden globe	Jack Henry;Daisy True	Harry Hyra	Susan Jordan	Bethel High School	Dennis Quaid;John Mellencamp
Cameron Diaz	The Mask;My Best Friend's Wedding;Something About Mary	golden globe	-	Emilio Diaz	Billie	Long Beach Polytechnic High School	Matt Dillon;Justin Timberlake;Alex Rodriguez
Alyssa Milano	who is the boss;charmed;Melrose Place	Young Artist Award	-	Thomas M. Milano	Lin	Buckley School	Cinjun Tate;David Bugliari ;Alex Rodriguez
Matt Damon	Good Will Hunting;Saving Private Ryan;The Talented Mr. Ripley	Golden Globe	Isabella; Gia Zavala; Stella Zavala	Kent Telfer Damon	Nancy Carlsson-Paige	Cambridge Alternative School	Luciana Bozán Barroso;Minnie Driver;Alex Rodriguez
Angela Merkel	physics;President of the European Council;honorary doctorate	Vision for Europe Award	Horst Kasner	Herlind	Blessed Trinity Catholic School	-	Ulrich Merkel;Joachim Sauer
Charlie Sheen	two and a half men;hot shots;spin city	Golden Globe	Martin Sheen	Janet Templeton	Santa Monica High School	Cassandra Jade Estevez;Sam Sheen;Bob Sheen	Paula Profit;Denise Richards;Brooke Mueller
Bill O'Reilly	The O'Reilly Factor;Inside Edition;news reporter	National Academy of Television Arts and Sciences	Madeline	William James, Sr.	Winifred Angela Drake	Harvard University	Maureen E. McPhilmy;Maureen McPhilmy

Ellen DeGeneres	The Ellen Show;Ellen;Ace Ventura	Academy Awards	Elliott Everett DeGeneres	Elizabeth Jane DeGeneres	-	Grace King High School	Portia de Rossi
Oprah Winfrey	The Oprah Winfrey Show;O, The Oprah Magazine;Oprah.com	Academy Award	Vernon Winfrey	Vernita Lee	-	Lincoln High School	Stedman Graham
Chuck Norris	Walker, Texas Ranger;Chuck Norris facts;martial arts	Veteran of the Year award	Mike;Dina	ray norris	Wilma	National Council on Bible Curriculum in Public School	Dianne Holechek;Gena O'Kelley
David Letterman	Late Show with David Letterman;Late Night with David Letterman;Everybody Loves Raymond	Emmy Award	Harry Joseph Letterman	Harry Joseph Letterman	Dorothy Letterman	Broad Ripple High School	Michelle Cook;Regina Lasko
Seth Green	Austin Powers;Buffy the Vampire Slayer;The Italian Job	Emmy Awards	Herbert	Barbara	Clare Grant	-	-
Britney Spears	...Baby One More Time;Oops!... I Did It Again;toxic	MTV Movie Awards	Sean Preston	James Parnell Spears	Lynne Irene	Parklane Academy	Kevin Federline;Jason Allen Alexander
Robbie Williams	Take That;Life Thru a Lens;Sing When You're Winning	Echo Awards	Peter	-	-	-	-
Brendan Fraser	George of the Jungle;Monkeybone;the mummy	Screen Actors Guild	Griffin Arthur Fraser;Holden Fletcher Fraser;Leland Francis Fraser	Peter	Carol	Upper Canada College	Afton Smith
Michael Caine	The Italian Job;Muppet Christmas Carol;The Dark Knight	Academy Award	Dominique	Maurice Joseph Micklewhite	Ellen Frances Marie	North Runcton	Patricia Haines;Shakira Baksh
Elizabeth II	Commonwealth of Nations;Trooping the Colour	Head of the Commonwealth	Prince Albert, Duke of York	Elizabeth	Charles, Prince of Wales;Anne, Princess Royal;Prince Andrew, Duke of York	-	-
James Hetfield	Metallica;The 100 Greatest Metal Guitarists;Hit Parader	Grammy Award	Cali;Castor; Marcella	Virgil	Cynthia	Downey High School	Francesca Tomasi
Hugh Laurie	house;Blackadder;Jeeves and Wooster	Golden Globe	Charles Archibald;William Albert;Rebecca Augusta	Ran Laurie	Patricia	Selwyn College	Jo Green
Rowan Atkinson	blackadder;Not The Nine O'Clock News;Mr. Bean	Variety Club Award	Benjamin;Lily	Eric Atkinson	Ella May	Durham Choristers School	Sunetra Sastry
Phil Collins	In the Air Tonight;Another Day in Paradise;Sussudio	Grammy Awards	Simon Collins;Lily Collins	Greville Collins	june	Chiswick Community School	Andrea Bertorelli;Orianne Cevey
John Lennon	yellow submarine;the beatles;All you need is love	Grammy Award	Julian;Sean	alfred	julia	Dovedale Primary School	Yoko ono;Cynthia Powell
Bob Marley	I Shot the Sheriff;buffalo soldier;Three Little Birds	Grammy Award	Sharon;Cedella	Norval Sinclair Marley	Cedella Booker	Dovedale Primary School	rita;Yvette

Phrasen im Detail

Konzept Mountain

M1:	it is located in [%county%]
M2:	[%name%] is the highest mountain in the [%gebirge%]
M3:	[%name%] is located in southcentral [%county%]
M4:	[%name%], located in the [%gebirge%], has a height of [%height%]
M5:	it is the [%first_ascent%]
M6:	[%name%]'s height is [%height%]
M7:	the coordinates of [%coordinates%]

Konzept Athlete

A1:	it is a [%sport%]
A2:	it is an american [%sport%]
A3:	[%name%] is a [%sport%] player
A4:	[%name%] is a professional [%sport%] player
A5:	[%name%] is a major league [%sport%]
A6:	[%name%] is an power forward for [%team%]
A7:	he won the [%achievement%]
A8:	it was the first [%achievement%]
A9:	it was born [%birthday%]
A10:	his spouse is [%spouse%]
A11:	his hobby is [%hobby%]

Konzept University

U1:	the university of [%city%] is the oldest university in [%country%]
U2:	[%city%] is the city of [%country%]
U3:	[%city%] is the largest city in [%country%]
U4:	[%city%] is the city in [%country%]
U5:	it is a [%type%]
U6:	the university of [%city%] is a [%type%] university
U7:	it was founded in [%established%]
U8:	[%name%] is a [%type%]
U9:	his motto is [%motto%]
U10:	his president is [%president%]

Konzept City

Ci1:	[%name%] is the capital of [%country%]
Ci2:	it is [%name%] [%country%] country
Ci3:	[%name%] is the capital city of [%country%]
Ci4:	[%name%] is the city of [%country%]
Ci5:	[%name%] is the centre of [%country%]
Ci6:	[%name%] is the largest city in [%country%]
Ci7:	[%name%] is the city in [%country%]
Ci8:	it is located in [%name%]
Ci9:	one place to visit is [%sight%]
Ci10:	the [%university%] is the university in [%country%]
Ci11:	the [%university%] is the largest university in [%country%]

Ci12: the population is [%population%]
Ci13: his transportation is [%transportation%]
Ci14: his mayor is [%mayor%]

Konzept Mobile
Mo1: it is a [%manufacturer%]
Mo2: it is the latest [%manufacturer%]
Mo3: it get the [%manufacturer%]
Mo4: it is from [%manufacturer%]
Mo5: it comes with [%manufacturer%]
Mo6: it is the first [%operating_system%]
Mo7: it is a [%operating_system%]
Mo8: it is an [%operating_system%]
Mo9: it is running on [%operating_system%]
Mo10: it is the new [%operating_system%]
Mo11: it is running [%operating_system%]
Mo12: it runs [%operating_system%]
Mo13: it has [%resolution%] pixels
Mo14: it comes with [%resolution%] pixels
Mo15: it comes with [%memory%]
Mo16: it has [%memory%] memory
Mo17: his weight is [%weight%]
Mo18: his sar is [%sar%]

Konzept Celebrity
Ce1: [%name%] is an american [%job%]
Ce2: [%name%] is an [%job%]
Ce3: it is an american [%job%]
Ce4: it is an [%job%]
Ce5: it was born on [%birthday%]
Ce6: [%name%] was born in [%birthplace%]
Ce7: [%name%] is the child of [%father%] and [%mother%]
Ce8: [%name%] born [%fullname%] is an american [%job%]
Ce9: it is a [%origin%]
Ce10: it is a [%origin%] actor
Ce11: it is an [%origin%]
Ce12: [%fullname%] was born in [%birthplace%]
Ce13: [%name%] was born on [%birthday%]
Ce14: it is the [%leistung%]
Ce15: it is married to [%spouse%]
Ce16: he met [%spouse%]
Ce17: it was married to [%spouse%]
Ce18: she won a [%award%]
Ce19: it be found here [%award%]
Ce20: he won a [%award%]
Ce21: it was awarded the [%award%]
Ce22: he won the [%award%]
Ce23: it has two children [%child%]
Ce24: his school is [%school%]

Anhang B – Gesamttext Evaluation

Generierte Texte:

Text 1 – Konzept „Celebrity“, Entity „Meg Ryan“

“Meg Ryan born Margaret Mary Emily Anne Hyra is an american actress and producer. She was born in Bethel and her school is Bethel High School. In addition, she has two children Jack Henry and Daisy True and she was awarded the golden globe. She met Dennis Quaid and John Mellencamp. Also noteworthy, she is the Sleepless in Seattle, City of Angels and When a Man Loves a Woman and was born on 19.11.1961. She is an american and the child of Harry Hyra and Susan Jordan.”

Text 2 – Konzept „Mountain“, Entity „Monte San Valentin“

“Monte San Valentin is located in southcentral chile. Its height is 4058 m and is the first ascent. The coordinates of 46°35'42"S 73°20'45"W and is the highest mountain in the Andes.”

Text 3 – Konzept „Mobile“, Entity „Samsung Wave Y“

“Samsung Wave Y is the latest Samsung. Its sar is 0,318 and its weight is 103 g. It comes with 320 x 480 pixels and has 150 MB memory. It is running on Bada.”

Text 4 – Konzept „Athlete“, Entity „Dirk Nowitzki“

“Dirk Nowitzki is a basketball player. His hobby is saxophone and his spouse is Jessica Olsson. In addition, he won the NBA Championship and was born June 19, 1978. He is a power forward for the Dallas Mavericks.”

Text 5 – Konzept „City“, Entity „Tokyo“

“Tokyo is the centre of japan. Its population is 13,185,502 and its transportation is rail. Furthermore, its mayor is Shintaro Ishihara and one place to visit is Tokyo Tower. The University of Tokyo is the largest university in japan.”

Text 6 – Konzept „University“, Entity „Humbold University of Berlin“

“Humboldt University of Berlin is public. Its motto is Veritas, Iustitia, Libertas and Berlin is the largest city in germany. Its president is Jan-Hendrik Olbertz and was founded in 1810.”

Text 7 – Konzept „City“, Entity „Berlin“

“Berlin is the largest city in germany. Its population is 3,471,756 and its transportation is metro. Furthermore, its mayor is Klaus Wowereit and one place to visit is alexander. The Humboldt University is the largest university in germany.”

Text 8 – Konzept „Celebrity“, Entity „Matt Damon“

“Matthew Paige "Matt" Damon was born in Cambridge, Massachusetts. He is an american Comedian, screenwriter and philanthropist and his school is Cambridge Alternative School. In addition, he has two children Isabella, Gia Zavala and Stella Zavala and he won a Golden Globe. He was married to Luciana Bozán Barroso, Minnie Driver and Alex Rodriguez. He is the Good Will Hunting, Saving Private Ryan and The Talented Mr. Ripley and Matt Damon was born on October 8th 1970. He is an American and Canadian. Also noteworthy, Matt Damon is the child of Kent Telfer Damon and Nancy Carlsson-Paige.”

Text 9 – Konzept „Mobile“, Entity „LG Optimus Pro“

“LG Optimus Pro is a LG Handy. Its sar is 0,55 and its weight is 129 g. In addition, it has 240 x 320 pixels and comes with 150 MB. It is an android.”

Text 10 – Konzept „Mountain“, Entity „Vinson Massif“

“Vinson Massif is located in southcentral Antarctica. Its height is 4896 m and is the Nicholas Clinch. Furthermore, the coordinates of 78°31'31.74"S 85°37'1.73"W and it is the highest mountain in the Sentinel Range.”

Auszüge aus Wikipedia:

Text 11 – Konzept „Celebrity“, Entity „Adam Sandler“
Quelle: http://en.wikipedia.org/wiki/Adam_Sandler

“Adam Richard Sandler (born September 9, 1966) is an American actor, comedian, singer, screenwriter, musician, and film producer. He is best known for his comedic roles, such as in the films Billy Madison (1995), Happy Gilmore (1996), Big Daddy (1999), and Mr. Deeds (2002). […]Adam Sandler was born in Brooklyn, New York to Jewish parents, Stanley, an electrical engineer, and Judy Sandler, a nursery school teacher. When he was five, his family moved to Manchester, New Hampshire, where he attended Manchester Central High School.

[…] Sandler married actress Jacqueline Samantha Titone, and they are the parents of two daughters, Sadie Madison Sandler and Sunny Madeline Sandler."

Text 12 – Konzept „Mountain", Entity „Aconcagua"
Quelle:(http://en.wikipedia.org/wiki/Aconcagua)

"Aconcagua is the highest mountain in the Americas at 6,962 m. It is located in the Andes mountain range, in the Argentine province of Mendoza. […] The coordinates are 32°39'12.35"S 70°00'39.9"W.[…] The first attempt on Aconcagua by a European was made in 1883 by a party led by the German geologist and explorer Paul Güssfeldt."

Text 13 – Konzept „City", Entity „Budapest"
Quelle: (http://en.wikipedia.org/wiki/Minsk)

"Budapest […] is the capital of Hungary. […] In 2011, Budapest had 1,733,685 inhabitants. The current mayor is István Tarlós. The neo-Gothic Parliament, containing amongst other things the Hungarian Crown Jewels. […]There are three main railway termini in Budapest, Keleti (eastbound), Nyugati (westbound), and Déli (southbound), operating both domestic and international rail services. Budapest is Hungary's main centre of education and home to […] the Budapest University of Technology and Economics."

Anhang C – Detaillierte Ergebnisse der Online Auswertung

Seite 2: Einschätzung Meg Ryan

[2.1]: Text (text/picture)

Text 1:

"Meg Ryan born Margaret Mary Emily Anne Hyra is an american actress and producer. She was born in Bethel and her school is Bethel High School. In addition, she has two children Jack Henry and Daisy True and she was awarded the golden globe. She met Dennis Quaid and John Mellencamp. Also noteworthy, she is the Sleepless in Seattle, City of Angels and When a Man Loves a Woman and was born on 19.11.1961. She is an american and the child of Harry Hyra and Susan Jordan."

[2.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v1	Ist dieser Text grammatikalisch korrekt?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[2.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v2	Finden sie Rechtschreibfehler?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[2.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v3	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	3.33 (3.33)

[2.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v4	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

[2.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v5	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[2.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v6	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	3.33 (3.33)

[2.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v7	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[2.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v8	Glauben sie, dass dieser Text von einem Computer	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

stammt?						

Seite 3: Einschätzung Monte San Valentin

[3.1]: Text (text/picture)

Text 2:

“Monte San Valentin is located in southcentral chile. Its height is 4058 m and is the first ascent. The coordinates of 46°35′42″S 73°20′45″W and is the highest mountain in the Andes.”

[3.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v9	Ist dieser Text grammatikalisch korrekt?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[3.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v10	Finden sie Rechtschreibfehler?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[3.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v11	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[3.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v12	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[3.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v13	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[3.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v14	Wie gefällt ihnen der	0 (0%)	0 (0%)	66.67%	33.33%	0 (0%)	3.33 (3.33)

Ausdrucksstil des Textes?			(2)	(1)		

[3.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v15	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[3.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v16	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	33.33% (1)	33.33% (1)	33.33% (1)	4 (4)

Seite 4: Einschätzung Samsung Wave Y

[4.1]: Text (text/picture)

Text **3:**

“Samsung Wave Y is the latest Samsung. Its sar is 0,318 and its weight is 103 g. It comes with 320 x 480 pixels and has 150 MB memory. It is running on Bada.”

[4.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v17	Ist dieser Text grammatikalisch korrekt?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[4.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v18	Finden sie Rechtschreibfehler?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[

4.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v19	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[4.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v20	Empfinden sie die	33.33%	66.67%	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

	Anordnung der Fakten gut strukuriert?	(1)	(2)				

[4.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v21	Sind die Inhalte des Textes einfach zu erfassen?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.33 (1.33)

[4.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v22	Wie gefällt ihnen der Ausdrucksstil des Textes?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[4.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v23	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[4.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v24	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

Seite 5: Einschätzung Dirk Nowitzki

[5.1]: Text (text/picture)

Text 4:

“Dirk Nowitzki is a basketball player player. His hobby is saxophone and his spouse is Jessica Olsson. In addition, he won the NBA Championship and was born June 19, 1978. He is a power forward for the Dallas Mavericks.”

[5.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v25	Ist dieser Text grammatikalisch korrekt?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.33 (1.33)

[5.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v26	Finden sie Rechtschreibfehler?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.33 (1.33)

[5.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v27	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[5.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v28	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

[5.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v29	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[5.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v30	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[5.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v31	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[5.9]: Natürlichkeit (rating/ranking)

	Von einem Menschen	Eher von Menschenhand	Kann ich nicht sagen	Eher von Computer	Von einem Computer	Mittelwert

		angefertigt (Wert: 1)	(Wert: 2)	(Wert: 3)	(Wert: 4)	angefertigt (Wert: 5)	
v32	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

Seite 6: Einschätzung Tokyo

[6.1]: Text (text/picture)

Text **5:**

“Tokyo is the centre of japan. Its population is 13,185,502 and its transportation is rail. Furthermore, its mayor is Shintaro Ishihara and one place to visit is Tokyo Tower. The University of Tokyo is the largest university in japan.”

[6.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v33	Ist dieser Text grammatikalisch korrekt?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[6.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v34	Finden sie Rechtschreibfehler?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.33 (1.33)

[6.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v35	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[6.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v36	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[6.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v37	Sind die Inhalte des Textes einfach zu erfassen?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[6.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v38	Wie gefällt ihnen der	0 (0%)	66.67%	33.33%	0 (0%)	0 (0%)	2.33 (2.33)

Ausdrucksstil des Textes?		(2)	(1)			

[6.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v39	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	0 (0%)	33.33% (1)	0 (0%)	2.67 (2.67)

[6.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v40	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

Seite 7: Einschätzung Humbold University of Berlin

[7.1]: Text (text/picture)

Text 6:

“Humboldt University of Berlin is public. Its motto is Veritas, Iustitia, Libertas and Berlin is the largest city in germany. Its president is Jan-Hendrik Olbertz and was founded in 1810.”

[7.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v41	Ist dieser Text grammatikalisch korrekt?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[7.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v42	Finden sie Rechtschreibfehler?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[7.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v43	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[7.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v44	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[7.6]: Verständnis (rating/ranking)

	Der Text war	Der Text war	Der Text war	Der Text ist teils	Der Text ist	Mittelwert

		sehr gut verständlich. (Wert: 1)	verständlich. (Wert: 2)	halbwegs verständlich. (Wert: 3)	unverständlich. (Wert: 4)	komplett unverständlich. (Wert: 5)	
v45	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[7.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v46	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	3.67 (3.67)

[7.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v47	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	3.33 (3.33)

[7.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v48	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	0 (0%)	66.67% (2)	33.33% (1)	4.33 (4.33)

Seite 8: Einschätzung Berlin

[8.1]: Text (text/picture)

Text **7:**

“Berlin is the largest city in germany. Its population is 3,471,756 and its transportation is metro. Furthermore, its mayor is Klaus Wowereit and one place to visit is alexander. The Humboldt University is the largest university in germany.”

[8.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v49	Ist dieser Text grammatikalisch korrekt?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[8.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v50	Finden sie Rechtschreibfehler?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[8.4]: Umfang (rating/ranking)

		Zu kurz (Wert:	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert

		1)						
v51	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[8.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v52	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[8.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v53	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[8.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v54	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[8.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v55	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[8.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v56	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

Seite 9: Einschätzung Matt Damon

[9.1]: Text (text/picture)

Text **8:**

“Matthew Paige "Matt" Damon was born in Cambridge, Massachusetts. He is an american Comedian, screenwriter and philanthropist and his school is Cambridge Alternative School. In addition, he has two children Isabella, Gia Zavala and Stella Zavala and he won a Golden Globe. He was married to Luciana Bozán Barroso, Minnie Driver and Alex Rodriguez. He is the Good Will Hunting, Saving Private Ryan and The Talented Mr. Ripley and Matt Damon was born on October 8th 1970. He is an American and Canadian. Also noteworthy, Matt Damon is the child of Kent Telfer Damon and Nancy Carlsson-Paige.”

[9.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v57	Ist dieser Text grammatikalisch korrekt?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[9.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v58	Finden sie Rechtschreibfehler?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.33 (1.33)

[9.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v59	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	3.67 (3.67)

[9.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v60	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[9.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v61	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[9.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v62	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[9.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v63	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[9.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v64	Glauben sie, dass	0 (0%)	0 (0%)	66.67%	33.33%	0 (0%)	3.33 (3.33)

dieser Text von einem Computer stammt?			(2)	(1)		

Seite 10: Einschätzung LG Optimus Pro

[10.1]: Text (text/picture)

Text 9:

“LG Optimus Pro is a LG Handy. Its sar is 0,55 and its weight is 129 g. In addition, it has 240 x 320 pixels and comes with 150 MB. It is an android.”

[10.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v65	Ist dieser Text grammatikalisch korrekt?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[10.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v66	Finden sie Rechtschreibfehler?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[10.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v67	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[10.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v68	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[10.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v69	Sind die Inhalte des Textes einfach zu erfassen?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[10.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v70	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[10.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v71	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[10.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v72	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

Seite 11: Einschätzung Vinson Massif

[11.1]: Text (text/picture)

Text **10:**

“Vinson Massif is located in southcentral Antarctica. Its height is 4896 m and is the Nicholas Clinch. Furthermore, the coordinates of 78°31′31.74″S 85°37′1.73″W and it is the highest mountain in the Sentinel Range.”

[11.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v73	Ist dieser Text grammatikalisch korrekt?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[11.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v74	Finden sie Rechtschreibfehler?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[11.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v75	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.67 (2.67)

[11.5]: Reihenfolge (rating/ranking)

	Sehr clever	Angemessen	Befriedigend	Schlecht	Keine Strukur	Mittelwert

		strukturiert (Wert: 1)	strukturiert (Wert: 2)	strukturiert (Wert: 3)	strukturiert (Wert: 4)	erkennbar (Wert: 5)	
v76	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[11.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v77	Sind die Inhalte des Textes einfach zu erfassen?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[11.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v78	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	3.33 (3.33)

[11.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v79	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[11.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v80	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	3.67 (3.67)

Seite 12: Einschätzung Adam Sandler

[12.1]: Text (text/picture)

Text **11:**

"Adam Richard Sandler (born September 9, 1966) is an American actor, comedian, singer, screenwriter, musician, and film producer. He is best known for his comedic roles, such as in the films Billy Madison (1995), Happy Gilmore (1996), Big Daddy (1999), and Mr. Deeds (2002). Adam Sandler was born in Brooklyn, New York to Jewish parents, Stanley, an electrical engineer, and Judy Sandler, a nursery school teacher. When he was five, his family moved to Manchester, New Hampshire, where he attended Manchester Central High School. Sandler married actress Jacqueline Samantha Titone, and they are the parents of two daughters, Sadie Madison Sandler and Sunny Madeline Sandler."

[12.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v81	Ist dieser Text grammatikalisch	100%	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

	korrekt?	(3)					

[12.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v82	Finden sie Rechtschreibfehler?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[12.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v83	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	0 (0%)	100% (3)	0 (0%)	4 (4)

[12.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v84	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[12.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v85	Sind die Inhalte des Textes einfach zu erfassen?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[12.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v86	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[12.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v87	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

[12.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v88	Glauben sie, dass dieser Text von einem Computer stammt?	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	0 (0%)	1.67 (1.67)

Seite 13: Einschätzung Aconcagua

[13.1]: Text (text/picture)

Text **12:**

"Aconcagua is the highest mountain in the Americas at 6,962 m. It is located in the Andes mountain range, in the Argentine province of Mendoza. The coordinates are 32°39′12.35″S 70°00′39.9″W. The first attempt on Aconcagua by a European was made in 1883 by a party led by the German geologist and explorer Paul Güssfeldt."

[13.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v89	Ist dieser Text grammatikalisch korrekt?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[13.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v90	Finden sie Rechtschreibfehler?	33.33% (1)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	1.5 (1)

[13.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v91	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[13.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v92	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[13.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v93	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[13.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v94	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[13.8]: Abwechslungsreichtum (rating/ranking)

	Sehr abwechslungsreich	Recht abwechslungsreich	Mittelmäßig (Wert: 3)	Hauptsächlich monoton	Sehr monoton	Mittelwert

		(Wert: 1)	(Wert: 2)		(Wert: 4)	(Wert: 5)	
v95	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[13.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v96	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

Seite 14: Einschätzung Budapest

[14.1]: Text (text/picture)

Text **13:**

"Budapest is the capital of Hungary. In 2011, Budapest had 1,733,685 inhabitants. The current mayor is István Tarlós. The neo-Gothic Parliament, containing amongst other things the Hungarian Crown Jewels. There are three main railway termini in Budapest, Keleti (eastbound), Nyugati (westbound), and Déli (southbound), operating both domestic and international rail services. Budapest is Hungary's main centre of education and home to the Budapest University of Technology and Economics."

[14.2]: Grammatik (rating/ranking)

		Fehlerlos (Wert: 1)	Vereinzelt Fehler (Wert: 2)	Mehrere Fehler (Wert: 3)	Sehr viele Fehler (Wert: 4)	Text unleserlich (Wert: 5)	Mittelwert
v97	Ist dieser Text grammatikalisch korrekt?	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	0 (0%)	2.33 (2.33)

[14.3]: Rechtschreibfehler (rating/ranking)

		Fehlerlos (Wert: 1)	1 bis 2 Fehler (Wert: 2)	3 bis 4 Fehler (Wert: 3)	5 bis 7 Fehler (Wert: 4)	Mehr als 7 Fehler (Wert: 5)	Mittelwert
v98	Finden sie Rechtschreibfehler?	100% (3)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1)

[14.4]: Umfang (rating/ranking)

		Zu kurz (Wert: 1)	Etwas zu kurz (Wert: 2)	Angemessen (Wert: 3)	Etwas zu lang (Wert: 4)	Zu lang (Wert: 5)	Mittelwert
v99	Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	0 (0%)	0 (0%)	100% (3)	0 (0%)	0 (0%)	3 (3)

[14.5]: Reihenfolge (rating/ranking)

		Sehr clever strukturiert (Wert: 1)	Angemessen strukturiert (Wert: 2)	Befriedigend strukturiert (Wert: 3)	Schlecht strukturiert (Wert: 4)	Keine Strukur erkennbar (Wert: 5)	Mittelwert
v100	Empfinden sie die Anordnung der Fakten gut strukuriert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[14.6]: Verständnis (rating/ranking)

		Der Text war sehr gut verständlich. (Wert: 1)	Der Text war verständlich. (Wert: 2)	Der Text war halbwegs verständlich. (Wert: 3)	Der Text ist teils unverständlich. (Wert: 4)	Der Text ist komplett unverständlich. (Wert: 5)	Mittelwert
v101	Sind die Inhalte des Textes einfach zu erfassen?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[14.7]: Ausdruck (rating/ranking)

		Gut gelungen. (Wert: 1)	Ganz gut (Wert: 2)	Befriedigend (Wert: 3)	Teils unglücklich formuliert (Wert: 4)	Unleserlich (Wert: 5)	Mittelwert
v102	Wie gefällt ihnen der Ausdrucksstil des Textes?	0 (0%)	66.67% (2)	33.33% (1)	0 (0%)	0 (0%)	2.33 (2.33)

[14.8]: Abwechslungsreichtum (rating/ranking)

		Sehr abwechslungsreich (Wert: 1)	Recht abwechslungsreich (Wert: 2)	Mittelmäßig (Wert: 3)	Hauptsächlich monoton (Wert: 4)	Sehr monoton (Wert: 5)	Mittelwert
v103	Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	0 (0%)	100% (3)	0 (0%)	0 (0%)	0 (0%)	2 (2)

[14.9]: Natürlichkeit (rating/ranking)

		Von einem Menschen angefertigt (Wert: 1)	Eher von Menschenhand (Wert: 2)	Kann ich nicht sagen (Wert: 3)	Eher von Computer (Wert: 4)	Von einem Computer angefertigt (Wert: 5)	Mittelwert
v104	Glauben sie, dass dieser Text von einem Computer stammt?	0 (0%)	33.33% (1)	66.67% (2)	0 (0%)	0 (0%)	2.66 (2.66)

Einschätzung Fließtexte für WebKnox

Belegarbeit Christian Hensel

29%

Text 3:

"Samsung Wave Y is the latest Samsung. Its sar is 0,318 and its weight is 103 g. It comes with 320 x 480 pixels and has 150 MB memory. It is running on Bada."

	Fehlerlos	Vereinzelt Fehler	Mehrere Fehler	Sehr viele Fehler	Text unleserlich
Ist dieser Text grammatikalisch korrekt?	○	○	○	○	○

	Fehlerlos	1 bis 2 Fehler	3 bis 4 Fehler	5 bis 7 Fehler	Mehr als 7 Fehler
Finden sie Rechtschreibfehler?	○	○	○	○	○

	Zu kurz	Etwas zu kurz	Angemessen	Etwas zu lang	Zu lang
Finden sie den Text zu kurz bzw. zu langschweifig formuliert in Hinblick auf die darin enthaltenen Fakten?	○	○	○	○	○

	Sehr clever strukturiert	Angemessen strukturiert	Befriedigend strukturiert	Schlecht strukturiert	Keine Strukur erkennbar
Empfinden sie die Anordnung der Fakten gut strukuriert?	○	○	○	○	○

	Der Text war sehr gut verständlich.	Der Text war verständlich.	Der Text war halbwegs verständlich.	Der Text ist teils unverständlich.	Der Text ist komplett unverständlich.
Sind die Inhalte des Textes einfach zu erfassen?	○	○	○	○	○

	Gut gelungen.	Ganz gut	Befriedigend	Teils unglücklich formuliert	Unleserlich
Wie gefällt ihnen der Ausdrucksstil des Textes?	○	○	○	○	○

	Sehr abwechslungsreich	Recht abwechslungsreich	Mittelmäßig	Hauptsächlich monoton	Sehr monoton
Sind die einzelnen Sätze gleichbleibend monoton oder eher abwechslungsreich formuliert?	○	○	○	○	○

	Von einem Menschen angefertigt	Eher von Menschenhand	Kann ich nicht sagen	Eher von Computer	Von einem Computer angefertigt
Glauben sie, dass dieser Text von einem Computer stammt?	○	○	○	○	○

Zurück Weiter

powered by eQuestionnaire™

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[Lid01] Brigitte Grote, Manfred Stede. 1998. ***Discourse Marker Choice in Sentence Planning.*** **Universität Berlin : s.n., 1998.**

[Lyn99] Cahill, Lynne. 1999. ***Lexicalisation in applied NLG systems.*** **University of Brighton : EPSRC, 1999.**

[Chr98] Chris Mellish, Alistair Knott, Jon Oberlander, Mick O'Donnell. 1998. ***Experiments using stochastic Search for text planning.*** **University of Edinburgh : s.n., 1998.**

[Dra98] Dragomri R. Radev, Kathleen R. McKeown. 1998. ***Generating Natural Language Summaries from Multiple On-Line Sources.*** **Columbia University : s.n., 1998.**

[Ehu 95] Ehud Reiter, Robert Dale. 1995. ***Building Applied Natural Language Generation Systems.*** **University of Aberdeen : Cambridge University Press, 1995.**

[ELI66] 1966. ***ELIZA—a computer program for the study of natural language communication between man and machine.*** **1966.**

[Her95] Hercules Dalianis, Eduard Hovy. 1995. ***Aggregation in Natural Language Generation.*** **Pisa, Italy : Springer Verlag, 1995.**

[Hov98] Hovy, Eduard. 1998. **Entry for MIT Encyclopedia of Computer Science (MITECS). Cambridge: MIT.** ***MIT Encyclopedia of Computer Science.*** **Cambridge : s.n., 1998.**

[Wik] **http://en.wikipedia.org/wiki/Natural_language_generation. [Online]**

[Joh93] Johanna D. Moore, Cécile L. Paris. 1993. ***Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information.*** **University of Pittsburgh : s.n., 1993.**

[Mir03] Lapata, Mirella. 2003. ***Probabilistic Text Structuring: Experiments with Sentence Ordering.*** **University of Sheffield : s.n., 2003.**

[Lid01] Liddy, E.D. 2001. ***Natural Language Processing.*** **New York : Marcel Decker, Inc, 2001.**

[Lid05] Liddy, Elizabeth D. 2005. ***Automatic Document Retrieval.*** **In Encyclopedia of Language and Linguistics. 2nd Edition. : Elsevier Press , 2005.**

[Mat98] Matthew Stone, Christine Doran. 1998. ***Sentence Planning as Description Using Tree Adjoining Grammar.*** **University of Pennsylvania, Philadelphia : s.n., 1998.**

[Kat84] McKeown, Kathleen R. 1984. ***Discourse Strategies for Generating Natural-Language Text.*** **Columbia University, New York : s.n., 1984.**

[Mic98] O'Donnell, Michael. 1998. ***A Markup Tool for Rhetorical Structure Theory.*** **University of Edinburgh : s.n., 1998.**

[Ell96] Riloff, Ellen. 1996. ***Automatically Generating Extraction Patterns from Untagged Text.*** **University of Utah : s.n., 1996.**

[Dan01] Vrajitoru, Dana. 2001. ***Evolutionary Sentence Building for Chatterbot.*** **South Bend : s.n., 2001.**

[Wil03] Williams, Sandra and Reiter, Ehud. 2003. ***A corpus analysis of discourse relations for Natural Language Generation.*** **Lancaster University, UK. : s.n., 2003.**

[Win80] Winograd, Terry. 1980. ***What does it mean to understand language.*** **Stanford University : s.n., 1980.**