Informationslinguistische Verfahren der Informationsselektion aus Textdokumenten

Jürgen Reischer
Informationslinguistik Regensburg
Linguistisches Forum 26.1.2011

Übersicht

- Informationslinguistik = Linguistik und Informationswissenschaft:
  - Sprachverarbeitung für informations-wissenschaftliche Fragestellungen und Anwendungen:
    - Informationsermittlung (-selektion/-extraktion):
      - automatische Indexierung;
      - automatisches Zusammenfassen (Summarizing);
      - Information-Retrieval (Frage-Antwort-Systeme, [Within-]Document-Retrieval).

Übersicht

- Sprachbasierte Mensch-Maschine-Interaktion:
  - Spracherkennung;
  - Dialogsysteme, Chatbots;
  - Maschinelle Übersetzung (Web, Smartphones).
- Wissensbasierte Systeme (semantische Modellierung):
  - (sprachbasierte) Expertensysteme;
  - (elektronische) Wortnetze, Thesauri, Lexika;
  - 'Semantic Web', Ontologien.

Übersicht

- Nutzung von Erkenntnissen aus Nachbar-Disziplinen:
  - Linguistik:
    - Textlinguistik;
    - Korpuslinguistik.
  - Computerlinguistik:
    - Sprach- und Texttechnologie;
    - Evaluation.
  - Künstliche-Intelligenz-Forschung/Informatik:
    - Repräsentationssysteme (OWL: 'Web Ontology Language');
    - Programmiersprachen (Java/C#, Prolog).

Übersicht

- Sprachverarbeitung im Detail:
  - Formale Verarbeitung:
    - Satzerkennung (Text-zu-Satz-Zerlegung);
    - Tokenisierung (Satz-zu-Wort/Token-Zerlegung);
    - Normalisierung (Deflexion, Dekomposition);
    - Mehrwortterm-Erkennung;
    - Eigennamen-Erkennung;
    - Neologismen-Erkennung;
    - POS-Tagging (Wortart-Annotation);
    - (Chunk-)Parsing.

Übersicht

- Inhaltliche Verarbeitung:
  - Ermittlung von Kohärenz und Informativität;
  - Ermittlung zentraler/wichtiger Wörter und Sätze;
  - Ermittlung des/der Textthemen und thematische Segmentierung;
  - Ermittlung von Textsorten;
  - Ermittlung von Term-Okkurrenzen (Begriffs-Spezifität) und Kookkurrenzen (semantische Nachbarschaft).

Einleitung

- Beispiel EXCERPT-System:
  - "Expert in Computational Evaluation and Retrieval of Passages of Text";
  - EXCERPT als integriertes System zur Selektion von Information aus Texten:
    - Within-Document-Retrieval zur Suche von Passagen in Texten (Passagen-Retrieval);
    - (informatives) Summarizing zur Aufbereitung und Ausgabe ermittelter Passagen ([Ab-]Sätze).

Einleitung

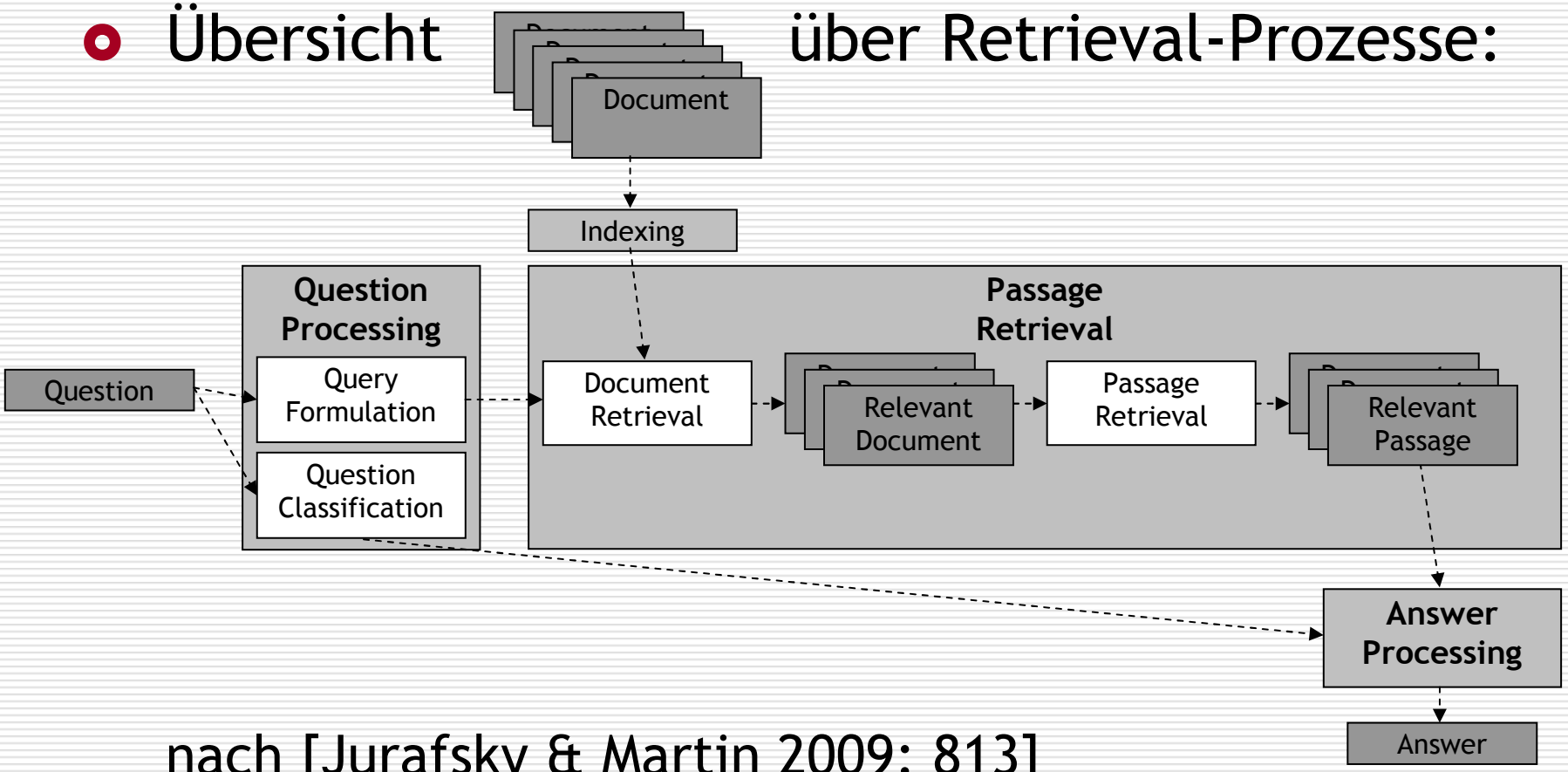
- Ziel/Motivation:
  - Unterstützung des Nutzers bei der Suche und Sichtung von Information in Texten:
    - 'Suchmaschine' für Texte mit automatischer Vorschau- oder Zusammenfassungs-Funktion (Snippets, Summarys);
    - Aufbereitung der Suchergebnisse durch automatische Bewertung und Sortierung der Fundstellen ('Rating' und 'Ranking' von Text-Einheiten).

Einleitung

- Hintergrund: Dokumentensuche liefert meist zu viele und zu lange Dokumente:
  - Längere Dokumente können aus Zeitgründen nicht intellektuell durchsucht werden;
  - einfache 'Suchen'/'Finden'-Kommandos können Nutzerbedürfnisse selten erfüllen:
    - nur 1 Suchbegriff möglich;
    - kein 'Rating' und 'Ranking' gefundener Textstellen;
    - keine systematische Aufbereitung und Ausgabe der Fundstellen.

Einleitung

- Übersicht über Retrieval-Prozesse:



nach [Jurafsky & Martin 2009: 813]

Einleitung

- EXCERPT setzt *nach* Dokumenten-Retrieval an:

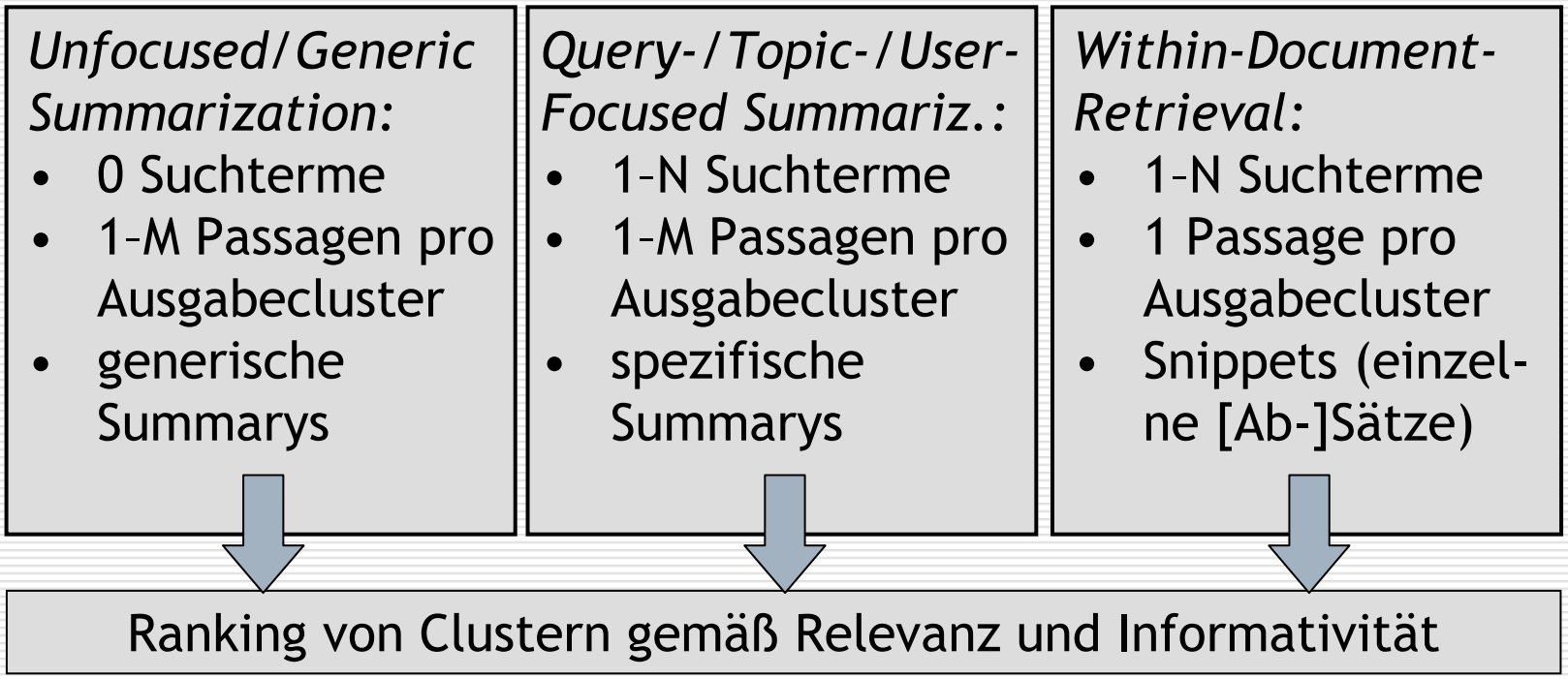
Dokumenten-Retrieval in Dokumentenkollektion K:
- Suchanfrage X an Dokumentenkollektion K
- gerankte Liste relevanter Dokumente D_1, D_2, …, D_N

Passagen-Retrieval in Dokument D_i:
- Suchanfrage Y (≠ X) an Dokument D_i
- (intern) gerankte Liste relevanter Passagen P_1, P_2, …, P_M

Informatives Summarizing von Dokument D_i:
- iteratives Clustern von Passagen $P_{1…K} \epsilon P_{1…M}$ zu Summarys S_j
- (extern) gerankte Liste informativer Summarys S_1, S_2, …, S_L

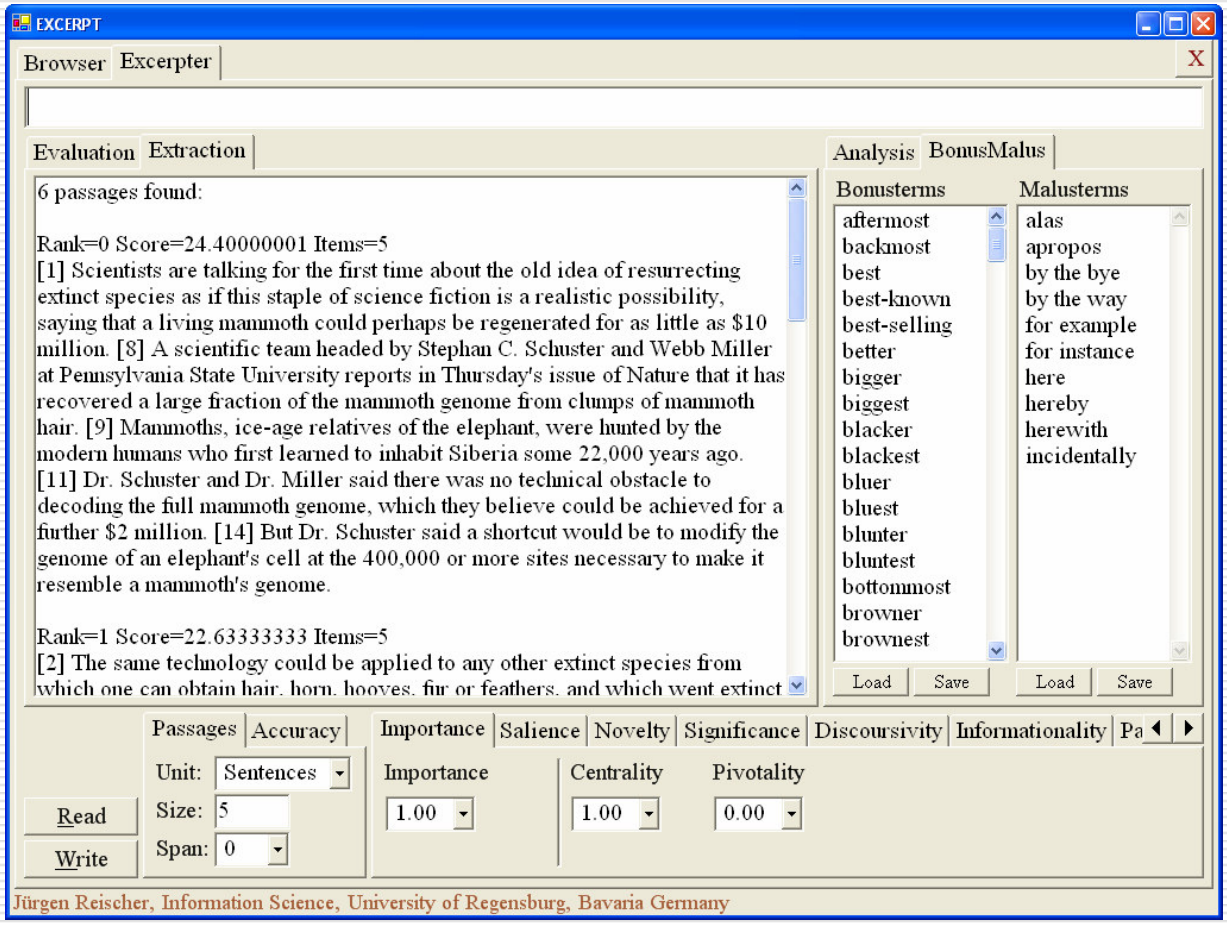
Einleitung

- Fließender Übergang zwischen Summarizing und Within-Document-Retrieval:



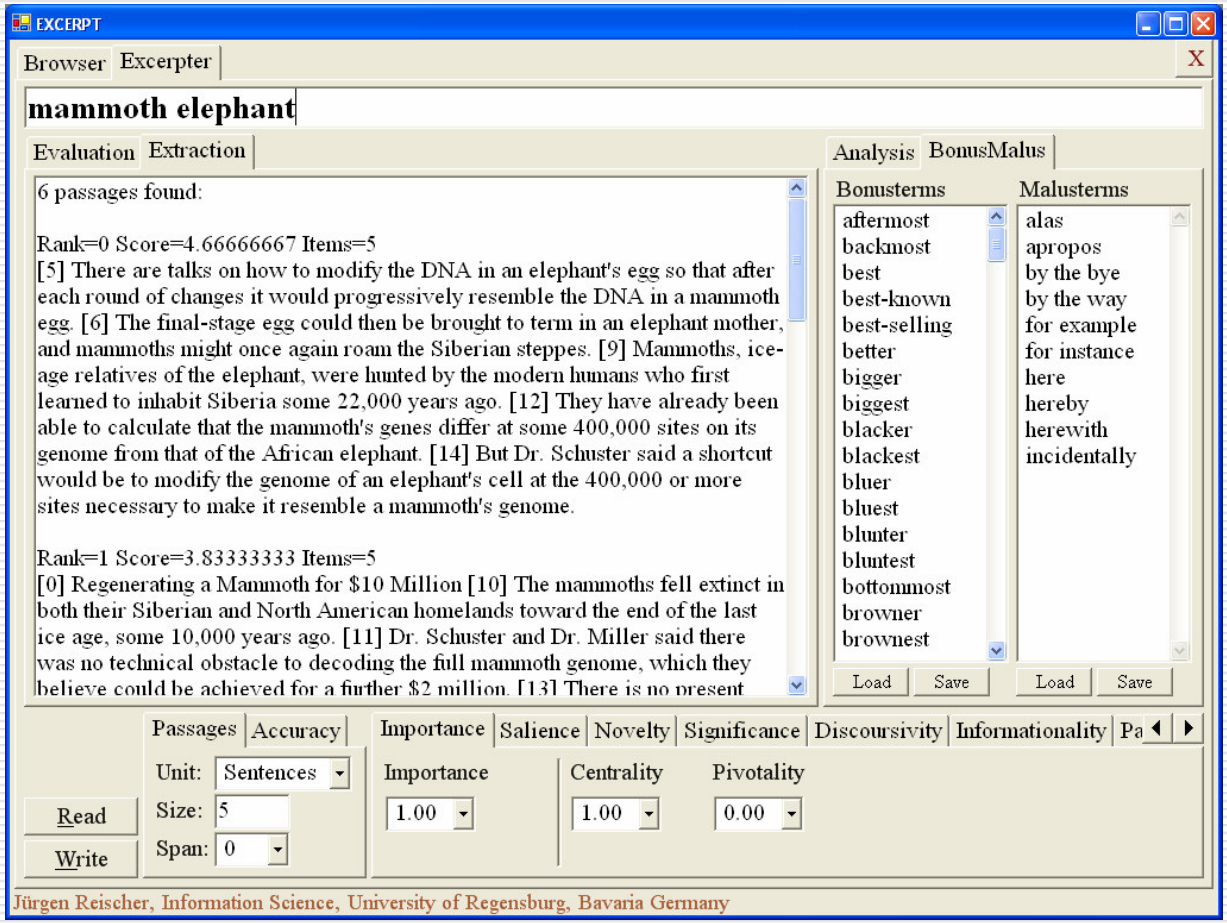
Einleitung

Unfocused/ Generic Summarys:



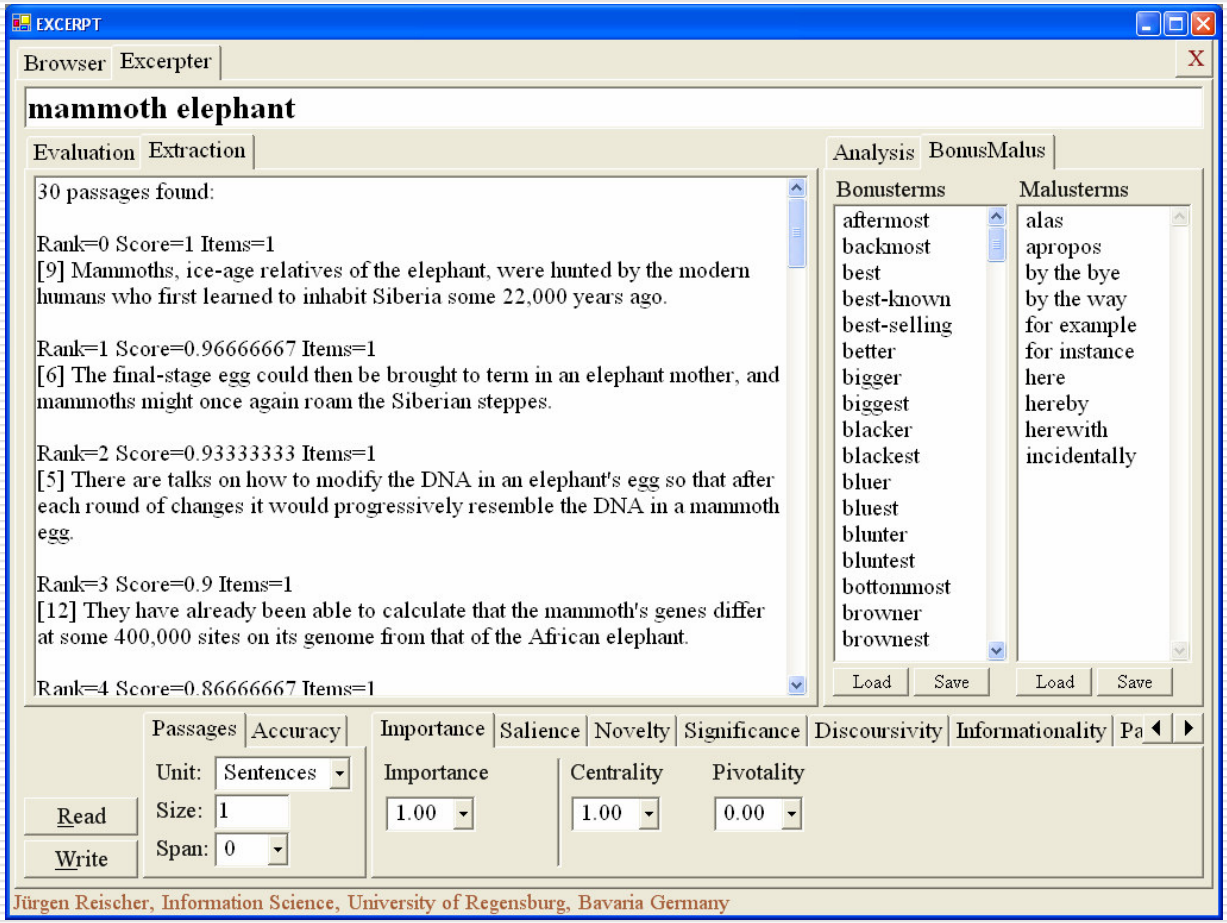
Einleitung

Query-/
Topic-/
User-
Focused
Summarys:



Einleitung

Within-Document-Retrieval-Snippets (Sätze/ Absätze):



Theorie—Ansätze

- Begriff 'Informativität' im Fokus des Interesses:
  - Informationswissenschaft:
    - informative vs. indikative Summarys [Borko & Bernier 1975];
    - Auftretenshäufigkeit/Spezifität von Termen: tf•idf ([Spärck Jones 1972]; [Salton & Buckley 1988]).
  - Linguistik:
    - Antwortinformation: thematische Relevanz und Informativität (Neuigkeitswert) einer Aussage [Grewendorf 1981];
    - Informativität als Textualitätskriterium: Ausmaß, in dem textuelle Einheiten (un)erwartet, (un)bekannt und/oder (un)vorhersagbar sind [Beaugrande & Dressler 1981].

Theorie—Informativitätskriterien

- Kriterien der Informativität:
  - Verständlichkeit als notwendige Bedingung für Informativität:
    - Unverständliches ist nicht informativ ([Nöth ²2000]; [Weizsäcker 1974a]);
    - Verständlichkeit erreichbar bspw. durch generelle Lesbarkeit (z. B. Terminologie), Kohäsion/Kohärenz.
  - Überblick über Informativitäts-Kategorien:
    - Thematizität, Spezifität, Novität, Faktizität (s. u.);
    - Kriterien basierend auf Erkenntnissen aus der Literatur und eigenen Untersuchungen (s. u.).

Theorie—Informativitätskriterien

- *Thematizität*: Welche Information vermittelt der Text ('aboutness')?
  - Frequenz von Inhaltstermen im Text (vs. in der Sprache bzw. einem Referenzkorpus wie z. B. WordNet) (vgl. [Nenkova & Vanderwende 2005]);
  - Topikalität von Termen im Satz oder Absatz (Subjekt-/Topik-Position in Passage) (vgl. [Baxendale 1958]; [Edmundson 1969]; [Lambrecht 1994]; [Marcu 1997]; [Strzalkowski & al. 1999]; [Krifka 2006ab]);

Theorie—Informativitätskriterien

- semantische Relationen zwischen Termen bzw. thematische Kohärenz/Konvergenz von Passagen (vgl. [Mihalcea 2004]; [Mihalcea & Tarau 2004]):
  - Synonymie, Antonymie, Hyp(er)onymie, Meronymie/ Holonymie usw.;
  - Derivation/Komposition, Entailment, Attribut u. a.
- Bonus-/Malus-Terme (z. B. Steigerungsformen, "therefore"; "incidentally", "by the way") ([Edmundson 1969]; [Mittal & al. 1999]; [Goldstein & al. 1999]).

Theorie—Informativitätskriterien

- *Spezifität*: Wie viel Information vermitteln Terme (bzw. die von ihnen signifizierten Konzepte)?
  - Verständlichkeit von Termen als Voraussetzung ihrer grundsätzlichen Interpretierbarkeit:
    - formal: häufige/ambige und vertraute Ausdrücke sind besser verständlich (vgl. [Jastrzembski 1981]; [Tengi 1998]);
    - inhaltlich: konkrete Begriffe der Basisebene sind einfacher verstehbar als abstraktere Begriffe anderer Ebenen (vgl. [Rosch & Mervis 1975]; [Rosch 1978]).

Theorie—Informativitätskriterien

- Inhalts- vs. Funktionsterme:
  - Funktionswörter sind syntaktisch formativ, nicht semantisch informativ (Frage nach dem potenziellen Informationsgehalt von Termen);
  - bestimmte Inhalts- bzw. Funktionswörter kommen in *Abstracts* häufiger/seltener vor als in Volltexten ([Goldstein & al. 1999]; [Mittal & al. 1999]):
    - häufiger: z. B. Steigerungsformen, semantisch relationierte Inhaltswörter;
    - seltener: z. B. die meisten Funktionswörter, vage Inhaltswörter ("several", "really").

Theorie—Informativitätskriterien

- Generizität vs. Spezifität von Termen (bzw. deren Konzepten):
  - Allgemeine Begriffe vermitteln weniger Information als spezielle ("Tier" vs. "Krake");
  - Partikularität von Termen (Eigennamen als Terme für maximal spezifische Konzepte):
    - Eigennamen sind verständlicher [Langer & al. 1974],
    - besitzen einen besonderen Interessantheitswert [Flesch 1948] und
    - stellen die signifikantesten Konzepte in Sätzen überhaupt dar [Paradis & Berrut 1996].

Theorie—Informativitätskriterien

- Basiskategorialität: Begriffe auf der mittleren Ebene der Begriffshierarchie bieten ein Optimum zwischen Informativität und Verständlichkeit ([Rosch & al. 1976]; [Rosch 1978]).
- Vagheit vs. Präzisheit von Begriffen:
  - scharfe vs. unscharfe Begriffsränder (Kategorie 'Primzahl' vs. 'Spiel') [Keller 1995];
  - semantische Relativität von Begriffen ('schnell' vs. 'schnell für ein Segelschiff/ Auto') [Pinkal 1985].

Theorie—Informativitätskriterien

- *Novität*: Welche und wie viel neue Information vermitteln die Terme?
  - Wortbildungen:
    - Unbekannte Terme geben Aufschluss über den inhaltlichen Neuigkeitswert einer Passage oder eines Textes (relativ zu einem Referenzlexikon, z. B. WordNet);
    - neue Komposita und Derivata (dekomponierbar) bzw. Neologismen (nicht dekomponierbar) besitzen meist nur eine einzige spezifische Bedeutung.

Theorie—Informativitätskriterien

- Type-Token-Verhältnis:
  - Die Anzahl neu eingeführter Einheiten pro Passage gibt Aufschluss über den Informationsfluss im Text [Wimmer 2005];
  - Passagen mit neu im Text auftretenden Termen führen neue Themen und Inhalte ein:
    - Gerade am Textanfang befinden sich wichtige Aussagen;
    - neue thematische Abschnitte oder Absätze weisen notwendig bislang ungenannte Terme auf.
  - Wiederholt auftretende Terme deuten auf inhaltliche Redundanz hin [Weizsäcker 1974b].

Theorie—Informativitätskriterien

- *Faktizität*: Wie wahr oder faktisch ist die vermittelte Information?
  - Wahre Aussagen sind informativer als nahezu wahre oder falsche Aussagen [Floridi 2004];
  - objektiv-faktenorientierte Aussagen sind informativer als subjektiv-meinungsorientierte [Heylighen & Dewaele 1999]:
    - Pronomen der 1. Person verraten subjektive Aussagen;
    - konjunktivische (vs. indikative) Konstrukte weisen auf subjektive Stellungnahmen (Wunsch) bzw. Irrealität/Kontrafaktizität (Möglichkeit) hin.

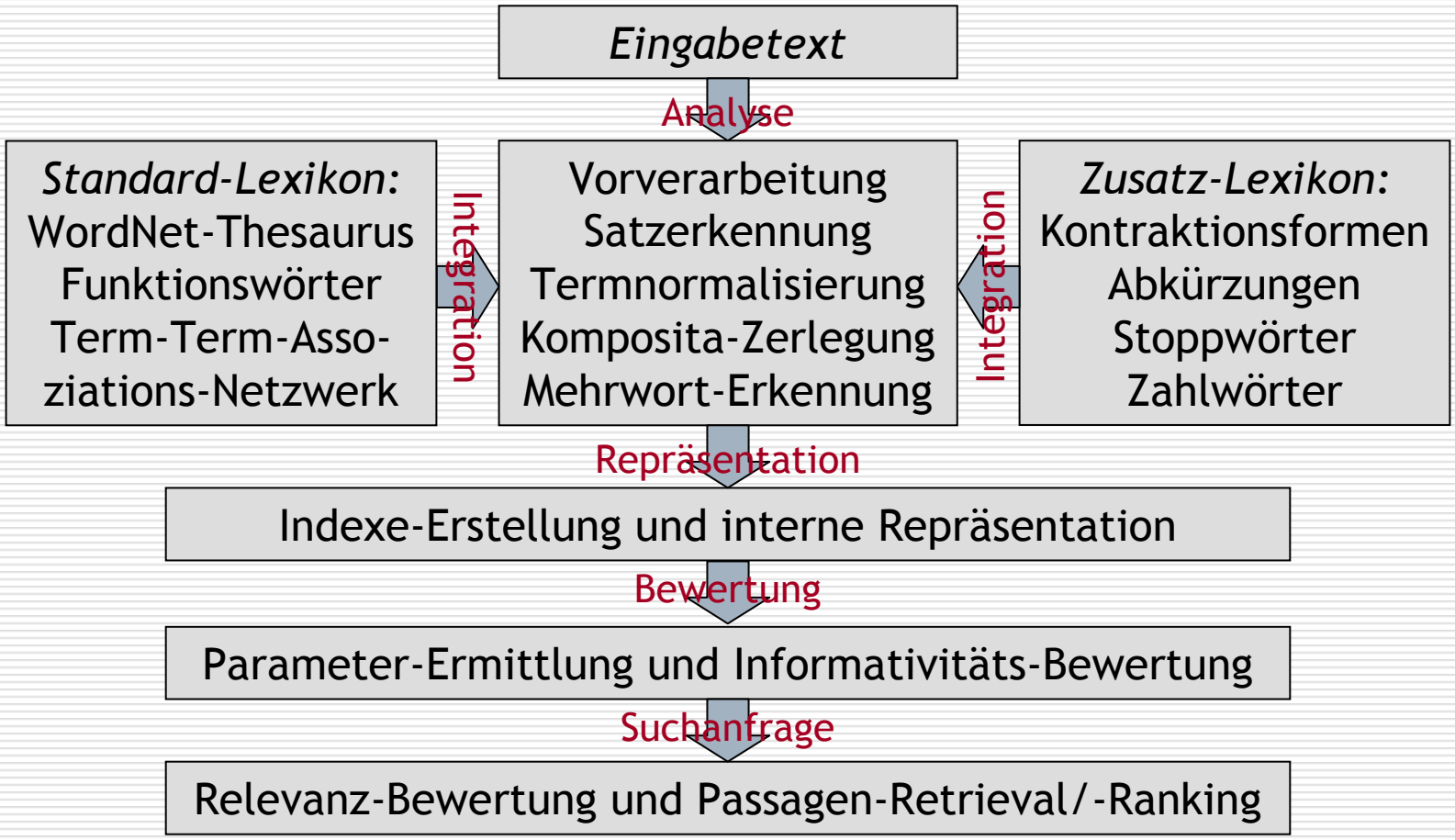
Theorie—Informativitätskriterien

- Sprechakttyp von Sätzen:
  - Nur Aussagen sind grundsätzlich wahrheitsfähig, d. h. können wahre Fakteninformation vermitteln;
  - Fragen und Aufforderungen/Anweisungen sind nicht wahrheitsfähig:
    - Information ist die Antwort auf eine Frage, nicht die Frage selbst (s. o.); d. h. Fragen sind nicht informativ [Alfonseca & Rodríguez 2003];
    - Anweisungen/Instruktionen wie in Rezepten/ Bedienungsanleitungen oder Aufforderungen sind zumindest informativ im Sinne von instruktiv (Information als Instruktion).

Praxis—Überblick

- Realisierung des EXCERPT-Systems:
  - Strukturen:
    - Gegeben: Thesauri/Wortlisten;
    - Erzeugt: Indexe/Textrepräsentation.
  - Prozesse:
    - Intern: Linguistische Textanalyse und Parameter-Ermittlung;
    - Extern: Retrieval und Ranking.
- Übersicht über Komponenten/Workflow:

Praxis—Überblick



Praxis—Textanalyse

- Textanalyse:
  - Vorverarbeitung zur Normalisierung von Zeichen:
    - Abbildung von `“´/`”´ auf `"´, `‘´/`’´ auf `'´;
    - Vereinheitlichung von `—´ oder `–´ zu `-´;
    - Reduktion von `...´ zu `…´, `!!´ zu `!´, `??´ zu `?´ usw.;
    - Auflösung von Abkürzungen wie "e.g." zu "eg", "you're" zu "you are".
  - ➡ Verbesserung der Satzerkennung.

Praxis—Textanalyse

- Absatz- und Satzerkennung:
  - Absatz-Erkennung und -Verarbeitung:
    - Einlesen von ASCII- bzw. UTF8-kodierten Textdateien;
    - einfache Segmentierung anhand von 'CarriageReturn'- und/oder 'LineFeed'-Zeichen;
    - Zerlegung des Absatzes in Tokens, getrennt durch Whitespace-Zeichen (ASCII-Kode ≤ 32).

Praxis—Textanalyse

- Satzerkennung:
  - Als Satzende-Zeichen gelten `.´, `!´, `?´ und das freistehende `…´, aber nicht `;´ und `:´ (keine Teilsätze als Ausgabe-Einheit!);
  - der Folgetoken muss mit einem Großbuchstaben beginnen (zu beachten ist, dass dieser auch in Anführungszeichen oder Klammern stehen kann!);
  - der aktuelle Token, an dem das Satzende-Zeichen hängt, muss mehr als 1 Zeichen umfassen (sonst wahrscheinlich abgekürzter Name oder Nummerierung);

Praxis—Textanalyse

- der mögliche Satztrenner darf nicht innerhalb einer offenen Zitation "…" oder Klammerung `(…)´ bzw. `… – … – …´ stehen, außer er ist das letzte Zeichen in einem solchen Konstrukt:
  - In einen nicht-zitierten Satz eingebettete Zitationen werden nicht zerlegt;
  - in einen nicht-zitierten Satz eingebettete Einschübe (z. B. Klammerstrukturen) werden ebenfalls nicht zerlegt.
- ➡ Vermeidung unselbständiger Sätze als Ausgabe-Einheit beim Retrieval und Ranking.

Praxis—Textanalyse

- Term-Erkennung und Normalisierung:
  - Terme werden auf ihre Grundformen reduziert:
    - Entfernung von Flexionsendungen (Deflexion z. B. anhand einer erweiterten Liste von WordNet-Flexiven);
    - Abbildung unregelmäßiger Formen auf regelmäßige (z. B. via WordNet-Liste englischer Ausnahme-Formen);
    - Abgleich der normalisierten Formen auf WordNet-Lexikon.

Praxis—Textanalyse

- Zerlegung zusammengeschriebener Komposita:
  - relativ zum Datenbestand von WordNet 3.0;
  - wenn nicht dekomponierbar, dann als Neologismus gewertet.
- Elimination bestimmter Terme:
  - Funktionswörter und Adverben;
  - hochfrequente Inhaltswörter wie "make";
  - zusätzlich 'void words' wie "Mr.", "Dr." usw.

Praxis—Textanalyse

- Mehrwortterm- (MWT-)Erkennung:
  - Erkennung von lexikalischen Termen, die aus mehr als einem Ausdruck bestehen (ca. 64.000 Lexeme in WordNet 3.0);
  - Vorteile:
    - Mehrteilige Funktionswörter können zusätzlich eliminiert werden ("in terms of", "as well as");
    - mehrteilige Inhaltswörter können mit ihren zusammengeschriebenen Pendants vereinheitlicht werden ("web site", "website"; auch "web-site");
    - ➡ Erhöhung der Analyse-Präzision.

Praxis—Textanalyse

- Probleme bei der MWT-Erkennung:
  - Interne Flexion: z. B. "kick[ed] the bucket[s]";
  - Überlappung von MWTs: z. B. enthält "Vice President of the United States" die zwei MWTs
    - "Vice President" und
    - "President of the United States".
  - Irrtümliche MWTs: "(Sharon) Stone's age" bezieht sich nicht auf "Stone Age";
  - verteilte MWTs: z. B. "turn the light on/off" (nicht-adjazente Einzelterme).

Praxis—Textanalyse

- Neologismen-Erkennung:
  - Ausdrücke, die nicht Teil des WordNet-Bestandes sind und nicht dekomponiert werden können, werden als Neologismen betrachtet;
  - auch dekomponierbare Ausdrücke in Bindestrich-Schreibung, die mindestens einen unbekannten Term enthalten, gelten als Neologismen;
  - Ziffernkombinationen (Zahlen) werden hingegen nicht als Neologismen gewertet.

Praxis—Parameter

- Parameter-Ermittlung:
  - Parameter lassen sich in verschiedenen Dimensionen klassifizieren (Beispiele):
    - formal:
      - wortbasiert: Frequenz/Polysemiegrad eines Terms;
      - satzbasiert: Position eines Satzes im Absatz/Text;
      - absatzbasiert: Länge eines Absatzes in Termen/Sätzen.
    - inhaltlich:
      - wortbasiert: semantische Relationiertheit von Termen;
      - satzbasiert: Sprechakttyp eines Satzes;
      - absatzbasiert: thematische Nähe von Absatz und Text.

Praxis—Parameter

- Die Operationalisierung der theoretischen Informativitäts-Parameter erfordert die Einschränkung auf
  - die maschinell grundsätzlich ermittelbaren Parameter (z. B. kann die Wahrheitsnähe einer Aussage nicht ermittelt werden);
  - die algorithmisch bzgl. Zeit- und Ressourcen-aufwand überhaupt berechenbaren Parameter (z. B. setzen Generizität/Spezifität oder Vagheit/Präzisheit eine vollständige und korrekte Disambiguierung aller Terme des Textes zu eindeutigen Konzepten voraus).

Praxis—Parameter

- Die im EXCERPT-System verwendeten Parameter sind zu sieben Oberkategorien zusammengefasst, die jeweils zwei Parameter beinhalten (engl. Begriffe; Details s. u.);
  - *Importance*: Centrality, Pivotality;
  - *Salience*: Prominency, Initiality;
  - *Novelty*: Innovativeness, Neologisms;
  - *Significance*: Bonus-/Malus-Words;
  - *Discoursiveness*: Pronouns, Conjunctions;
  - *Informationality*: Declarativeness, Enumerativeness;
  - *Particularity*: Individuals, Numbers.

Praxis—Parameter

- Kategorie 'Importance':
  - Parameter 'Centrality':
    - thematische Wichtigkeit eines Satzes durch semantische Verknüpftheit (Kohärenz) der Terme eines Satzes mit anderen Termen;
    - stark verknüpfte Sätze als wichtiger erachtet.
  - Parameter 'Pivotality':
    - Summe aller Vorwärts- und Rückwärts-Relationen der Terme eines Satzes zeigt an, ob Satz am Anfang oder Ende eines thematischen Abschnitts steht;
    - thematische Dreh- und Angelsätze am Beginn eines Abschnitts als wichtiger bewertet.

Praxis—Parameter

- Kategorie 'Salience':
  - Parameter 'Prominency':
    - formale Initialität oder Finalität eines Satzes im übergeordneten Absatz;
    - ein-/überleitende Sätze am Anfang und Ende von Absätzen höher bewertet.
  - Parameter 'Initiality':
    - formale Initialität eines Satzes innerhalb des ersten Paragraphen nach einer Überschrift;
    - thematisch einleitende Sätze am Anfang eines Abschnitts höher bewertet.

Praxis—Parameter

- Kategorie 'Novelty':
  - Parameter 'Innovativeness':
    - Anzahl neu erwähnter Terme pro Satz gibt Aufschluss über die thematische Progression im Text (Informationsfluss);
    - Sätze mit hohem Anteil neuer Terme als informativer erachtet.
  - Parameter 'Neologisms':
    - Neologismen (relativ zu WordNet 3.0) als Ausdruck neuer und spezifischer Konzepte;
    - erhöhte Anzahl deutet auf informativere Aussagen im Text hin.

Praxis—Parameter

- Kategorie 'Significance':
  - Parameter 'Bonus words':
    - Liste vom System vorgegebener oder nutzerdefinierter Terme (z. B. Komparative und Superlative; "sexy", "extremely", "in summary" usw.);
    - Auskunft über Interessantheit/Wichtigkeit von Sätzen.
  - Parameter 'Malus words':
    - Nebeninformation einleitende Phrasen ("by the way", "incidentally");
    - Auskunft über Uninteressantheit/Uninformativität von (Neben-)Sätzen.

Praxis—Parameter

- Kategorie 'Discoursiveness':
  - Parameter 'Conjunctions':
    - adversative und kausative Konnektoren deuten evtl. interessierende Gegensätze oder Begründungen an ("but", "because");
    - Sätze mit solchen Konjunktionen als wichtig erachtet.
  - Parameter 'Pronouns':
    - 1.-Person-Pronomen deuten persönlich-subjektive oder meinungs-orientierte statt objektiv-neutraler und fakten-orientierter Sätze an;
    - Sätze mit Pronomen der 1. Person als weniger objektiv-/faktisch-informativ erachtet.

Praxis—Parameter

- Kategorie 'Informationality':
  - Parameter 'Declarativeness':
    - Deklarativität/Assertivität statt Interrogativität von Sätzen: Interrogative Sätze taugen nicht als Antwort auf Frage (Information ist *Antwort* auf eine Frage);
    - interrogative Sätze werden als uninformativ gewertet.
  - Parameter 'Enumerativeness':
    - Sätze/Absätze innerhalb einer Aufzählung explizit ausgezeichnet und inhaltlich meist stichpunktartig zusammenfassend;
    - aufgezählte Sätze als inhaltlich wichtig erachtet.

Praxis—Parameter

- Kategorie 'Particularity':
  - Parameter 'Individuals':
    - Eigennamen, die inhaltlich maximal spezifische Konzepte denotieren;
    - Individualkonzepte als interessant und informativ gewertet.
  - Parameter 'Numbers':
    - Maß-/Jahreszahlen und Daten als Hinweis auf spezifische Größen(ordnungen) (anstelle von "some", "several" usw.);
    - nummerische Angaben als wichtig erachtet.

Evaluation

- Durch die Evaluation soll die Leistung des EXCERPT-Systems gemessen werden:
  - im Within-Document-Retrieval-Modus;
  - im informativen Summarizing-Modus.
- Grundlage hierfür sind
  - Evaluations-Maße;
  - Evaluations-Korpora.

Evaluation—Maße

- Evaluationsmaß für Information-Retrieval und Summarizing:
  - Anforderungen an ein Evaluationsmaß:
    - Es muss sich gleichermaßen für Within-Document-Retrieval und Summarizing eignen (sonst keine Vergleichbarkeit);
    - es muss einfach anzuwenden und aussagekräftig sein.
  - *R-Precision-Maß* erfüllt diese Bedingungen:

$$\text{R-Prec} = \frac{\text{Anzahl ermittelter relevanter Items}}{\text{Anzahl ermittelter=relevanter Items}}$$

Evaluation—Maße

R-Precision stellt ein sinnvolles Maß zur Berechnung der Leistung eines Retrieval- bzw. Summarizing-Systems dar:

- Vom System sollen genau so viele Items der Treffermenge ermittelt bzw. untersucht werden, wie in der vorab bekannten Relevanzmenge enthalten sind;
- R-Precision kann sowohl für den Within-Document-Retrieval- als auch den Summarizing-Modus verwendet werden.

Evaluation—Probleme

- Evaluations-Korpora:
  - Probleme I:
    - Keine existenten Evaluations-Korpora für Within-Document-Retrieval;
    - wenige frei und vollständig verfügbare Korpora für extraktiv-informatives Summarizing im Englischen [Hasler & al. 2003: 309];
    - verfügbare Summarizing-Korpora zielen nicht explizit auf den Aspekt der Informativität bzw. Interessantheit ab.

Evaluation—Probleme

- Probleme II:
  - Verfügbare Korpora sind oft nur durch 1 oder 2 Bewerter beim Passagen-'Rating' entstanden;
  - es ist unklar, was optimales/ideales Referenz-Summary oder allgemein Referenz–Set von Passagen ist [Mani 2001: 222];
  - für verlässliche Ergebnisse manuelle Erstellung von Korpora und Auswertung der Systemleistung notwendig (Vermeidung von Artefakten durch maschinelle Korpus-Erstellung- und Ergebnis-Auswertung).

Evaluation—Probleme

- Abhilfe:
  - Manuelle Erstellung und Auswertung eigener Korpora mit Fokus auf Informativität und Interessantheit für extraktive Summarys;
  - Erzeugung von 'Kompromiss'- oder 'Konsens'-Extracts aus mehreren Bewertern anstelle von 'idealen' Extracts:
    - 'Übereinanderlegen' mehrerer Bewerter-Extracts zu einem bestmöglichen Kompromiss-/Konsens-Extract;
    - Mehrheitsmeinung bezüglich Passagen-Selektion als Referenz-Extract soll möglichst viele potenzielle Nutzer zufrieden stellen.

Evaluation—Korpora

- Für die Evaluation verwendete existente Korpora:
  - [Zechner 1995]:
    - 6 Texte aus dem 'Daily Telegraph Corpus' mit effektiv je 13 Bewertern;
    - Fragestellung: Ermittlung von 5 bis 7 Sätzen, die am *zentralsten oder relevantesten* für den Inhalt des Textes waren;
    - Annahme: Bei Nachrichtentexten, die der Vermittlung aktueller und neuer Information dienen, sind die zentralsten und relevantesten Sätze gerade die informativsten und interessantesten.

Evaluation—Korpora

- [Hasler & al. 2003]: Korpus des CAST-Projekts mit insgesamt 163 Texten des Reuters- und BNC-Korpus, davon
  - 7 Texte mit 3 Bewertern (s. u.);
  - Fragestellung: Ermittlung der *essenziellen und wichtigen* Sätze eines Textes;
  - Annahme auch hier: Bei Nachrichtentexten sind die essenziellen und wichtigen die informativen und interessanten Sätze.

Evaluation—Korpora

- Erstellung eigener Korpora:
  - Zwei Mal 13 Texte in zwei unabhängigen Untersuchungen (WS 2008/09, SS 2009):
    - Informations-Texte aus verschiedenen Online-Quellen und mit verschiedenen Inhalten:
      - 'bbc.co.uk', 'cnn.com', 'nytimes.com', 'readwriteweb.com', 'reuters.com', 'spiegel.de', 'telegraph.co.uk', 'usatoday.com', 'washingtonpost.com', 'wikipedia.org', 'wikisource.org';
      - Nachrichten-Artikel zu Themen aus Politik, Gesellschaft, Technik, Wissenschaft; Buch-Kapitel über Themen Reichtum und Yoga; Brief, Proklamation.

Evaluation—Korpora

- Zusammensetzung der 26 Texte:
  - 16 (7+9) Texte im Within-Document-Retrieval-Modus (mit 1 bis 3 Suchtermen);
  - 10 (6+4) Texte im Summarizing-Modus (ohne Suchterme).
- Minimaler und maximaler Umfang der Texte in Sätzen:
  - Within-Document-Retrieval-Modus: 15 bis 68 Sätze;
  - Summarizing-Modus: 15 bis 82 Sätze.

Evaluation—Korpora

- Bewertung:
  - Studierende und Wissenschaftler(innen) vorwiegend der Informationswissenschaft Regensburg mit (sehr) guten Englisch-Kenntnissen;
  - Alter der Personen von 20 bis 73 Jahren.
- Durchführung:
  - selbständige Bearbeitung der Texte gemäß der Fragestellung (s. u.);
  - 1. Untersuchung ohne Aufwandsentschädigung (überwiegend Wissenschaftler[innen]),
    2. Untersuchung mit Aufwandsentschädigung (überwiegend Studierende).

Evaluation—Korpora

- Grundlegende Fragestellung für alle Texte nach einer Idee von [Hovy 2004: 594]:

  "Ask experts to underline and extract the most interesting or informative fragments of the text. Measure recall and precision of the system's summary against the human's extract …".

  - Welche Passagen interessieren den Leser eigentlich an einem Text dahingehend, dass sie in ein informatives Summary gehören?
  - Welche Stellen findet der Rezipient darüber hinaus grundsätzlich informativ/interessant?

Evaluation—Korpora

- Beobachtungen:
  - Für jeden Text wurden in beiden Modi ausreichend viele Sätze selektiert:
    - 3/4 der Bewerter haben mindestens 2 bis maximal 11 bzw. 13 Sätze selektiert (Untersuchung 1 bzw. 2);
    - 1/2 der Bewerter haben mindestens 5 bzw. 6 bis maximal 29 bzw. 20 Sätze selektiert (Untersuchung 1 bzw. 2).
  - Im Summarizing-Modus ergibt sich eine Kompressionsrate zwischen 8% bis 50% (im sinnvollen Bereich gemäß [Hovy 2004]).

Evaluation—Durchführung

- Leistungsmessungen:
  - Within-Document-Retrieval-Modus:
    - Keine aktuellen Vergleichssysteme verfügbar:
      - nicht (mehr) zugänglich: TOP(OGRAPH)IC [Hahn & Reimer 1986], TileBars [Hearst 1995], ProfileSkim [Harper & al. 2004];
      - Google-Buch-Suche: keine eigenen Texte auswertbar.
    - Leistungsmessung allein des EXCERPT-Systems:
      - Bei 50% Nutzerübereinstimmung (mehr Sätze in Referenz-Menge, einfacher für System);
      - bei 75% Nutzerübereinstimmung (weniger Sätze in Referenz-Menge, schwieriger für System).

Evaluation—Durchführung

- Leistung auf Korpus I und II (nur eigene Daten):

Korpora→ ↓Überein.	Korpus I (7 Texte)	Korpus II (9 Texte)	Gesamt-schnitt
50% Über-einstimm.	0.72	0.84	0.78
75% Über-einstimm.	0.70	0.78	0.74
Gesamt-schnitt	0.71	0.81	0.76

Evaluation—Durchführung

- Interpretation:
  - Im Schnitt werden 3 von 4 Passagen bezüglich der Referenz-Menge korrekt ermittelt, d. h. 3 von 4 ermittelten Passagen sind relevant und informativ für den Nutzer;
  - die Performance bei Passagen mit 75% Über-einstimmung ist kaum schlechter als bei 50%:
    - d. h. obgleich die Referenz-Menge bei 75% geringer ist als bei 50%, hält das System die Performance;
    - d. h. auch bei weniger ermittelten Passagen sind viele relevante und informative (Ab-)Sätze dabei.

Evaluation—Durchführung

- Summarizing-Modus:
  - Quervergleich mit anderen freien oder kommerziellen Summarizern (30-Tage-Testversionen von 5/2009):
    - Copernic-Summarizer [Copernic 2009];
    - Intellexer-Summarizer [EffectiveSoft 2009];
    - SubjectSearch-Summarizer [Kryloff 2009].
    - ➡ Einige weitere Summarizer nicht installierbar oder herunterladbar, andere konnten nur Webseiten zusammenfassen usw. (z. B. MEAD, QuickJist, Sinope u. a.).

Evaluation—Durchführung

- Zusätzlicher Vergleich mit Baseline:
  - Wahl der ersten N Sätze eines Textes inkl. Überschrift (Satz [0]) gemäß der Länge N des jeweiligen Referenz-Summarys;
  - Differenz zur Baseline erlaubt Aussage über Fähigkeit des Systems,
    - wichtige Information in allen Teilen des Textes positions-unabhängig ausfindig zu machen,
    - nicht unter eine bestimmte Performance-Grenze zu rutschen, bei der der Nutzer keinen Mehrwert mehr von einem maschinellen Summarizer hat.

Evaluation—Durchführung

- Ergebnisse:

	BL	CS	IS	SSS	EXC	Ø \ BL
Zech.	0.43	0.59	0.40	0.35	0.64	0.50
EK I	0.33	0.40	0.43	0.39	0.58	0.45
EK II	0.33	0.36	0.45	0.37	0.46	0.41
CAST	0.15	0.31	0.33	0.28	0.43	0.34
Ø	0.31	0.42	0.40	0.35	0.53	0.43

Legende: EK = Eigenes Korpus; BL = Baseline, CS = Copernic-, IS = Intellexer-, SSS = Subject-Search-Summarizer, EXC = EXCERPT.

Evaluation—Durchführung

- Interpretation:
  - Die Leistung nimmt tendenziell bei allen Systemen mit den verschiedenen Korpora gemeinsam zu oder ab;
  - Beobachtung:
    - Alle Systeme außer EXCERPT liegen mindestens einmal unterhalb des Durchschnitts;
    - der SubjectSearch-Summarizer liegt nahe an der Baseline und wird partiell von ihr geschlagen.

Evaluation—Durchführung

- Evaluation einiger Textmerkmale:
  - Im Hinblick auf die Frage nach Kriterien der Informativität oder Interessantheit von Texten wurden die eigenen Korpora auf 3 Textmerkmale hin ausgewertet:
    - Steigerungsformen (Komparative und Superlative);
    - Pronomen der 1. Person;
    - indefinite Determinierer.
  - Betrachtet wurden alle Sätze, die von mindestens 50% der Bewerter selektiert wurden (markiert mit '+', sonst '–' für Sätze mit < 50% Zustimmung).

Evaluation—Durchführung

- Ergebnisse:

	Korpus I	Korpus II	Schnitt
Steigerungs-formen	0.48 (+) 0.17 (-)	0.28 (+) 0.18 (-)	0.38 (+) 0.18 (-)
Pronomen 1. Person	0.07 (+) 0.36 (-)	0.04 (+) 0.19 (-)	0.06 (+) 0.28 (-)
indefinite Determin.	0.49 (+) 0.51 (-)	0.58 (+) 0.56 (-)	0.54 (+) 0.54 (-)

Alle Vorkommnisse eines Merkmals in selektierten (+) vs. nicht-selektierten (-) Sätzen des Textes wurden auf Vorkommnisse pro Satz umgerechnet.

Evaluation—Durchführung

- Interpretation:
  - Beispiel Steigerungsformen:
    - In Texten des Korpus I treten 0.48 Steigerungsformen pro Satz auf, der von mehr als 50% der Bewerter für informativ und/oder interessant beurteilt wurde;
    - hingegen treten in nicht-selektierten Sätzen nur 0.17 Steigerungsformen pro Satz auf (d. h. Sätze mit einer Zustimmungsrate der Bewerter von weniger als 50%).

Evaluation—Durchführung

- Zusammenfassend:
  - Steigerungsformen kommen in selektierten Sätzen doppelt so häufig vor wie in nicht-selektierten Sätzen ⇨ Indikator für zu selektierende Sätze;
  - Pronomen der 1. Person treten in nicht-selektierten Sätzen mehr als 4.5 Mal so häufig auf wie in selektierten ⇨ Indikator für *nicht* zu selektierende Sätze;
  - indefinite Determinierer treten praktisch gleich häufig auf ⇨ keinerlei Indikatorkraft für (nicht) zu selektierende Sätze.

Fazit

- Ausblick:
  - Suchmaschinen für Textpassagen sind der nächste Schritt nach Web- und Desktop-Suchmaschinen;
  - fortschreitende Digitalisierung von Texten wird Problem der Informationsfülle weiter verschärfen (vgl. Google-Buch-Suche; EBook-Reader).
- Prognose:
  - Integrierte Within-Document-Retrieval- und Summarizing-Systeme werden kommen und sich durchsetzen;
  - Within-Document-Retrieval und Summarizing werden als notwendige Ergänzung zum (Web-)Document-Retrieval vorausgesetzt werden.

Anhang—Beispieltext

- Beispiel: Text mit 29 Sätzen, bewertet von 13 Personen auf Relevanz und Informativität bzgl. der Suchbegriffe "mammoth" und "elephant" (Satz [0] ist Überschrift; ≥10 Nennungen [≥75% Zustimmung] orange, ≥7 Nennungen [≥50% Zustimmung] rot):

  [0] Regenerating a Mammoth for $10 Million

  [1] Scientists are talking for the first time about the old idea of resurrecting extinct species as if this staple of science fiction is a realistic possibility, saying that a living mammoth could perhaps be regenerated for as little as $10 million.

  [2] The same technology could be applied to any other extinct species from which one can obtain hair, horn, hooves, fur or feathers, and which went extinct within the last 60,000 years, the effective age limit for DNA.

Anhang—Beispieltext

[3] Though the stuffed animals in natural history museums are not likely to burst into life again, these old collections are full of items that may contain ancient DNA that can be decoded by the new generation of DNA sequencing machines.

[4] If the genome of an extinct species can be reconstructed, biologists can work out the exact DNA differences with the genome of its nearest living relative. [5] There are talks on how to modify the DNA in an elephant's egg so that after each round of changes it would progressively resemble the DNA in a mammoth egg. [6] The final-stage egg could then be brought to term in an elephant mother, and mammoths might once again roam the Siberian steppes.

[7] The same would be technically possible with Neanderthals, whose full genome is expected to be recovered shortly, but there would be several ethical issues in modifying modern human DNA to that of another human species.

Anhang—Beispieltext

[8] A scientific team headed by Stephan C. Schuster and Webb Miller at Pennsylvania State University reports in Thursday's issue of Nature that it has recovered a large fraction of the mammoth genome from clumps of mammoth hair. [9] Mammoths, ice-age relatives of the elephant, were hunted by the modern humans who first learned to inhabit Siberia some 22,000 years ago. [10] The mammoths fell extinct in both their Siberian and North American homelands toward the end of the last ice age, some 10,000 years ago.

[11] Dr. Schuster and Dr. Miller said there was no technical obstacle to decoding the full mammoth genome, which they believe could be achieved for a further $2 million. [12] They have already been able to calculate that the mammoth's genes differ at some 400,000 sites on its genome from that of the African elephant.

Anhang—Beispieltext

[13] There is no present way to synthesize a genome-size chunk of mammoth DNA, let alone to develop it into a whole animal. [14] But Dr. Schuster said a shortcut would be to modify the genome of an elephant's cell at the 400,000 or more sites necessary to make it resemble a mammoth's genome. [15] The cell could be converted into an embryo and brought to term by an elephant, a project he estimated would cost some $10 million. [16] "This is something that could work, though it will be tedious and expensive," he said.

[17] There have been several Russian attempts to cultivate eggs from frozen mammoths that look so perfectly preserved in ice. [18] But the perfection is deceiving since the DNA is always degraded and no viable cells remain. [19] Even a genome-based approach would have been judged entirely impossible a few years ago and is far from reality even now.

Anhang—Beispieltext

[20] Still, several technical barriers have fallen in surprising ways. [21] One barrier was that ancient DNA is always shredded into tiny pieces, seemingly impossible to analyze. [22] But a new generation of DNA decoding machines use tiny pieces as their starting point. [23] Dr. Schuster's laboratory has two, known as 454 machines, each of which costs $500,000.

[24] Another problem has been that ancient DNA in bone, the usual source, is heavily contaminated with bacterial DNA. [25] Dr. Schuster has found that hair is a much purer source of the host's DNA, with the keratin serving to seal it in and largely exclude bacteria.

[26] A third issue is that the DNA of living cells can be modified only very laboriously and usually at one site at a time. [27] Dr. Schuster said he had been in discussion with George Church, a well-known genome technologist at Harvard Medical School, about a new method Dr. Church has invented for modifying some 50,000 genomic sites at a time.

Anhang—Beispieltext

[28] The method has not yet been published, and until other scientists can assess it they are likely to view genome engineering on such a scale as being implausible. [29] Rudolph Jaenisch, a biologist at the Whitehead Institute in Cambridge, said the proposal to resurrect a mammoth was "a wishful-thinking experiment with no realistic chance for success."

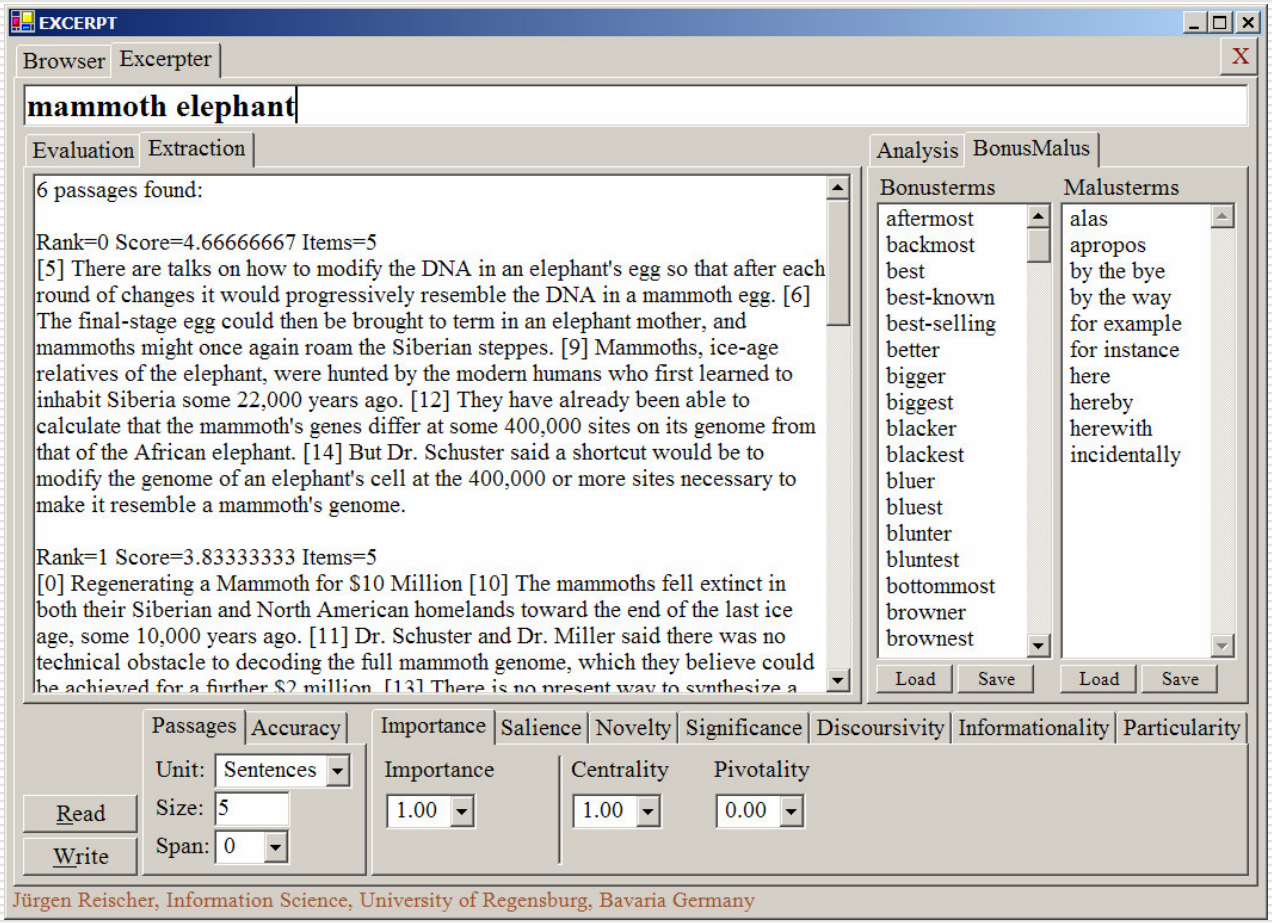
- Bestes topik-zentriertes Kompromiss-Summary bezüglich der Suchbegriffe "mammoth" und "elephant":
  - 5 Sätze mit mehr als 50% und 75% Zustimmung (Original-Anzahl an Sätzen, die von Versuchspersonen selektiert wurden);
  - vor allem Sätze selektiert, in denen beide Suchbegriffe gemeinsam auftreten (kursiv).

Anhang—Beispieltext

[5] There are talks on how to modify the DNA in an *elephant's* egg so that after each round of changes it would progressively resemble the DNA in a *mammoth* egg. [6] The final-stage egg could then be brought to term in an *elephant* mother, and *mammoths* might once again roam the Siberian steppes. [9] *Mammoths*, ice-age relatives of the *elephant*, were hunted by the modern humans who first learned to inhabit Siberia some 22,000 years ago. [12] They have already been able to calculate that the *mammoth's* genes differ at some 400,000 sites on its genome from that of the African *elephant*. [14] But Dr. Schuster said a shortcut would be to modify the genome of an *elephant's* cell at the 400,000 or more sites necessary to make it resemble a *mammoth's* genome.

Anhang—Beispieltext

- Suche mittels des EXCERPT-Systems:



Literatur

- [Alfonseca & Rodríguez 2003] Alfonseca, E. & Rodríguez, P. (2003): Generating Extracts with Genetic Algorithms. *Proceedings of ECIR 2003*, S. 511–519.
- [Baxendale 1958] Baxendale, P. B. (1958): Machine-Made Index for Technical Literature – An Experiment. *IBM Journal of Research and Development*, 2(4), S. 354–361.
- [Beaugrande & Dressler 1981] Beaugrande de, R.-A. & Dressler, W. (1981): *Introduction to Text Linguistics*. London & New York: Longman.
- [Borko & Bernier 1975] Borko, H. & Bernier, C. L. (1975): *Abstracting Concepts and Methods*. New York u. a.: Academic Press.
- [Copernic 2009] http://www.copernic.com/ (20.5.2009).
- [Edmundson 1969] Edmundson, H. P. (1969): New Methods in Automatic Extracting. *Journal of the American Society for Information Science*, 16(2), S. 264–285.

Literatur

- [EffectiveSoft 2009] http://summarizer.intellexer.com/ (20.5.2009).
- [Flesch 1948] Flesch, R. (1948): A New Readability Yardstick. *Journal of Applied Psychology*, 32(3), 221–233.
- [Floridi 2004] Floridi, L. (2004): Information. In Floridi, L. (2004; Hrsg.): *The Blackwell Guide to the Philosophy of Computing and Information*. Malden MA u. a.: Blackwell Publishing, S. 40–61.
- [Goldstein & al. 1999] Goldstein, J. & Kantrowitz, M. & Mittal, V. & Carbonell, J. (1999): Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of SIGIR'99*, S. 121–128.
- [Grewendorf 1981] Grewendorf, G. (1981): Pragmatisch sinnvolle Antworten. Ein entscheidungstheoretischer Explikationsvorschlag. In Krallmann, D. & Stickel, G. (1981): *Zur Theorie der Frage*. Forschungsberichte des Instituts für deutsche Sprache. Tübingen: Narr, S. 95–118.

Literatur

- [Hahn & Reimer 1986] Hahn, U. & Reimer, U. (1986): TOPIC Essentials. *Proceedings of the 11th Conference on Computational Linguistics*, S. 497–503.
- [Harper & al. 2004] Harper, D. J. & Koychev, I. & Yixing, S. & Pirie, I. (2004): Within-Document-Retrieval: A User-Centred Evaluation of Relevance Profiling. *Information Retrieval*, 7, S. 265–290.
- [Hasler & al. 2003] Hasler, L. & Orasan, C. & Mitkov, R. (2003): Building better corpora for summarisation. *Proceedings of Corpus Linguistics 2003*, S. 309–319.
- [Hearst 1995] Hearst, M. A. (1995): TileBars: Visualization of Term Distribution Information in Full Text Information Access. *Proceedings of CHI'95*, S. 56–66.
- [Heylighen & Dewaele 1999] Heylighen, F. & Dewaele, J.-M. (1999): *Formality of Language: definition, measurement and behavioral determinants*. Freie Universität Brüssel: Interner Report.

Literatur

- [Hovy 2004] Hovy, E. (2004): Text Summarization. In Mitkov, R. (2004; Hrsg.): *The Oxford Handbook of Computational Linguistics*. Oxford: University Press, S. 583–598.
- [Jastrzembski 1981] Jastrzembski, J. E. (1981): Multiple Meanings, Number of Related Meanings, Frequency of Occurrence, and the Lexicon. *Cognitive Psychology*, 13, S. 278–305.
- [Jurafsky & Martin 2009] Jurafsky, D. & Martin, J. H. (^{2}2009): *Speech and Language Processing*. London & al.: Pearson Education.
- [Keller 1995] Keller, R. (1995): *Zeichentheorie*. Tübingen & Basel: Francke.
- [Krifka 2006a] Krifka, M. (2006): Functional Similarities between Bimanual Coordination and Topic/Comment Structure. http://www.sfb632.uni-potsdam.de/publications/A2/A2_Krifka_2006.pdf (2.6.2008).

Literatur

- [Krifka 2006b] Krifka, M. (2006): Basic Notions of Information Structure. http://amor.rz.hu-berlin.de/~h2816i3x/Publications/Krifka_Information Structure.pdf (23.8.2009).
- [Kryloff 2009] http://www.kryltech.com/ (13.5.2009).
- [Lambrecht 1994] Lambrecht, K. (1994): *Information Structure and Sentence Form*. Cambridge: University Press.
- [Langer & al. 1974] Langer, I. & Schulz von Thun, F. & Tausch, R. (1974): *Verständlichkeit*. München & Basel: Reinhardt Verlag.
- [Mani 2001] Mani, I. (2001): *Automatic Summarization*. Amsterdam & Philadelphia: Benjamins.
- [Marcu 1997] Marcu, D. (1997): The Rhetorical Parsing of Natural Language Texts. *Proceedings of the ACL-97*, S. 96–103.
- [Mihalcea 2004] Mihalcea, R. (2004): Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. http://www.cs.unt.edu/~rada/papers/mihalcea.acl2004.pdf (17.10.2008).

Literatur

- [Mihalcea & Tarau 2004] Mihalcea, R. & Tarau, P. (2004): TextRank – bringing order into texts. http://www.cs.unt.edu/~rada/papers/mihalcea.emnlp04.pdf (17.10.2008).
- [Mittal & al. 1999] Mittal, V. & Kantrowitz, M. & Goldstein, J. & Carbonell, J. (1999): Selecting Text Spans for Document Summaries: Heuristics and Metrics. *Proceedings of AAAI-99*, S. 467–473.
- [Nenkova & Vanderwende 2005] Nenkova, A. & Vanderwende, L. (2005): The Impact of Frequency on Summarization. http://research.microsoft.com/apps/pubs/default.aspx?id=67448 (17.6.2009).
- [Nöth ²2000] Nöth, W. (²2000): *Handbuch der Semiotik*. Stuttgart & Weimar: Metzler.
- [Pinkal 1985] Pinkal, M. (1985): *Logik und Lexikon – Die Semantik des Unbestimmten*. Berlin & New York: de Gruyter.

Literatur

- [Paradis & Berrut 1996] Paradis, F. & Berrut, C. (1996): Experiments with Theme Extraction in Explanatory Texts. In Ingwersen, P. & Pors, N. O. (1996; Hrsg.): *Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Kopenhagen: The Royal School of Librarianship, S. 433–446.
- [Rosch 1978] Rosch, E. (1978): Principles of Categorization. In [Rosch & Lloyd 1978] Rosch, E. & Lloyd, B. B. (1978; Hrsg.): *Cognition and Categorization*. Hillsdale: Erlbaum, S. 27–48.
- [Rosch & al. 1976] Rosch, E. & Mervis, C. B. & Gray, W. D. & Johnson, D. M. & Boyes-Braem, P. (1976): Basic Objects in Natural Categories. *Cognitive Psychology*, 8, S. 382–439.
- [Rosch & Mervis 1975] Rosch, E. & Mervis, C. B. (1975): Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, S. 573–605.

Literatur

- [Salton & Buckley 1988] Salton, G. & Buckley, C. (1988): Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5), S. 513–523.
- [Spärck Jones 1972] Spärck Jones, K. (1972): A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), S. 11–21.
- [Strzalkowski & al. 1999] Strzalkowski, T. & Stein, G. & Wang, J. & Wise, B. (1999): A Robust Practical Text Summarizer. In Mani, I. & Maybury, M. T. (1999; Hrsg.): *Advances in Automatic Text Summarization*. Cambridge & London: MIT Press. S. 137–154.
- [Tengi 1998] Tengi, R. I. (1998): Design and Implementation of the WordNet Lexical Database and Searching Software. In Fellbaum, C. (1998; Hrsg.): *WordNet – An Electronic Lexical Database*. Cambridge & London: MIT Press, S. 105–127.
- [Weizsäcker 1974a] Weizsäcker, C. F. von (1974): *Die Einheit der Natur*. München: dtv.

Literatur

- [Weizsäcker 1974b] Weizsäcker, E. U. von (1974): Erstmaligkeit und Bestätigung als Komponenten der pragmatischen Information. In Weizsäcker, E. U. von (1974): *Offene Systeme I*. Stuttgart: Klett, S. 82–113.
- [Wimmer 2005] Wimmer, G. (2005): The type-token relation. In Köhler, R. & Altmann, G. & Piotrowski, R. G. (2005; Hrsg.): *Quantitative Linguistik. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*, Bd. 27. Berlin & New York: de Gruyter, S. 361–368.
- [Zechner 1995] Zechner, K. (1995): Automatic Text Abstracting by Selecting Relevant Passages. Edinburgh: M.Sc. Dissertation. http://www.cs.cmu.edu/~zechner/abstr.pdf (23.8.2009).