Sommersemester 2006 • Institut für Germanistik I

Vorlesung
Computerphilologie

Themenfeld
Textuelle Repräsentation

„Wie kann man formale Eigenschaften eines Textes oberhalb der Satzgrenzen mit dem Computer systematisch, repräsentieren, finden und verarbeiten?“

Textbegriff

- In der Computerphilologie wird meistens unter einem Text ausschließlich ein geschriebener Text verstanden, während der linguistische Textbegriff geschriebene wie gesprochene Realisierung umfaßt.
- Davon muß man unterscheiden, ob der Text geschriebene oder gesprochene Sprache (gesprochenen Stil) darstellt.

CP

CL

Stil \ Medium	Sprechen	Schreiben
gesprochen	Gesprochene Sprache	Sprechsprache
geschrieben	Vortragssprache	Text i. e.S. Schriftsprache

Text und Metainformation (-text)

- In der CP unterscheidet man häufig zwischen dem eigentlichen (linguistisch beschreibbaren) Text mit seinen Wörtern und Sätzen einerseits und andererseits der
- Metainformation über diese Einheiten, wie Auszeichnungen, Fonts, Farben, Strukturierung (Zeile, Absatz, Kapitel, etc.) oder Dokumentinformation wie Autorenname, Verlagsname, Kurztitel, ISBN-Nr, etc.)

Text- / Diskursrepräsentation Schwerpunkt Kohärenz

- Kohärenzgraphen haben als Knoten
  - Weltobjekte, Konzepte, Wörter, Wortgruppen, Teilsätze, Sätze, Teiltexte

  Und als Kanten
  - Kohärenzklassen, wie Deixis, Generalisierung, Instantiierung, Rekurrenz, Bedeutungspostulat
- Es gibt zur Zeit noch keine Tools für Kohärenzgraphen oder Indexverwaltung

Hans braucht Geld. Er zerschlägt sein Sparschwein und

kauft sich davon ein Fahrrad

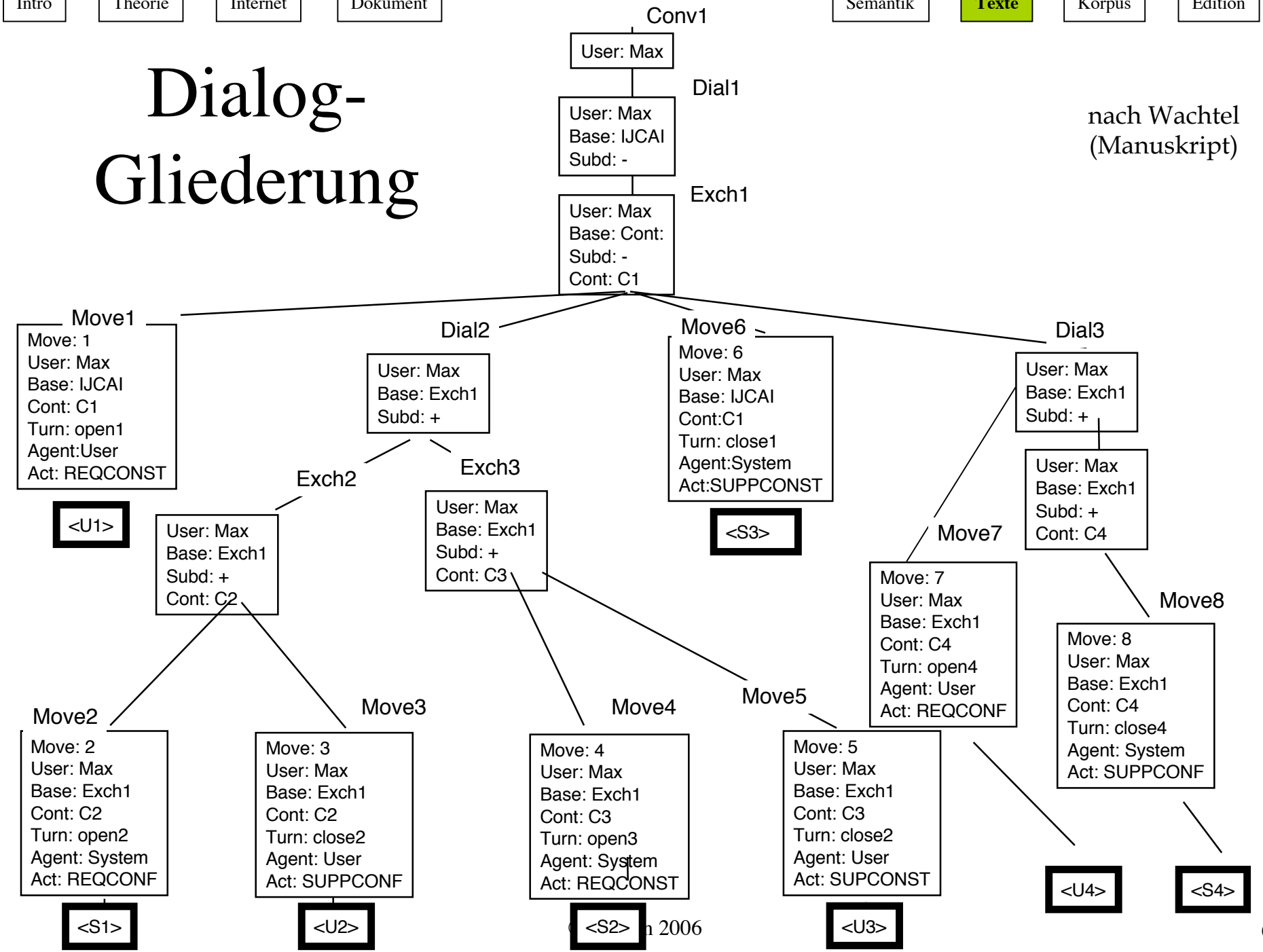
Textrepräsentation Schwerpunkt Dialog

- Zur Darstellung von Dialogstrukturen gibt es eine große Zahl von Formaten (z.B. Dialogakte, Dialoghandlungen).
- Dafür gibt es keine dedizierten Tools, falls man nicht ein Partitursystem auch dafür benutzt.
- Man könnte sich ein (interaktives) Tool (Editor) vorstellen, das die Architektur eines Dialogs nach den Parametern
  - Dialog, Zug, Sprecher, Dialogakt, Thema, Inhalt,Vorgänger, Nachfolger, Signalreferenz Anfang/Ende

  Beschreibt.
- Ein Dialogakt-Assigner ist im Evaluationstool (Plattform:Unix) von Verbmobil enthalten, arbeitet aber nur auf VM-Formaten und VM-Files.

Dialog-Gliederung

nach Wachtel (Manuskript)



Textrepräsentation Schwerpunkt Diskurs

- Es gibt weder eine generelle Diskursrepräsentationstheorie noch eine computergestützte Erstellung derselben
- Die logische Modellierung von Diskursen wird von DRT geleistet. Hierzu gibt es eine graphische Form, aber keine Computerunterstützung.

DRT

- Diskursrepräsentationssprachen stellen ein (logikbasiertes) Mittel dar, Zusammenhänge innerhalb und außerhalb eines Diskurses darzustellen. Dazu gehört z.B.
- Die Kohärenz und der
- Wechsel der Situation
- Die Repräsentationssprache DRT erzeugt je eine DRS (Discourse Representation Structure)

Rhetorical Structure Theory

Die Hauptfrage, die man mit RST -Graphen beantworten möchte, ist:

> *Welche Funktion oder welche Aufgabe spielt ein bestimmter Textabschnitt in seinem und für seinen umgebenden Kotext?*

Durch die Frage nach dem „Wie“ der Textanordnung erhält man interessante Aufschlüsse über die Kommunikation in diesem Text und die Kommunikation überhaupt

Rhetorical Structure Theory (RST)

RST ist eine graphische Darstellungsform (eine Sprache) für den stilistisch/logisch/argumentativen Zusammenhang von Teiltexten.

Die Darstellung ist statisch, deklarativ und kontextfrei,
Betrachtungsaspekte sind:

- Textbeschreibung, Generierungsschablonen,
- "Rhetorische" Wohlgeformtheit,
- erwartbarer Effekt auf den Rezipienten und die Lokalisierung des Effekts,
- Informations-Anordnung,
- Textsortenunterschiede,
- einzelsprachliche Unterschiede (?),
- Indifferent gegen Sozialgruppen und Domänen.

Bei der folgenden Darstellung folge ich Folien von Vaut und Wiebelitz

RST Entwicklung

Entwickelt von Mann, Matthiessen und Thompson an der Univ. of South.Cal. im Zusammenhang mit intelligenter Generierung.

RST ist daher auch vor allem geeignet, um (englische) Texte zu beschreiben und Texte zu generieren, weniger, um Textverstehen zu analysieren.

Untersucht sind die Textsorten Zeitung, Werbung, persönliche Briefe, Leserbriefe, politische Stellungnahmen, Rezepte, Artikel und wissenschaftliche Abstracts.

RST: Grundlegende Annahmen

Grundlegende Annahmen:

1. Organisation: Texte bestehen aus funktional bedeutsamen Segmenten, die zu größeren Segmenten zusammengefügt werden können und diese wiederum zu ganzen Texten.
2. Geschlossenheit und Kohärenz: Der Text, zu dem jeder Teiltext beiträgt, muß als ganzes in sich geschlossen sein.
3. Der Text muß zielgerichtet sein. Der Autor muß einen Effekt erzielen wollen.
4. Der Text sollte hierarchisch gegliedert sein.
5. Der Text sollte relativ homogen sein
6. Die einzelnen Segmente eines Textes sollten durch Relationen verbunden sein.

RST: Basiseinheiten, Nukleus und Sateliten

Zunächst muß der Text in Basiseinheiten zerlegt werden. Umfang und Größe der Basiseinheiten ist prinzipiell willkürlich. Meist sind es Teilsätze oder Phrasen, in Ausnahmefällen auch Paragraphen.

Gefundene überlappungsfreie Basiseinheiten werden anschließend durch Relationen miteinander verbunden.

Der Nukleus spielt dabei die Rolle des Motiv- und Informationsträgers, der Satellit besteht aus Erläuterungen zum Verständnis oder bessere Akzeptanz beim Hörer.

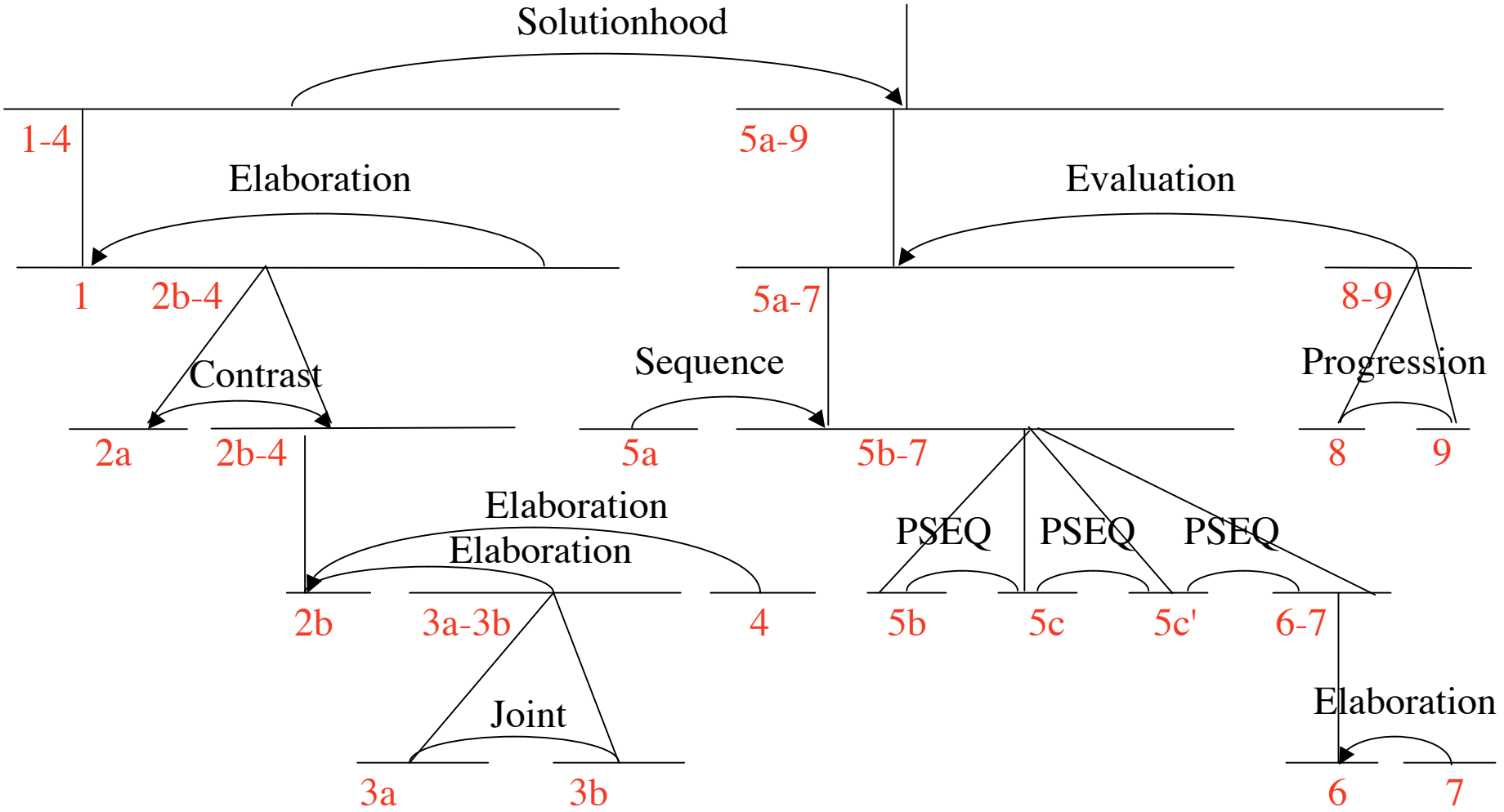
Abschnitte werden in Schemata angeordnet, die die Zahl und Arten der Relationen festlegen.

RST: Relationen

Elaboration
Circumstance
Solutionhood
Volitional Result
Non-volitional Result
Non-volitional Cause
Purpose
Condition
Otherwise
Interpretation

Evaluation
Restatement
Summary
Motivation
Antithesis
Background
Enablement
Evidence
Justify
Concession

RST-Beispiel (nach Rösner/Stede)



Der Beispieltext

1 In unserer mobilen Gesellschaft ist geistige Mobilität gefordert.
2a Es geht nicht mehr nur darum, möglichst schnell von A nach B zu kommen
2b es geht zunehmend um die Frage: Wie?
3a Welches Verkehrsmittel ist wann das beste,
3b welches ist gesellschaftlich zu verantworten?
4 Zur Wahl stehen Auto, Flugzeug und Bahn.
5a Die Fakten:
5b Die Bahn benötigt bei gleicher Verkehrsleistung 1/3 der Fläche einer Autobahn,
5c Sie verbraucht 71,5% weniger Energie,
5c' Und produziert 87,9% weniger Schadstoffe als das Auto.
6 Flüge unter 400 km halten selbst Experten der Airlines für ökonomisch wenig sinnvoll.
7 (Das ist fast ganz Deutschland, von Frankfurt aus gesehen!)
8 Alles Antworten auf brennende Fragen.
9a Wenn wir sie nicht stellen,
9b Unsere Kinder bestimmt!

Spezifikation der Relationen

(Beispiel)

- Relationsname: SOLUTIONHOOD
- Bedingungen an N: keine
- Bedingungen an S: Präsentiert ein Problem
- Bedingungen an die Kombination von N + S:
  Die in N präsentierte Situation ist eine Lösung
  für das in S präsentierte Problem
- Effekt: L erkennt die Situation in N als Lösung
  für das in S präsentierte Problem
- Ort des Effekts: N+S

RST: Anwendung

Leider gibt es keinen

- RST-Tagger, noch einen
- Parser oder einen
- graphischen Browser.

Man kann sich allerdings gut vorstellen, daß RST-Tags in XML formuliert werden und dann

- nach den Schemata (automatisch) RST-Graphen gezeichnet werden und
- aus einem RST-getaggten Text ausgewertet wird.

RST- Literatur

Mann, William, Thompson, Sandra A., Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. In Text 8.3, 1988. Seite 243 - 281.

Rösner, Dietmar, Stede, Manfred, Zur Struktur von Texten. Eine Einführung in die Rhetorical Structure Theory. In: KI. Künstliche Intelligenz 2, 1993. Seite 14 - 21

Automatische Zusammenfassung

(3) Gliederung

Die Einteilung einer Arbeit in Abschnitte und Unterabschnitte ist keine Frage einer gefälligen graphischen Form, sondern sie ist Ergebnis der inneren Logik einer Arbeit bzw. der Aufgabe. Daher muß die Gliederung den Gang der Argumentation spiegeln. (Siehe Anlage)

Machen Sie Sich rechtzeitig Gedanken über die Umfangsverhältnisse der einzelnen Teile Ihrer Arbeit, damit nicht nachher die Literaturübersicht 2/3 ausmacht und die eigentliche eigene Arbeit nur 1/4 abbekommt.

(5) Umgang mit Literatur

In den meisten Arbeiten stellt man im Abschnitt "Stand der Forschung" (oder ähnlich) wesentliche Literatur kurz dar. Die dann noch verbleibende Literatur stellen Sie knapp und allein unter dem Aspekt Ihrer Arbeit dar.

6. Zusammenfassung
7. Literaturverzeichnis

(11) Weitere Literatur

Wer es genauer wissen will, kann lesen:

Poenicke, Klaus, Die schriftl. Arbeit.

Automatische Zusammenfassung

(3) Gliederung

Die Einteilung einer Arbeit in Abschnitte und Unterabschnitte ist keine Frage einer gefälligen graphischen Form, sondern sie ist Ergebnis der inneren Logik einer Arbeit bzw. der Aufgabe. Daher muß die Gliederung den Gang der Argumentation spiegeln. (Siehe Anlage)

Machen Sie Sich rechtzeitig Gedanken über die Umfangsverhältnisse der einzelnen Teile Ihrer Arbeit, damit nicht nachher die Literaturübersicht 2/3 ausmacht und die eigentliche eigene Arbeit nur 1/4 abbekommt.

(5) Umgang mit Literatur

In den meisten Arbeiten stellt man im Abschnitt "Stand der Forschung" (oder ähnlich) wesentliche Literatur kurz dar. Die dann noch verbleibende Literatur stellen Sie knapp und allein unter dem Aspekt Ihrer Arbeit dar.

6. Zusammenfassung
7. Literaturverzeichnis

(11) Weitere Literatur

Wer es genauer wissen will, kann lesen:

Poenicke, Klaus, Die schriftl. Arbeit.

Zusammenfassung von MS Word

(5) Umgang mit Literatur
In den meisten Arbeiten stellt man im Abschnitt "Stand der Forschung" (oder ähnlich) wesentliche Literatur kurz dar. Die dann noch verbleibende Literatur stellen Sie knapp und allein unter dem Aspekt Ihrer Arbeit dar.

6. Zusammenfassung
7. Literaturverzeichnis

(11) Weitere Literatur
Wer es genauer wissen will, kann lesen:
Poenicke, Klaus, Die schriftl. Arbeit.

Summarizing

Es gibt

- „blinde“ Verfahren. Diese orientieren sich an formalen Eigenschaften, wie Überschriften, Auszeichnungen oder expliziten „Zusammenfassungen“ im Text, gesucht wird häufig am Anfang und am Ende eines Textes,
- Lexikalische Verfahren, die nach lexikalischen Hinweisen suchen, wie abstrahierenden Wörter („allgemein“, „Definition“, „zusammenfassen“)

und

- wissensbasierte Verfahren, die z.B. mithilfe von Ontologien allgemeinere oder definitorische Abschnitte suchen oder eine Verbalisierung des genannten Teilgraphen vornehmen.

oder

- Kombinationen davon

Es gibt keine ausgereiften Verfahren