UNIVERSITÄT DES SAARLANDES
GRUNDLAGEN- UND GESCHICHTSWISSENSCHAFTEN
FACHRICHTUNG 5.5 - INFORMATIONSWISSENSCHAFT

Computergestütztes Texterschließungssystem
KURZBESCHREIBUNG DES SYSTEMS

PROF. DR. HARALD H. ZIMMERMANN 6600 SAARBRÜCKEN 11 TELEFON: 0681-302-3537
PROJEKT TRANSIT: GEFÖRDERT VOM BMFT KENNZEICHEN: PT 200.10/ IDA 0160-6

CTX
COMPUTERGESTÜTZTES TEXTERSCHLIESSUNGSSYSTEM

---

1. EINFÜHRUNG

Das Computergestützte Texterschließungssystem CTX stellt das Ergebnis langjähriger universitärer Forschungen im Bereich der Informationssoftware dar. Die in grundlagenorientierter Forschung entwickelten Modelle zur Sprachdatenverarbeitung wurden anwendungsorientiert zu Verfahrensbausteinen für die Texterschließung fortentwickelt. Aufbauend auf einer wort- und satzorientierten Verarbeitung von Texten werden dabei zu einem deutschsprachigen Text/Dokument formal-inhaltliche Stichwörter (Deskriptoren) erstellt.

CTX wurde erstmals an juristischen Dokumenten im Bereich Datenschutz getestet. Gegenwärtig wird das System anwendungsnah in unterschiedlichen Bereichen erprobt. CTX kann aufgrund seiner modularen Struktur und seiner rechnerunabhängigen Schnittstellen in verschiedene Information-Retrieval-Systeme eingebunden bzw. als Indexierungskomponente angeschlossen werden.

2. FUNKTIONEN

Im Rahmen einer formal-inhaltlichen Erschließung natürlichsprachiger Texte erfüllt CTX fol-

gende Funktionen:

- Die aus dem Text extrahierten sinntragenden Wortformen werden maschinell (mittels eines allgemeinsprachlichen Stammlexikons mit über 130.000 Stammeinträgen) auf ihre jeweilige Grundform reduziert.

  Beispiele:

Textwortform	Grundform
Vorzüge	VORZUG
trat	TRETEN
trifft ... zu	ZUTREFFEN

- ZusammengesetzteWörter (Komposita) werden zusätzlich in sinntragende, ggf. vereindeutigte Bestandteile zerlegt.

  Beispiel:

  Persönlichkeitssphäre

  Teil: PERSÖNLICHKEIT (Individuum)
  Teil: SPHÄRE

- Die Mehrdeutigkeit von Textwörtern wird aufgezeigt; ggf. werden mehrdeutige Wörter aufgrund des Textzusammenhangs computergestützt vereindeutigt.

  Beispiel:

  ...in der Praxis der Datenverarbeitung ...

  PRAXIS (prakt. Vorgehen) - im Gegensatz zu Arztpraxis

- Mehrwortige Begriffe werden aufgrund eines entsprechenden Regelsystems erkannt.

  Beispiele:

tritt ... in Kraft	IN KRAFT TRETEN
personenbezogene, durch das Gesetz geschützte Daten	PERSONENBEZOGENE DATEN

- Fachgebietsorientierte Einfache und Komplexe Deskriptoren /1/ werden durch Nutzung entsprechender (z.T. vom Anwender pflegbarer) Lexika ermittelt.

  Beispiele:

... modernen Industriestaaten ..

(einfach) INDUSTRIESTAAT
(komplex) MODERNER INDUSTRIESTAAT

- Nicht-sinntragende Wörter ("Funktionswörter": DAS, ABER, OBWOHL, UND etc.) werden ausgefiltert.

Besonderer Wert wurde auf die Realisierung einer leistungsfähigen Lexikonkomponente und auf die weitgehend automatisierte Lexikonpflege gelegt. Allgemeinsprachliche Daten werden durch ein morpho-syntaktisches /2/ und ein semantisches /3/ Lexikon erfasst. Fachsprachliche Daten werden durch ein fachspezifisches Lexikon sowie ein semantisches Relationenlexikon ("Thesaurus" /4/) identifiziert und beschrieben. Im fachsprachlichen Bereich kann der Anwender ggf. selbst Begriffsstruktur und Begriffsverwendung bestimmen.

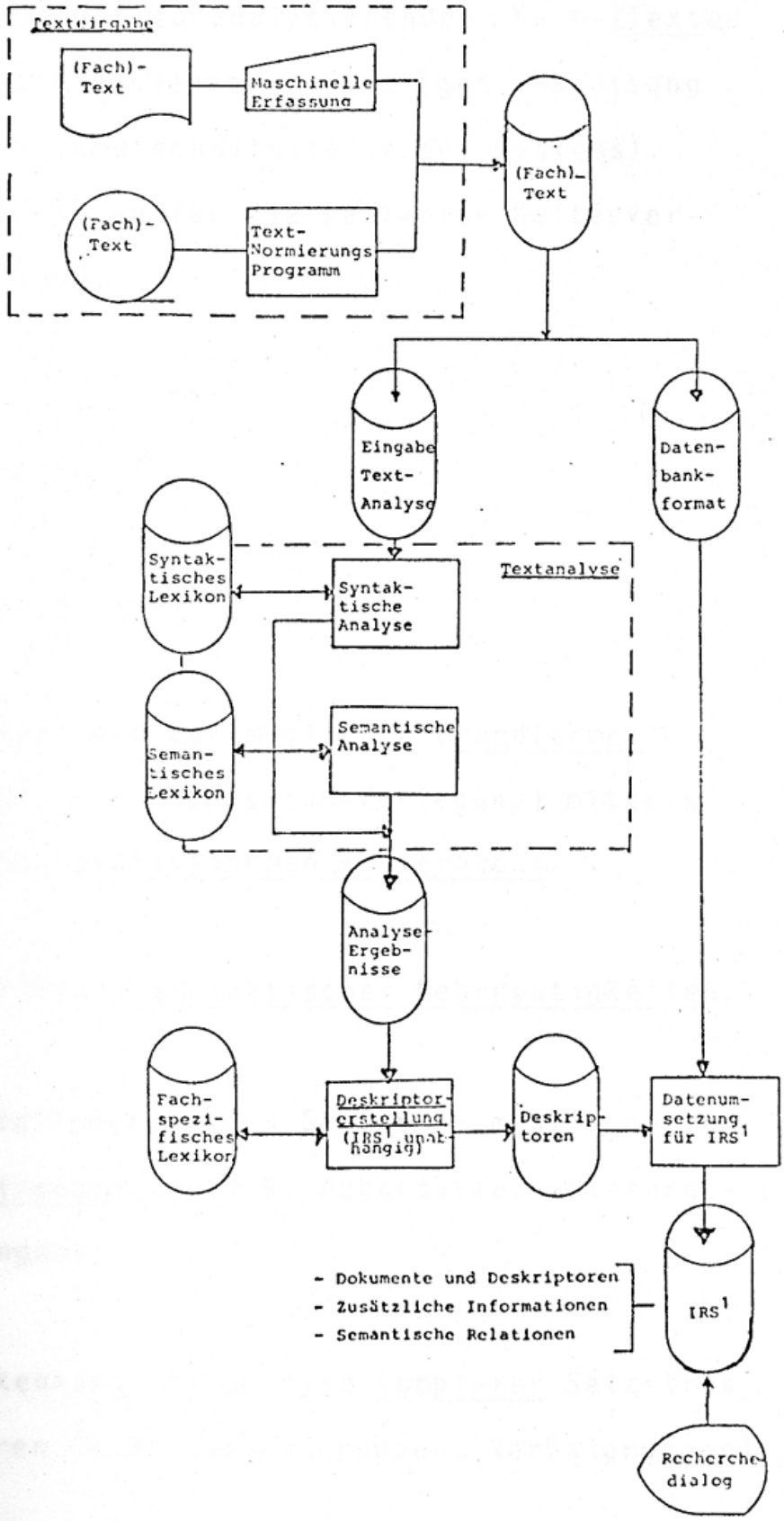
3. VERFAHREN

Die Grundlage der Texterschließungskomponente bildet das Saarbrücker Übersetzungssystem (SUSY), dessen Analysekomponente in CTX eingebracht wurde. Die linguistische Analyse ist vorwiegend satz- und syntaxorientiert.

Das System kann ggf. in rechnerspezifische Retrievalkomponenten integriert werden. Auf diese Weise wird die Analyse und Bearbeitung einer natürlichsprachigen Problembeschreibung während des Retrievalvorgangs nach den gleichen Regeln möglich, wie sie bei der Texterschließung (Indexierung der Dokumente) verwendet werden. Dies erlaubt eine einfache formale Anpassung der Suchbegriffe an die indexierten Termini der Texte/Dokumente.

Der Verfahrensablauf der Texterschließungskomponenten "Textanalyse" und "Deskriptorerstellung" kann hier nur grob skizziert werden.

ABLAUFSCHEMA VON CTX


Texteingabe
(Fach)-Text
Maschinelle Erfassung
(Fach)-Text
Text-Normierungs Programm
(Fach)-Text
Eingabe Text-Analyse
Daten-bank-format
Textanalyse
Syntak-tisches Lexikon
Syntak-tische Analyse
Seman-tisches Lexikon
Semantische Analyse
Analyse-Ergeb-nisse
Fach-spezi-fisches Lexikon
Deskriptor-erstellung (IRS unabhängig)
Deskrip-toren
Datenum-setzung für IRS
- Dokumente und Deskriptoren
- Zusätzliche Informationen
- Semantische Relationen
IRS
Recherche-dialog

Begriffserklärung: [1] IRS = Information-Retrieval-System

TEXT-EINGABE

- Eingabe des zu analysierenden (Fach-)Textes in maschinenlesbarer Form (ggf. Anpassung an die Inputschnittstelle des Systems). Aufbereitung für die satzweise Weiterverarbeitung.

TEXT-ANALYSE

Syntaktische Analyse

- Ermittlung der möglichen Grundformen (ggf. mit Kompositum-Zerlegung) mittels eines syntaktischen Wörterbuchs.
- Auflösung syntaktischer Mehrdeutigkeiten.
- Aufgliederung des Satzes in mögliche Satzsegmente (z.B. Nebensätze, Satzanreihungen).
- Erkennung und Analyse komplexer Satzstrukturen (z.B. Nominalgruppen, Verbalgruppen).

Semantische Analyse

- Reduktion vorliegender Mehrdeutigkeiten von Wortbedeutungen mittels eines bedeutungsorientierten Regelsystems (semantisches Wörterbuch).

Ergebnis der Textanalyse:

Die Textwortformen sind auf z.T. schon eindeutige Grundformen zurückgeführt, die Satzstrukturen sind ermittelt.

DESKRIPTOR-ERSTELLUNG

- Stichwörter und komplexe Ausdrücke sowie Informationen über deren syntaktische Strukturen werden zur Verfügung gestellt. Mit Hilfe eines fachspezifischen Lexikons werden außerdem noch nicht reduzierte Mehrdeutigkeiten über eine fachgebietsorientierte Gewichtung aufgelöst.

Ergebnis der Deskriptor-Erstellung:

Formal-inhaltliche Deskriptoren mit strukturellen Zusatzinformationen in einer systemunabhängigen Schnittstelle.

4. ANWENDUNGSBEREICHE

Für CTX bieten sich vielfältige Einsatzmöglichkeiten an:

- CTX-SERVICE und CTX-DATENBANKEN
  - Verwendung von CTX als Texterschließungsverfahren
  - Verwendung als computergestütztes Indexierungsverfahren zum Aufbau von Informationsbanken
  - Verwendung als Texterschließungsmittel im Bereich der computerunterstützten Inhaltsanalyse
- CTX-IRS (Information-Retrieval-System)
  - Integration von CTX-Funktionen in ein Information-Retrieval-System
- CTX-MULTILINGUAL
  - Retrieval unter Integration fremdsprachiger Synonyme in den fachgebietsspezifischen Thesaurus
- CTX-REGISTER
  - Aufbau von Grundformen-Registern zu deutschsprachigen Texten
- CTX/SUSY

Die Ausweitung des CTX-Systems auf Übersetzungen (Deutsch-Englisch, Deutsch-Französisch) unter Verwendung der Übersetzungskomponente des Saarbrücker Übersetzungssystems (SUSY) ist in Vorbereitung. Diese Erweiterungen haben zum Ziel:

- Automatische Sprachübersetzung
- Dokumentation und Indexierung mit Übersetzung von Titeln oder Abstracts in ausgewählten Bereichen.

***

Mit CTX liegt inzwischen ein praktikables Softwaresystem zur Lösung von Problemstellungen im Bereich der natürlichsprachigen Texterschließung vor. Es ist zugleich ein System, das sich durch große Modularität und damit Anpassungsfähigkeit an wechselnde Anforderungen der Anwender und ein breites Spektrum von Anwendungsmöglichkeiten auszeichnet.