Das Inforum
Nr. 11, August 1981

Harald Zimmermann

Das Projekt JUDO-DS

'Juristische Dokumentanalyse' im Bereich Datenschutz' (JUDO-DS)

Gefördert vom Bundesminister für Forschung and Technologie unter dem Kennzeichen PT 131.08

Gliederung:

0. Übersicht

Förderung:	BMFT Projektträger: GID
Förderungszeitraum:	01.07.77 - 31.12.79 (JUDO-Grundlagensystem) (FKZ: PT 131.04) 01.01.80 - 31.12.81 (JUDO-DS; Anw. Datenschutz)
Personalausstattung:	8 wiss. Mitarbeiter, 5 stud. Hilfskräfte
Adresse:	Universität des Saarlandes Prof. Dr. Harald H. Zimmermann 5.5 Informationswissenschaft 6600 Saarbrücken 11 Tel. (0681) 302-3537
Projektmitarbeiter:	F.W. Felzmann, U. Hahl, D. Hai, Th. Klein, W. Kopelent, E. Kroupa, N. Lang, M. Line, G. Tham, D. Viets, Th. Wagner, H. Werner M. Werner, H. Whippey, H.H. Zimmermann
Zielsetzung:	Modellentwicklung eines Software-Systems zur computergestützten Indexierung am Beispiel der Analyse JUristischer Dokumente (JUDO) unter Anwendung im Bereich Datenschutz (JUDO-DS)

1. Das System JUDO

JUDO umfasst die Modellentwicklung eines Software-Systems zur computergestützten Indexierung auf linguistischer Grundlage am Beispiel juristischer Dokumente, v.a. zum Datenschutzrecht (Abb. 1). Unter Indexierung wird hier die Ermittlung und Darstellung von in einem Dokument vorhandenen sinntragenden Wörtern verstanden, die das jeweilige Dokument inhaltlich beschreiben. Zu lösende Aufgaben sind dabei u.a.

- die Ermittlung der korrekten Grundform (FAELLE --- FALL/FAELLEN)
- die Vereindeutigung von mehrdeutigen Wörtern (ANLAGE --- (1) Computer, (2) Schriftstück ...)
- die Ermittlung mehrwortiger (Fach-) Begriffe (JURISTISCHE PERSON)
- die Verknüpfung von Begriffen bei der Indexierung (Präkoordination) (ÄNDERUNG K LÖSCHUNG)
- die sinnvolle Zerlegung von Komposita
  (Datenschutzrecht --- Recht, Datenschutz; Zweigniederlassung --- Niederlassung (nicht aber: Zweig)).

Im Rahmen des Projekts wird die Frage der (automatischen) Selektion von Begriffen (oder der Gewichtung) bezüglich der Relevanz für ein Dokument gegenwärtig nicht behandelt.

Wesentliche Merkmale des Verfahrens sind:

- Einsatz linguistischer Verfahren
  Durch die Verwendung der Saarbrücker Automatischen Text-Analyse (SATAN) werden die zu verarbeitenden Texte einer morphologischen, syntaktischen und ansatzweise einer semantischen Analyse unterzogen, die über eine kontextunabhängige Einzelwortanalyse hinausgeht und eine Strukturbeschreibung der Texteinheit 'Satz' liefert. Die Ergebnisse einer derartigen Analyse erlauben u.a.:

  - eine automatische Bereitstellung von Textwörtern (Substantive, Verben, Adjektive), die auf ihre Grundform zurückgeführt sind, als Einfache Deskriptoren (Abb.2),
  - eine automatische Bereitstellung von Komplexen Deskriptoren. Dabei handelt es sich um die Kombination von im Text belegten Einzeldeskriptoren, die in bestimmten syntaktischen Relationen zueinander stehen (Abb. 2).

- Fachgebietsorientierte Deskribierung von Dokumenten unter Verwendung von Wörterbüchern

In der Entwicklungsphase orientiert sich JUDO am Datenschutzrecht. Als Texte sollen in Auswahl Gesetze, Judikate, Berichte, Fachlexika, Kommentare, Zeitungsartikel und sonstige Meldungen verarbeitet werden. Diese Texte sind in Dokumente zerlegt (bei Gesetzen z.B. die Paragraphen).

Zur Ermittlung der Deskriptoren werden (fachtextspezifische) Lexika aufgebaut. Diese Lexika enthalten unter anderem das für das Fachgebiet typische Wortmaterial, etwa auch mehrwortige

Begriffe, die speziell auf dem Gebiet Datenschutz von Bedeutung sind (Bsp.: RECHT AUF AUSKUNFT; PERSONENBEZOGENE DATEN).

- Thesauruserstellung, -erweiterung und -pflege

Eine weitere Hilfe zur inhaltlichen Erschließung eines Dokumentes stellt der fachspezifische Thesaurus dar. Semantische Relationen zwischen den Deskriptoren sollen - zuschaltbar vom jeweiligen Benutzer des Systems - bei der Abfrage eine 'Anpassung' zwischen dem Vokabular des Anfragenden und den auf die Dokumente verweisenden Deskriptoren ermöglichen (Abb.4).

- Einbringung der Ergebnisse in ein Retrievalsystem

Die Daten (Wörterbücher, Analyseergebnisse, Thesaurusinformationen) werden auf einer Schnittstelle ausserhalb eines konkreten Anwendungssystems aufbewahrt und gepflegt. Sie können im Prinzip - durch Implementierung verschiedener Anpassungssoftware - in beliebige Informationssysteme überführt werden.

Die Indexierungsergebnisse des Systems JUDO werden im Rahmen des Projekts JUDO-DS in eine mit dem System GOLEM erstellte Informationsbank eingebracht. Die äußere Form eines Dokuments, insbesondere die aspektgebundene Deskribierung, orientiert sich dabei weitgehend an den JURIS-Konventionen.

- Natürlichsprachige Problembeschreibung

Im Rahmen des Systems JUDO wurde ein Verfahren entwickelt, das es dem Benutzer erlaubt, sich aus einer von ihm formulierten natürlichsprachigen Problembeschreibung automatisch Deskriptoren ermitteln zu lassen, die dann beim Retrieval benutzt werden können. Dieses Vorgehen soll den Benutzer bei der Erstellung systemkonformer Deskriptoren unterstützen (also bei der Verwendung der Grundformen, insbesondere aber bei der Erstellung der Komplexen Deskriptoren (Abb. 4)). Diese Systemvariante ist gegenwärtig nur auf dem Entwicklungsrechner von JUDO (Telefunken TR 440) verfügbar, eine Migration auf Siemens (7.760) soll noch im Laufe von 1981 - in Zusammenarbeit mit dem SFB 100 Elektronische Sprachforschung an der Universität des Saarlandes - erfolgen.

2. Linguistische Grundlagen

Die Qualität der Indexierungsergebnisse ist entscheidend für die Akzeptanz eines Dokumentationssystems. Ausgehend von der Annahme, dass linguistische Verfahrensweisen

- die Indexierungsergebnisse verbessern (vgl. die Erstellung Komplexer Deskriptoren sowie die Ermittlung mehrwortiger Begriffe) und
- den Retrievalvorgang für den Benutzer komfortabler gestalten (vgl. die natürlichsprachige Problembeschreibung),

sind die Entwicklung und Implementierung linguistischer Methoden die zentrale Aufgabenstellung des Projekts JUDO.

Im folgenden sollen kurz zwei Schwerpunkte der Arbeit umrissen werden:

- Semantische Disambiguierung

Für das Ziel, im Rahmen von JUDO bei der Dokumentindexierung semantisch eindeutige Deskriptoren (z.B. PROZESS im Sinne von "Gerichtsverfahren") zur Verfügung zu stellen, werden u.a. drei Verfahren erprobt:

a) Semantische Disambiguierung im Rahmen der Satzanalyse

Die Saarbrücker Automatische Textanalyse - entwickelt vom SFB 100 - stellt ein Verfahren zur Auflösung von Mehrdeutigkeiten zur Verfügung. Erste Auswertungen dieses (noch im Aufbau befindlichen) Programmteils haben gezeigt, dass bisher etwa die Hälfte der (im Text) belegten mehrdeutigen Wörter aufgrund des Satzkontexts erfolgreich vereindeutigt werden.

b) Wahrscheinlichkeitsorientierte Vereindeutigung

Unter Verwendung einer fachgebietsorientierten Angabe, die die Wahrscheinlichkeit widerspiegelt, mit der (hier aus fachlich-juristischer Sicht) das Vorkommen der einzelnen Bedeutungen (Bedeutungsvarianten) eines mehrdeutigen Wortes in einem bestimmten Fachgebiet zu erwarten ist, werden niedrig gewichtete Bedeutungsvarianten ausgeschlossen und so eine Vereindeutigung bzw. eine Reduktion möglicher Bedeutungen erreicht. (Diese Markierung muss natürlich von Fachgebiet zu Fachgebiet ggf. neu festgelegt - u.U. auch empirisch ermittelt - werden.)

c) Lexikalische und satzübergreifende Vereindeutigungen

Zu einer ergänzenden Methode, bei der im Prinzip alle zur Verfügung stehenden Informationen (Fachgebietszugehörigkeit, semantische Begriffsbeziehungen im Thesaurus, Kompositazerlegung und der Kontext ausserhalb eines Satzes) einbezogen werden, liegen erste Lösungsansätze vor.

- Syntaktisch verschiedene (Satz-)Strukturen mit gleichem Informationsgehalt (Paraphrasen)

Gleiche Informationen können in verschiedenen (syntaktischen) Strukturen wiedergegeben sein. Ein Mensch ist normalerweise in der Lage, syntaktischen Varianten ein- und denselben Informationsgehalt zuzuordnen. Innerhalb eines computergestützten Informationssystems stellt sich die Aufgabe, oberflächensyntaktisch unterschiedliche Strukturen als "informationsgleich" zu erkennen, zusammenzuführen und für den Retrievalvorgang nutzbar zu machen.

> Beispiele:
> Das Syntagma ÜBERMITTLUNG VON DATEN AN DRITTE bildet denselben / ähnlichen Informationsgehalt ab wie die verbhaltige Paraphrase DATEN AN DRITTE ÜBERMITTELN.
>
> Das Kompositum DATENSPEICHERUNG hat (im jetzigen Textbestand) Belegstellen für Strukturen der Art:
> "Daten werden gespeichert"; "gespeicherte Daten";
> "Daten speichern"; "Speicherung von Daten"; "Speicherung der Daten".

Im Rahmen des linguistischen F&E-Teils des Forschungsprojekts sollen für diese Probleme Lösungen (algorithmisch) entwickelt und so weit wie möglich erprobt werden.

3. Anwendungsmöglichkeiten

Das System ist für Pilotanwendungen für deutschsprachige Texte verfügbar. Gegenwärtig wird das Basissystem SATAN/ SUSY durch den SFB 100 auf eine Siemens-Anlage migriert (z.Zt. TR 440). Im Rahmen von JUDO sind eine Reihe von konkreten Anwendungen geplant, zu denen Pilotpartner gesucht werden:

- JUDO-DS-SERVICE

Aufbau eines Dokumentationssystems zum Bereich Datenschutz (Wunschpartner: JURIS (Datenbank) und Verlage (SDI-Dienste))

- JUDO-MULTILINGUAL

Mehrsprachiges Retrieval unter Integration fremdsprachiger Synonyme in den Thesaurus (Wunschpartner: EG bzw. ein FIZ)

- JUDO-IRS

Integration von JUDO in ein bestehendes IRS (GOLEM oder DIRS-GRIPS) unter Realisierung einer natürlichsprachigen "Anfrage"-Komponente - genauer: eines Bausteins zur Extraktion von systemrelevanten (semantisch vereindeutigten) Begriffen aus einer natürlichsprachigen Problembeschreibung, ggf. auch unter Aufbau eines entsprechend abgestimmten automatischen Retrievalverfahrens (Wunschpartner: DIMDI und/oder Siemens)

- JUDO-REGISTER

Anwendung von JUDO (insbesondere der Registerfunktionen) zum Aufbau von (lemmatisierten) Registern zu (deutschsprachigen) Texten (Wunschpartner: Verlage)

- JUDO-DB

Aufbau von fachgebietsunabhängigen Text-Datenbanken bzw. Informationssystemen ohne Verwendung (bzw. unter Einschränkung) des Thesaurus-Teils. Hierbei wird eine gewisse Ungenauigkeit der Ergebnisse (z.B. im Bereich der semantischen Vereindeutigung) in Kauf genommen, der Pflegeaufwand reduziert sich dementsprechend. Diese Variante - wenn man so will eine Variation von Verfahren wie PASSAT - eignet sich v.a. zur Verarbeitung grosser Textmengen im Medienbereich (Wunschpartner: ein PresseVerlag).

Technische Daten zum System:

Die "Saarbrücker Automatische Textanalyse" (SATAN) und das Anwendungssystem JUDO sind auf einer TelefunkenRechenanlage TR 440 implementiert. Eine Migration auf einen Rechner Siemens 7.760 (BS 2000) ist in Entwicklung.

Kernspeicherbedarf TR 440: 60 K Worte
Analysezeit je Satz eines Textes: ca. 10 CPU-Sekunden (TR 440) einschl. der JUDO-Aufbereitung, jedoch ohne Umsetzung in (GOLEM-)-IRS.

Grundwortschatz SATAN: z.Zt.: rd. 100.000 Wortstämme
Ergänzungswortschatz JUDO im Bereich Datenschutz (Wörter, die fachgebietsspezifisch aufbereitet bzw. ergänzt werden): ca. 5 - 10.000.

4. Literatur (Auszug)

Werner, H.: Die Anwendungsseite von JUDO (Vortrag, gehalten bei der 6. Sitzung des LDV-Beirats der GID am 22.06.1979 in Regensburg) Regensburg, Juni 1979

Zimmermann, H.: Automatische Textanalyse und Indexierung. In: Krallmann, D. (ed.): Kolloquium zur Lage der Linguistischen Datenverarbeitung, Essen, 1978, S. 20-33

Zimmermann, H.: Strategien zur Texterschließung und -retrieval in der maschinellen Dokumentation. Regensburg, Oktober 1979

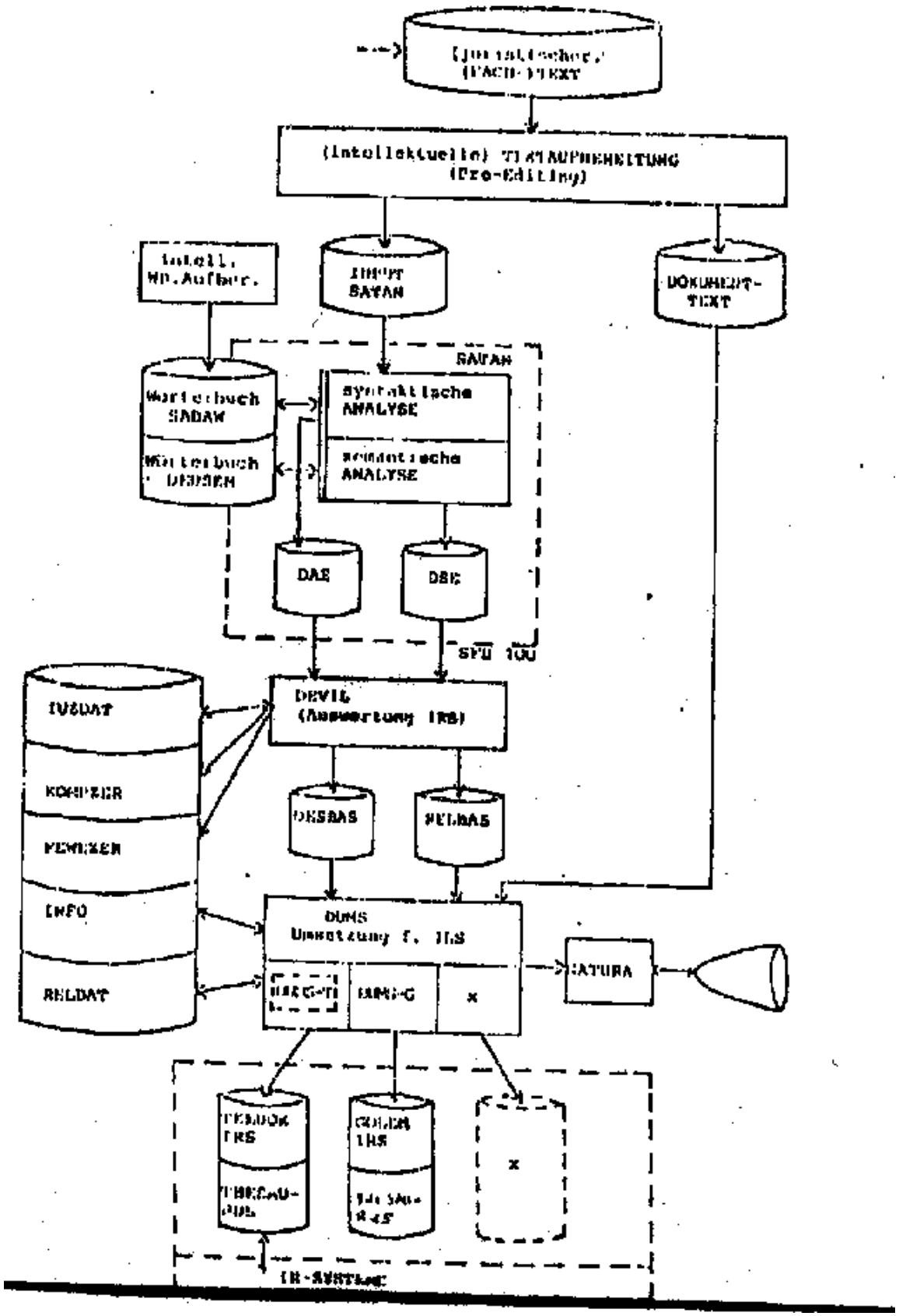
Zimmermann, H.: Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren. In: Kuhlen, R. (ed.): Datenbasen, Datenbanken, Netzwerke. München, 1979, S. 311-338

Zimmermann, H.: Introduction to the experimental System JUDO. Tokyo, Sept. 1980 (COLING 80)

Zimmermann, H.: Bürgernahe Informationsvermittlung am Beispiel des Modellsystems 'Juristische Dokumentanalyse im Bereich Datenschutz' (JUDO-DS). In: Informationssysteme für die 80er Jahre. Fachtagung 1980. Linz, 1980, Bd. 1, S. 143-170

5. Abbildungen

Abb. 1 Komponenten des JUDO-Systems:


(Intellektuelle) TEXTAUFBEREITUNG
(Pre-Editing)
SATAN
Syntaktische ANALYSE
DAE
DBE
DEVIL
(Auswertung IRS)
DUMS
Umsetzung f. IRS
x
x

Begriffserklärungen:

SATAN:	Saarbrücker Automatische Text-Analyse
SADAW:	Saarbrücker Deutsches Arbeits-Wörterbuch (Syntax-WB)
DEUSEM:	Deutsches Semantik-Wörterbuch (Semantische Disambiguierungsregeln)
IUZDAT:	Identifikations- und Zerlegungs-Daten (Types)
KOMPZER:	Komposita-Zerlegungen
FEWEZER:	Feste-Wendungen-Zerlegungen
INFO:	Informationsdaten zu Homonymen und Homographen
RELDAT:	Relationen-Daten(Thesaurus)
DESBAS:	Deskriptoren-Basis-Daten
RELBAS:	Relationen-Basis-Daten
DAE:	Deutsche (Syntaktische) Analyse-Ergebnisse
DSE:	Deutsche Semantische (Analyse-)Ergebnisse
DUMS:	Daten-Umsetzungs-System (für konkrete Retrieval-Systeme)
IRS:	Informations-Retrieval-System
NATURA:	Analyse Natürlichsprachiger Anfragen
DEVIL:	Deskriptor-Erstellungs-Verfahren mittels Identifikations- und Zerlegungs-Informationen und Linguistischer Verfahren

Abb. 2: JUDO-Deskriptoren:

"Die Vorschriften dieses Abschnittes gelten fuer natuerliche und juristische Personen, Gesellschaften und andere Personenvereinigungen des privaten Rechts, soweit sie geschuetzte personenbezogene Daten als Hilfsmittel fuer die Erfuellung ihrer Geschaeftszwecke oder Ziele verarbeiten." (§ 22 BDSG)

Folgende Deskriptoren werden für diesen Satz erstellt:

Einfache Deskriptoren

VORSCHRIFT	RECHT
ABSCHNITT	SCHUETZEN
GELTEN	PERSONENBEZOGEN
NATUERLICH	DATUM
JURISTISCH	HILFSMITTEL
PERSON	ERFUELLUNG
GESELLSCHAFT	GESCHAEFTSZWECK
ANDER	ZIEL
PERSONENVEREINIGUNG	VERARBEITEN
PRIVAT	

Komplexe Deskriptoren

VORSCHRIFT	G ABSCHNITT
PERSON	K GESELLSCHAFT
GESELLSCHAFT	K PERSON
PERSON	K PERSONENVEREINIGUNG
PERSONENVEREINIGUNG	K PERSON
GESELLSCHAFT	K PERSONENVEREINIGUNG
PERSONENVEREINIGUNG	K GESELLSCHAFT
GESELLSCHAFT	G RECHT
HILFSMITTEL	P ERFUELLUNG
ERFUELLUNG	G GESCHAEFTSZWECK
GESCHAEFTSZWECK	K ZIEL
ERFUELLUNG	G ZIEL

G = die Einzelwörter der Relation stehen in der Beziehung GENITIV
K = die Einzelwörter der Relation stehen in der Beziehung KONJUNKTION
P = die Einzelwörter der Relation stehen in der Beziehung PRÄPOSITION

Komposita und ihre Zerlegungen:

PERSONENVEREINIGUNG:	PERSON
	VEREINIGUNG
HILFSMITTEL:	HILFE
	MITTEL
GESCHAEFTSZWECK:	GESCHAEFT
	ZWECK

Feste Wendungen:

JURISTISCHE PERSON
NATUERLICHE PERSON
PERSONENVEREINIGUNG DES PRIVATEN RECHTS PERSONENBEZOGENE DATEN
GESCHUETZTE PERSONENBEZOGENE DATEN

Vereindeutigte Deskriptoren (vgl. Abb.3):

DATUM1	(im Sinne von 'Information')
GELTEN1	(im Sinne von 'gültig sein')
NATUERLICH2	(im Sinne von 'existierend')

Abb. 3: Ausschnitt aus der Datei 'INFO'

L	S	DATUM
01	9	Information: personenbezogene Daten
02	0	Zeitpunkt: das Datum des nächsten Tages
L	V	GELTEN
01	5	gültig sein: Diese Bestimmung gilt für alle.

02 2 bestimmt sein: Der Vorwurf gilt Max.
03 4 angesehen werden: Max gilt als der wichtigste Mitarbeiter.

L NATUERLICH
01 U 6 adv: selbstverständlich: Max steht natürlich zur Verfügung.
02 A 4 real existierend: natürliche Person (versus juristische Person)

(L= Lemmaname S= Substantiv V= Verb U= Adverb A= Adjektiv)

01, 02, ... = Zählung der Bedeutungsvarianten ("Bedeutungsnummern")

Zahlen 1 - 9 = 'Gewichtungsmarkierung', die die Vorkommenswahrscheinlichkeit der betreffenden Bedeutungsvariante im bearbeiteten Fachtext signalisiert).

Abb. 4 Übersicht über die Thesaurusrelationen:

Anzahl Paare(*) (9282=100%)	%	Kürzel d.Rel.	Bezeichung der Relation	Kurzcharakterisierung (A für Ausgangsbegriff) B für Relatum)	Beispiele
2646	28.5	ASS	Assoziation mit Gewichtung (10-90)	A ist frei assoziiert zu B	AUSKUNFT ASS 90 AUSKUNFTSERTEILUNG
83	0,9	IUS	Regelung-Regelungs-Gegenstand	A ist Regelung zu B	AUSKUNFTSPFLICHT IUS AUSKUNFT1
83	0,9	REG	Regelungsgegenstand Regelung	A ist Regelungsgegenstand zu B	AUSKUNFT1 REG AUSKUNFTSPFLICHT
244	2,6	GAN	Ganzes/Teil	A enthält B	DATENSICHERUNG GAN AUFTRAGSKONTROLLE
244	2,6	TEI	Teil/Ganzes	A ist Teil von B	AUFTRAGSKONTROLLE TEI DATENSICHERUNG
524	5,7	OBR	Oberbegriff	A ist Oberbegriff zu B	DATENVERARBEITUNG OBR AUTOMATISCHE DATENVERARBEITUNG
524	5,7	UNT	Unterbegriff	A ist Unterbegriff zu B	AUTOMATISCHE DATENVERARBEITUNG UNT DATENVERARBEITUNG
336	3.6	NEB	Nebenbegriff	A ist Nebenbegriff zu B	BENUTZERKONTROLLE NEB SPEICHERKONTROLLE
436	4,6	GEG	Gegenbegriff	A ist gegensätzlich oder komplementär zu B	AUSKUNFTSANSPRUCH GEG VERSCHWIEGENHEITSPFLICHT
810	8,7	QUA	Quasisynonymie mit Gewichtung(10-90)	A ist quasisynonym zu B	ANZEIGE1 QUA 20 NACHRICHT ANZEIGE1 QUA 50 MELDUNG BEHOERDE QUA 80 ÖFFENTLICHE STELLE
146	1,6	SYN	(strenge) Synonymie (aber keine der unten aufgeführten Synonymierelationen)	A ist synonym zu B	PARLAMENT SYN ABGEORDNETENHAUS
124	1,3	SYS	Synonymie durch abgewandelte Schreibweise	A ist synonym zu B, aber ähnlich geschrieben	BUNDESDATENSCHUTZGESETZ SYS BUNDES-DATENSCHUTZGESETZ
48	0,5	ABK	Abkürzung	A ist Abkürzung von B	BDSG ABK BUNDESDATENSCHUTZGESETZ
40	0,5	LNF	Langform	A ist Langform zu B	BUNDESDATENSCHUTZGESETZ LNF BDSG
868	9,3	Derivationsrelationen		A und B sind aus einem	STRUKTUR DSA STRUKTURELL

2074	22,3	Zerlegungssynonymie-relationen	Wort abgeleitet A und B sind verschiedene synonyme Darstellungen (als Kompositum, mehrwortiger Begriff oder syntaktische Relation) mit (mophologisch) gleichen Bestandteilen	DATENSCHUTZBEAUF-TRAGTER SKF BEAUFTRAGTER FÜR DEN DATENSCHUTZ

* Die Anzahl der Paare enthält die invertierten Begriffsbeziehungen, nicht jedoch aufgrund von Transitivitäts- und Synonymieeigenschaften generierte Begriffsbeziehungen.

Abb. 5 Deskriptorermittlung aus natürlichsprachiger Problembeschreibung:

SATZ 1
Gesetzliche Massnahmen, die dem Missbrauch der Daten durch unbefugten Zugriff natuerlicher oder juristischer Personen vorbeugen*

********** Deskriptoren zu Satz 0001

DATUM
DATUM1
GESETZLICH
GESETZLICHE MASSNAHME
JURISTISCH
JURISTISCHE PERSON
MASSNAHME
MISSBRAUCH
MISSBRAUCH DER DATEN
MISSBRAUCH G DATUM
MISSBRAUCH P ZUGRIFF
NATUERLICH
NATUERLICHE PERSON
PERSON
UNBEFUGT
UNBEFUGT A ZUGRIFF
VORBEUGEN
ZUGRIFF
ZUGRIFF G PERSON

Anschrift des Autors:

Prof. Dr. Harald Zimmermann
Universität des Saarlandes
5.5. Informationswissenschaft
6600 Saarbrücken 11

Als Vorlage für den Druck wurde das Manuskript des Autors verwendet.