Sprachtechnologie und Informationsextraktion

Dr. Günter Neumann
LT-lab, DFKI
neumann@dfki.de

Themen

- Motivation
  - Informationsextraktion (IE)
  - Textuelle Frage-/Anwortsysteme (QA)
- DFKI Sprachtechnologie (LT – Language Technology)
  - XML-Server
  - Robuste und flache Systeme
- Anwendungen
  - Machinelles Lernen von Template Elements
  - IE-basierte Suchmaschine (Google-Aufsatz)

LT Anwendungsfelder

(1) Information Retrieval (IR)

(2) Passage Retrieval

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

LT Anwendungsfelder

(1) Information Retrieval (IR)

Identifiziere und extrahiere Dokumente als Antwort einer Informationsanfrage.

(2) Passage Retrieval

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

LT Anwendungsfelder

(1) Information Retrieval (IR)

Identifiziere und extrahiere Dokumente als Antwort einer Informationsanfrage.

(2) Passage Retrieval

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Texte so wie der Mensch verstehen: Artificial Intelligence

LT Anwendungsfelder

(1) Information Retrieval (IR)

Identifiziere und extrahiere Dokumente als Antwort einer Informationsanfrage.

(2) Passage Retrieval

Identifiziere und extrahiere Dokumentenausschnitte als Antwort einer Informationsanfrage.

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Texte so wie der Mensch verstehen: Artificial Intelligence

LT Anwendungsfelder

(1) Information Retrieval (IR)

Identifiziere und extrahiere Dokumente als Antwort einer Informationsanfrage.

(2) Passage Retrieval

Identifiziere und extrahiere Dokumentenausschnitte als Antwort einer Informationsanfrage.

(3) Information Extraction (IE)

Identifiziere und extrahiere relevante textuelle Ausschnitte zum Füllen von vordefinierten Datenschemata/Templates.

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Texte so wie der Mensch verstehen: Artificial Intelligence

LT Anwendungsfelder

(1) Information Retrieval (IR)

Identifiziere und extrahiere Dokumente als Antwort einer Informationsanfrage.

(2) Passage Retrieval

Identifiziere und extrahiere Dokumentenausschnitte als Antwort einer Informationsanfrage.

(3) Information Extraction (IE)

Identifiziere und extrahiere relevante textuelle Ausschnitte zum Füllen von vordefinierten Datenschemata/Templates.

(4) textuelle Frage-/Anwortsysteme

Beantworte beliebige Fragen durch Verwendung von Texten als Wissensbasis: Fact retrieval, Kombination von IR und IE.

(5) Textverstehen

Texte so wie der Mensch verstehen: Artificial Intelligence

Interpretation von Texten

(1) Information Retrieval (IR)

(2) Passage Retrieval

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Interpretation von Texten

(1) Information Retrieval (IR)

Benutzer

(2) Passage Retrieval

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Interpretation von Texten

(1) Information Retrieval (IR)

Benutzer

(2) Passage Retrieval

Benutzer

(3) Information Extraction (IE)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Interpretation von Texten

(1) Information Retrieval (IR)

Benutzer

(2) Passage Retrieval

Benutzer

(3) Information Extraction (IE)

System (statisch, vordefiniert)

(4) textuelle Frage-/Anwortsysteme

(5) Textverstehen

Interpretation von Texten

(1) Information Retrieval (IR)

Benutzer

(2) Passage Retrieval

Benutzer

(3) Information Extraction (IE)

System (statisch, vordefiniert)

(4) textuelle Frage-/Anwortsysteme

System (dynamisch, einfache Fakten/Relationen)

(5) Textverstehen

Interpretation von Texten

(1) Information Retrieval (IR)

Benutzer

(2) Passage Retrieval

Benutzer

(3) Information Extraction (IE)

System (statisch, vordefiniert)

(4) textuelle Frage-/Anwortsysteme

System (dynamisch, einfache Fakten/Relationen)

(5) Textverstehen

System (vollständig)

Informationsextraktion

- Robuste Extraktion von relevanten Begriffen, Phrasen, Aussagen aus Texten.
- Erfolgsraten (Vollständigkeit und Präzision) hängen von der Aufgabe und vom Gegenstandsbereich ab.
- Bereits eingesetzt in verschiedenen Anwendungen, z.B.
  - Firmennamenerkennung,
  - Übersichten zu Firmenindikatoren (Umsatz, Gewinn, Kurse)
  - Nachrichtenübersichten zu speziellen Themen

Beispiel: Informationsextraktion (1)

In der IE werden gezielt relevante Informationen aus Texten herausgesucht und strukturiert.

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwal-tungssoftware Store Age vorgestellt..
Die neue Version ermöglicht jetzt auch ...

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

...

Beispiel: Informationsextraktion (1)

In der IE werden gezielt relevante Informationen aus Texten herausgesucht und strukturiert.

Bremen, 14. 10. 1997, wiwo: Lagersoftware weiter im Aufwind

Die Bremer Firma Trade Consult hat auf einer Pressekonferenz in Hannover die Version 2.0 ihrer erfolgreichen Lagerverwal-tungssoftware Store Age vorgestellt..
Die neue Version ermöglicht jetzt auch ...

Auf der Pressekonferenz gab Geschäftsführer Franz Merleback auch die Umsatzzahlen der Softwareschmiede für das 3. Quartal bekannt. Wurden im zweiten Quartal bereits über 30 Millionen Mark umgesetzt, so konnte Merleback jetzt das stolze Ergebnis von 42,5 Millionen verkünden.

...

Beispiel: Informationsextraktion (2)

Firma	96Q4	1996	97Q1	97Q2	97Q3	97Q4	1997	Diff.
ComSoft		120Mio					110Mio	
Trade Consult				30 Mio	42,5Mio			
Z&M					71,0Mio			

Textuelle Frage/Antwortsysteme (QA)

Beantworte eine beliebige Frage durch Bestimmung eines kleinen Textausschnittes, in dem die Antwort tatsächlich vorkommt.

- Betrachte sehr grosse Mengen von on-line Dokumenten
- Betrachte nur kleinste relevante Textausschnitte
- Die in natürlicher sprache (NL) ausgedrückten Fragen sind nicht beschränkt bezüglich
  - Gegenstandsbereich oder
  - Typ der Frage (d.h., nicht nur W-Fragen, wie wer, was, wem, wo, wie, warum)

QA Beispiel

What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

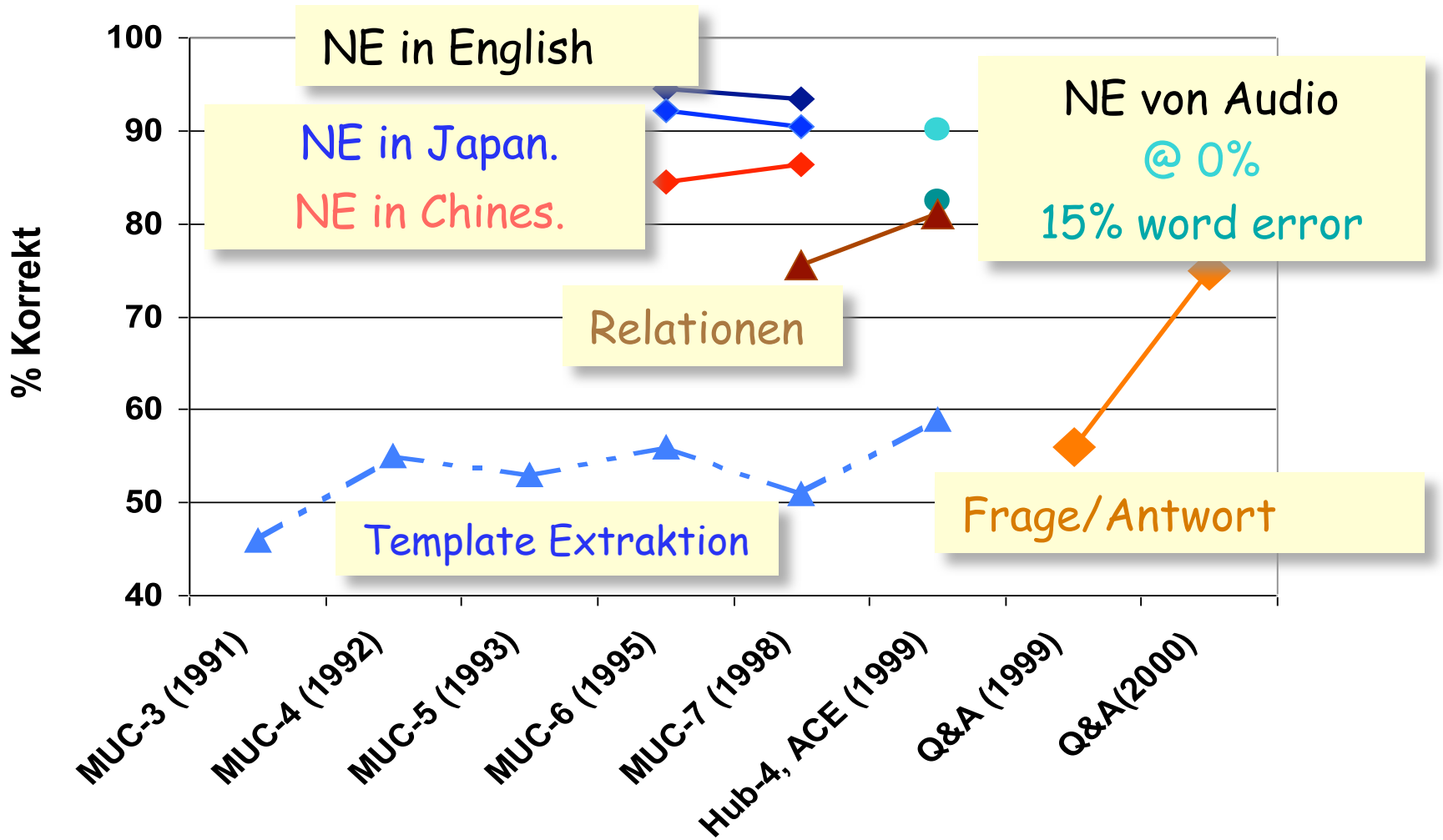
Antwort (50 bytes Textausschnitt aus einem Webdokument):

...who said she has both <Answer> Tourette‘s Syndrome </Answer> and...

Teilaufgaben für IE & QA

- Named Entities NE (Namen, Zeit/Datum, Maße, Fachausdrücke etc.)
  - Martin Marietta Corp, von 12.3. bis 15.3.02
- Relationen (unär, binär)
  - Nachfolger(person), joint-venture(firma1,firma2)
- Ereignisse
  - Umsatz(Firma,Ort,Betrag,Tendenz)
- Aktuelle Forschungsschwerpunkte in QA:
  - Fragetypen: Fakten, Aufzählungen, Kontextuelle, etc...
  - auf viele (multilinguale) Dokumente verteilte Antworten
  - Paraphrasierung von Fragen
  - Integration von Ontologien
  - Klärungsdialoge, Validierung von Antworten

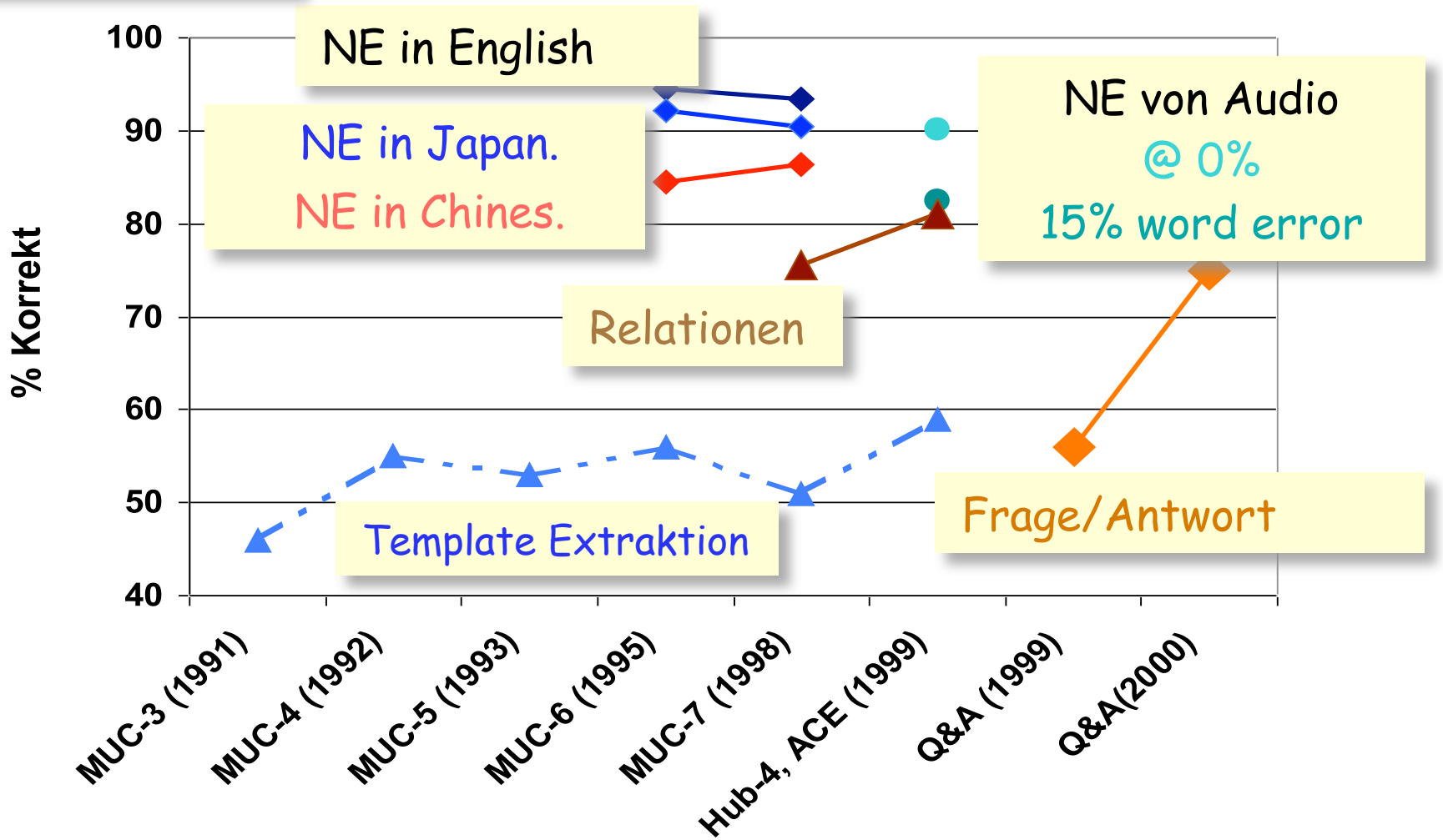
Fortschritte in IE



IE und QA werden systematisch evaluiert …

Fortschritte in IE

NE in Deutsch (DFKI)

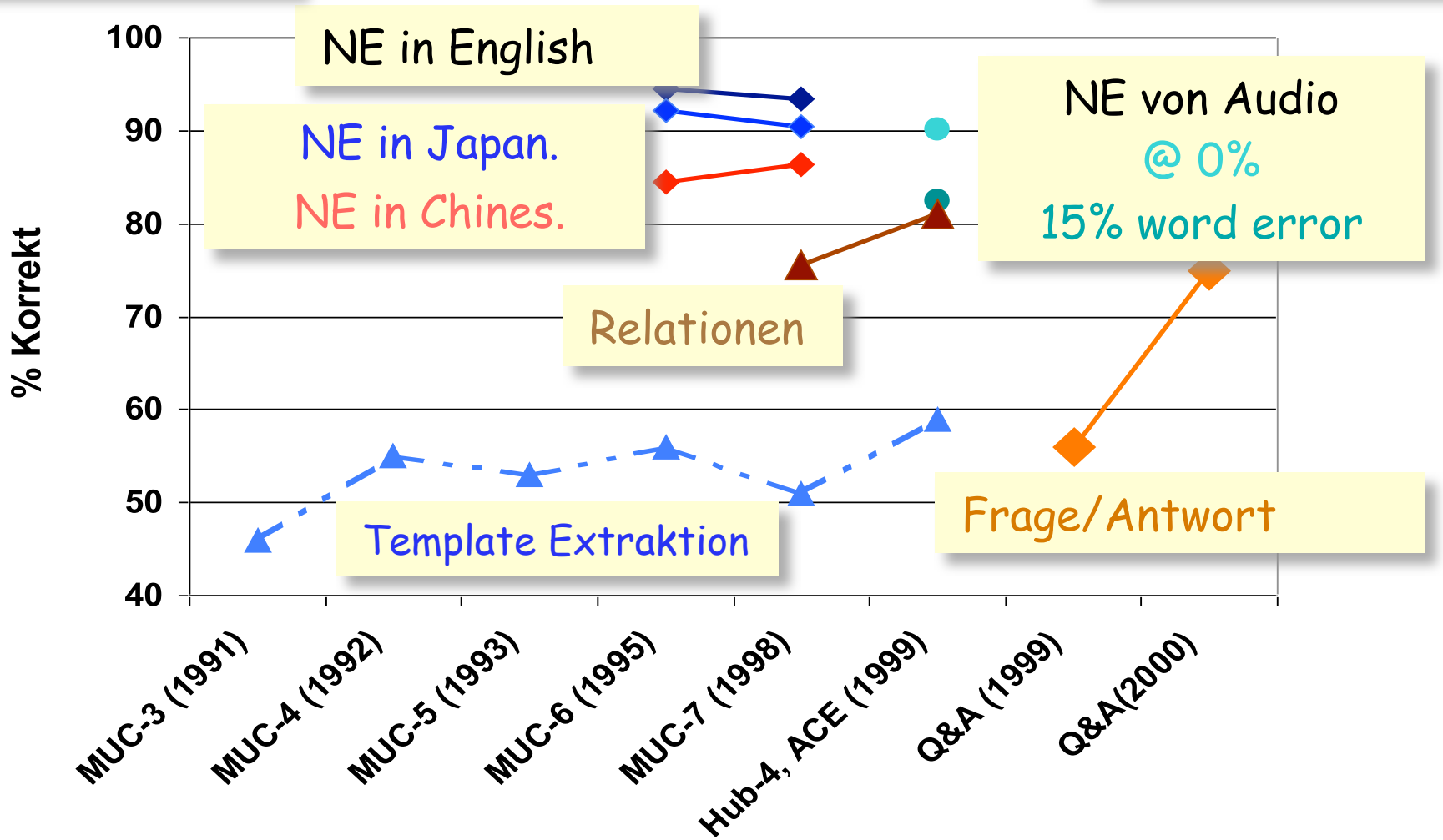


IE und QA werden systematisch evaluiert …

Fortschritte in IE

NE in Deutsch (DFKI)

Unäre Relationen (DFKI)

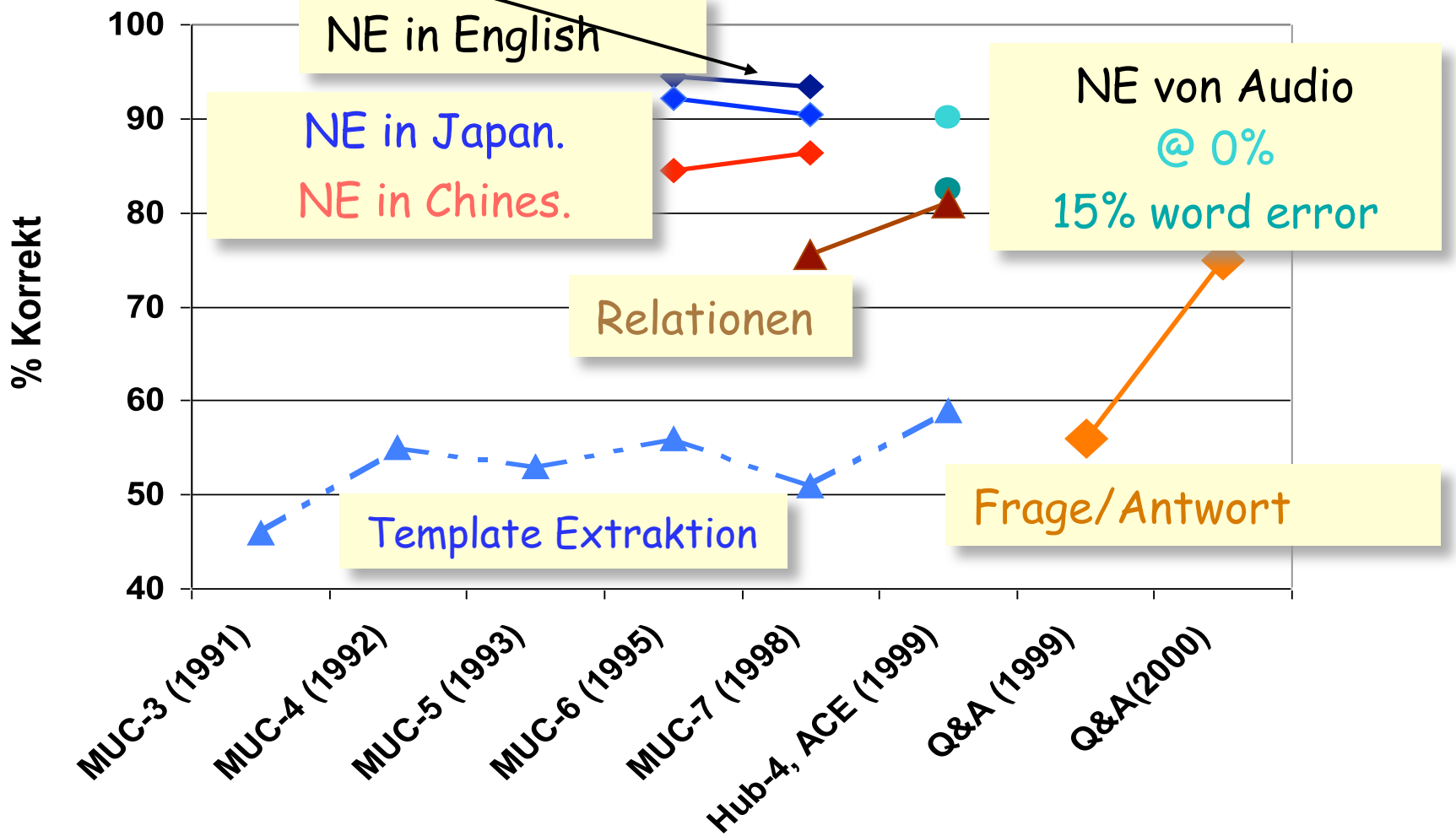


IE und QA werden systematisch evaluiert …

Fortschritte in IE

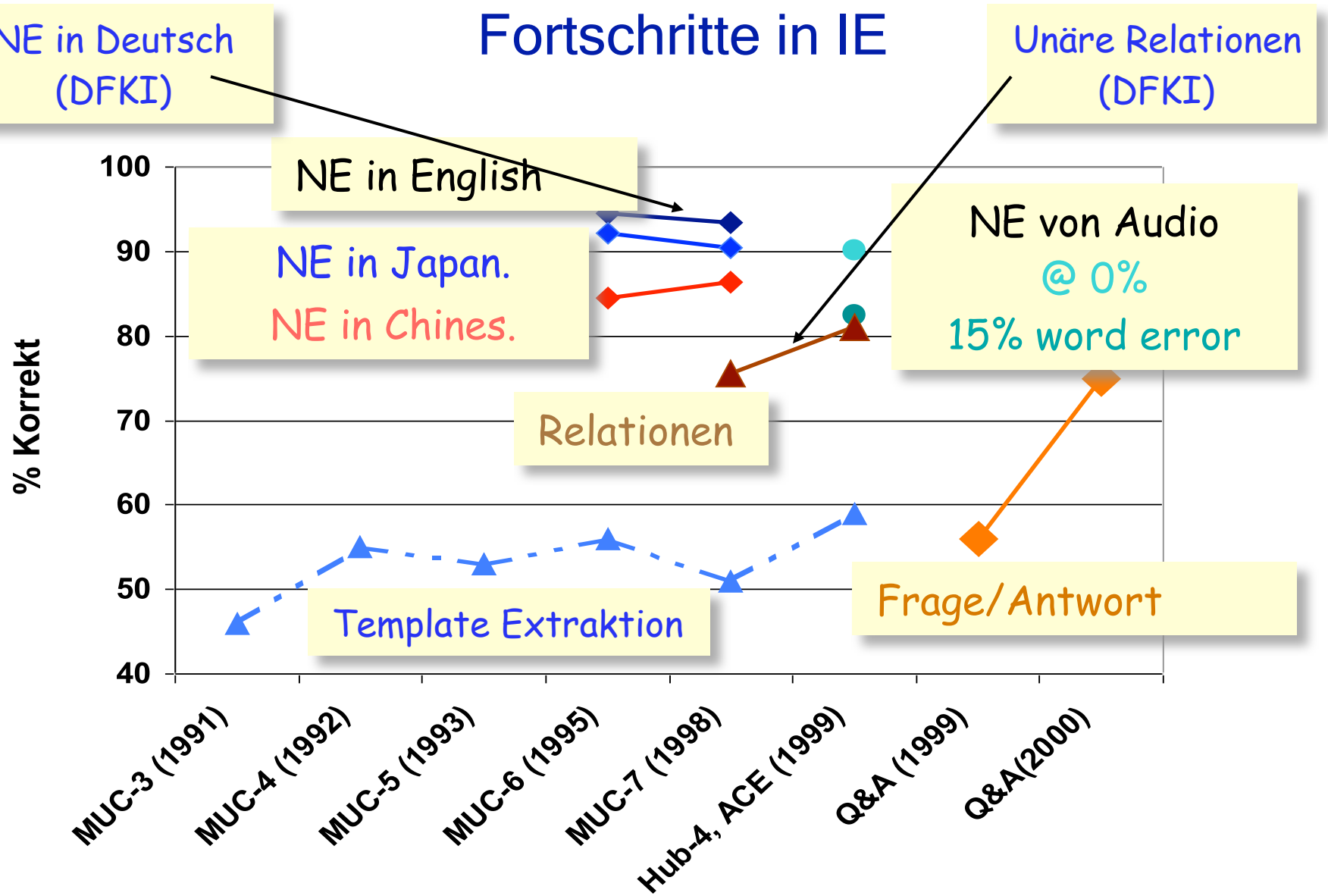
NE in Deutsch (DFKI)

Unäre Relationen (DFKI)



IE und QA werden systematisch evaluiert …

Fortschritte in IE



IE und QA werden systematisch evaluiert …

Inhaltliche Analyse von Texten ist komplex

- Schier unendliche Formulierungsmöglichkeiten
- Mehrdeutigkeiten, referentielle Ausdrücke
- Organismus Sprache (Kreativität/Produktivität)
- In den letzen Jahren:
  ingenieursmässige, problem- und datenorientierte (bottom-up) Herangehensweise vielversprechender als eine stark top-down orientierte Computerlinguistische

Linguistische Analyse als schrittweise Normalisierung

- morphologische Analyse:
  - Bestimmung von lexikalischen Stämmen:
    Flexionsanalyse, Komposita (Häusern - haus; Schiffskoch - Schiff koch)

- spezielle Phrasen:
  - Datums- und Zeitausdrücke: kanonische Darstellung, z.B.
    18.12.98 und Freitag, der achtzehnte Dezember 1998
    ⇨ <type=date, year=1998, month=12, day=18, weekday=5>
  - Eigennamen: Personen, Institute, Firmen, Orte
  - Zahlausdrücke, Adressen, etc.

Linguistische Analyse als schrittweise Normalisierung

- allgemeine Phrasen:
  - Nominalphrasen, Präpositonalphrasen, Verbgruppen

    Für die deutsche Wirtschaft

    ⇨ <head=für, comp=<head=wirtschaft,
    quant=def, mod=deutsch>>

- komplexe flache Satzstrukturen

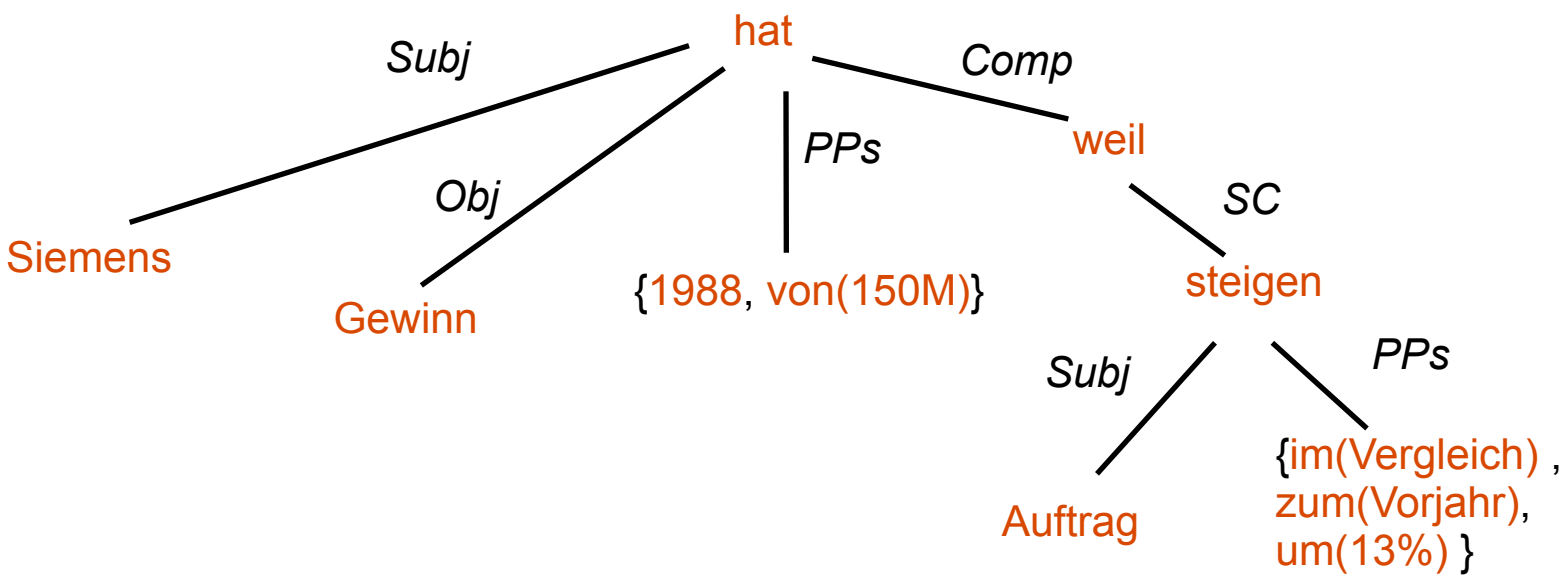
- domänenspezifische Templates (Integration von Ontologie)

Ein linguistisch analysierter Text wird als Sequenz von unterspezifizierten (partiellen) funktionalen Strukturen UFDs repräsentiert

UFD: flache dependenz-basierte Struktur, nur obere Schranken für Attachment und Skopus

[$_{PN}$Die Siemens GmbH] [$_{V}$hat] [$_{year}$1988][$_{NP}$einen Gewinn] [$_{PP}$von 150 Millionen DM],

[$_{Comp}$weil] [$_{NP}$die Auftraege] [$_{PP}$im Vergleich] [$_{PP}$zum Vorjahr] [$_{Card}$um 13%] [$_{V}$gestiegen sind].



DFKI Language Technology

- XML Text-Annotation-Server
- Robuste und effiziente NL Komponenten
- Maschinelle Lernverfahren
  - Automatische Adaptation von Grammatik an Domänen (zB. EBL, Data-oriented Parsing/ Generierung)
  - Statistisch-basierte Verfahren (Tagging, PCFG)
  - IE: NE/Slot-recognition, n-äre Relationen
- Domänenoffene IE-basierte Web-Suchstrategien

Multilevel Annotation for Dynamic Free Text Processing

BMBF-Förderung (DFKI-Rahmenvertrag)

Laufzeit März 2000 bis März 2003

~ 5 Informatiker und Computerlinguisten

Flache und tiefe Analysesysteme

Flache Sprachanalyse (SNLP)

- J Große Eingabe-Textmengen
- J Robust bezüglich Abdeckung, Fehlern im Text
- K Ausgabe präferierter Lesarten
- L Grobstrukturierte linguistische Analyse
- L Ungenügende Präzision

Tiefe Sprachanalyse (DNLP)

- L Satzorientiert
- L Nicht robust bezüglich Abdeckung, Fehlern im Text
- K Potentiell hochambige Analyseergebnisse
- K Feinstrukturierte linguistische Analyse
- J Präzisionsorientiert

Ziel: **Verbinden beider Paradigmen**

Ist SNLP nicht ausreichend?

- Einfache NL Aufgaben benötigen wahrscheinlich nur SNLP
  - E.g., text clustering, Named Entity recognition, term extraction
- Information extraction: 60% f-measure Barriere im Falle von scenario template extraction (zB. Firmenumsatzmeldungen). Schwierigkeiten:
  - Template-Elemente (slots) sind über den Artikel verteilt (Analyse von mehreren, zusammenhängenden Sätzen nötig)
  - Notwendig, grammatikalische Funktionen zu identifizieren (domänenspezifische Kasusrahmen)
  - Referenzresolution im Falle von Templatemerging (Objekt-Identität)
- Inhaltsorientierte Web-Services erfordern immer mehr die Extraktion von komplexen Strukturen („semantic web“)
  - Ontology extraction, Web-oriented question/answering systems

Zentrale Projektziele

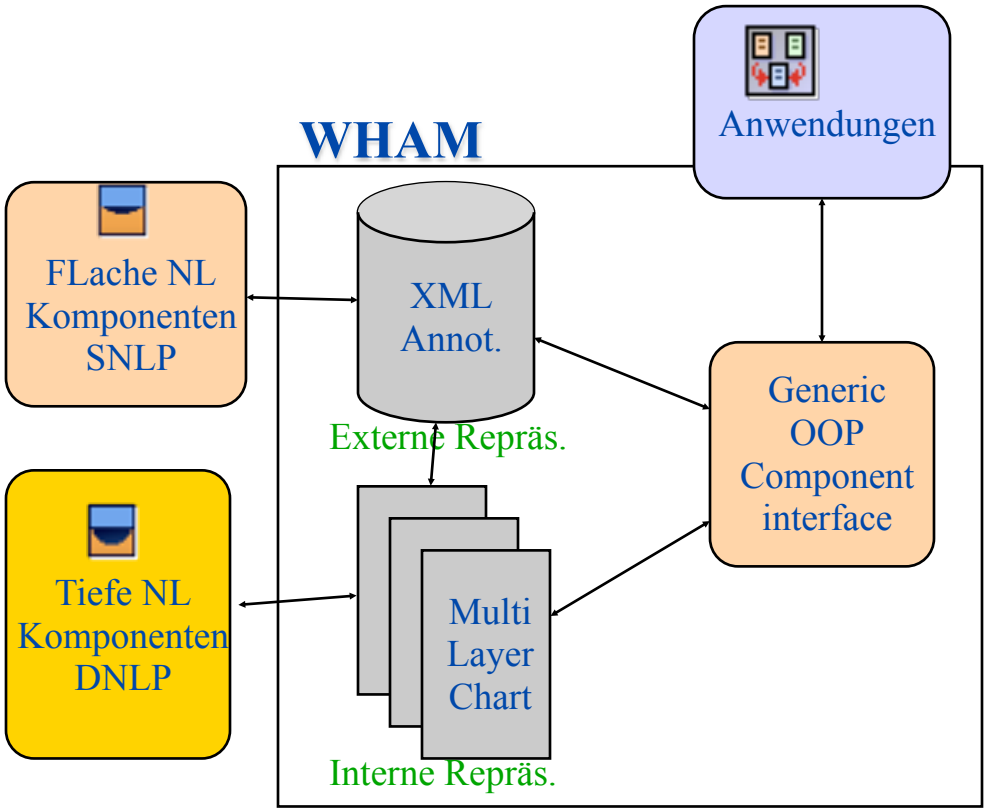
- Integration von flacher & tiefer NLP
  - Verarbeitung von freien Texten „beliebiger“ Domäne
  - Parametrisierbare Textanalysetiefe
- XML-basierte Systemplattform
  - Uniforme Art der Repräsentation und Verwaltung aller Ergebnisse der verschiedenen Verarbeitungskomponenten
  - Transparente Software-Infrastruktur für LT-basierte Anwendungen
- Intelligente Informationsextraktion
- Web-basierte Frage/Antwortsysteme

XML Text-Annotation

- XML als lingua franca, Business-Standard
  - Natürliches Konzept der Dokument-Annotation (SGML)
  - Modular, flexibel, erweiterbar
  - Standardisierte Verarbeitungswerkzeuge

- Herausforderung: Einige interessante sprachliche Phänomene können nicht direkt in XML-Klammerstruktur ausgedrückt werden:
  - Überlappende Analysen / Lesarten
  - Diskontinuierliche Konstituenten (Nehme am Termin teil.)
  - Koreferenzen (Microsoft .... Der grösste Softwarehersteller ...)

Annotationsbasierte Architektur
(WHAM, Whiteboard Annotation Machine)

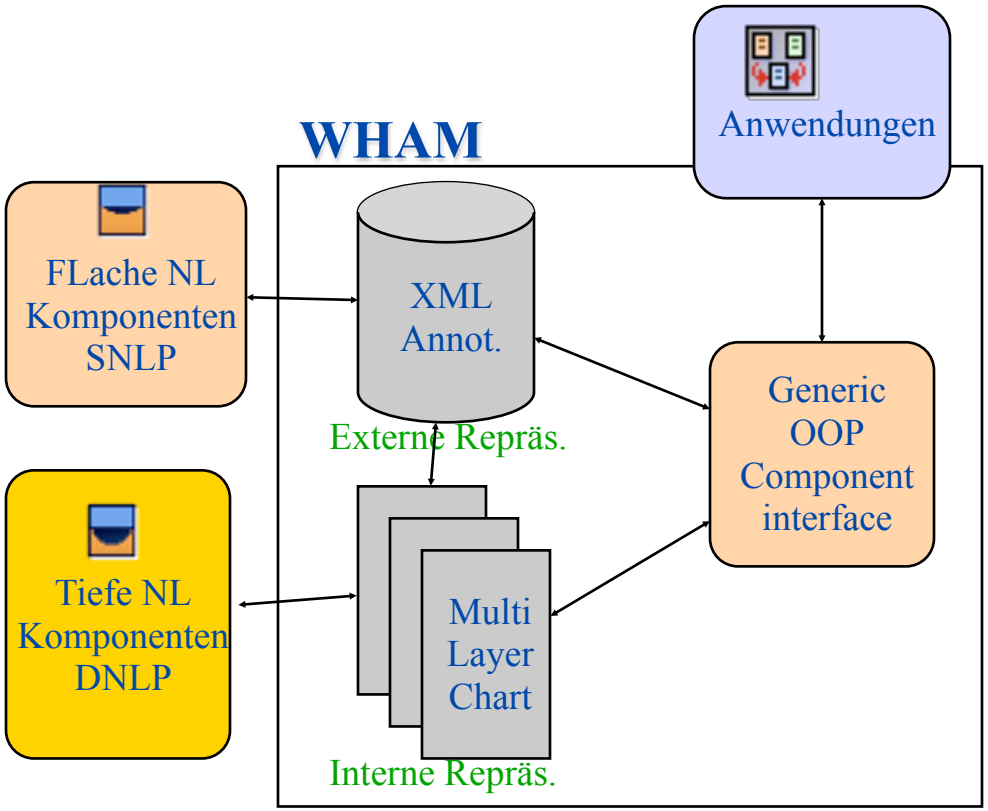


- **XML-Speicherung**
  - **Externe offline info**
  - **IO level für SNLP**
  - **Grosse Textkorpora**

  Mehrebenen Chart
  - **Interne online info**
  - **IO level für DNLP**
  - **Nur relevante XML-Textausschnitte**

  Generische Interfacedesign
  - **NL Komponenten als Unterklassen, definiert via DTD und /oder Chart-Methoden**
  - **Iterator-Objecte zur Travserierung der XML-Layer**
- **WHAM-Kern implementiert in Java**

Annotationsbasierte Architektur
(WHAM, Whiteboard Annotation Machine)



- **XML-Speicherung**
  - **Externe offline info**
  - **IO level für SNLP**
  - **Grosse Textkorpora**

  Mehrebenen Chart
  - **Interne online info**
  - **IO level für DNLP**
  - **Nur relevante XML-Textausschnitte**

  Generische Interfacedesign
  - **NL Komponenten als Unterklassen, definiert via DTD und /oder Chart-Methoden**
  - **Iterator-Objecte zur Travserierung der XML-Layer**
- **WHAM-Kern implementiert in Java**

Aktuell entwickeln wir ein XSLT-basiertes Komponenten-Interface

Flache Analysekomponenten für große Textmengen

- DFKI-Kern-Technologien für Deutsch (System SMES-SPPC, cf. Neumann&Piskorski, Journal of Computational Intelligence, 2002):
  - Tokenisierung, Morphologie (inkl. online Kompositaanalyse)
  - Erkennung von Eigennamen und Phrasen (NP, PP, VG) DEMO
  - Robustes Satzparsing (Topologische Struktur, grammatikalische Funktion) Beispiel

- Zentrale Technologien
  - Dynamische Tries für lexikalische Analyse
  - Gewichtete Automaten für Eigennamen und Phrasen
  - Einfache, effiziente Unifikation für Agreement- und Subkategorisierungschecks (Satzparsing)

Hohe Effizienz und Abdeckung

- Basis

  Korpus „Wirtschaftswoche“ (1,2 MB, 197.118 tokens)

- Performanz (ohne clause parsing – mit ca. 4mal langsamer)

  ~10 Sec. (~12.000 Wörter/Sek.; Pentium III, 700 MHz, 256 MB RAM)

- Evaluation (20.000 tokens)

	Erkennung	Korrektheit
- Kompositaanalyse:	98.53 %	99.29 %
- Wortklassen-Filterung:	74.50 %	96.36 %
- Namenerkennung	85 %	95.77 %
- Phrasen (NPs, PPs):	76.11 %	91.94 %
- Topo. Parser:	87.14 % f-measure	

➔ **weitere Verbesserung durch Integration mit tiefer Analyse**

Aktuelle Anwendungsfelder im Whiteboard project)

- Maschinelles Lernen von Eigennamen und unären Relationen
- Web-basiertes Frage-/Antwort-System

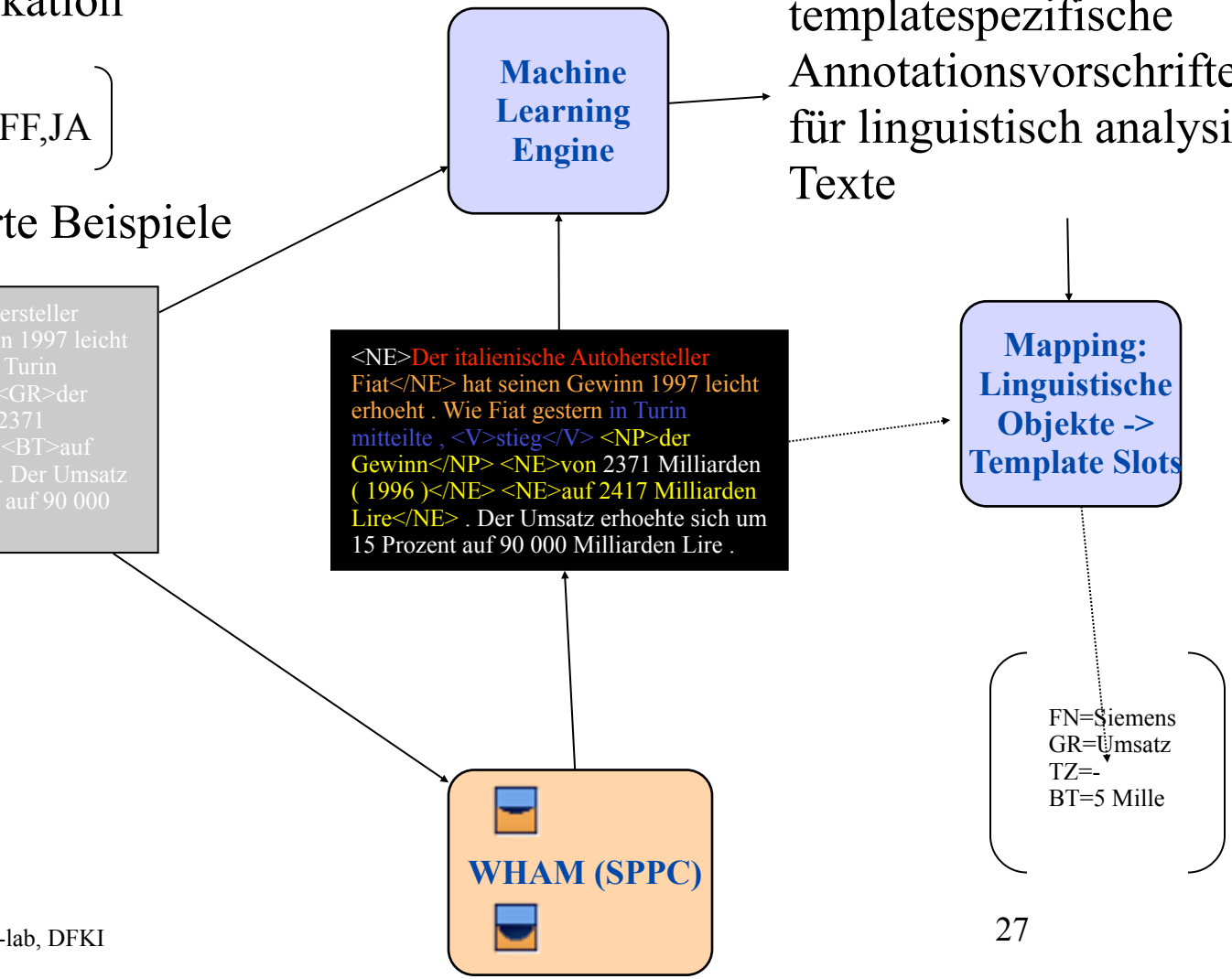
WHAM-Machinelles Lernen für IE

Input:
Templatespezifikation

FN,GR,TZ,BT,DIFF,JA

XML-Annotierte Beispiele

<FN>Der italienische Autohersteller Fiat</FN> hat seinen Gewinn 1997 leicht erhoeht . Wie Fiat gestern in Turin mitteilte , <TZ>stieg</TZ> <GR>der Gewinn</GR> <DIFF>von 2371 Milliarden (1996)</DIFF> <BT>auf 2417 Milliarden Lire</BT> . Der Umsatz erhoehte sich um 15 Prozent auf 90 000 Milliarden Lire .



Output:
templatespezifische Annotationsvorschriften für linguistisch analysierte Texte

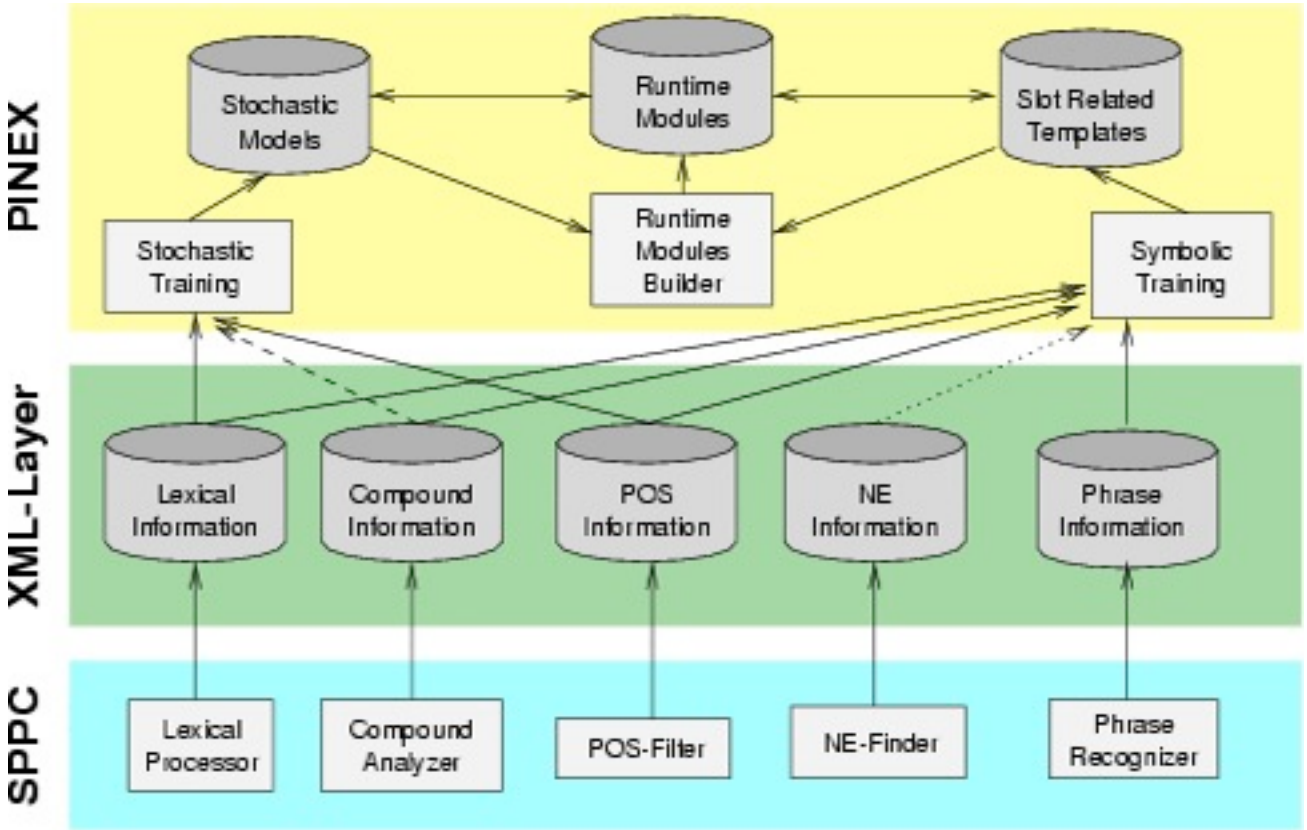
Mapping: Linguistische Objekte -> Template Slots

FN=Siemens
GR=Umsatz
TZ=-
BT=5 Mille

Unser Ansatz: hybrides Lernverfahren

- Gemischtes Bottom-up/top-down Verfahren
  - Bottom-up: lexikalisch
    - Statistisches Model zur Bestimmung relevanter lexikalischer Constraints (Maximum Entropy Modelling)
  - Top-down: syntaktisch (Phrasen)
    - Lokale, syntaktische Patterns
    - Komprimierung relevanter Syntaxbäume

Architektur



Experimente – verwendeter Korpus

- Domain: Umsatzmeldungen
- 6 semantische Typen (Slots)
- 60 annotierte Pressemeldungen für die Lernphase (4850 Tokens)
- 15 (neue) Pressemeldungen für die Testphase (1000 Tokens)

Nur symbolische Komponente

	PREC	REC	FME
BT	040.7407 %	025.5814 %	031.4286 %
DIFF	031.0345 %	036.0000 %	033.3333 %
FN	070.0000 %	025.0000 %	036.8421 %
GR	055.1020 %	071.0526 %	062.0690 %
JA	050.0000 %	045.8333 %	047.8261 %
TZ	092.8571 %	092.8571 %	092.8571 %
Task-TE	055.1515 %	048.9247 %	051.8519 %

Nur stochastische Komponente

	PREC	REC	FME
BT	072.0000 %	083.7209 %	077.4194 %
DIFF	060.6061 %	080.0000 %	068.9655 %
FN	081.2500 %	046.4286 %	059.0909 %
GR	092.6829 %	100.0000 %	096.2025 %
JA	066.6667 %	075.0000 %	070.5882 %
TZ	092.8571 %	092.8571 %	092.8571 %
Task-TE	077.4359 %	081.1828 %	079.2651 %

Symbolisch + Stochastisch

	PREC	REC	FME
BT	084.7826 %	090.6977 %	087.6404 %
DIFF	088.8889 %	096.0000 %	092.3077 %
FN	081.2500 %	046.4286 %	059.0909 %
GR	092.6829 %	100.0000 %	096.2025 %
JA	066.6667 %	075.0000 %	070.5882 %
TZ	092.8571 %	092.8571 %	092.8571 %
Task-TE	085.4054 %	084.9462 %	085.1752 %

WHAM- Machinelles Lernen für IE

- Diplomarbeit bei Volker Morbach (~Juni, 02)
- Implementation in Java
- XML-Layer für LT-tools
- Hoch modular
- MUC-kompatible Evaluationssoftware

WHAM-SearchEngine

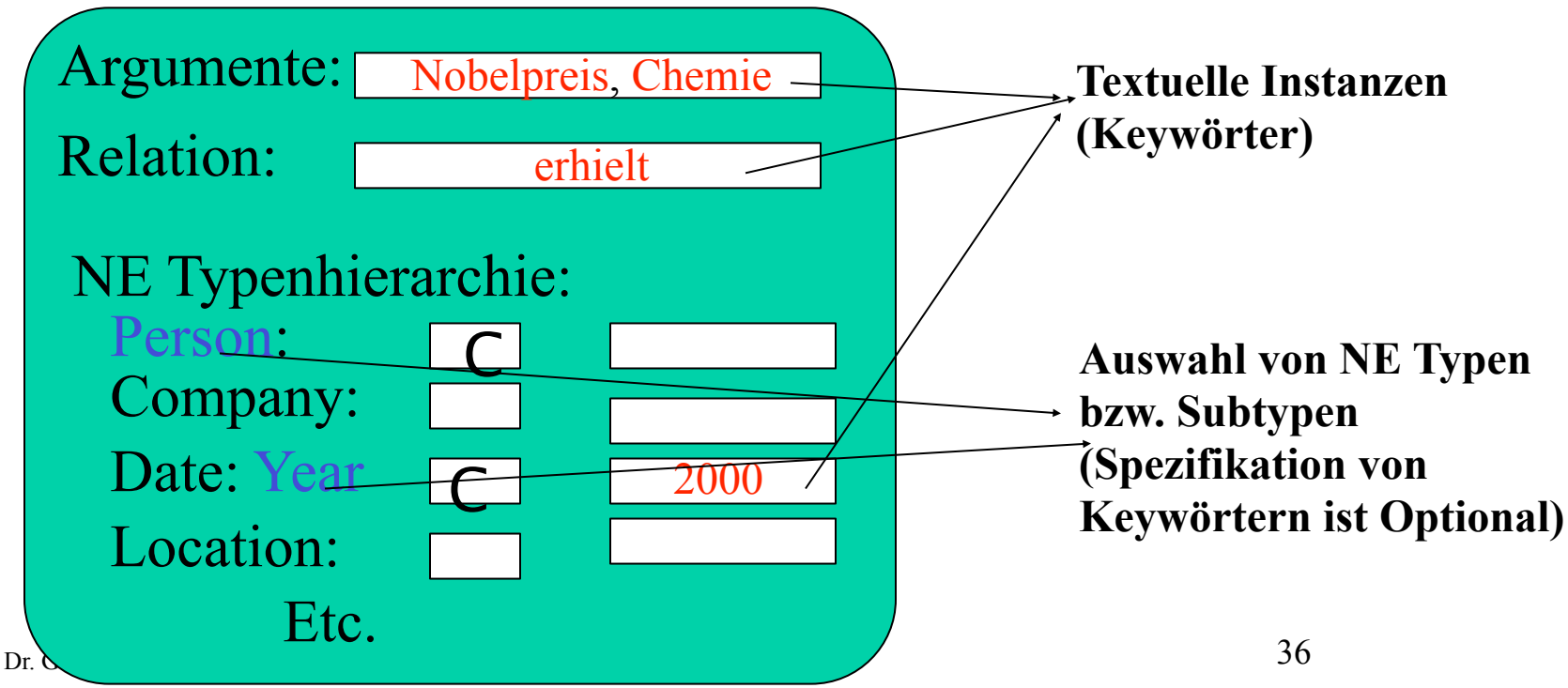
- Kombination von Standard-Suchmaschinen mit DFKI IE Technologie
  - Neumann & Xu, „Mining Answers from the Web using Information extraction & fusion“, forthcoming
- Praktische Erfahrung in der Enwicklung von intelligenten Web-basierten Suchmaschine erhalten
- Als Grundlage für zukünftige Forschungs-/ Entwicklungsprojekte:
  - Multi-linguales Frage-/Antwortsysteme (DFKI Projekt in Planung)

Unser Ansatz (1): Semi-strukturierte Anfrage

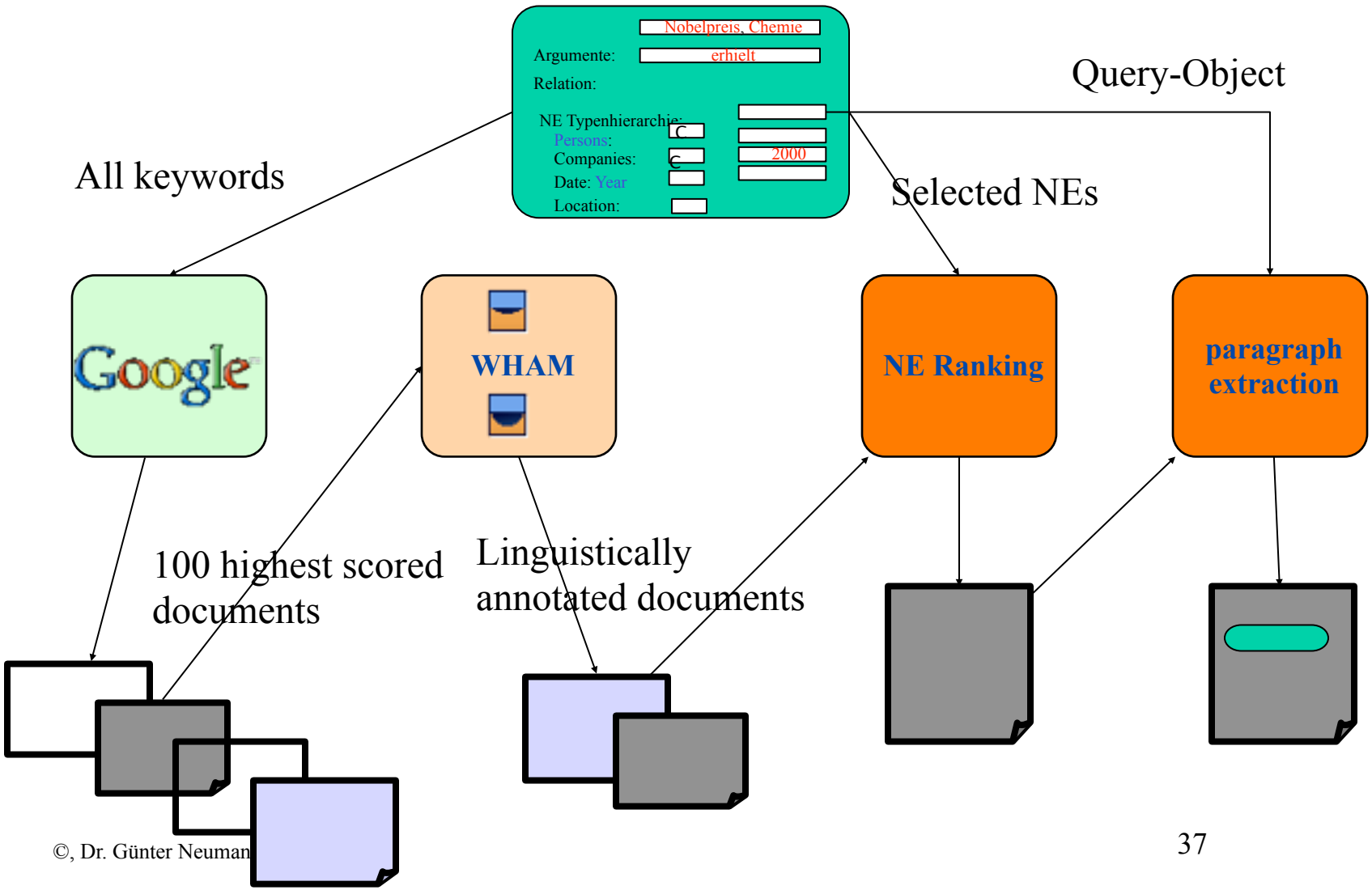
- Benutzer formuliert Fragen in Form von partiell instantiierten Relationen

QueryInterface

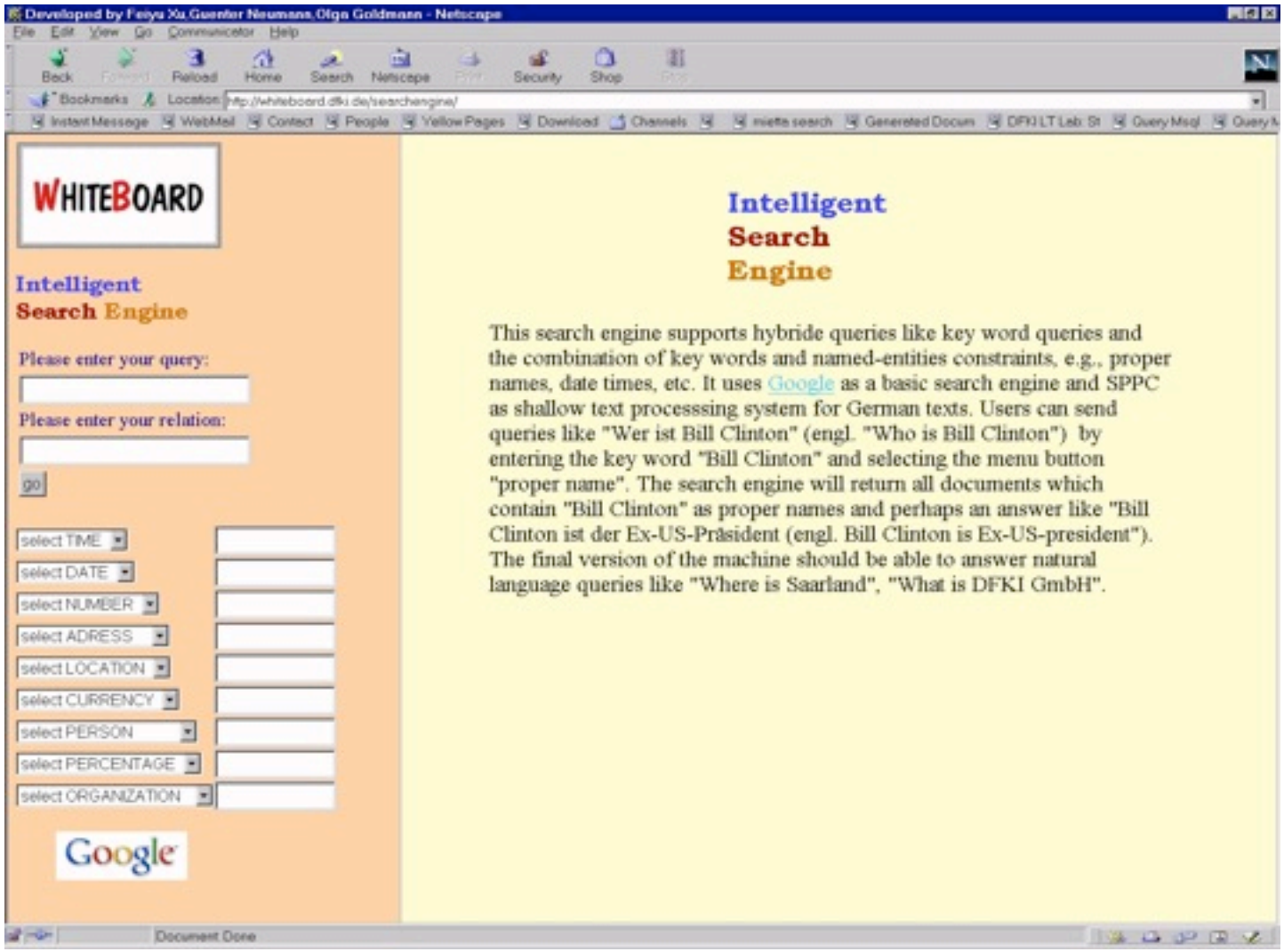
Welche Person erhielt im Jahr 2000 den Nobelpreis in Chemie?



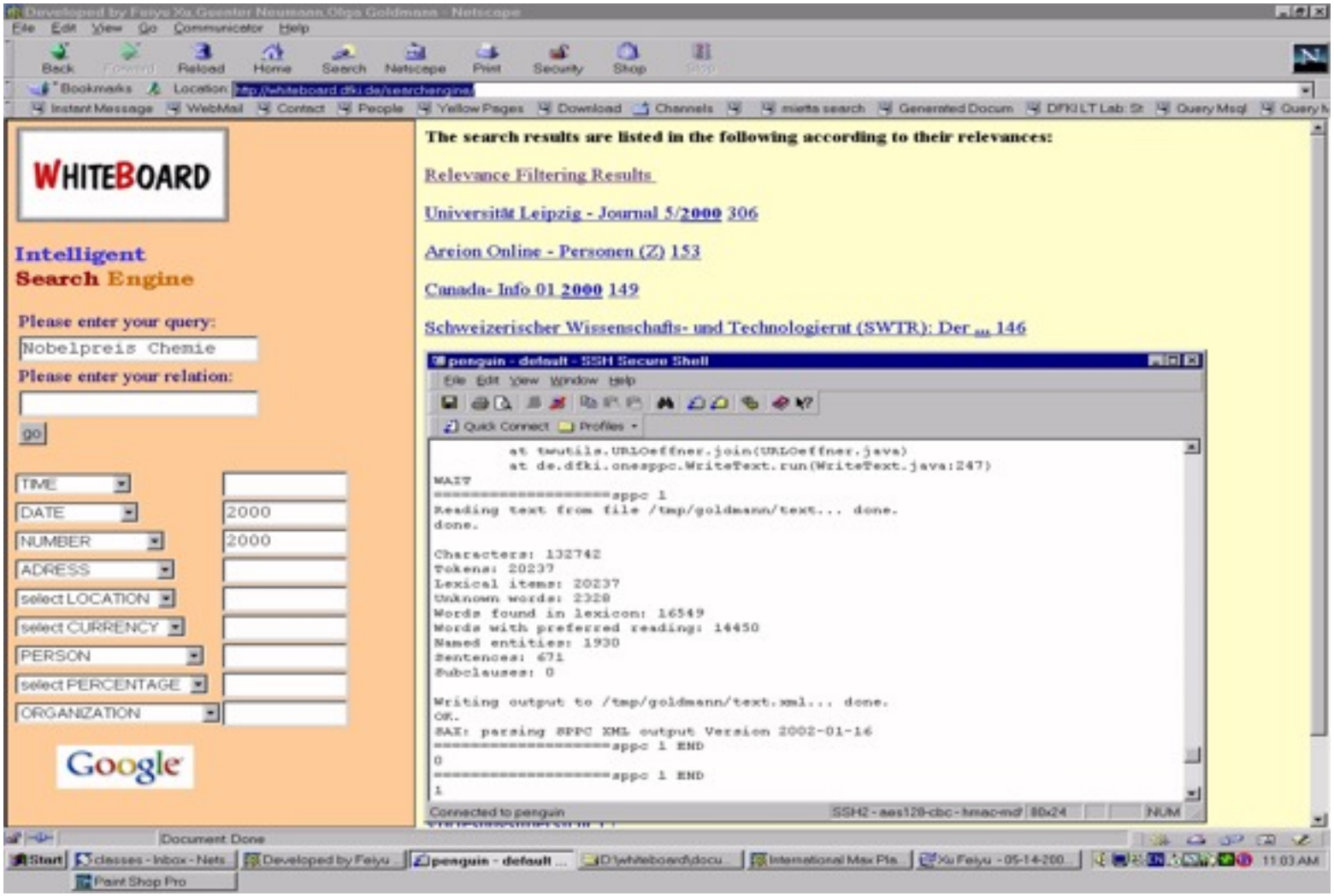
Unser Ansatz (2)
Querysuche als Textzooming



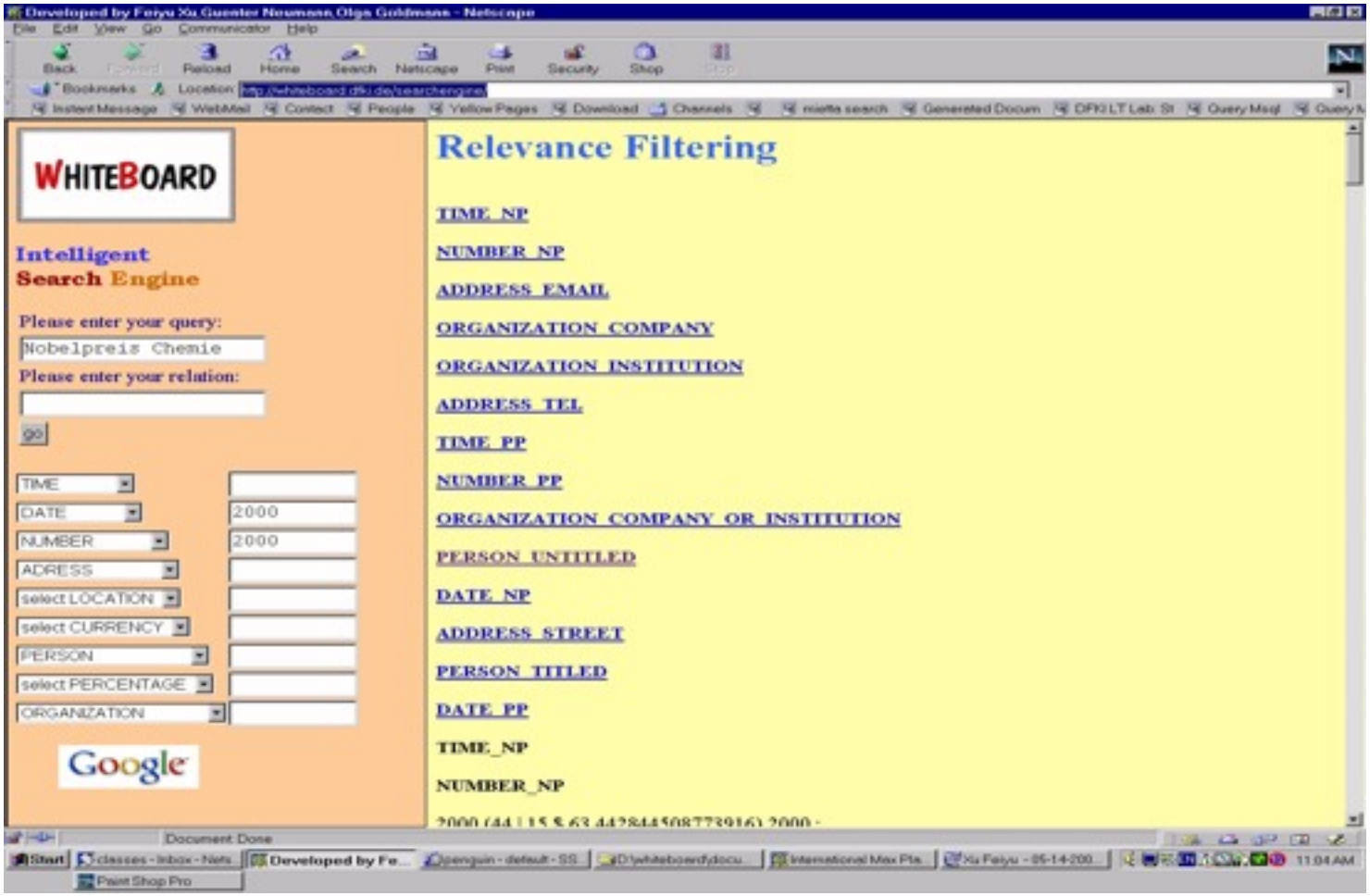
Aktueller Prototyp



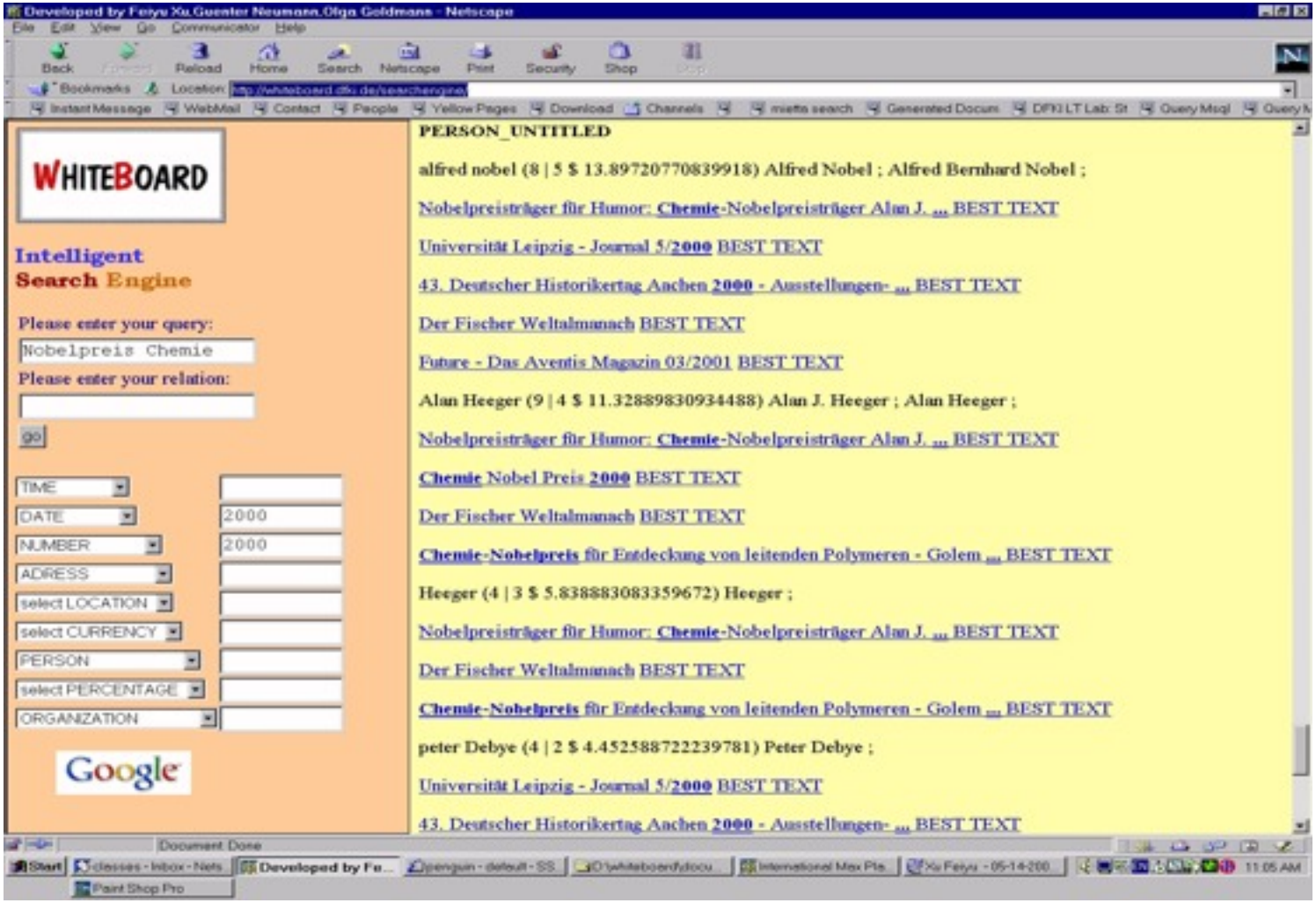
Linguistische Analyse der Google-Ergebnisse



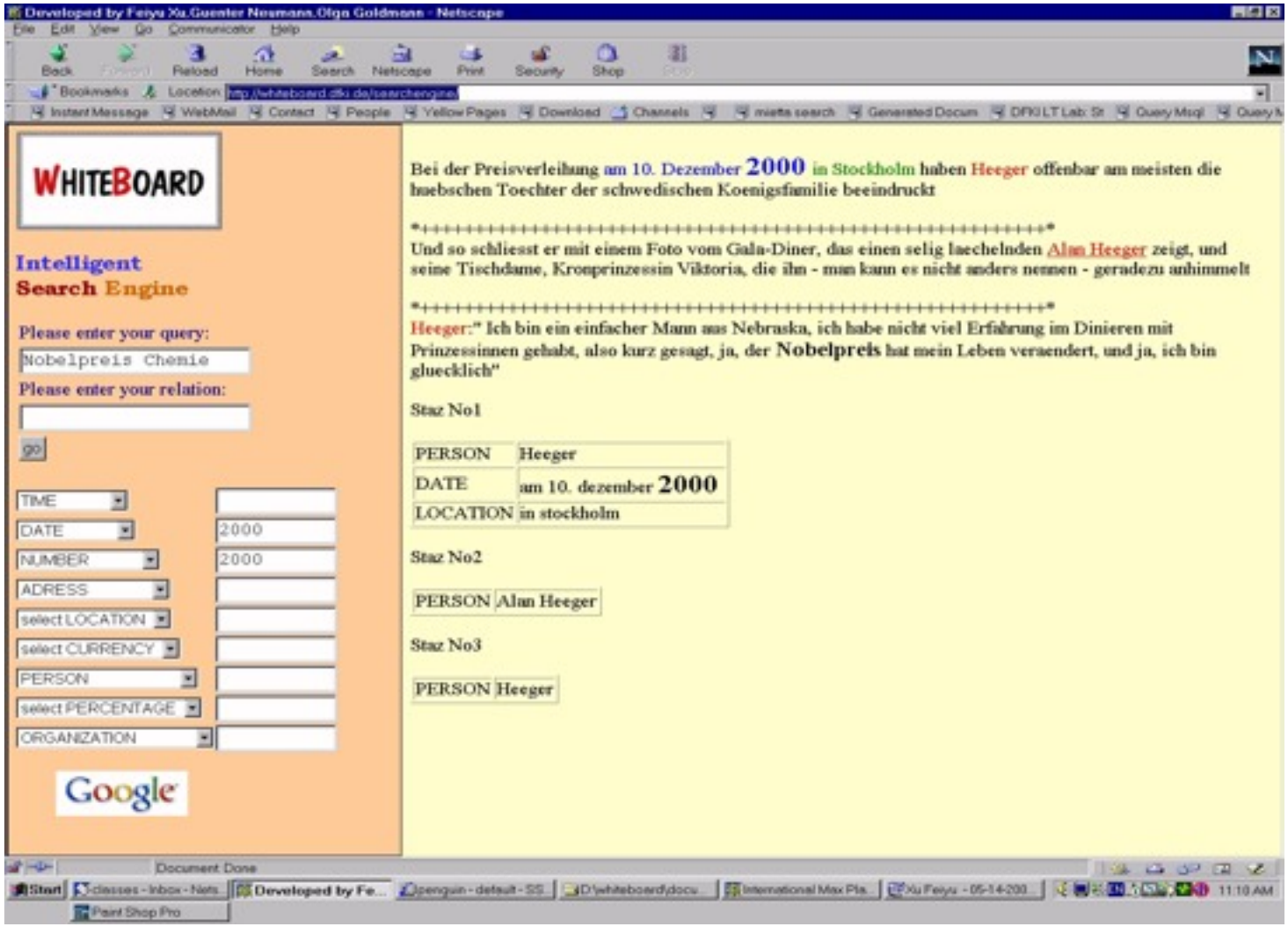
NE-basiertes Ranking



Topikalisierung



Answer Extraction



Next Steps

- Template merging
  - Unify all partial template instances found in paragraph window (3 Sentences)
  - Per paragraph one template (but different paragraphs can be found in one document)
- Template fusion
  - Combine all merged template instances
  - Choose n-best candidates as answers
  - Multi-document answer extraction

Outline: QA-Architektur