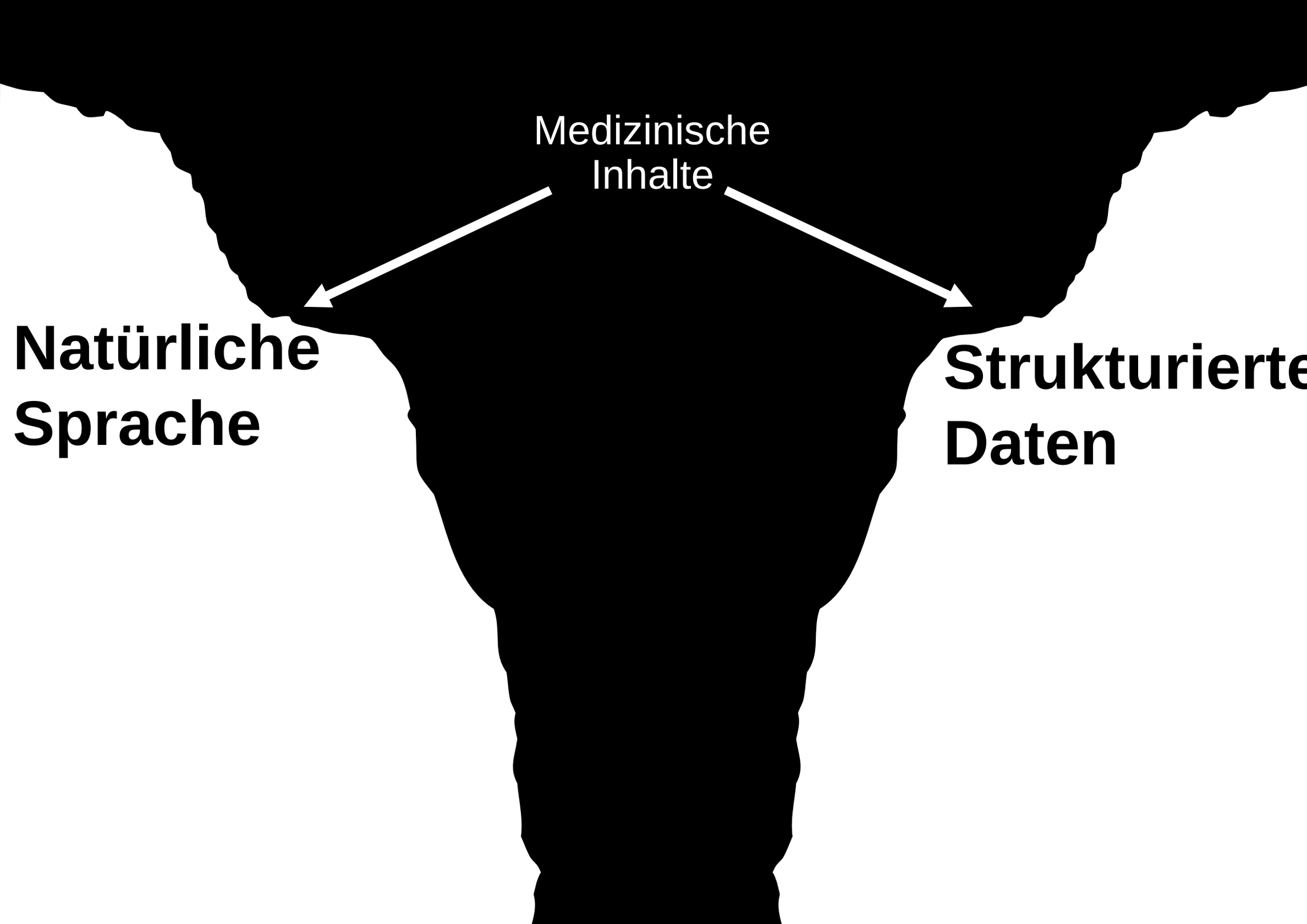
Projektgruppe
Standardisierte Terminologien in der Medizin (STM)
Deutsche Gesellschaft für Medizinische Informatik,
Biometrie und Epidemiologie e.V.

The Tower of Babel
Pieter Bruegel the Elder (about 1525 - 1569)

Automatische Verarbeitung medizinischer Sprache

Stefan Schulz

Abteilung Medizinische Informatik
Universitätsklinikum Freiburg


Medizinische
Inhalte
Natürliche
Sprache
Strukturierte
Daten

Natürliche Sprache

Datum	Uhrz.	Pflegebericht – Verlaufsbeschreibung Krankenbeobachtung	Hz.
22.4.	19^{30}	Pat. kam mit einer [illegible], hat keine Bettruhe, war schon zur Toilette und hat sich selbstständig gewaschen	
22.4	5^{30}	Pat. hat in d. Nacht sehr wenig geschlafen, sie gab also Beschwerden an	
22.4	14^{00}	Pat. lt. Pflegeplan versorgt	
	21^{00}	Pat. lt. Plan versorgt; ist nach Belastung sehr geschafft	
23.4.	5^{00}	Pat saß fast die ganze Nacht im Bett und konnte nicht im liegen schlafen	
	11^{00}	Pat hatte 3x breiig bis dünnflüssigen Stuhlgang (braun)	
	14^{00}	Versorgg. lt. Plan	
	21^{00}	Pat. lt. Plan versorgt, hält sich nicht an ihre Bettruhe	
24.4.	2^{30}	Pat beim Toilettengang erwischt, hält sich nicht an Bettruhe (sieht keinen Sinn darin)	
25	3^{30}	Pat wollte aus dem Zimmer [illegible]	

Familienanamnese:
Vater verstorben an Bronchial-Karzinom, Mutter verstorben an den Folgen einer Pneumonie. Mutter Diabetes mellitus. 5 gesunde Kinder.

Systemanamnese:
Derzeit Appetitlosigkeit, Trockengewicht um 75 Kg, derzeit 80 Kg. Miktio: gelegentlich Harn-verhalt, gehäuft Harnwegsinfekte, derzeit keine Algurie. Vor Dialyse keine Rest-Diurese. Stuhlgang obstipiert, benutzt regelmäßig Abführmittel. Vor NTX starker Juckreiz, Seit NTX deutlich rückläufig. Kein Husten/Auswurf. Noxen: Nichtraucherin, kein Alkohol.

Soziale Anamnese:
Früher Arbeiterin in der Elektronikbrache, dann Hausfrau, verheiratet, lebt mit dem Ehemann zusammen.
Allergien. Keine bekannt.

Medikation bei Aufnahme:
Ulcogant 1-1-1, Pepdul mit 0-0-0-1, Cellcept 2x1 g, Bayotensin 3 x 1, Cynt 0,2 1x1, Ludiomil 50 mg 1 x 1, Sandimmun 2 x 150 mg, Clexane 0,4 ml 1 x täglich s.c.

Status bei Übenahme:
58-jährige Patientin in vorgealtertem, reduziertem Allgemein- und adipösem Ernährungszu-stand (80 Kg Gewicht bei 160 cm Körpergröße). RR 170/80 mm Hg, Puls 66/Minute, regelmäßig. Punktförmige Depigmentierungen an beiden Unterarmen bei Zustand nach heftigem Kratzen wegen Juckreiz. Keine zervikalen Lymphome. Mundschleimhaut trocken, Zunge weißlich belegt. Rachenschleimhaut reizlos, Tonsillen schlecht einsehbar. Schilddrüse nicht vergrößert. Pulmo: Sonorer Klopfschall und vesikuläres Atemgeräusch. Cor: Spitzenstoß nicht tastbar, leise, reine Herztöne. 3/6. spindelförmiges Systolikum und 1-2/6. Decrescendo-Sofort-dialstolikum über der Aorta mit Fortleitung in die Karotis. Kein abdominales und inguinales Strömungsgeräusch. Abdomen: Bei Adipositas Organgrenzen schlecht beurteilbar, Leber/Milz nicht vergrößert. Reizlose Narbe im Bereich des rechten Unterbauches bei Zustand nach NTX. Dort leichte Druckdolenz. Wirbelsäule nicht klopfschmerzhaft. Bds. Unterschen-kelödeme. Feinschlägiger Tremor beim Arm-Vorhalte-Versuch. Pupillen isokor, Lichtreaktion prompt. Finger-Nase-Versuch bds. unsicher, ataktisch. Reflexe seitengleich.

Burden of infectious diseases in South Asia

Anita K M Zaidi, Shally Awasthi, H Janaka deSilva

Infectious diseases are a major cause of death in South Asia, with children incurring a disproportionate share of the burden. This review discusses the underlying causes of some of the more common diseases and strategies to improve their detection and control

Preventable infections are a major cause of deaths and disabilities in South Asia. Over two thirds of the estimated 3.7 million deaths in children in South Asia in the year 2000 were attributable to infections such as pneumonia, diarrhoea, and measles.[1 2] India now has the second largest population with AIDS and HIV infection in the world, and tuberculosis and chronic hepatitis continue to threaten the lives of millions. Of the overall burden of deaths related to infectious disease in the region, around 63% are in children aged under 5 years.[3] Serious effort should be devoted to the control of infectious disease if South Asian countries are to meet their millennium development goal of two thirds reduction in child mortality by 2015.

Sri Lanka alone among South Asian countries has made remarkable progress in reducing the burden of infectious disease, despite civil war and meagre resources.

This review describes the burden of infectious

Summary points

- Acute respiratory infections, diarrhoea, and neonatal infections remain major child killers
- India has the second highest burden of HIV and AIDS in the world, with 4.58 million people infected with HIV
- Antibiotic misuse has resulted in high rates of antimicrobial resistance
- Only half of all South Asian children receive routine immunisations, and many new vaccines have not been introduced in mass immunisation programmes
- Lack of surveillance systems and poorly

NetDoktor.de
Startseite
Aktuell
Nachrichten
Features
Newsletter
Lexikon
Krankheiten
Symptome
Untersuchungen
Eingriffe
Laborwerte
Medikamente
Themen
Allergie
Asthma
Diabetes
Erektile Dysfunktion
Herz
Reizdarm
Rheuma
Transplantation
Verhütung
Alle Themen
Service
Teste Dich Selbst

Diabetes mellitus Typ-1 (Zuckerkrankheit)

Dr. med. Ingo Röhrig, Facharzt für Innere Medizin, Angiologie, Diabetologie

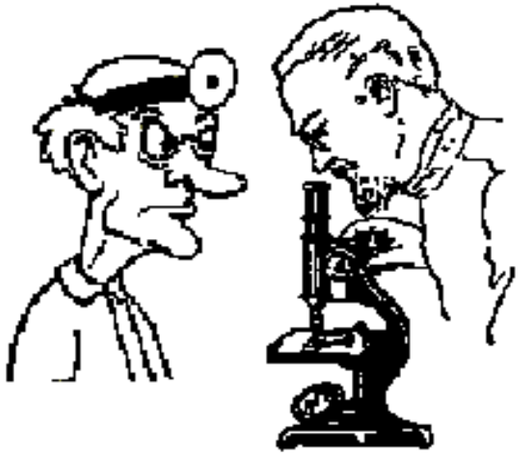
Was ist Diabetes Typ-1?

Diabetes mellitus ist ein Überbegriff für verschiedene Stoffwechselkrankheiten. Die Gemeinsamkeit ist, dass sie zu erhöhten Blutzuckerwerten führen. In Deutschland leben rund sechs Millionen Diabetiker - davon sind etwa 200.000 Menschen Typ-1-Diabetiker. Schätzungsweise 15.000 Menschen erkranken jährlich in Deutschland neu.

Typ-1 Diabetes: Lebenslang Insulin spritzen

Typ-1-Diabetes wird durch den absoluten Mangel am Hormon Insulin verursacht. Dieser Diabetestyp heißt deshalb auch insulinabhängiger Diabetes mellitus. Meistens beginnt die Erkrankung schon im Kindes- und Jugendalter, aber auch im fortgeschrittenen Alter kann sich Typ-1-Diabetes entwickeln. Die Krankheit ist derzeit noch nicht heilbar, lässt sich aber gut mit Insulin behandeln. Allerdings müssen die Patienten das lebenslange Spritzen von Insulin in Kauf nehmen. Für jeden Diabetiker ist es wichtig, den Blutzucker optimal einzustellen. Nicht nur, um akute Entgleisungen des Stoffwechsels wie eine Unterzuckerung zu verhindern, sondern auch um diabetische Folgeerkrankungen zu vermeiden oder hinauszuzögern.

Natürliche Sprache

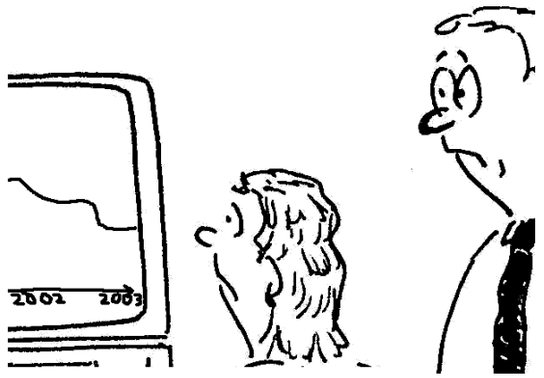


- … unverzichtbar für
- Kommunikation zwischen Mitarbeitern des Gesundheitswesens
- Wissenschaftliche Kommunikation
- Klinische Dokumentation
- Wissenschaftliches Publikationswesen
- Vermittlung kanonischen Fachwissens an Fachleute, Studierende und Laien

…unverzichtbar für

- Kodierung von Diagnosen und Prozeduren, DRGs
- Klinische und epidemiologische Studien
- Gesundheitsbericht-erstattung
- Krankheitsspezifische Register
- Qualitätssicherung, Controlling
- Dokumentenindexierung und Retrieval
- etc.

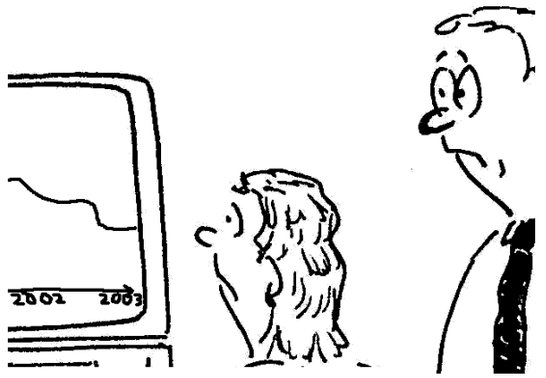
Strukturierte Daten



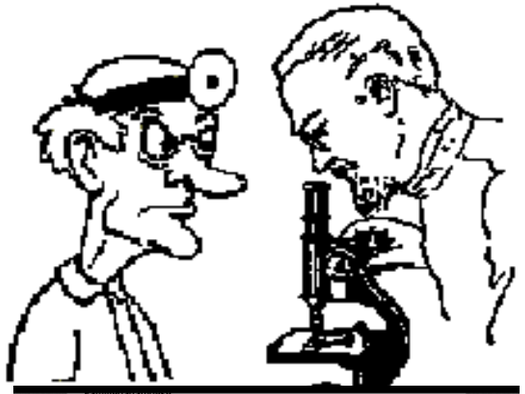
…erfordern medizinische Terminologiesysteme

- ICD
- OPS 301
- SNOMED
- LOINC
- MeSH
- etc., etc.

Strukturierte Daten

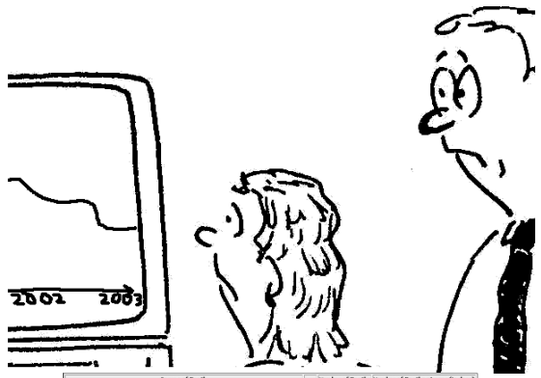


Natürliche Sprache



Familienanamnese:
Vater verstorben an Bronchial-Karzinom, Mutter verstorben an den Folgen einer Pneumonie. Mutter Diabetes mellitus. 5 gesunde Kinder.
Systemanamnese:
Derzeit Appetitlosigkeit, Trockengewicht um 75 Kg, derzeit 80 Kg. Miktio: gelegentlich Harn-verhalt, gehäuft Harnwegsinfekte, derzeit keine Algurie. Vor Dialyse keine Rest-Diurese. Stuhlgang obstipiert, benutzt regelmäßig Abführmittel. Vor NTX starker Juckreiz, Set NTX deutlich rückläufig. Kein Husten/Auswurf. Noxen: Nichtraucherin, kein Alkohol.
Soziale Anamnese:
Früher Arbeiterin in der Elektronikbrache, dann Hausfrau, verheiratet, lebt mit dem Ehemann zusammen.
Allergien. Keine bekannt.
Medikation bei Aufnahme:
Ulcogant 1-1-1, Pepdul mit 0-0-0-1, Cellcept 2x1 g, Bayotensin 3 x 1, Cynt 0,2 1x1, Ludiomil 50 mg 1 x 1, Sandimmun 2 x 150 mg, Clexane 0,4 ml 1 x täglich s.c.
Status bei Übenahme:
58-jährige Patientin in vorgealtertem, reduziertem Allgemein- und adipösem Ernährungszu-stand (80 Kg Gewicht bei 160 cm Körpergröße). RR 170/80 mm Hg, Puls 66/Minute, regelmäßig. Punktförmige Depigmentierungen an beiden Unterarmen bei Zustand nach heftigem Kratzen wegen Juckreiz. Keine zervikalen Lymphome. Mundschleimhaut trocken, Zunge weißlich belegt. Rachenschleimhaut reizlos, Tonsillen schlecht einsehbar. Schilddrüse nicht vergrößert. Pulmo: Sonorer Klopfschall und vesikuläres Atemgeräusch. Cor: Spitzenstoß nicht tastbar, leise, reine Herztöne. 3/6. spindelförmiges Systolikum und 1-2/6. Decrescendo-Sofort-dialstolikum über der Aorta mit Fortleitung in die Karotis. Kein abdominales und inguinales Strömungsgeräusch. Abdomen: Bei Adipositas Organgrenzen schlecht beurteilbar, Leber/Milz nicht vergrößert. Reizlose Narbe im Bereich des rechten Unterbauches bei Zustand nach NTX. Dort leichte Druckdolenz. Wirbelsäule nicht klopfschmerzhaft. Bds. Unterschen-kelödeme. Feinschlägiger Tremor beim Arm-Vorhalte-Versuch. Pupillen isokor, Lichtreaktion prompt. Finger-Nase-Versuch bds. unsicher, ataktisch. Reflexe seitengleich.

Strukturierte Daten



Causes of Death			Number of Deaths	Number of Deaths at ages (in days)			
ICD 10 Codes	Cause Groupings	Sex	< 1 year	< 1	1-6	7-27	28-364
-	All causes	M	2485	168	703	416	1198
		F	1885	138	446	279	1020
		U	0	0	0	0	0
A00-B99	Infectious and parasitic diseases	M	95	0	1	9	85
		F	93	0	2	9	82
A00-A09	Intestinal infectious diseases	M	65	0	0	6	59
		F	63	0	0	6	57
A37	Whooping cough	M	1	0	0	0	1
		F	0	0	0	0	0
E40-E64	Nutritional deficiencies	M	1	0	0	0	1
		F	0	0	0	0	0
G00-G98	Diseases of the nervous system	M	92	0	2	26	64
		F	61	0	4	16	41
G00,G03	Meningitis	M	37	0	2	18	17
		F	30	0	4	12	14
J00-J98	Diseases of the respiratory system	M	734	0	3	66	665
		F	609	1	3	51	554
J12-J18	Pneumonia	M	705	0	3	65	637
		F	587	1	3	50	533
J10,J11	Influenza	M	0	0	0	0	0
		F	0	0	0	0	0
Q00-Q99	Congenital anomalies	M	454	38	133	100	183
		F	375	35	94	61	185
Q03,Q05	Spina bifida and hydrocephalus	M	71	5	16	5	45
		F	74	2	6	17	47
Q20-Q28	Congenital anomalies of heart and circulatory system	M	222	8	67	57	90
		F	149	9	42	23	75
P00-P96	Certain conditions originating in the perinatal period	M	679	127	169	187	6
		F	571	97	340	131	3
P10-P15	Birth trauma	M	56	11	35	10	0
		F	34	4	25	5	0

Schwerpunkt:
Erfassung von Daten

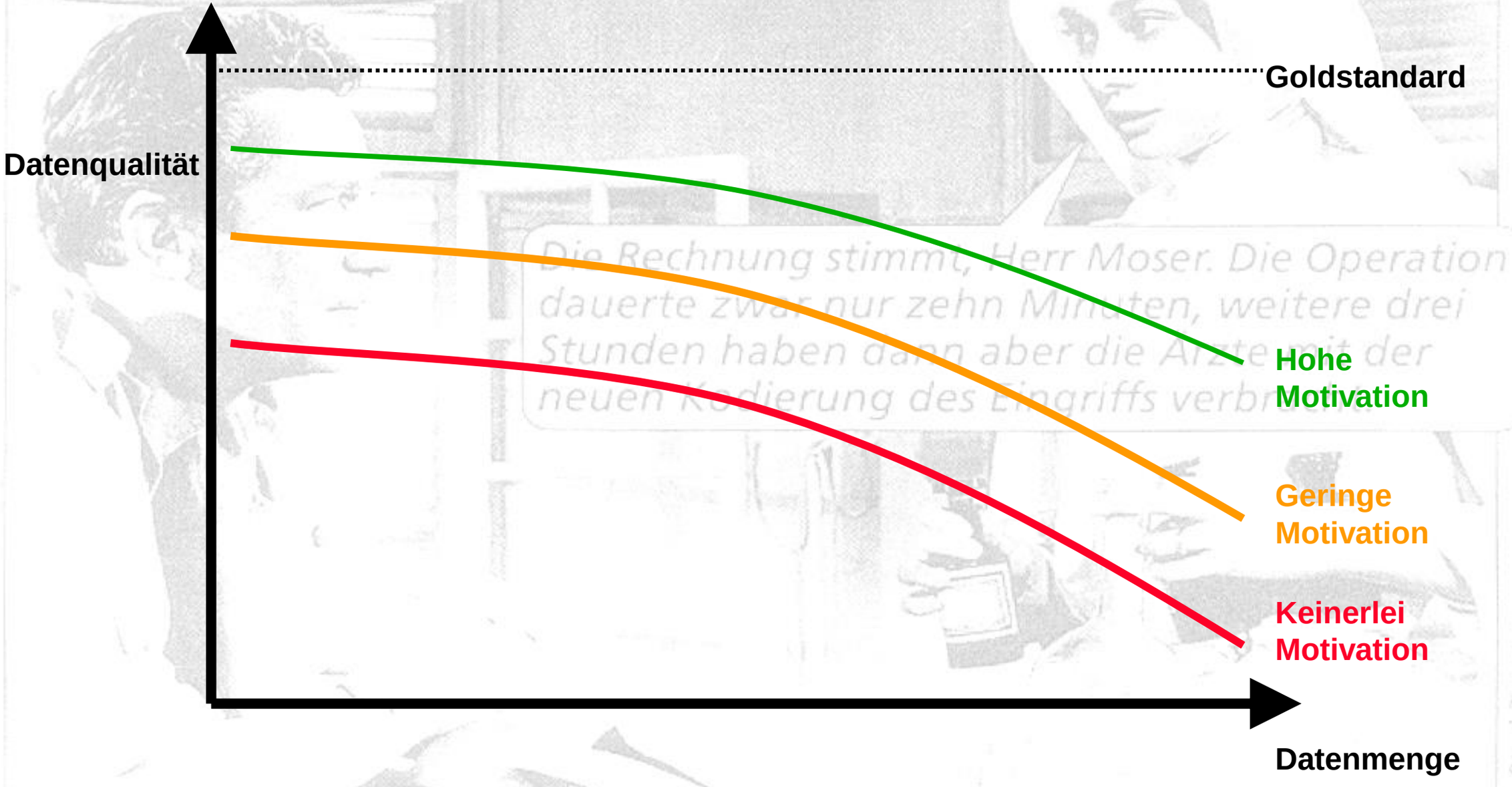
Natürliche Sprache

Strukturierte Daten

+	Qualität	-
-	Kosten	+


Die Rechnung stimmt, Herr Moser. Die Operation dauerte zwar nur zehn Minuten, weitere drei Stunden haben dann aber die Ärzte mit der neuen Kodierung des Eingriffs verbracht.

Abhängigkeit:
Datenmenge – Datenqualität - Motivation



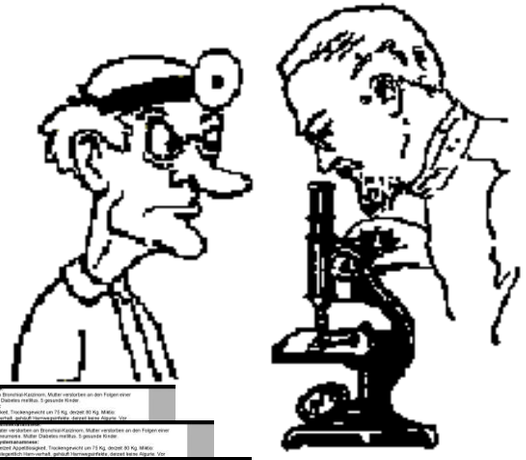
Schwerpunkt:
Auswertung von Daten

Natürliche Sprache

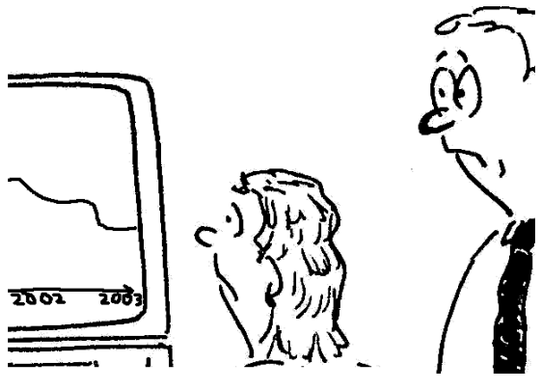
Strukturierte Daten

-	Qualität	+
+	Kosten	-

Natürliche Sprache



Strukturierte Daten



Automatische Verarbeitung medizinischer Sprache

Natürliche Sprache

Strukturierte Daten

Projektgruppe
Standardisierte Terminologien in der Medizin (STM)
Deutsche Gesellschaft für Medizinische Informatik,
Biometrie und Epidemiologie e.V.

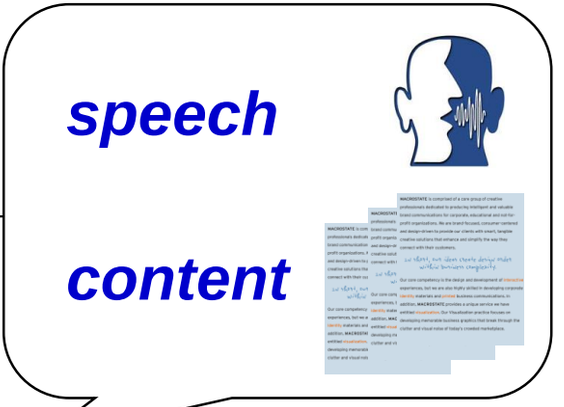
The Tower of Babel
Pieter Bruegel the Elder (about 1525 - 1569)

Automatische Verarbeitung medizinischer Sprache

Stefan Schulz

Abteilung Medizinische Informatik
Universitätsklinikum Freiburg

Projektgruppe
Standardisierte Terminologien in der Medizin (STM)
Deutsche Gesellschaft für Medizinische Informatik,
Biometrie und Epidemiologie e.V.



The Tower of Babel
Pieter Bruegel the Elder (about 1525 - 1569)

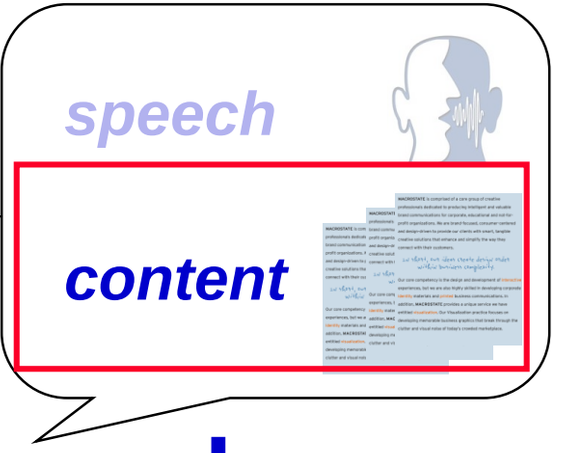
Automatische Verarbeitung medizinischer Sprache

Stefan Schulz

Abteilung Medizinische Informatik
Universitätsklinikum Freiburg

Projektgruppe
Standardisierte Terminologien in der Medizin (STM)
Deutsche Gesellschaft für Medizinische Informatik,
Biometrie und Epidemiologie e.V.

The Tower of Babel
Pieter Bruegel the Elder (about 1525 - 1569)



Automatische Verarbeitung medizinischer Sprache

Stefan Schulz

Abteilung Medizinische Informatik
Universitätsklinikum Freiburg

Auffälligkeiten der Medizinsprache

- Sprachmix: Deutsch / English / Lateinisch
- Unterschiedliche Sprachebenen: Ärzte- vs. Laiensprache
- Griechisch/Lateinische Wordstämme, Lateinische Flexionen:
  Thyreoglobulin, Ulzera, E.coli, Kolibakterien
- Hohe lexikalische Produktivität:
  - Komposita: *Bypassoperation, Kaliumüberdosierung*
  - Eponyme: *Parkinsonsche Erkrankung, M. Alzheimer*
  - Akronyme, Wortneubildungen: *SARS, AIDS, ARDS, 5-FU, HWI, psbAI, GGDEF, WDWN*
- Paragrammatikalität / Jargon:
  Kein Anhalt für Malignität. (Unvollständiger Satz)
- Agrammatikalität (Diktier-, Schreibfehler)
- Extragrammatikalität: *Gewebe wurde lymphoztyär infiltriert*

Zwei Hauptszenarien der medizinischen Textanalyse

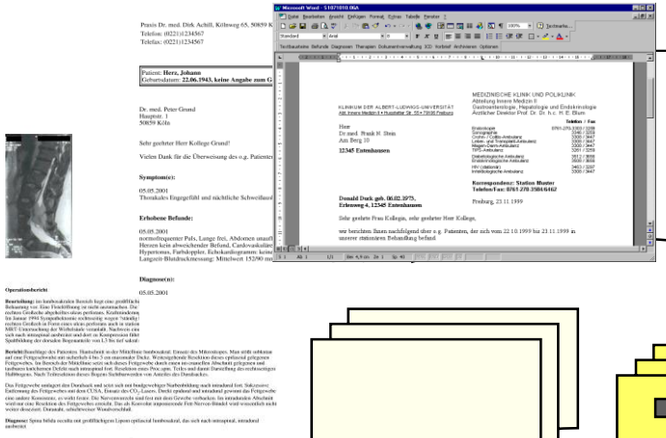
- **Information Retrieval:**

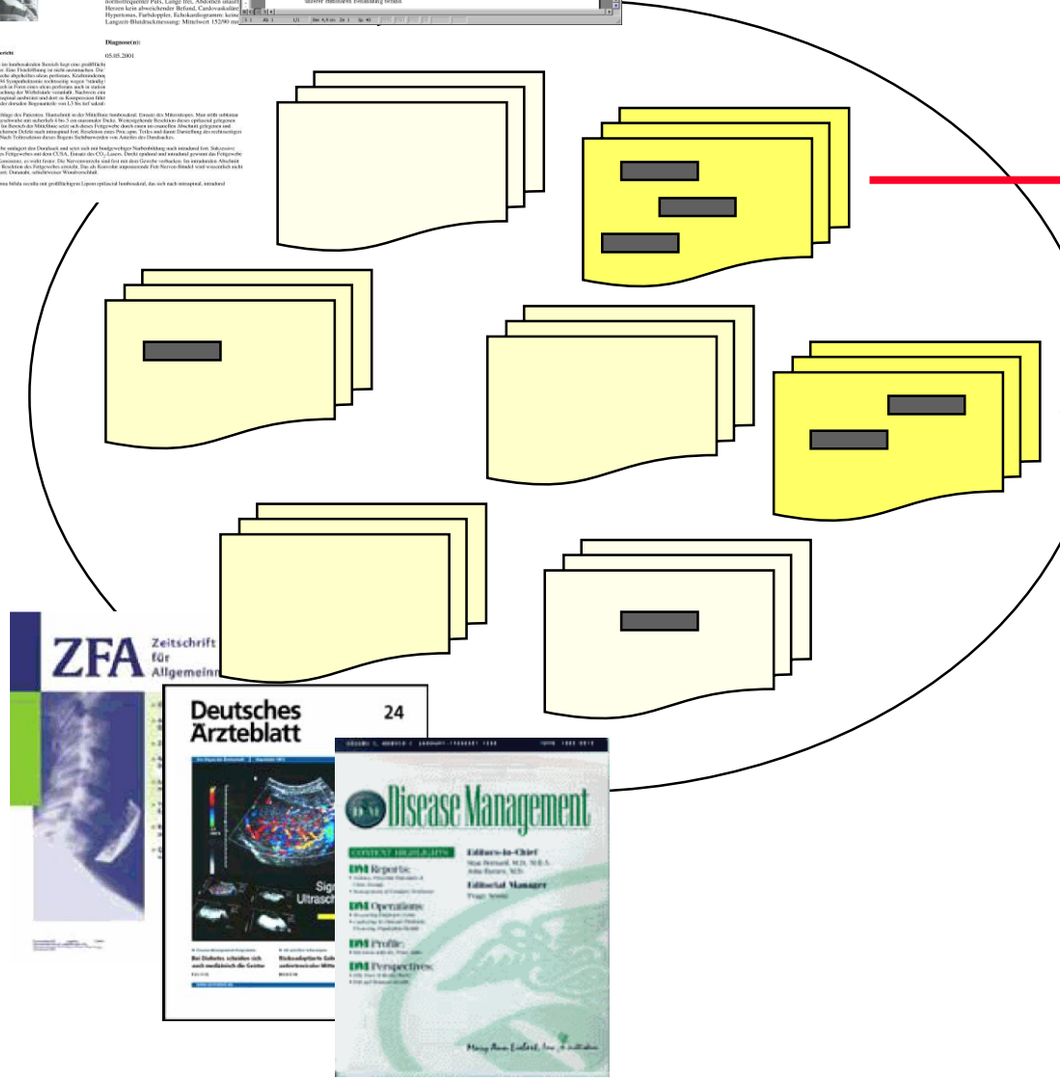
  gezieltes Suchen nach Informationen in einer oder mehreren großen Informationssammlungen.

- **Text Mining:**

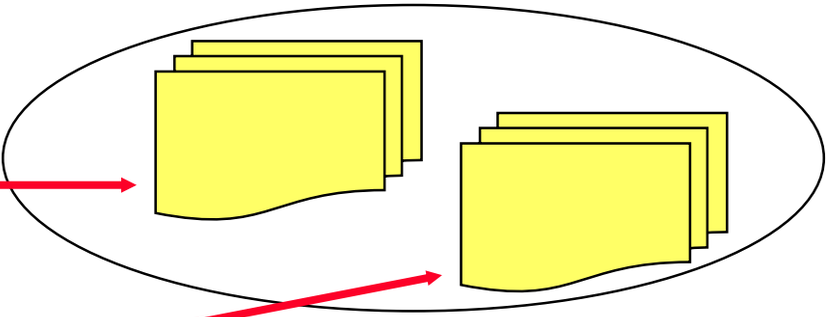
  Technologien, die es ermöglichen, relevante und „neue“ Information in unstrukturierten Texten automatisch zu erkennen und zu extrahieren

IR vs. TM

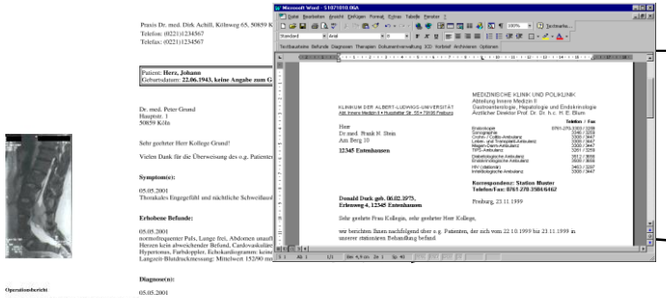


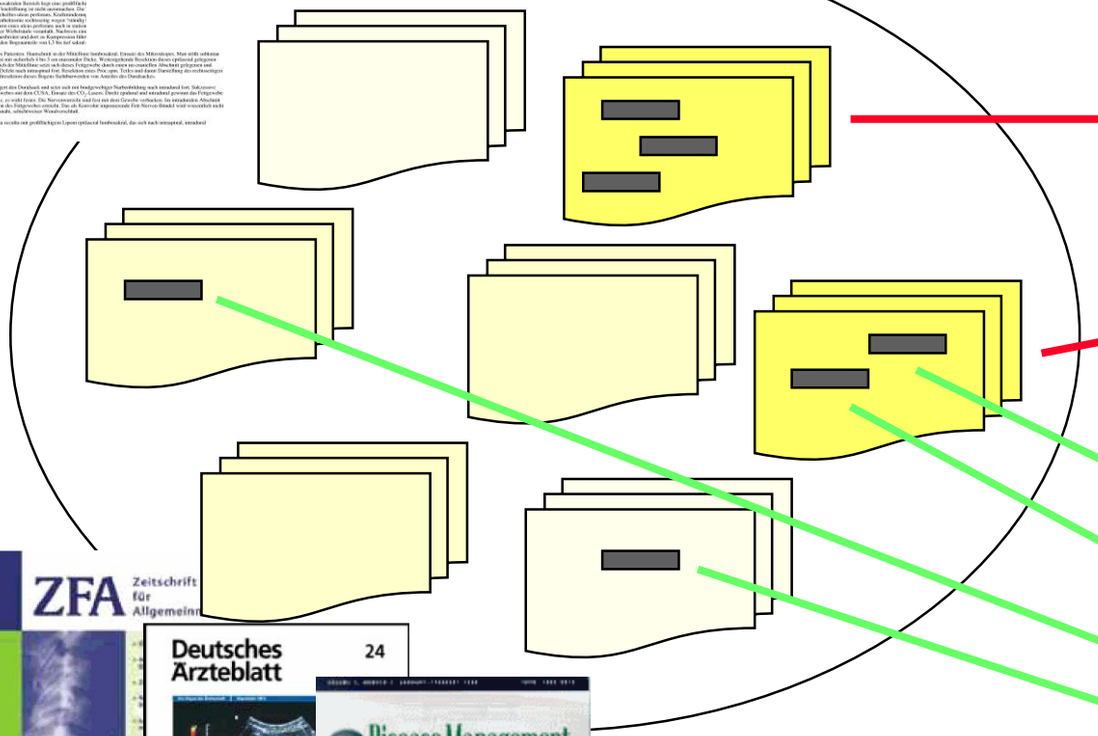


Information Retrieval

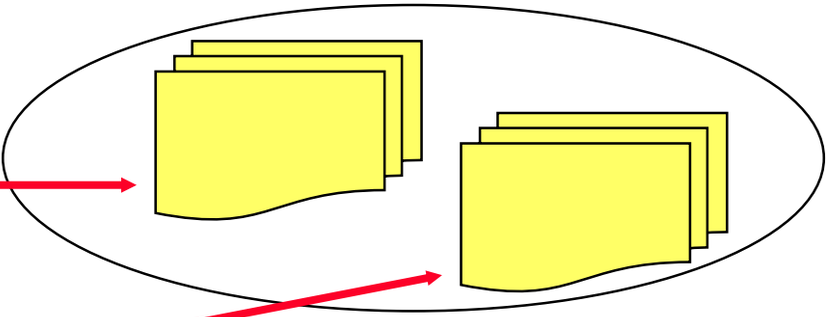


IR vs. TM

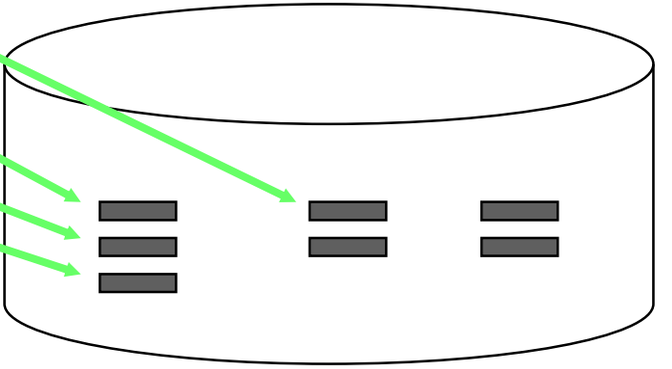




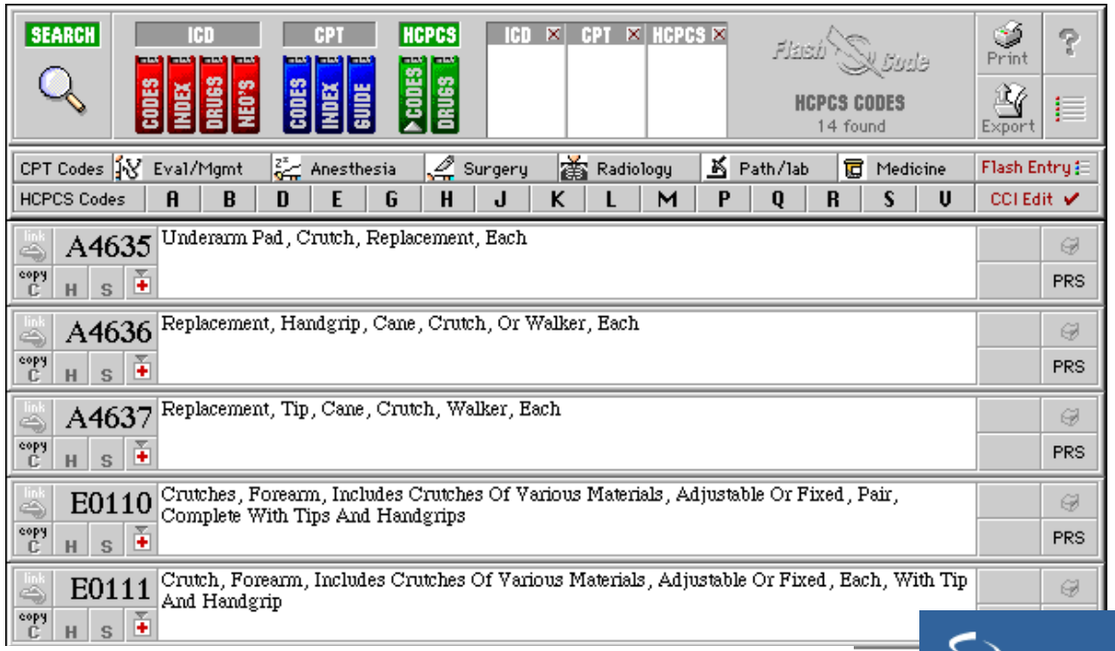
Information Retrieval

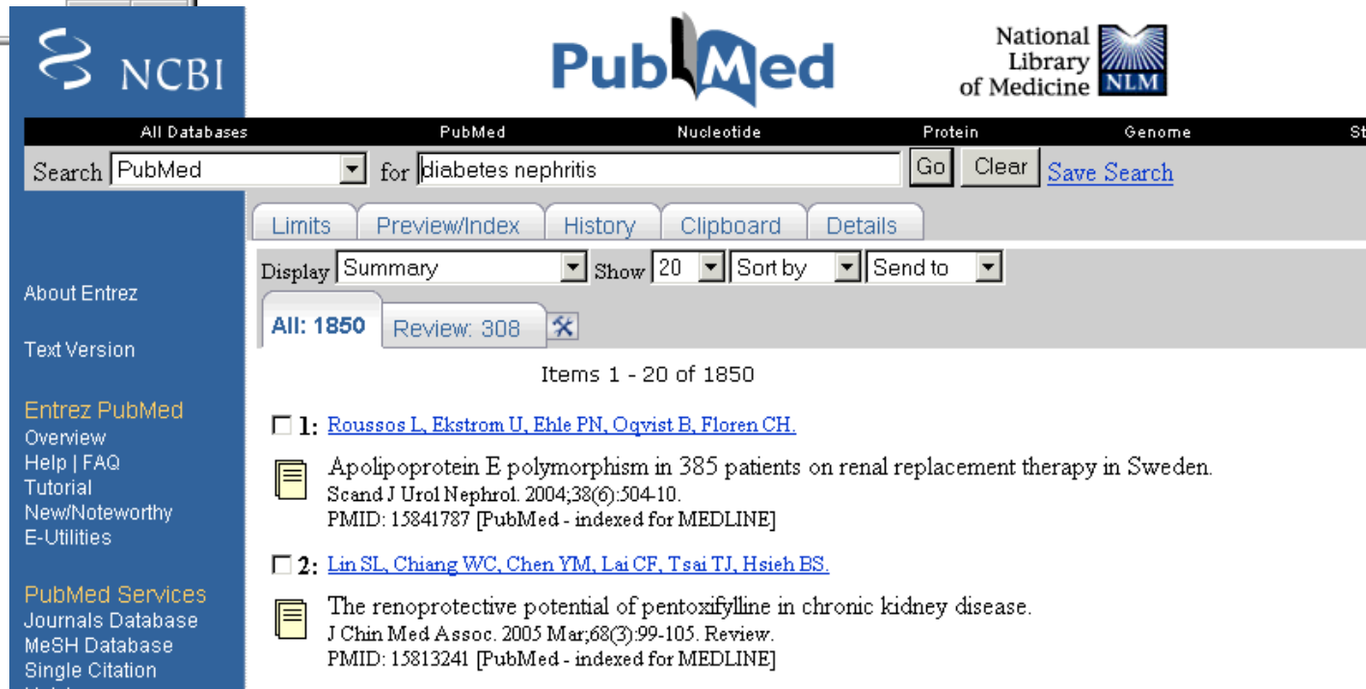


Text Mining



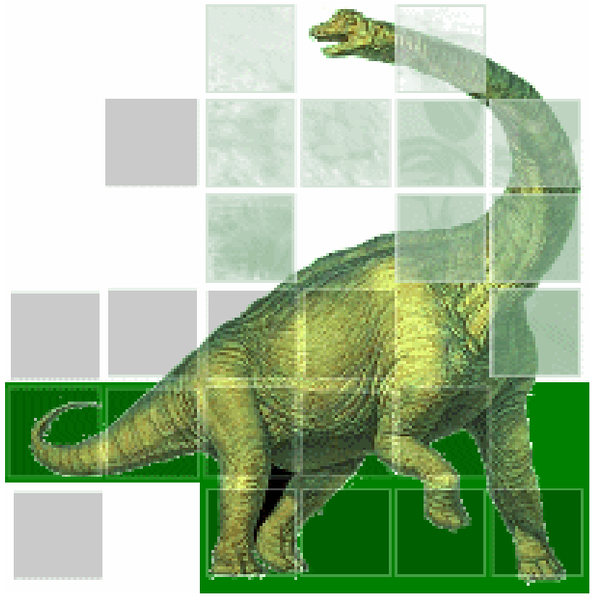
Information Retrieval





Beispiel: Sprachübergreifendes Dokumentenretrieval: MorphoSaurus

- Subwort-Lexikon:
  - Organisiert bedeutungstragende Wortstämme und Affixe in mehreren Sprachen
- Subwort-Thesaurus:
  - Gruppiert synonyme Lexikoneinträge (auch sprachübergreifend)
- Zerlegungsalgorithmus:
  - Extraktion von Subwörtern und Zuweisung von *Bedeutungsklassen*

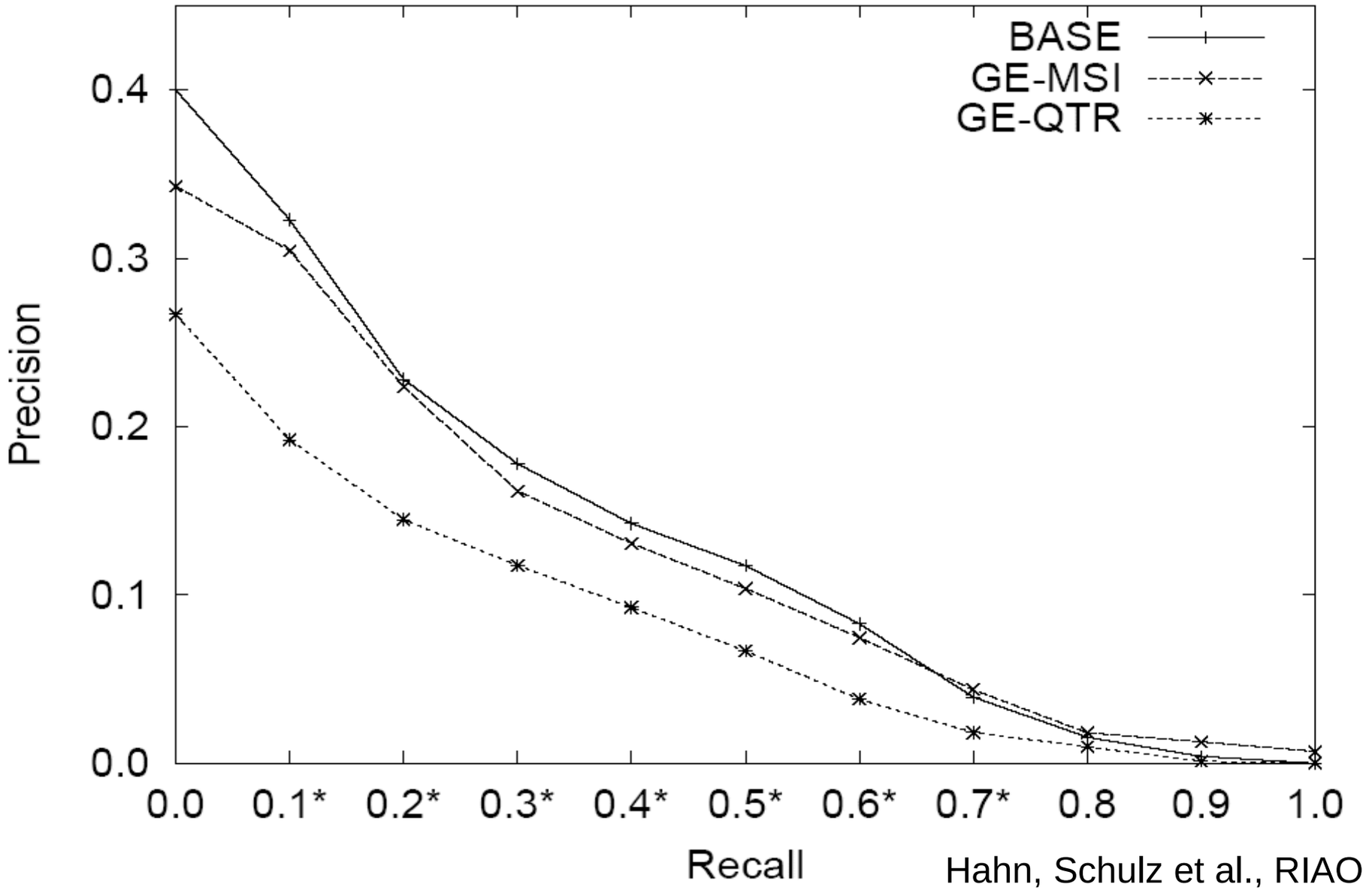


Morphosaurus

→ **Morphosaurus-Identifier (MID)**

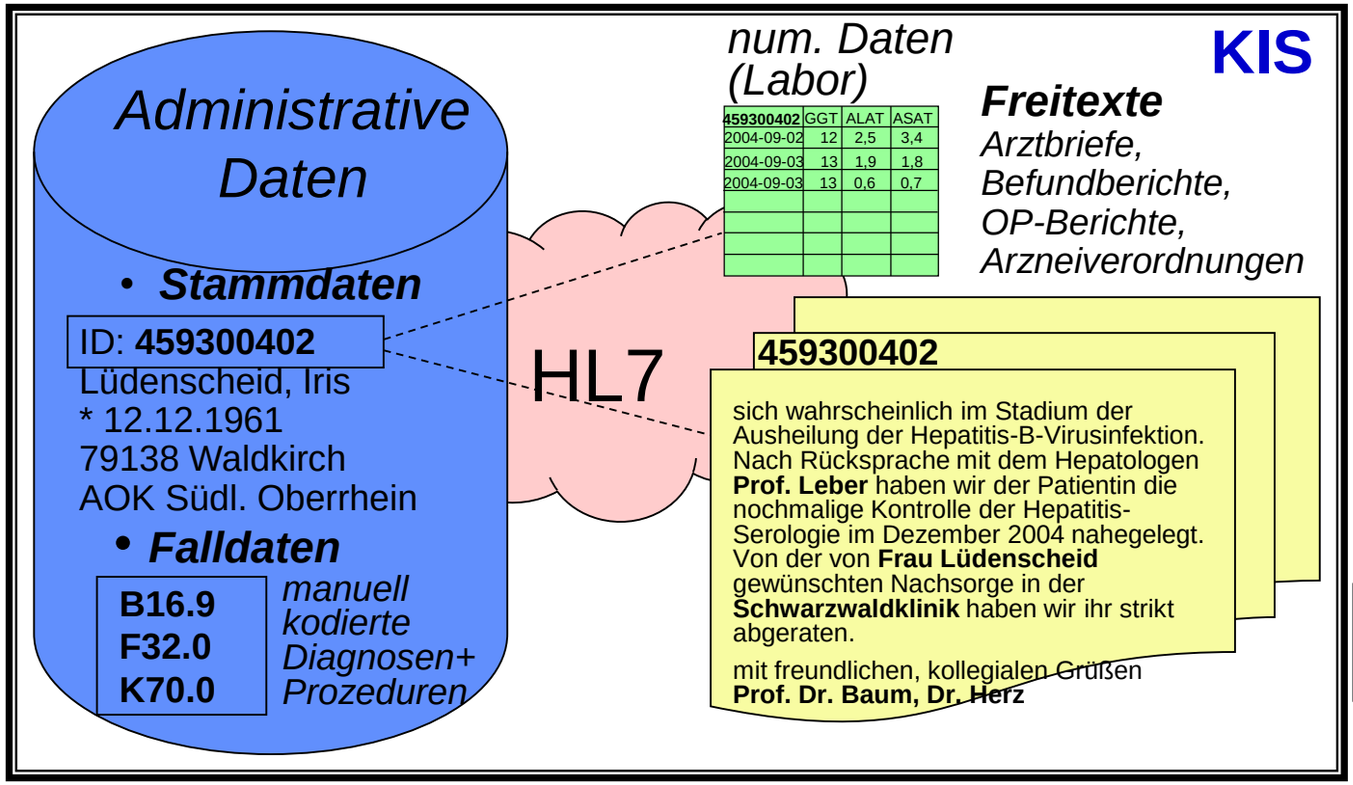
Original Document	Orthographic Normalization	Morphological Segmentation	Semantic Normalization MID
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value#** #suggest# **#diagnost# #primar# #small# #thyre# #suppress# tsh #nivell#** #suggest# **#up# #thyre#** .
Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose, ein supprimierter TSH-Spiegel spricht dagegen für eine Schilddrüsenüberfunktion.	erhoehte tsh-werte erlauben die diagnose einer primaeren hypothyreose, ein supprimierter tsh-spiegel spricht dagegen fuer eine schilddruesenueberfunktion.	er hoeh te tsh - wert e erlaub en die diagnos e einer primaer en hypo thyre ose, ein supprim iert er tsh - spiegel spricht dagegen fuer eine schilddrues en ueber funktion.	**#up# tsh - #value#** #permit# **#diagnost# #primar# #small# #thyre# , #suppress# tsh -** {#mirrorr# **#nivell#**} #speak# **#thyre# #up#** #function# .
A presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# **#value# #up# tsh** #suggest# **#diagnost# #small# #thyre# #primar# , #nivell# #suppress# tsh** #suggest# **#up# #thyre#** .

MorphoSaurus: Sprachübergreifendes med. Dokumentenretrieval (Deutsch / Englisch)



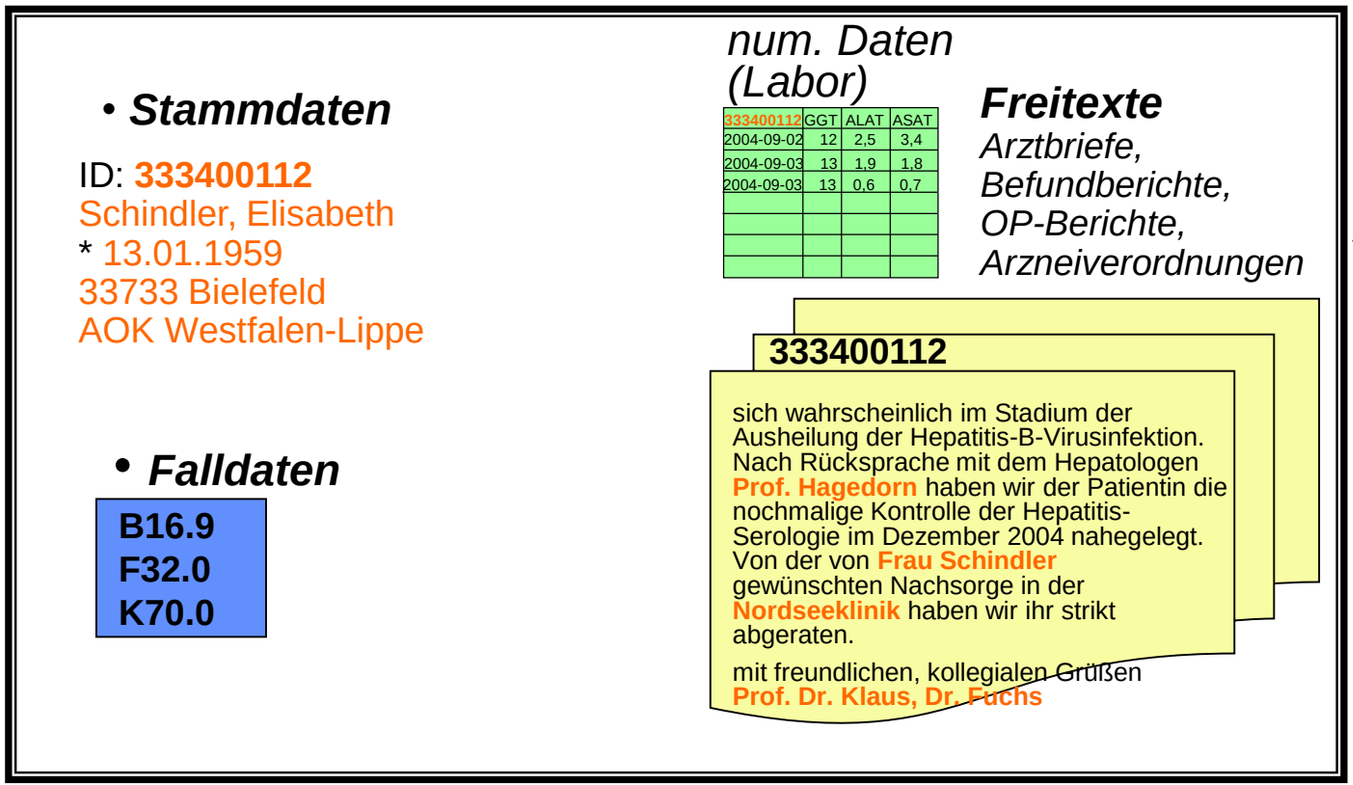
Hahn, Schulz et al., RIAO 2004

Text Mining



Text-Mining-System (klinikintern)

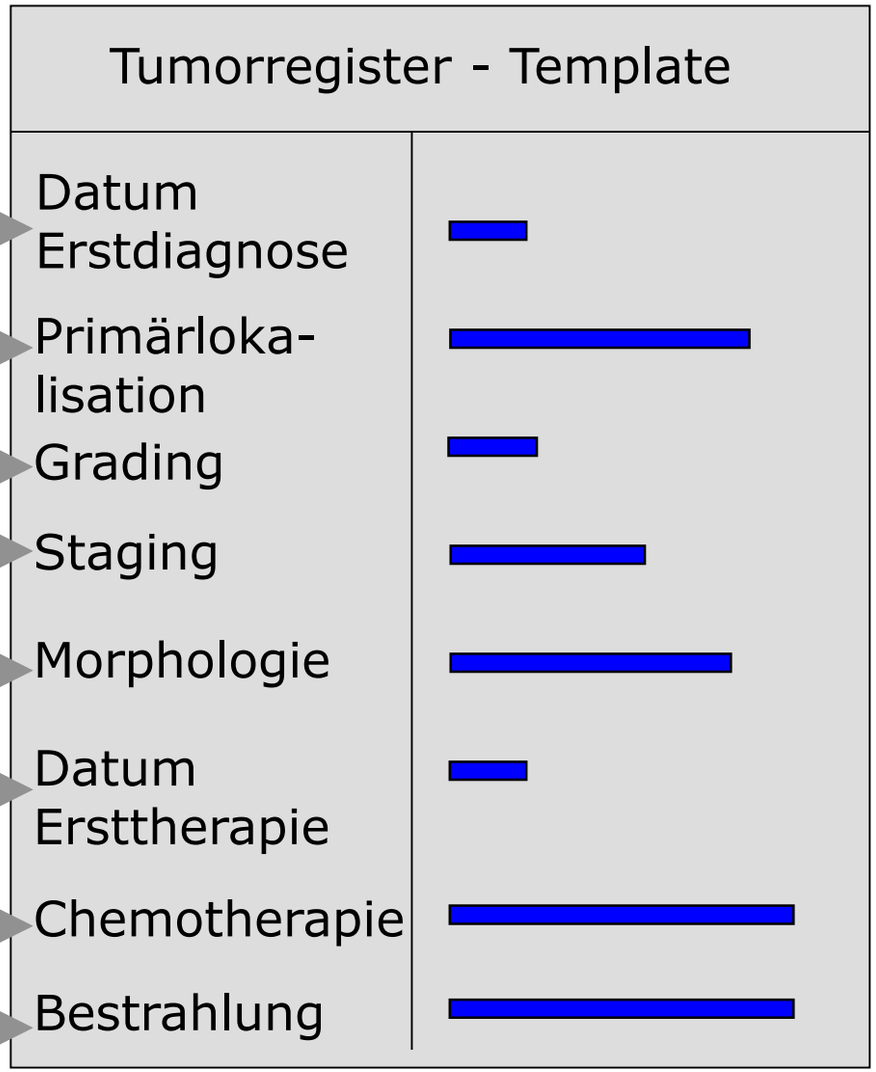
(semi)automatische Pseudonymisierung



Text-Mining-System (klinikextern)

Text Mining: Anwendungsszenario I

shadow was pointed out on a routine chest X-ray film, but she had no further examination. Physical examination on admission revealed purpura of the upper and lower extremities, swelling of the gums and tonsils, but no symptoms showing the complication of myasthenia gravis. Hematological tests revealed leucocytosis: WBC count 68 700/µl (blasts 11.5%, myelocytes 0.5%, bands 2.0%, segments 16.0%, monocytes 65.5%, lymphocytes 4.0%, atypical lymphocytes 0.5%), Hb 7.1 g/dl (reticulocytes 12%) and a platelet count of 9.1×10^4/µl. Further laboratory examination revealed elevated serum lactic dehydrogenase (589 U/l), vitamin B_{12} (2010 pg/ml) and ferritin (650. 0 ng/ml). Human chorionic gonadotropin and [alpha]-fetoprotein levels were normal. A bone marrow aspiration revealed hypercelllar bone marrow with a decreased number of erythroblasts and megakaryocytes and an increased number of monoblasts that were positive for staining by [alpha]-naphthyl butyrate esterase and negative for staining by naphthol ASD chloroacetate esterase. Chest X-ray upon admission revealed a mediastinal mass and an elevated left diaphragm . Computed tomography (CT) of the chest showed a left anterior mediastinal mass . Based on these findings, the patient was diagnosed with a mediastinal tumor accompanied by AMoL. First, in June 1991, the patient was treated with DCMP therapy: daunorubicin (DNR) (25 mg/m^2, days 1, 2, 3, 4, 6 and 8), cytosine arabinoside (Ara-C) (100 mg/m^2, days 1-9), 6MP-riboside (6-MP) (70 mg/m^2, days 1-9) and prednisolone (PSL) (20 mg/m^2, days 1-9), followed by five courses of consolidation chemotherapy [1, DCMP; 2, ID-Ara-C:adriacin (ADR), vincristine (VCR), Ara-C, PSL; 3, DCMP; 4, ID-Ara-C; 5, A-triple V: Ara-C, VP-16, VCR, vinblastine (VBL)]. After induction chemotherapy, a hematological examination and bone marrow findings had improved to normal, and complete remission was attained. Chest CT scan after chemotherapy in November 1991 revealed regression of the mediastinal tumor . An invasive thymic tumor was suspected and surgery was undertaken in January 1992. The tumor (50 × 45 × 45 mm), located mainly in the anterior mediastinum, was strongly adhered to the adjacent tissues. Resection of the tumor included the left upper lobe of the lung, the phrenic nerve and pericardium. The histological finding was that the tumor cells have large, vesicular nuclei and prominent nucleoli, but keratinazation was unclear . The results of immunohistochemical finding of anti-TdT was negative. From these findings, we diagnosed poorly or moderately differentiated squamous cell carcinoma of the thymus. The postoperative course was uneventful. The patient underwent radiation therapy of the mediastinum and left hilum at doses of 4000 cGy delivered over 4 weeks. She was discharged in March 1992. After the first AMoL remission, the patient suffered a relapse six times and was repeatedly admitted for chemotherapy. During these periods, chest X-ray and CT revealed no recurrence of the mediastinal tumor. During her tenth admission, the patient developed pneumonia during chemotherapy and died in October 1996. No autopsy was performed.



Text Mining: Anwendungsszenario II

Milde und Schwere Verlaufsformen: EB simplex (EBS), EB dystrophica (EBD)

links: Serumblasen
oben: Blutblase

Risikoabschätzung von Tumorentstehung durch Genotyp-Phänotyp-Korrelationen bei *Epidermolysis bullosa dystrophica*

- *Epidermolysis bullosa*: Gruppe von genetischen Hautkrankheiten mit Mutationen in Genen für Strukturproteine der dermo-epidermalen Basalmembranzone. Inzidenz: 1 / 100.000 Geburten.
- Minimale Traumata führen zu Blasenbildung an Haut und hautnahen Schleimhäuten, Abheilung der dadurch entstandenen Wunden führt oft zur Narbenbildung und ggf. zu Verwachsungen, die auch Kontrakturen bedingen können.

Text Mining: Anwendungsszenario II

EB dystrophica (EBD)

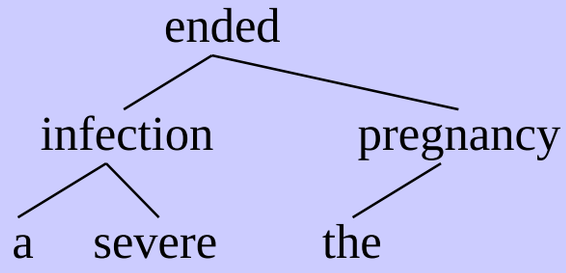
- mehr als 300 unterschiedliche Mutationen des Kollagen VII-Gens publiziert und/oder in den Mutations-Datenbanken, mehrere Hundert weitere, noch nicht bekannte Mutationen.
- Ziel des Text Minings: Verbesserung der Prognosestellung — Auffinden bislang unentdeckter Korrelationen zwischen Art und Lokalisation der Genmutation und des klinischen Langzeitverlaufs sowie der Erkennung maligner Entartungen
- Abgleichen der Daten
  - in der Literatur,
  - in Mutations-Datenbanken,
  - in eigenen Laborbefunden etc.
  - in internen und externen klinischen Dokumenten

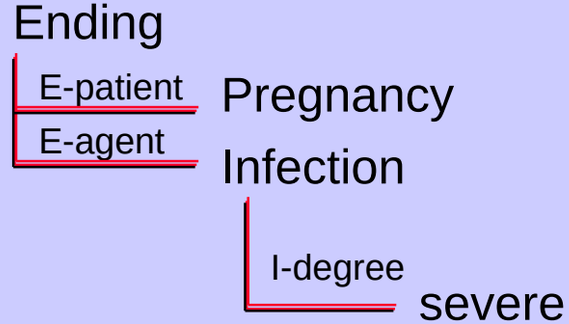
Architektur eines Biomedizinischen Textanalyse-Kernsystems

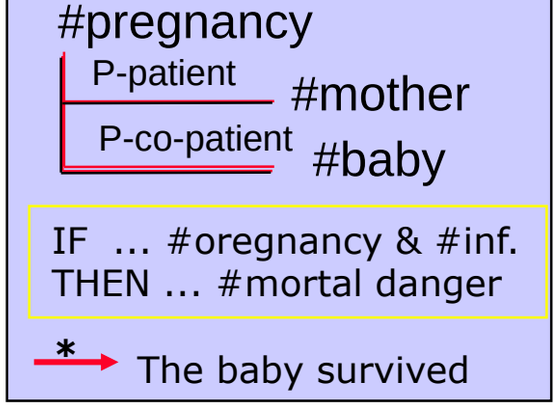
Architektur eines Biomedizinischen Textanalyse-Kernsystems

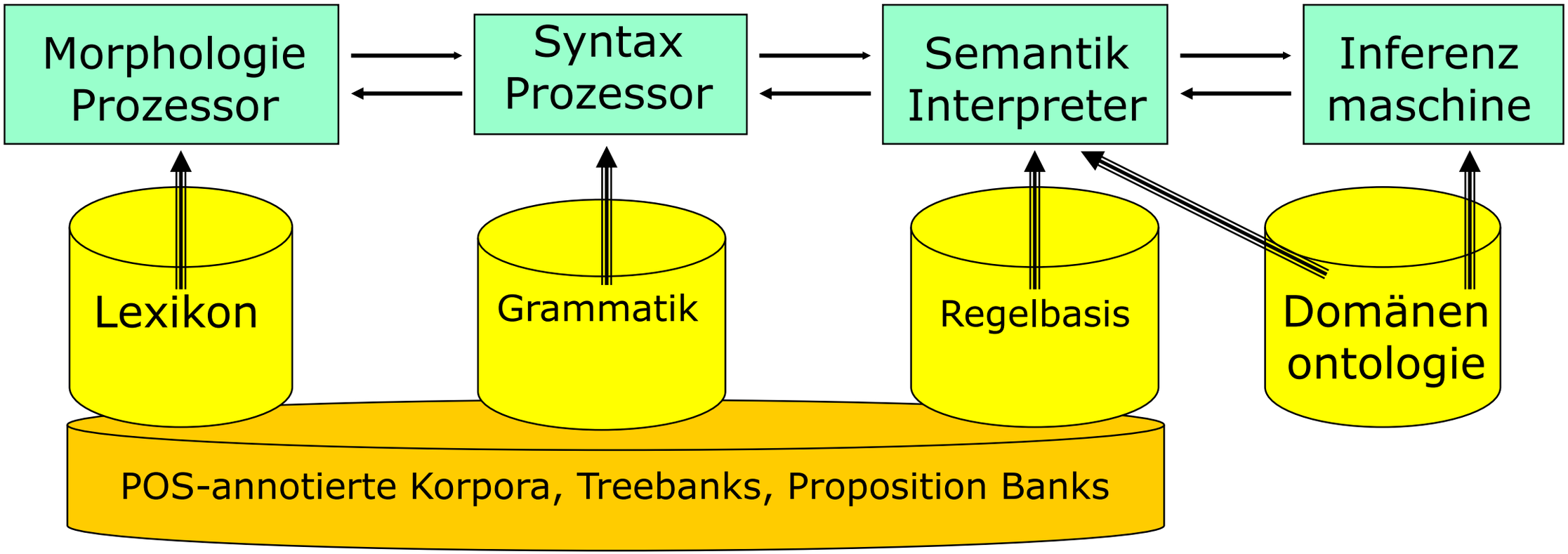
„A severe infection ended the pregnancy"

end + $ed_{PastTense}$









Methoden, Werkzeuge und Ressourcen

- Morphologiewerkzeuge (Stemmer)
- POS (part-of-speech) Tagger
- Chunker (NP), (shallow) Parser
- Lexika, Endliche Automaten, Grammatiken, Ontologien
- Namenserkenner (NE recognition)
- Große Textkorpora (annotiert, nichtannotiert)
- Machine learning – Verfahren, e.g. SVM
- Evaluationsstandards

Methoden, Werkzeuge und Ressourcen

- Morphologiewerkzeuge (Stemmer)
- POS (part-of-speech) Tagger
- Chunker (NP), (shallow) Parser
- Lexika, Endliche Automaten, Grammatiken, Ontologien
- Namenserkenner (NE recognition)
- Große Textkorpora (annotiert, nichtannotiert)
- Machine learning – Verfahren, e.g. SVM
- Evaluationsstandards

Paradigmenwechsel in der Computerlinguistik

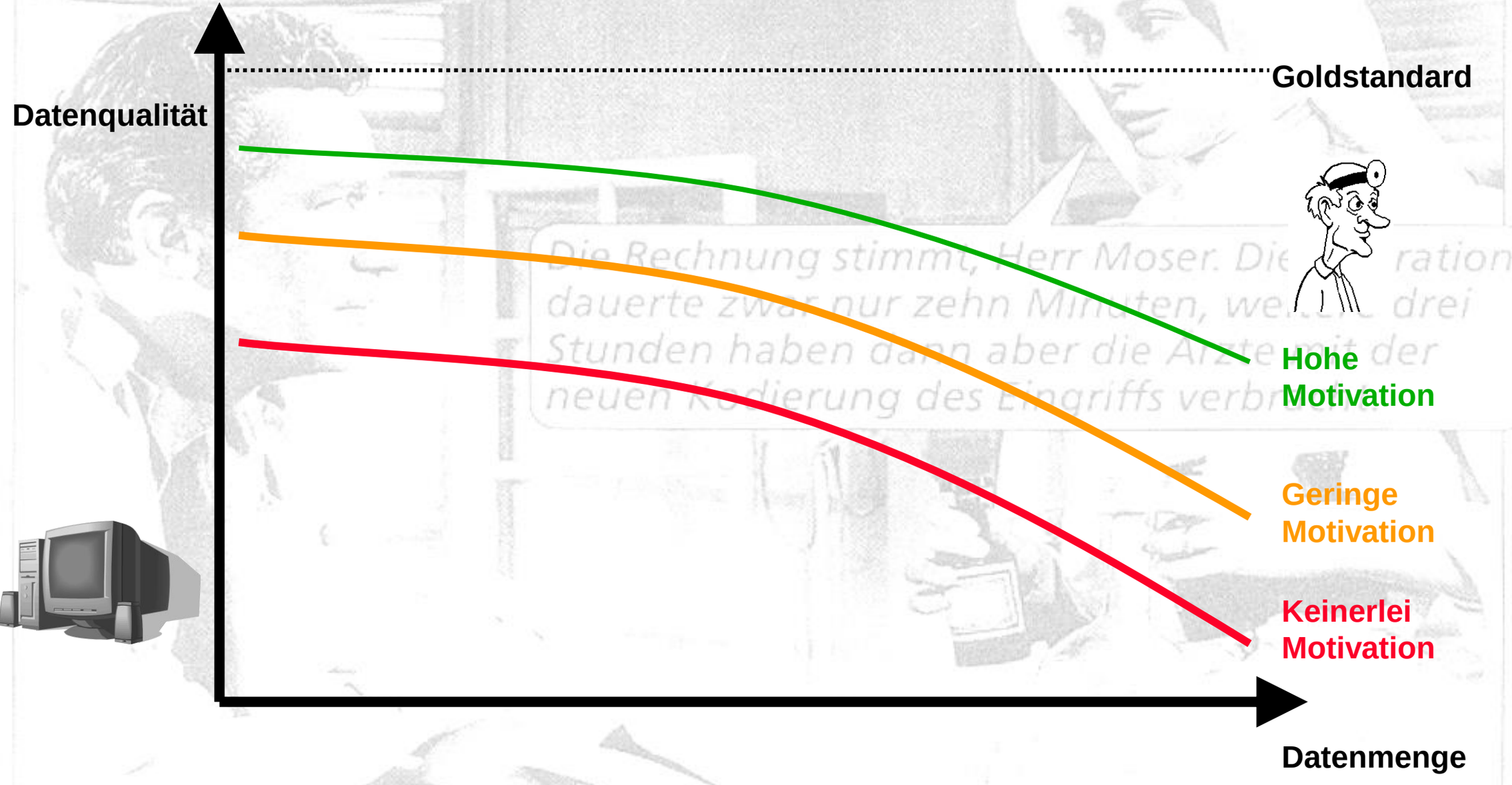
Regelbasiert, KI

Stochastisch, ML

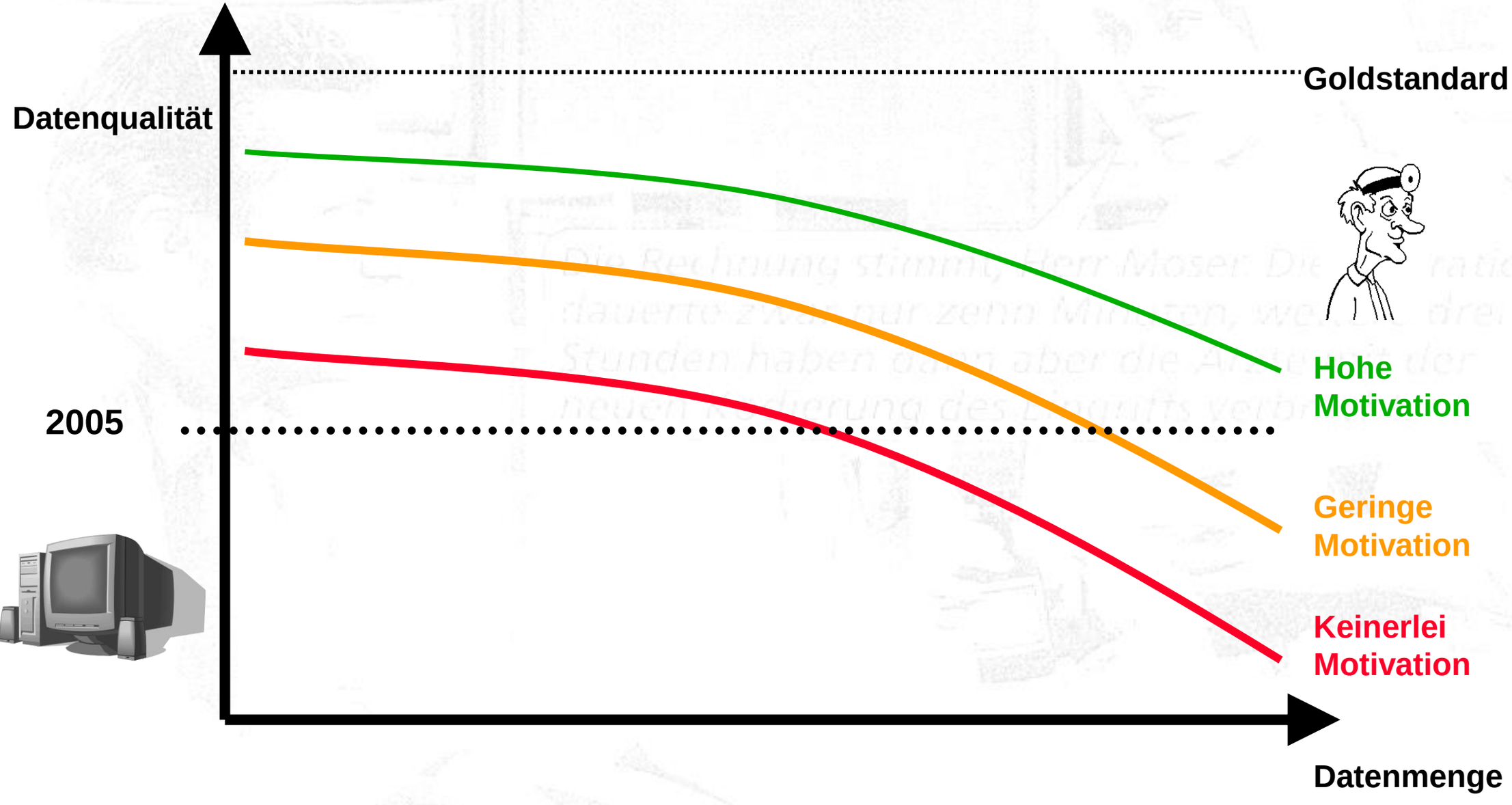
Herausforderung für medizinische Sprachverarbeitung

- Ambiguität
  - Lexikalisch: Bruch (Hernie) vs. Bruch (Fraktur)
  - Syntaktisch: z.B. Anbindung von PPs
    - extraction *[of the transplant [with a scalpel]]*
    - *[extraction]* {of the transplant} *[with a scalpel]*
  - Semantisch, z.B. Skopus von Quantoren, Negationen, Koordinationen, Gradaussagen
    - *each* sample showed an increased Ph value
- Komplexität, Berechenbarkeit, z.B.
  - Dependenzgrammatiken: NP-complete
  - Prädikatenlogik höherer Ordnung, Modallogik: unentscheidbar
- Kombination mit Ontologien und medizinischen Terminologiesystem
- Kombination von symbolischen und stochastischen Ansätzen

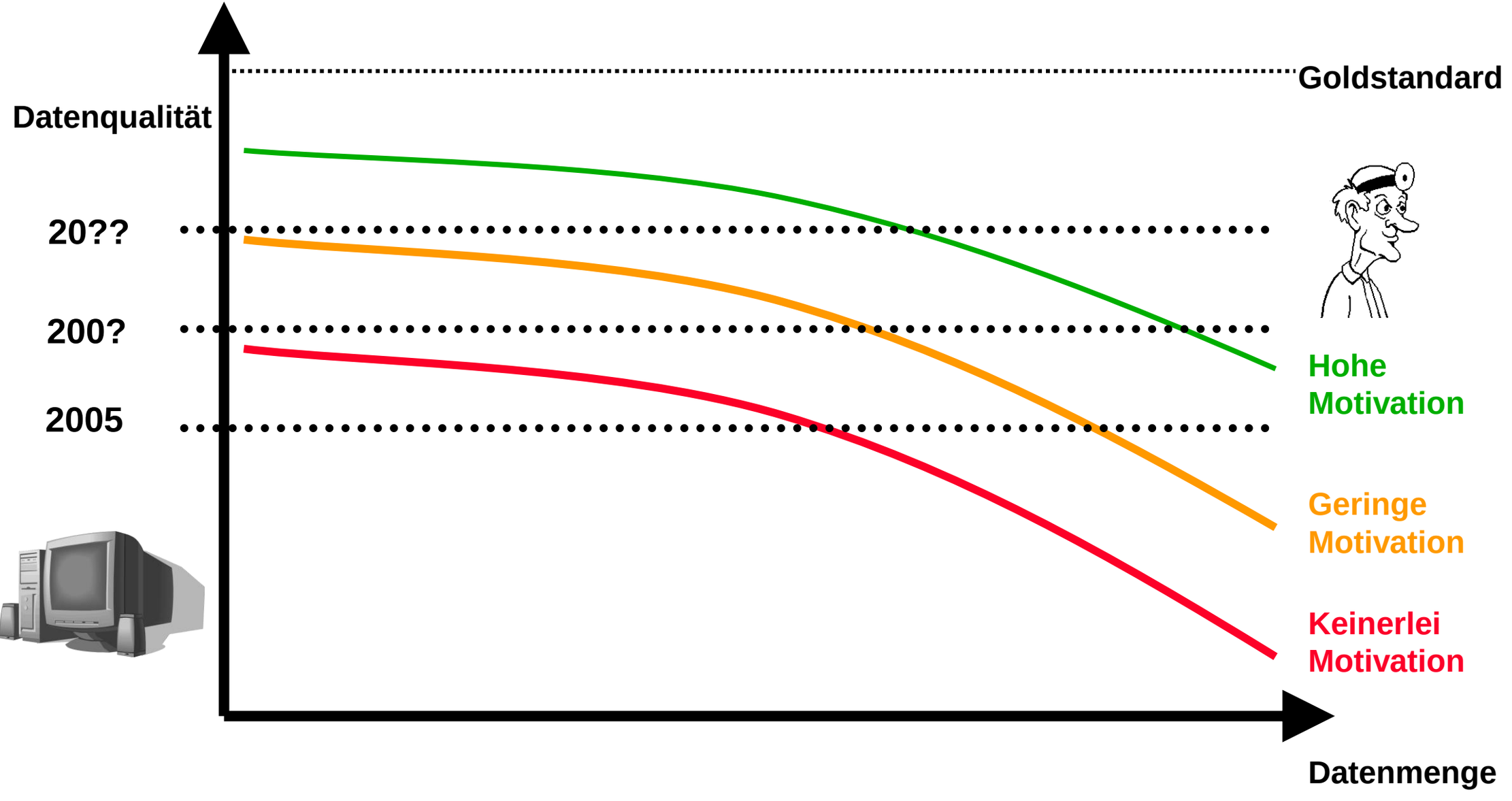
Ausblick: Menschliche vs. Maschinelle Sprachverarbeitung



Ausblick: Menschliche vs. Maschinelle Sprachverarbeitung



Ausblick: Menschliche vs. Maschinelle Sprachverarbeitung



Aktivitäten MI Freiburg

- EU 6th Framework :
  Network of Excellence “SemanticMining”
  (Semantic Interoperability and Data Mining in Biomedicine):
  2004 – 2006, 25 Partner
  www.semanticmining.org
- Gründung: AMIA Working Group Group KR-SIG
  “Formal (Bio)medical Knowledge Representation”, 2003
- Veranstalter: Workshop KR-MED 2004 in Whistler/Canada, Juni 2004
- Initiative BioTem (Zentrum für biomedizinisches Text Mining)
- Veranstalter: Konferenz SMBM 2005 (Semantic Mining in Biomedicine), in Cambridge UK, April 2005
- Wichtiger Partner: Udo Hahn, Computerlinguistik Universität Jena (bis 2004 in Freiburg)

Medical Terminology: Poor retrieval performance

Frequency of synonymous German Word forms in *Google* Searches

Spelling Variants Synonyms Inflections		
Kolonkarzinom	2070	1780
Colonkarzinom	248	135
Coloncarcinom	111	73
Colon-Ca	203	169
Kolon-Ca	66	46
Dickdarmkrebs	4000	3610
Dickdarmkarzinom	288	175
Dickdarmcarcinom	13	10
Kolonkarzinoms	471	253
Kolonkarzinome	275	139
Kolonkarzinomen	265	166

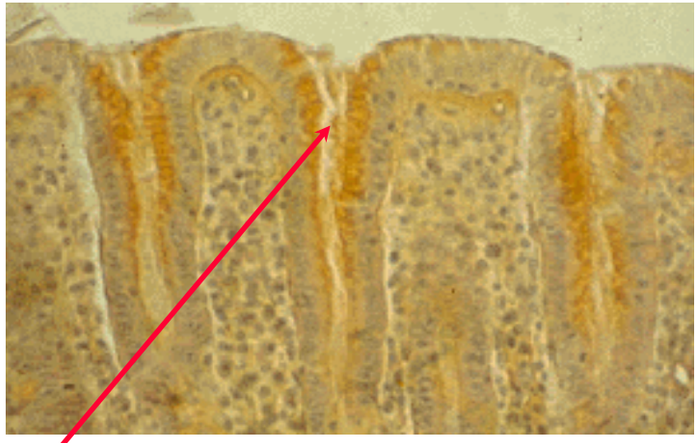
Number of Hits

Number of exclusive hits (no other form matches)

Neue Anwendungsszenarien

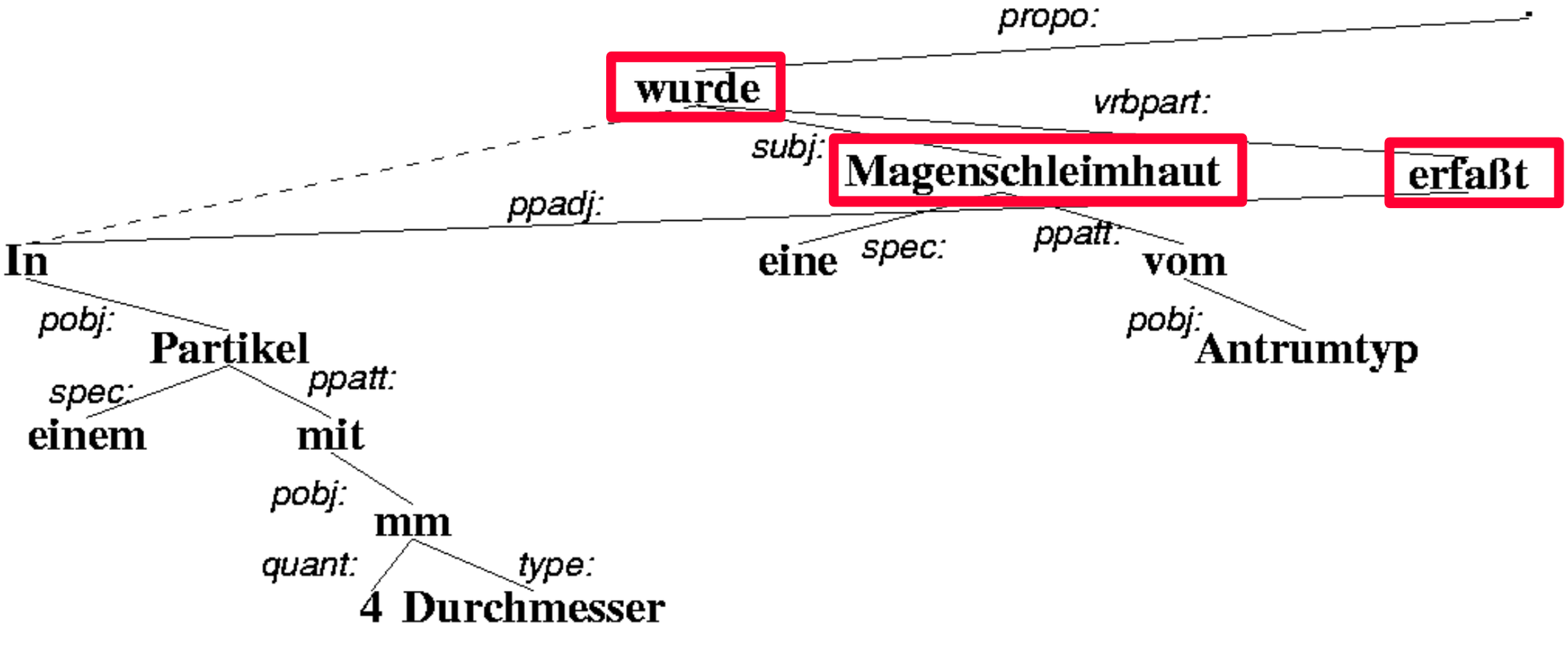
- The Semantic Electronic Health Record
- Named entity recognition challenged by the deluge of new proper names from the bio domain
- Use huge (Terabyte !) medical corpora (from all sources including anonymized EHR data) for the discovery of domain and linguistic knowledge
- Use content technologies to match genotype information (Bio-DBs) with phenotype information (EHR).

(1.) In einem Partikel mit 4 mm Durchmesser wurde eine Magenschleimhaut vom Antrumtyp erfaßt.



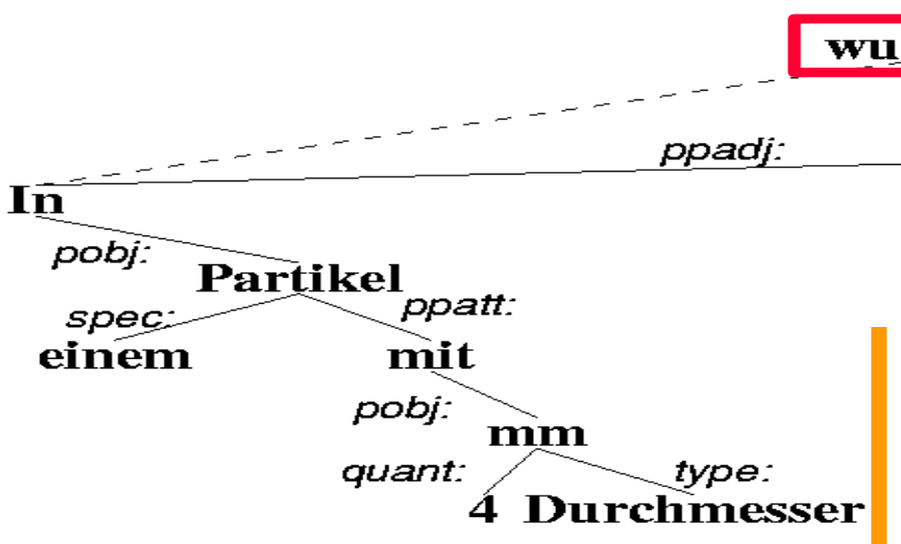
(2.) Das ödematöse Stroma wird massiv von Lymphozyten infiltriert.

In einem Partikel mit 4 mm Durchmesser wurde eine Magenschleimhaut vom Antrumtyp erfaßt.

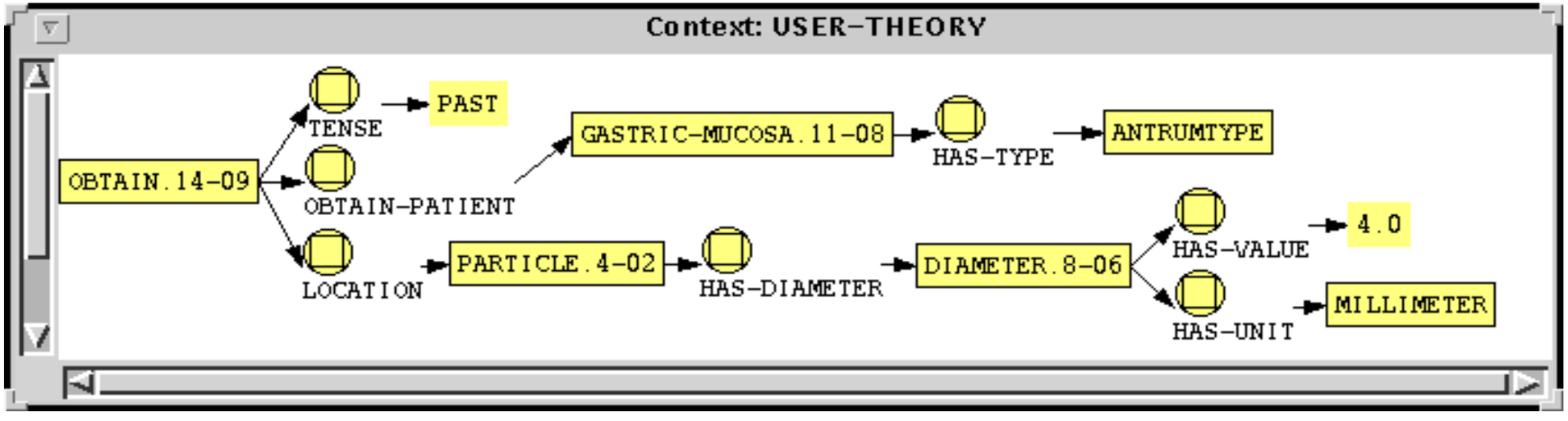


Semantische Interpretation

In einem Partikel mit 4 mm Durchmesser wurde eine Magenschleimhaut vom Antrumtyp erfaßt.

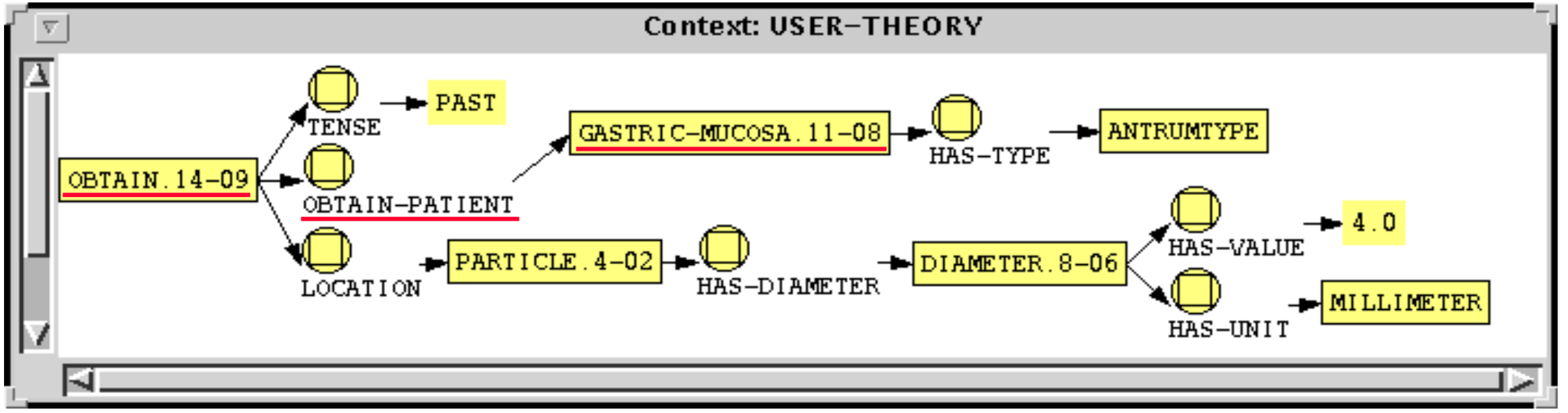


$$\exists\, x \in \text{PassivAux} : \text{Subj}(x, y) \wedge \text{Vrbpart}(x, z)$$
$$\Rightarrow \exists Con(z), Con(y) \in CONCEPTS :$$
$$Patient(Con(z), Con(y))$$



Konzept-Graph von Satz 1

In einem Partikel mit 4 mm Durchmesser wurde eine Magenschleimhaut vom Antrumtyp erfaßt.

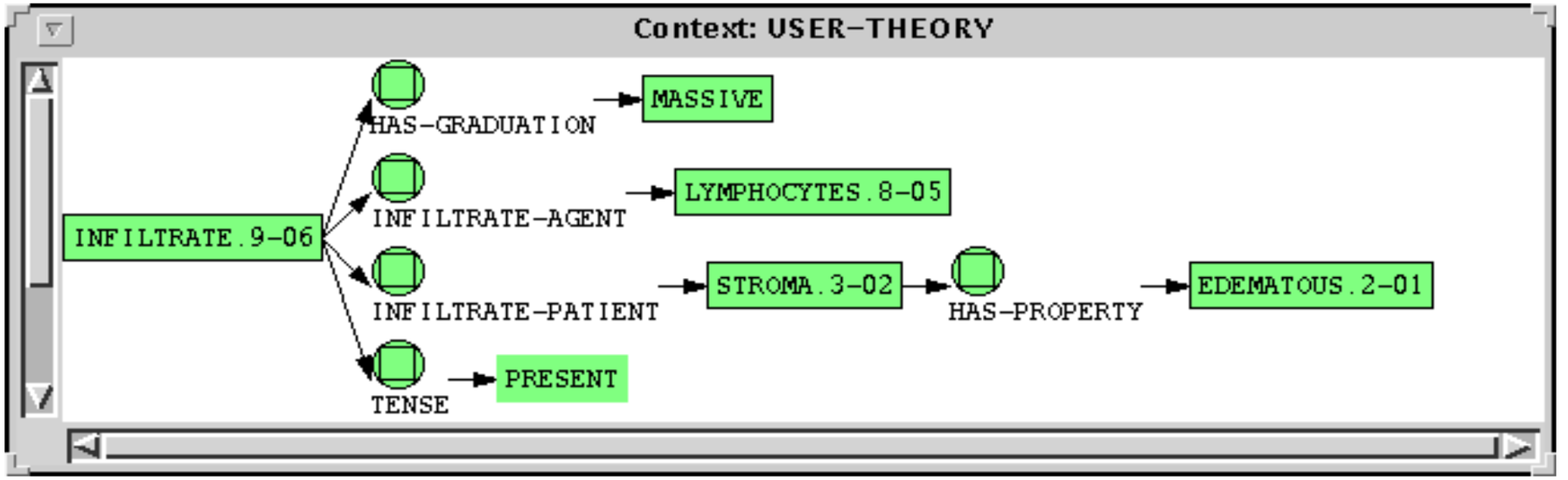


Semantische Interpretation:

- Berechnung der konzeptuellen Relationierungen
- “Normalisierung” des Passivs
- Korrekte Anbindung der Präpositionalphrasen (in, mit, vom)

Konzept-Graph von Satz 2

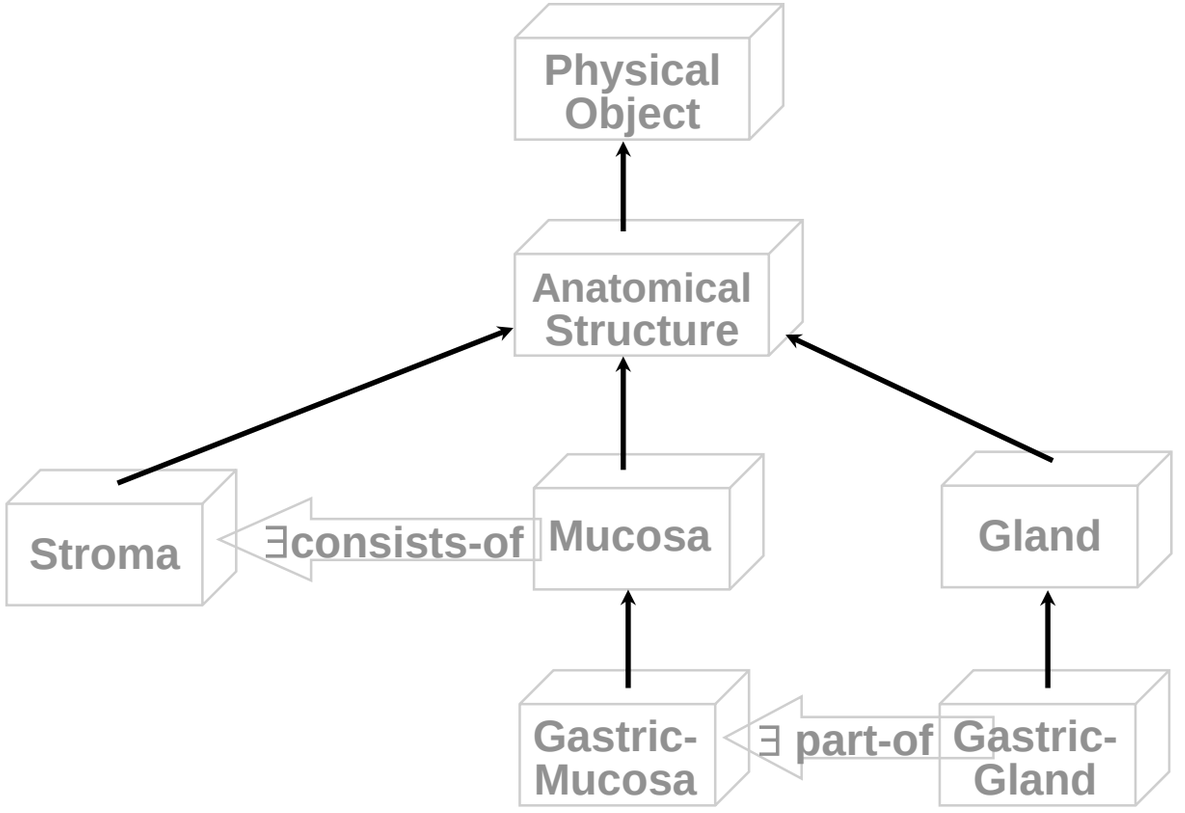
Das ödematöse Stroma wird massiv von Lymphozyten infiltriert.



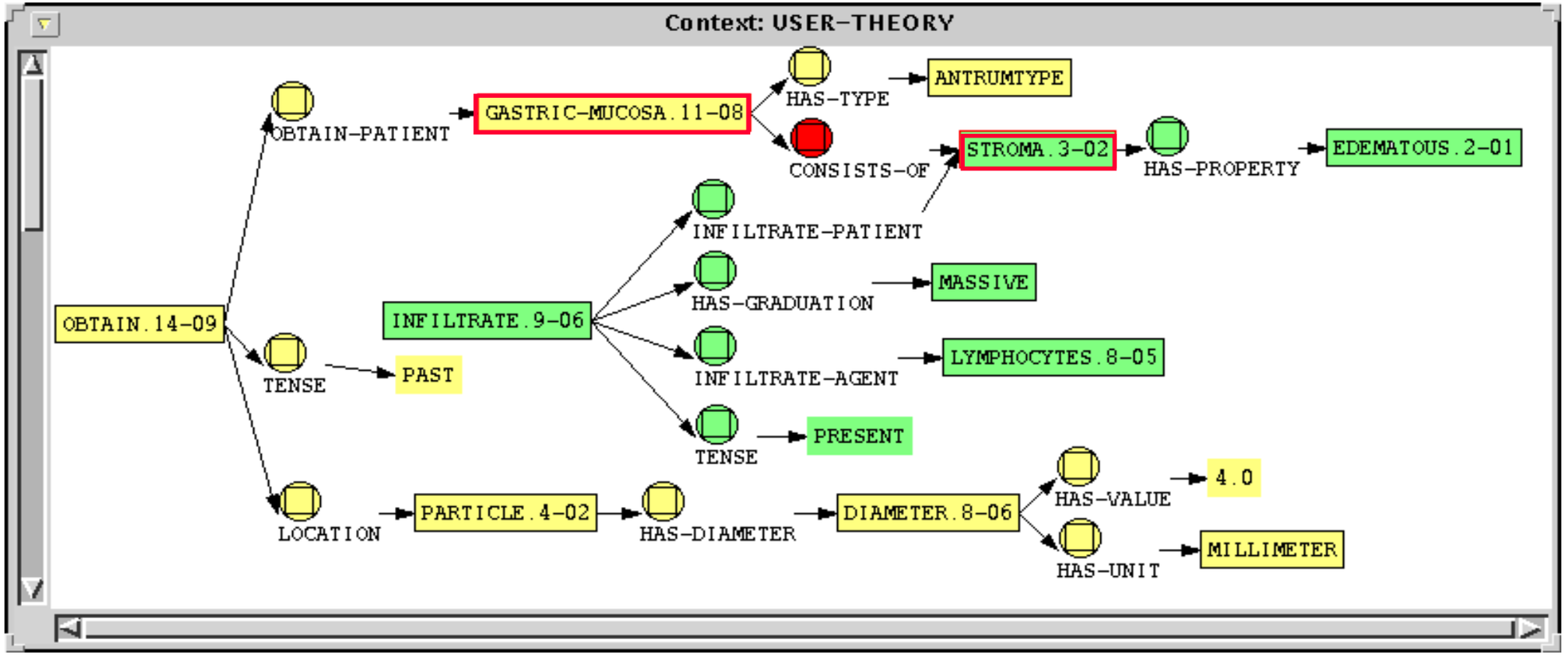
Ergebnis der satzorientierten Analyse:

- Pro Satz ein **isolierter** Konzeptgraph
- Ist dies eine adäquate Inhaltsrepräsentation des **Texts**?

Medizinisches Domänenwissen



Kombinierter Konzept-Graph von Satz 1 und 2



Formale Rekonstruktion des impliziten textuellen Bezugs:

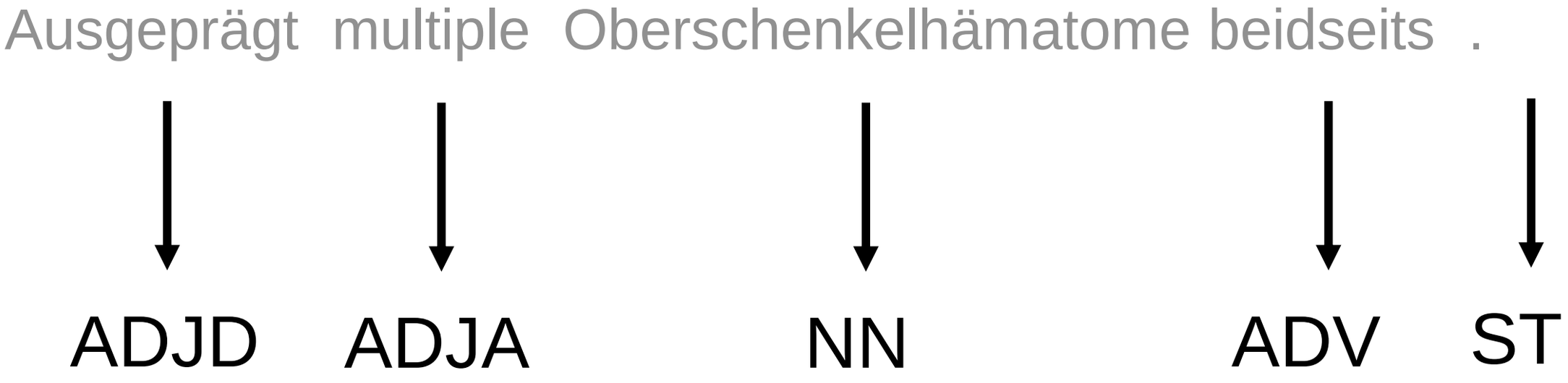
Konzeptgraph von Satz 1 (gelb) und Konzeptgraph von Satz 2 (grün) werden über die inferierte Rolle (rot) relationiert.

Stand der Kunst

- Umfassende Dokumentenkollektionen:
  - KIS, EPA, WWW, ...
- Computerlinguistische Engpässe
  - unvollständige (manuell erstellte) Lexika, Grammatiken, Domänenmodelle
- 1. Ausweg:
  - robuste Textanalyse mit unvollständigen Ressourcen
- 2. Ausweg:
  - automatisches Lernen sprachlichen Wissens (Vervollständigung linguistischer Ressourcen)

Zentrale sprachtechnologische Methode:

Part-of-Speech (POS) Tagging



Was aber bleibt (gleich) …?

- Analyse geschriebener Texte
- Fachsprache (IT, Medizin/Biologie)

Architektur eines Textanalysesysteme

end + ed$_{PastTense}$

ended

infection — pregnancy

a — severe — the

Ending

E-patient — Pregnancy

E-agent — Infection

I-degree — severe

Pregnancy

P-patient — Female

P-co-patient — Foetus

P-affected-by — Infection

IF ... Pregnancy & infect

THEN ... mortal danger

* → The baby survived