

Text Mining deutscher medizinischer Texte

i:DSem Workshop, 14.7.2017 @ Humboldt Universität zu Berlin 2017



**Stefan
Schulz**

Medical
University
of Graz
(Austria)



purl.org/steschu

Die Notwendigkeit der Unterscheidung zwischen Interface-Terminologien und Referenzterminologien

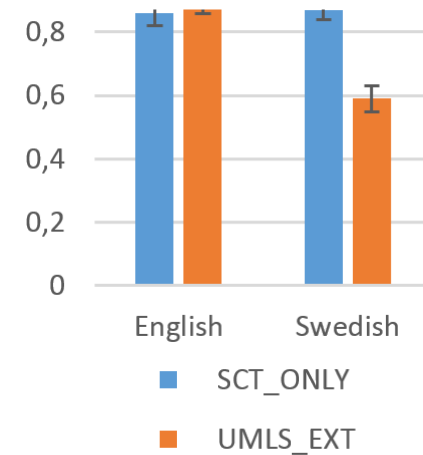
"Wie geeignet ist SNOMED CT als europäische Referenzterminologie?"

- Manuelle Annotation klinischer Texte
(Parallelkorpus) mit SNOMED CT vs.
UMLS-Extrakt
- Messung:
 - konzeptuelle Abdeckung
 - Term-Abdeckung

"Wie geeignet ist SNOMED CT als europäische Referenzterminologie?"

- Manuelle Annotation klinischer Texte (Parallelkorpus) mit SNOMED CT vs. UMLS-Extrakt
- Messung:
 - konzeptuelle Abdeckung
 - Term-Abdeckung

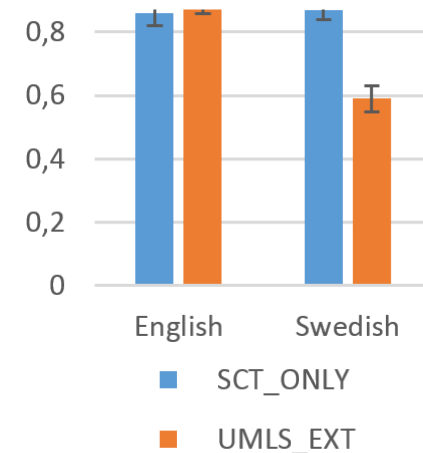
Concept coverage



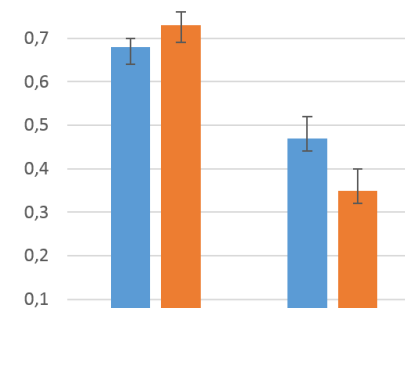
"Wie geeignet ist SNOMED CT als europäische Referenzterminologie?"

- Manuelle Annotation klinischer Texte (Parallelkorpus) mit SNOMED CT vs. UMLS-Extrakt
- Messung:
 - konzeptuelle Abdeckung
 - Term-Abdeckung
- Unterschiede SNOMED CT Schwedisch – Englisch
 - Schwedisch: ein (Vorzugs-)Term pro Konzept
 - Englisch: durchschnittlich 2,3 Terme pro Konzept (Vorzugsterme, Synonyme)

Concept coverage



Term coverage



Beispiel: Terme in PubMed [tiab]

Vorzugsterm (SNOMED CT)	Anzahl	Synonym	Anzahl
Primary malignant neoplasm of lung	0	Lung cancer Bronchial carcinoma	120682 3452
Cerebrovascular accident	3819	Stroke	191559
Block dissection of cervical lymph nodes	1	Neck dissection	7512
Electrocardiographic procedure	1	Electrocardiogram ECG	33670 55120
Backache	3489	Back pain	38132
Capillary blood specimens	32	Capillary blood samples	574

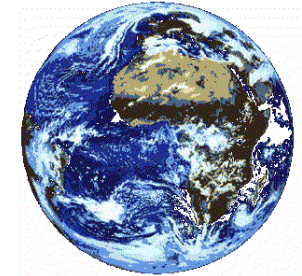
Beispiel: Terme in Kardiologie-Arztbriefen

Vorzugsterm (ICD, OPS)	Anzahl	Synonym	Anzahl
Aortenklappenstenose	3749	Aortenstenose	3126
Hirnfarkt	7	Schlaganfall	65
Elektrokardiogramm	0	EKG	12208
Koronare Herzerkrankung	331	KHK	18455
Nicht-ST-Hebungsinfarkt	498	NSTEMI	3839
Magnetresonanztomographie	2	NMR	17

Zwei Aspekte von Terminologien

■ Normativ

- Codes + Labels (Namen) bezeichnen wohldefinierte Gegenstände eines Diskursbereichs.
- "sprechende" Labels, z.B. "*Primary malignant neoplasm of lung (disorder)*".
- Erklärende Texte (*scope notes*) zur zusätzlichen Präzisierung der Repräsentationseinheit ("Konzept")
- Bedeutung durch formale Beschreibungen expliziert: → formale Ontologie



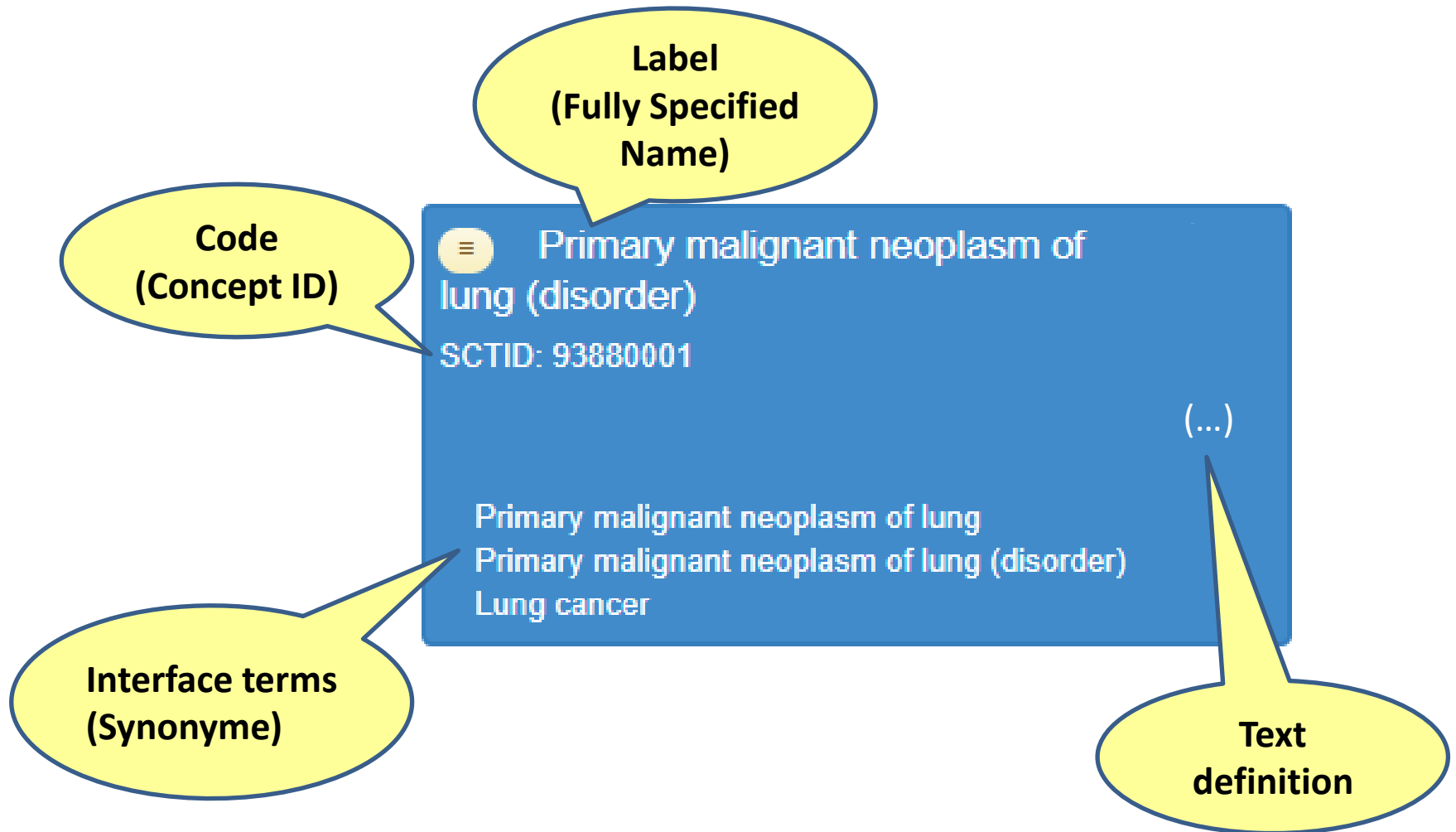
■ Deskriptiv

- Tatsächlicher Sprachgebrauch: Lexikon
- Erweiterung zu einem Thesaurus durch semantische Relationen (Synonymie, Hypernymie,...)
- Gängige Terminologiesysteme decken diese beiden Aspekte in unterschiedlichen Maß und meist unsystematisch ab



Beispiel SNOMED CT

SNOMED CT als Terminologie



SNOMED CT als formale Ontologie

Parents

- ≡ Malignant tumor of lung (disorder)
- ≡ Primary malignant neoplasm of intrathoracic organs (disorder)
- ≡ Primary malignant neoplasm of respiratory tract (disorder)

**Code
(Concept ID)**

≡ Primary malignant neoplasm of lung (disorder)
SCTID: 93880001

**Logische Axiome
(DL)**

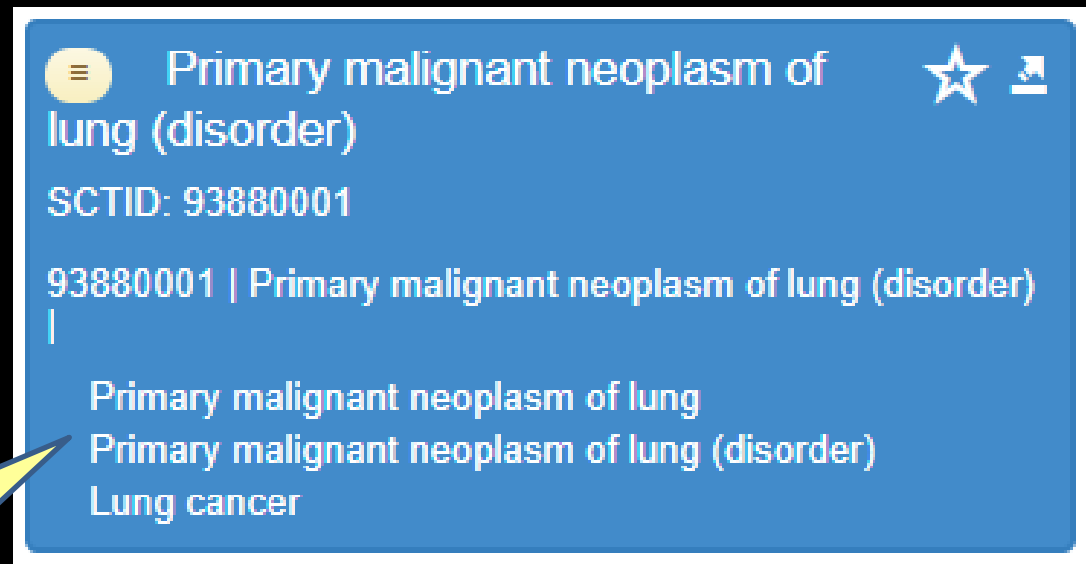
Finding site → Lung structure
Associated morphology → Malignant neoplasm, primary

Children (32)

- ▼ ● Carcinoma of lung parenchyma (disorder)
 - ≡ Carcinoma in situ of lung parenchyma (disorder)
 - ▼ ≡ Large cell carcinoma of lung (disorder)
 - ≡ Giant cell carcinoma of lung (disorder)
 - ● Large cell carcinoma of lung, TNM stage 1 (disorder)
 - ● Large cell carcinoma of lung, TNM stage 2 (disorder)
 - ● Large cell carcinoma of lung, TNM stage 3 (disorder)
 - ● Large cell carcinoma of lung, TNM stage 4 (disorder)
 - ▼ ≡ Small cell carcinoma of lung (disorder)
 - ● Extensive stage primary small cell carcinoma of lung (disorder)
 - ● Oat cell carcinoma of lung (disorder)

**Taxonomie
(is-a-
Hierarchie)**

Interface-Terme → Interface-Terminologie



The screenshot shows a blue rectangular box with a white border. Inside the box, at the top left, is a yellow circle containing three horizontal lines. To its right is the text "Primary malignant neoplasm of lung (disorder)" in white. To the right of this text are a white star icon and a white icon of a document with a magnifying glass. Below this is the text "SCTID: 93880001" in white. Further down is the text "93880001 | Primary malignant neoplasm of lung (disorder)" in white, followed by a vertical line. Below the line are three lines of text in white: "Primary malignant neoplasm of lung", "Primary malignant neoplasm of lung (disorder)", and "Lung cancer".

Primary malignant neoplasm of lung (disorder)

SCTID: 93880001

93880001 | Primary malignant neoplasm of lung (disorder)

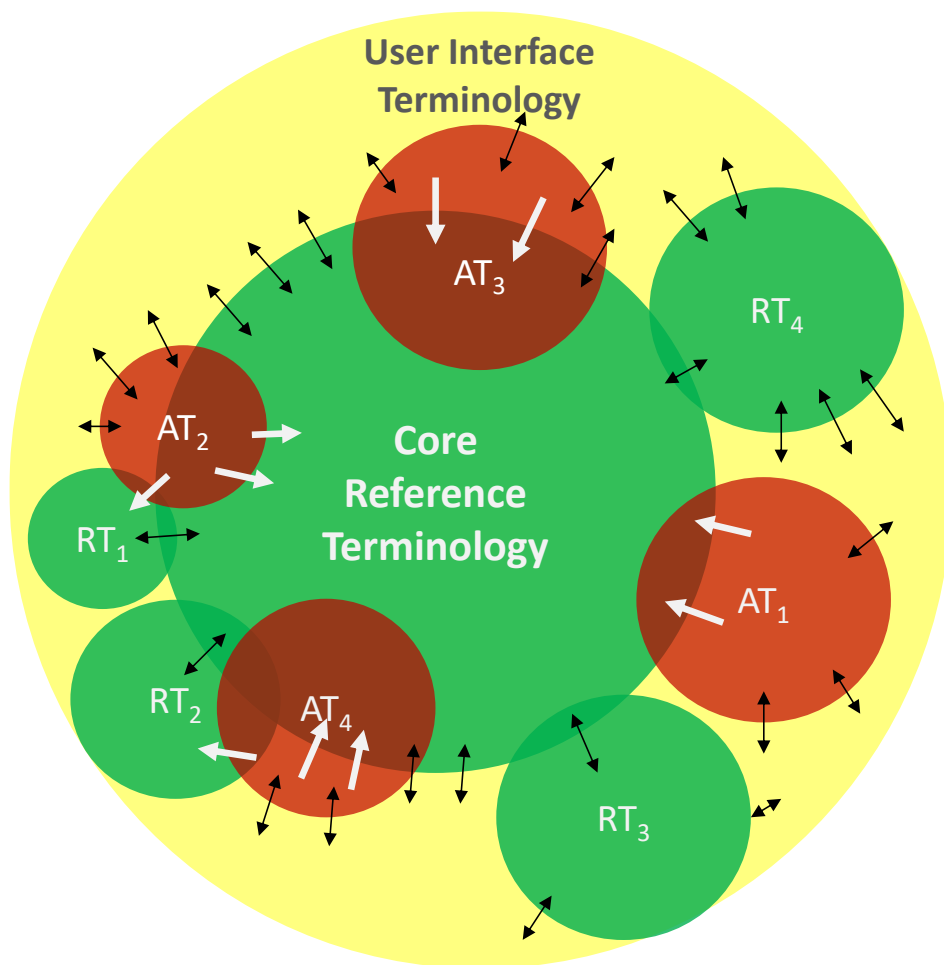
|

Primary malignant neoplasm of lung

Primary malignant neoplasm of lung (disorder)

Lung cancer

Interface terms
(Synonyme)



"SNOMED CT should be part of an ecosystem of terminologies, including international aggregation terminologies (e.g., the WHO Family of Classifications), and user interface terminologies, which address multilingualism in Europe and clinical communication with multidisciplinary professional language and lay language"

AT: aggregation terminology, RT: reference terminology

Unterscheidung Referenzterminologie - Interfaceterminologie

- Referenzterminologien:
 - Primär sprachunabhängige Repräsentationseinheiten (Konzepte):
Eigenschaften der Objekte, die von diesen denotiert werden
 - Maximal eindeutige Labels, unterstützt durch textliche und / oder
formale (ontologische) Definitionen
- Interfaceterminologien:
 - Sammlungen von sprachlichen Ausdrücken, die in schriftlicher und
mündlicher Kommunikation verwendet werden.
 - Fundierung durch Verknüpfung zu Referenzterminologien
 - Interfaceterme sind häufig ambig. Ihre Bedeutung hängt ab von
Sprachbenutzern, Sprachregister, sowie thematischen, dialektalen
und zeitlichen Kontexten
- Separate Erstellung / Pflege von Referenzterminologien und
Interfaceterminologien

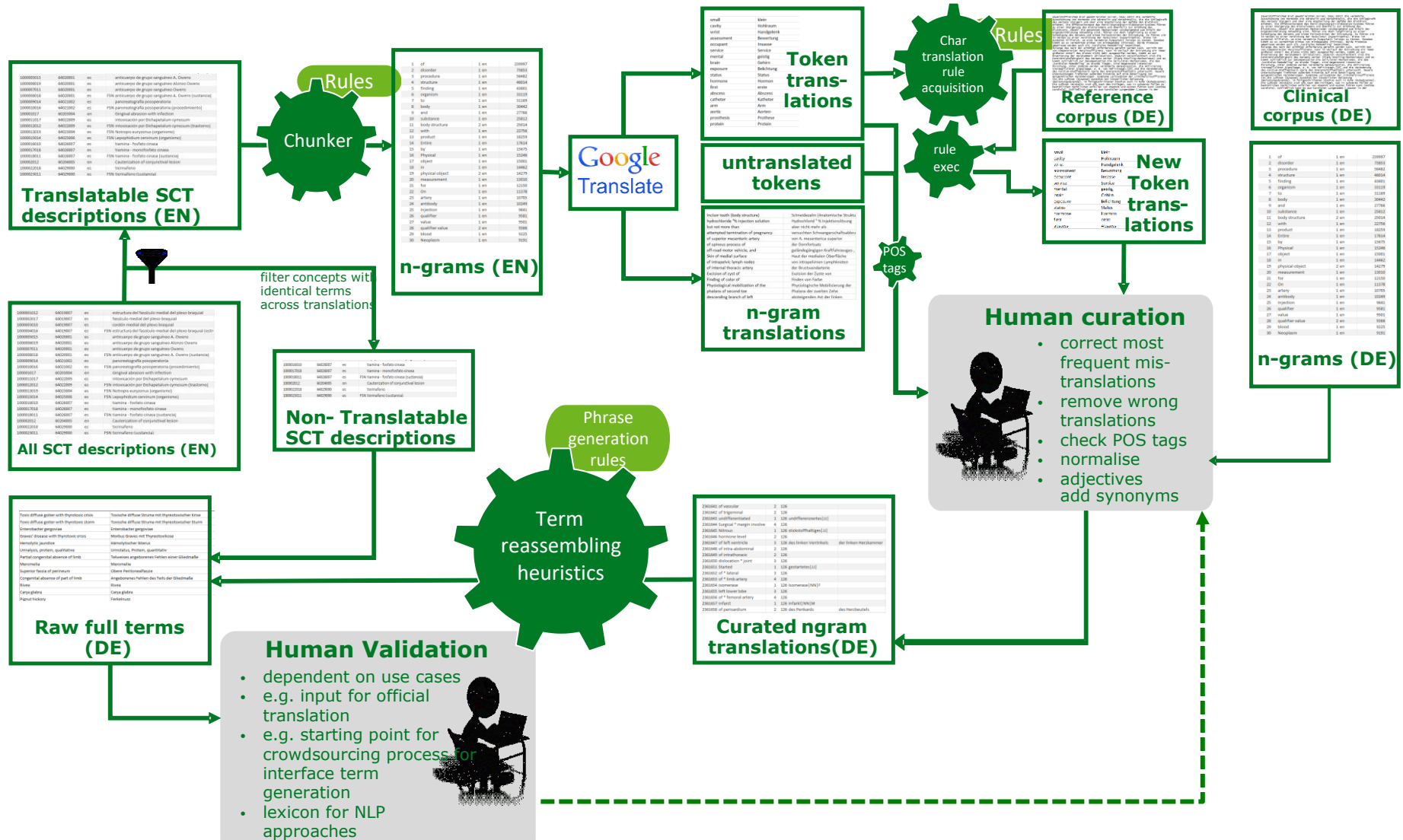
Bedeutung für manuelle Annotation und klinisches Text Mining

- Interface-Terme kommen vor:
 - als Synonyme in Referenzterminologien
 - als separate Interface-Terminologie, gemappt auf Referenzterminologie
- Terminologie-Lokalisierung via Label-Übersetzung
 - hoher Aufwand (Erfahrungen Dänemark + Schweden)
 - schlechte Benutzerakzeptanz*
 - geringer Recall beim Text Mining
- Besser: Akquisition von Interface-Termen und Verlinkung mit Inhalten der Referenzterminologie
- Hypothese: Bottom-up / Crowdsourcing
- Durchgängiges Problem: Ambiguitäten, Kurzformen

MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (I)

- MUG-GIT (*Medical University of Graz – German Interface Terminology*) zur maschinellen Annotation deutscher Kliniktexte → semantische Extrakte für Projekt IICCAB*
- limitierte Ressourcen, inkrementelles Vorgehen
- Modularisierung (Zerlegung in N-Gramme, meist NPs und PPs), Editieren eines abgeleiteten Kernvokabular, motiviert durch hoch repetitive Teilphrasen, z.B.
 - "Magnetic resonance imaging" in 627 SNOMED -Termen
 - "second degree burn" in 166 SNOMED-Termen
- Priorisierung nach Häufigkeit:
 - Manuelle Revision der NP-Liste
 - Anreicherung durch Terme aus anderen deutschen Terminologien und klinischen Corpora

MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (II)



MUG-GIT: Erstellung einer deutschen Interface-terminologie für SNOMED CT (III)

- Inkaufnahme "schlechter" Übersetzungen und ausufernde Permutationen bei langen SNOMED-CT-Termen
- Kernvokabular:
 - gepflegt durch zwei Medizinstudentinnen unter Aufsicht
 - Richtlinien, z.B. keine Schreibvarianten (c/k/z - Problem), Akronyme nur im Kontext (kein Eintrag für "CT", aber für "Schädel-CT"), keine Übersetzung unvollständiger N-Gramme, Umgang mit elliptischen Ausdrücken ("Verstopfung")
- Derzeitiger Stand:
 - 1,87 Millionen Interface-Terme
 - Automatisch generiert aus einem Kernvokabular mit deutschen 92,500 N-Grammen, verknüpft mit 85,400 englischen N-Grammen
 - Benchmark: MEDLINE extrahierter Parallelkorpus, bestehend aus Publikationstiteln in deutscher [tt] und englischer Sprache [ti]. Aktuelle Term-Abdeckung 33,1% für Deutsch gegenüber 55,4%

ngram - Kernvokabular

vaginal	1	1478	vaginales JJ	Scheiden-	
fluoroscopic guidance	2	1477	Durchleuchtungskontrolle NN F		
disc	1	1476	Scheibe NN F		
lower limb	2	1473	unteres JJ Extremität NN F	Bein NN N	
brain	1	1468	Gehirn NN N	Hirn NN N	Encephalon NN N
preparation	1	1464	Zubereitung NN F	Aufbereitung NN F	Präparation NN F
method	1	1463	Verfahren NN N	Methode NN F	
of bone	2	1462	des Knochens	_Knochen_	
Red	1	1455	rotes JJ		
Monitoring	1	1453	Überwachung NN F	Monitoring NN N	
Computed	1	1453	berechnetes JJ	Computer-	
phalanx	1	1449	Phalanx NN F		
subsp.	1	1449			
anastomosis	1	1447	Anastomose NN F	Anastomosierung NN F	
vessel	1	1446	Blutgefäß NN N	Gefäß NN N	
Computed tomography	2	1443	Computertomographie NN F		
uterus	1	1436	Uterus NN M	Gebärmutter NN F	
difficulty	1	1432	Schwierigkeit NN F		
elbow	1	1429	Ellbogen NN M	Cubitus NN M	Ellbogengelenk NN N
high	1	1429	hohes JJ		
food	1	1423	Lebensmittel NN N	Speise NN F	Nahrungsmittel NN N
Observation	1	1423	Beobachtung NN F		
using fluoroscopic	2	1422			
unable	1	1421	unfähiges JJ		
Peripheral	1	1419	peripheres JJ		
unable to	2	1418	unfähig zu		
Vascular	1	1417	vasculäres JJ	Gefäß-	
using fluoroscopic guidance	3	1416	mit Durchleuchtungskontrolle		
Benign neoplasm	2	1415	gutartiges JJ Neubildung NN F	gutartiges JJ Neoplasie NN F	benignes JJ Neoplasie NN F

Automatische generierte Interfaceterminologie

20170315_240011_002	126952004	Neoplasm of brain	Gehirnneubildung
20170315_240011_003	126952004	Neoplasm of brain	Neubildung des Hirns
20170315_240011_004	126952004	Neoplasm of brain	Hirnneubildung
20170315_240011_005	126952004	Neoplasm of brain	Neoplasie des Gehirns
20170315_240011_006	126952004	Neoplasm of brain	Gehirnneoplasie
20170315_240011_007	126952004	Neoplasm of brain	Neoplasie des Hirns
20170315_240011_008	126952004	Neoplasm of brain	Hirnneoplasie
20170315_240011_009	126952004	Neoplasm of brain	Neoplasma des Gehirns
20170315_240011_010	126952004	Neoplasm of brain	Gehirnneoplasma
20170315_240011_011	126952004	Neoplasm of brain	Neoplasma des Hirns
20170315_240011_012	126952004	Neoplasm of brain	Hirnneoplasma
20170315_241010_001	126953009	Neoplasm of cerebrum	Neubildung des Großhirns
20170315_241010_002	126953009	Neoplasm of cerebrum	Neoplasie des Großhirns
20170315_241010_003	126953009	Neoplasm of cerebrum	Neoplasma des Großhirns
20170315_242015_001	126954003	Neoplasm of frontal lobe	Neubildung des Frontallappens
20170315_242015_002	126954003	Neoplasm of frontal lobe	Neubildung des Lobus frontalis
20170315_242015_003	126954003	Neoplasm of frontal lobe	Neoplasie des Frontallappens
20170315_242015_004	126954003	Neoplasm of frontal lobe	Neoplasie des Lobus frontalis
20170315_242015_005	126954003	Neoplasm of frontal lobe	Neoplasma des Frontallappens
20170315_242015_006	126954003	Neoplasm of frontal lobe	Neoplasma des Lobus frontalis
20170315_243013_001	126955002	Neoplasm of temporal lobe	Neubildung des Temporallappens
20170315_243013_002	126955002	Neoplasm of temporal lobe	Neubildung des Lobus temporalis
20170315_243013_003	126955002	Neoplasm of temporal lobe	Neoplasie des Temporallappens
20170315_243013_004	126955002	Neoplasm of temporal lobe	Neoplasie des Lobus temporalis
20170315_243013_005	126955002	Neoplasm of temporal lobe	Neoplasma des Temporallappens

Ko-operative Entwicklung einer deutschen Interface-Terminologie

Ko-operative Entwicklung einer deutschen Interface-Terminologie

- Günstige Rahmenbedingungen
 - Datenintegration / Sekundärnutzung / semantische Suche: Thema in geförderten Großprojekten (D: BMBF-MI, A: CBmed, ELGA)
 - Wachsendes Interesse an internationalen Terminologien (SNOMED CT, LOINC, RadLex...) und Ontologien (GO, HPO, ...)
- Synergieeffekte vs. Ressourcenverschwendung
- Idee: Crowdsourcing-Plattform für Entwicklung deutschsprachigen Interface-Terminologien: GIT-CP

Erste Spezifikationen für GIT-CP

- Web-basierte Crowdsourcing-Plattform
- Registrierung als User → Commitment zu Kooperation
- Zentrales Datenelement:
Mapping Interface Term – Externer Code
"DM" - 81827009 / *Diameter (qualifier value)*
- Wichtige Attribute:
 - Ersteller, Erstellungsart, Datum, klinisches Fachgebiet, Nutzergruppe
Max Muster, manuell, 20170803, Dermatologie Graz, Ärzte
 - Beispielannotation, z.B.
"ein 3 cm im DM haltender Tumor"
 - Validierung / Kommentierung durch andere User
Sven Piefke, 20180912, ★★★★★
„Hä?? Beispiel unverständlich!“

GIT-CP – Offene Fragen

- Technisch
 - Versionierung (GIT – Zielsysteme)
 - Schnittstellen zu lokalen Annotationsplattformen
 - Intelligente Tools (z.B. recommender services)
- Rechtlich / Organisatorisch
 - Koordination
 - Nachhaltige Finanzierung
 - Qualitätssicherung
 - Eigentumsrechte
 - Verwertung
 - Datenschutz

Fazit

- Text Mining deutscher medizinischer Texte benötigt Interfaceterminologien, die den alltäglichen Sprachgebrauch abbilden
- Die Verlinkung von Interfaceterminologien mit Referenzterminologien hat Priorität gegenüber der Übersetzung von Referenzterminologien
- Terminologiemanagement
 - Referenzterminologien: top-down, zentralisiert
 - Interfaceterminologien: bottom-up, dezentral
- Zeit ist reif für die kooperative, verteilte Erstellung einer medizinischen Interfaceterminologie für den deutschsprachigen Raum

Text Mining deutscher medizinischer Texte

i:DSem Workshop, 14.7.2017 @ Humboldt Universität zu Berlin 2017



**Stefan
Schulz**

Medical
University
of Graz
(Austria)



purl.org/steschu

Kontakt:

stefan.schulz@medunigraz.at