Stefan Schulz
Medizinische Universität Graz

purl.org/steschu

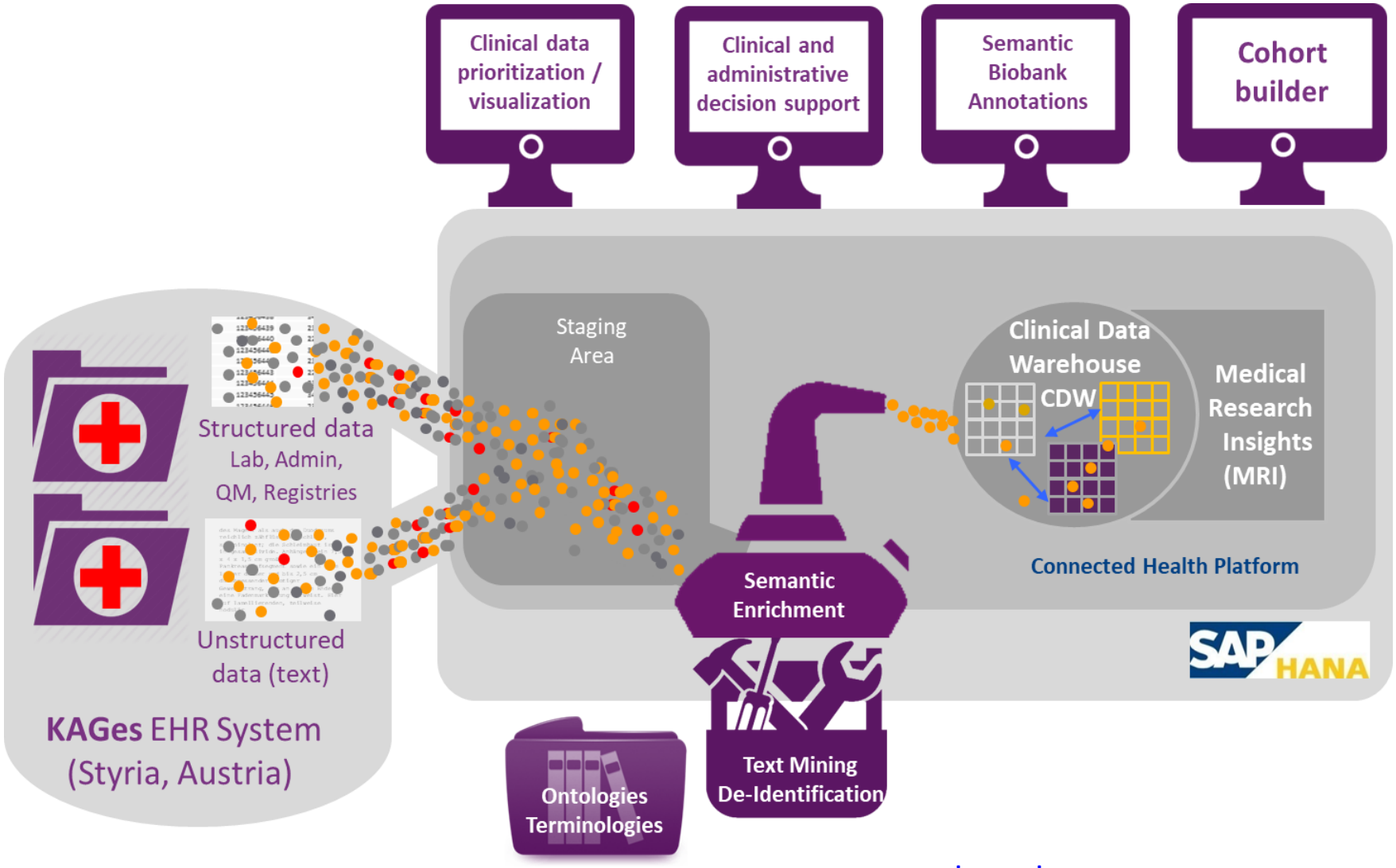
MEDIZIN INFORMATIK INITIATIVE

Workshop „Datenqualität“
TMF, Berlin, 03.05.2018

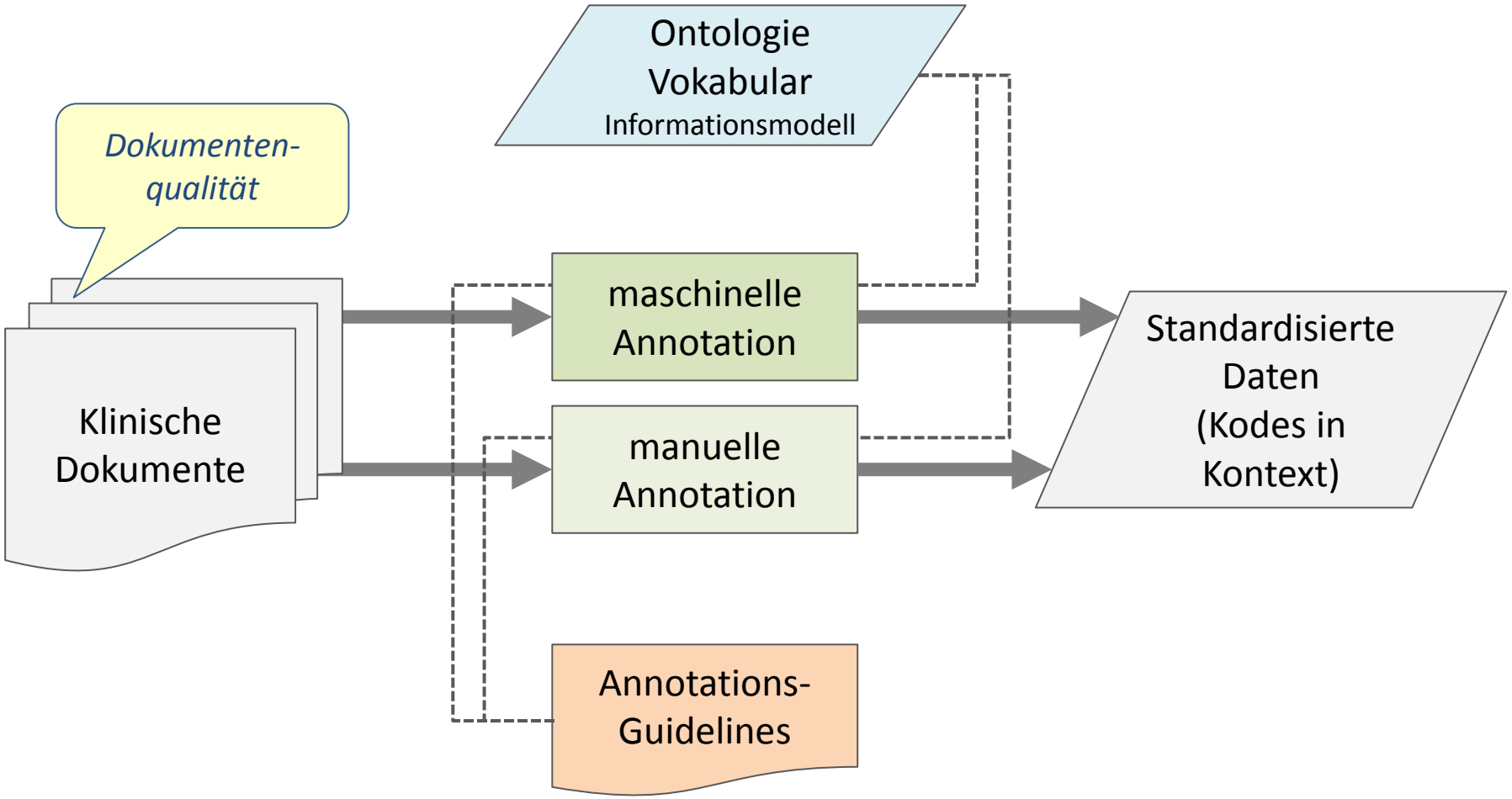
Natural Language Processing und Ontologie-Mapping Möglichkeiten und Grenzen bezüglich der Datenqualität

Kontext: CBmed Graz - Projekt

"Digital biomarkers for precision medicine"



NLP und Ontologie-Mapping: Prozesse, Ressourcen, Daten



Dokumentenqualität

Adip. Pat. mit DM Typ2	Adipöse Patientin mit Diabetes Mellitus Typ 2
St. p. TE eines exulc. sek.knot. SSM li US dors. Level IV	Zustand nach Totalexzision eines exulzerierenden, sekundär knotigen, superfiziell spreitenden Melanoms, Dorsalseite des linken Unterschenkels, Level 4
2,4 mm Tumor DM	2,4 mm Tumordurchmesser
Sentinnel LK ing. li. tumorfr.	Sentinel-Lymphknoten linke Leiste tumorfrei
Gepl. NTx bei term. NINS	Geplante Nierentransplantation bei terminaler Niereninsuffizienz
Euthyorx 75 1-0-0	Euthyrox (Levothyroxin-Na 75µg) 1-0-0

Dokumentenqualität

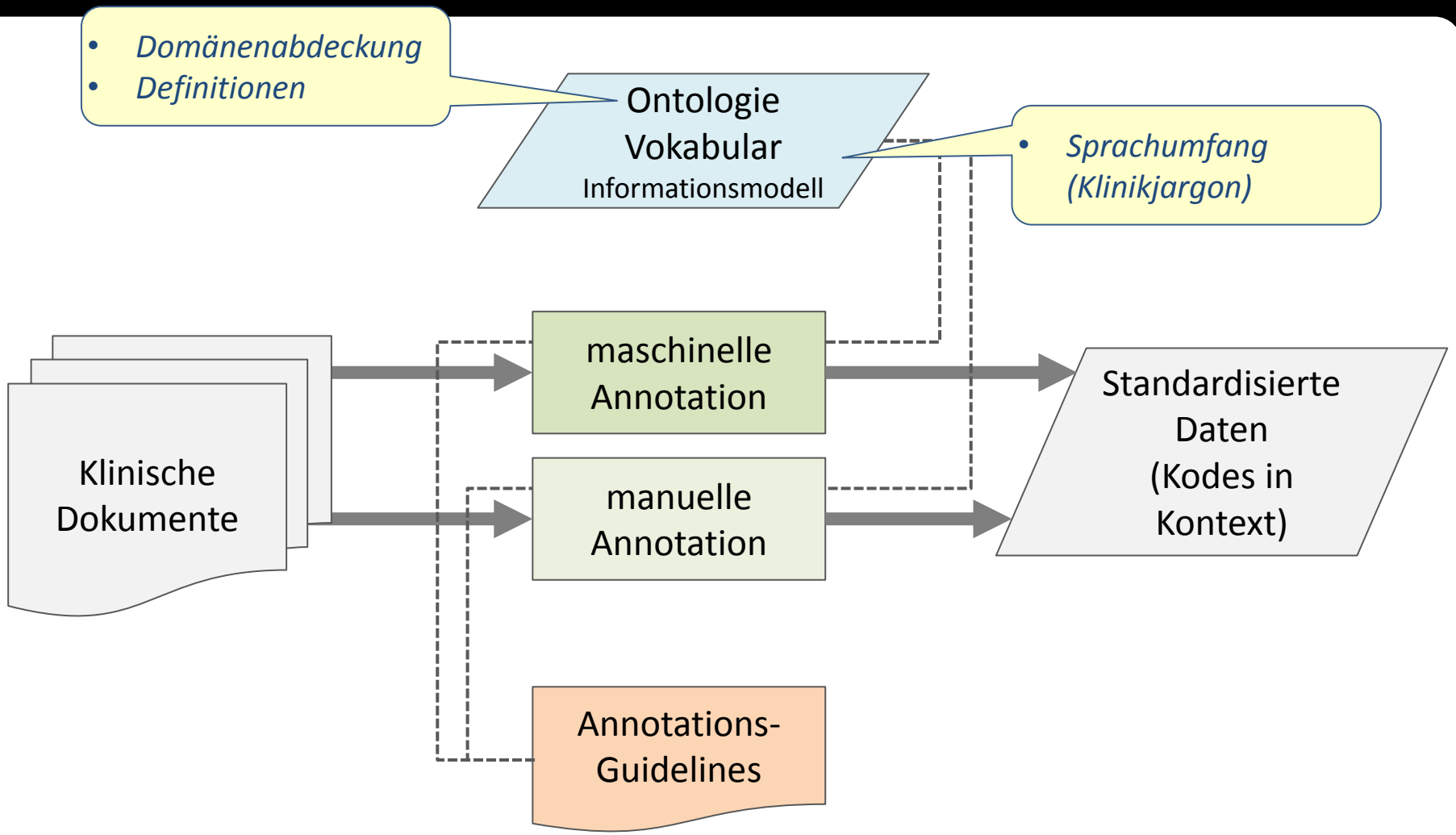
Adip. Pat. mit DM Typ2	Adipöse Patientin mit Diabetes Mellitus Typ 2
St. p. TE eines exulc. sek.knot. SSM li US dors. Level IV	Zustand nach Totalexzision eines exulzerierenden, sekundär knotigen, superfiziell spreitenden Melanoms, Dorsalseite des linken Unterschenkels, Level 4
2,4 mm Tumor DM	2,4 mm Tumordurchmesser
Sentinnel LK ing. li. tumorfr.	Sentinel-Lymphknoten linke Leiste tumorfrei
Gepl. NTx bei term. NINS	Geplante Nierentransplantation bei terminaler Niereninsuffizienz
Euthyorx 75 1-0-0	Euthyrox (Levothyroxin-Na 75µg) 1-0-0

- Zeitdruck, Sprachökonomie: Abkürzungen, Schreibfehler, Transkriptionsfehler, Unvollständigkeit, Redundanz, mangelnde Korrekturen
- Texteingabe in KIS-Systemen ohne zeitgemäßen"Komfort" (Schreibkorrektur, Auto-Vervollständigung, Wortvorhersage, Spracherkennung)
- Verbesserung der Dokumentenqualität eher durch Technologien als durch Veränderung der "Dokumentationskultur"

"Innovation all around – but it ain't in healthcare,
Internet and apps for you, but we get ancient software"

#LetDoctorsBeDoctors

Qualität Ontologie / Vokabular

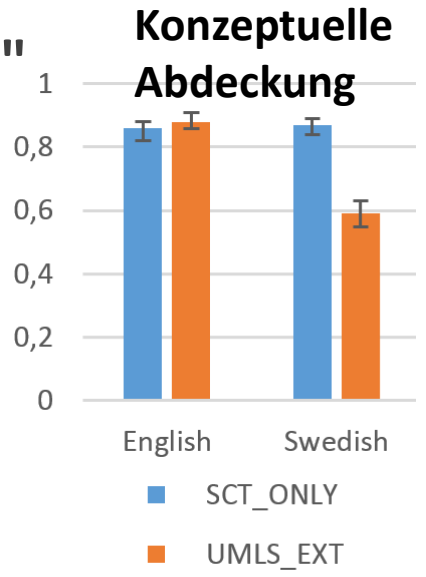


ASSESS CT

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

"SNOMED CT als europäische Referenzterminologie?"

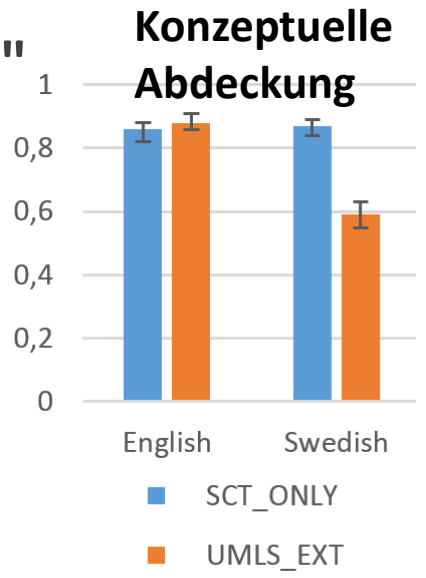
- Manuelle Annotation klinischer Texte (Parallelkorpus) mit SNOMED CT vs. UMLS-Extrakt
- Messung:
  - Konzeptuelle Abdeckung
  - Abbildung des Wortschatzes
- Unterschiede SNOMED CT Schwedisch – Englisch

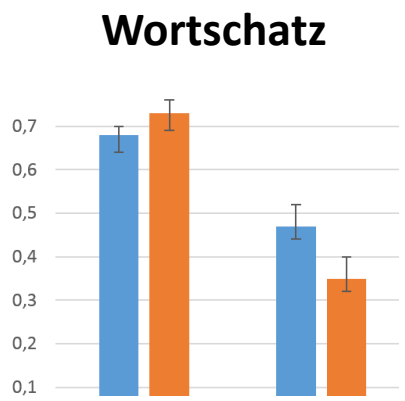


H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU

"SNOMED CT als europäische Referenzterminologie?"

- Manuelle Annotation klinischer Texte (Parallelkorpus) mit SNOMED CT vs. UMLS-Extrakt
- Messung:
  - Konzeptuelle Abdeckung
  - Abbildung des Wortschatzes
- Unterschiede SNOMED CT Schwedisch – Englisch
  - Schwedisch:
    ein (Vorzugs-)Term pro Konzept
  - Englisch:
    durchschnittlich 2,3 Terme Vorzugs- und Interface-Terme





ASSESS-CT Recommendations: http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf

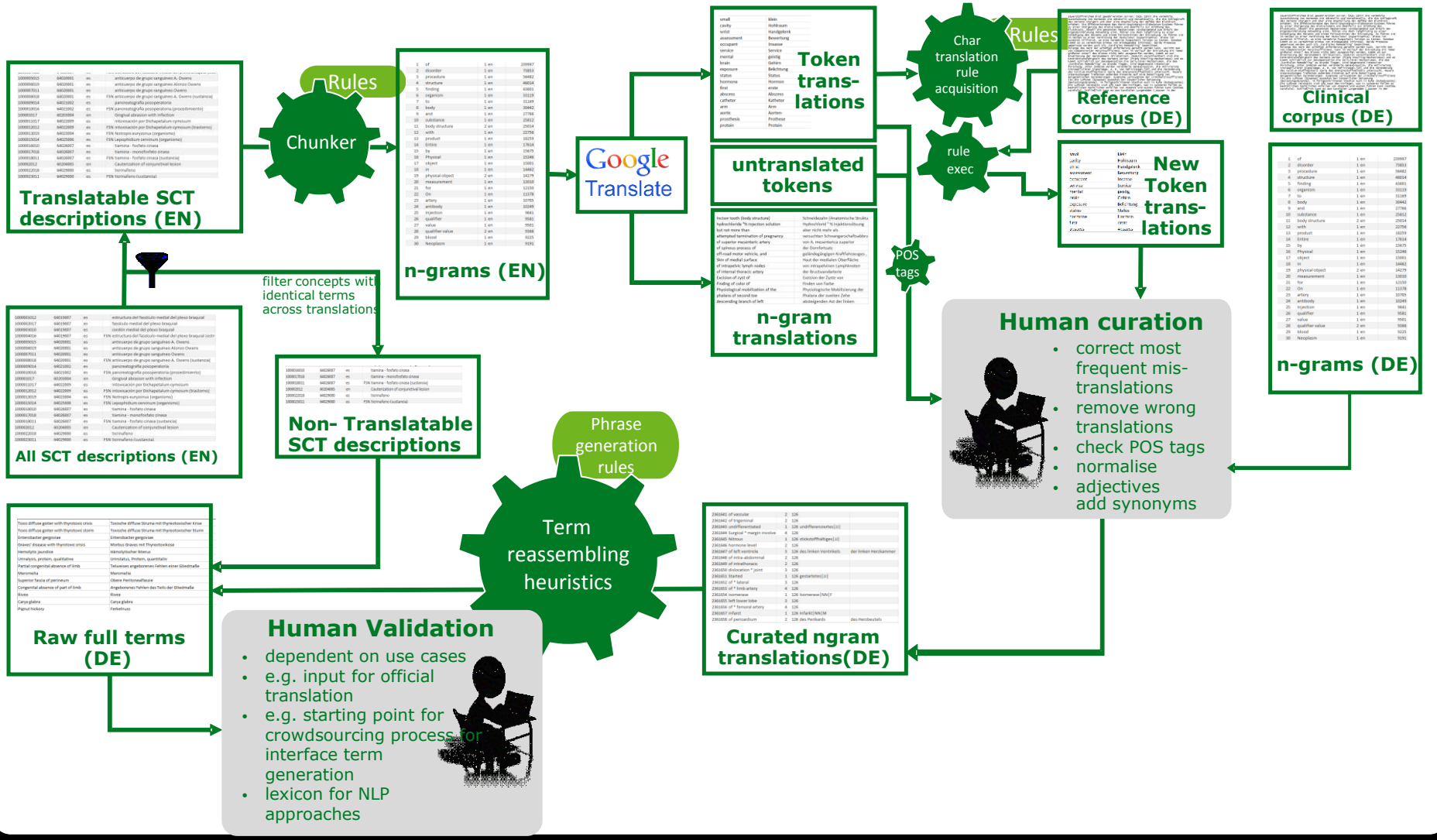
Interface-Vokabulars: Relevanz

- Termhäufigkeiten in Kardiologie-Korpus

  (30.000 Arztbriefe – Quelle: KAGes - Steiermärkische Krankenanstalten GmbH)

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

MUG-GIT: Deutsches Interface-Vokabular für SNOMED CT, derzeit ca. 2,3 Mio Terme



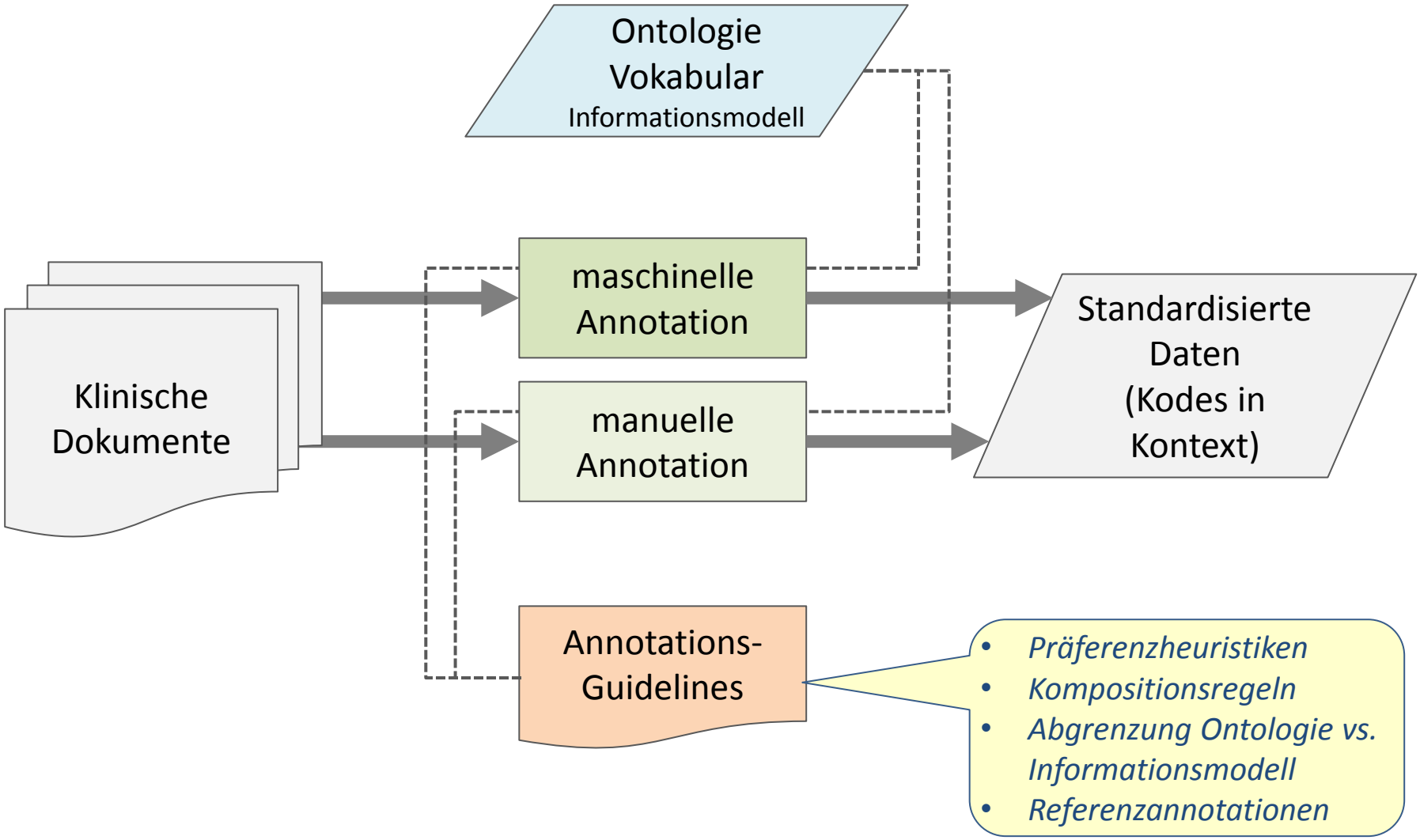
Schulz S. Building an experimental German user interface terminology linked to SNOMED CT. Proc. SNOMED Expo 2017, Bratislava

Automatisch generierte deutsche SNOMED CT - Interface-Terme

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

Kann vom Autor für Validierungszwecke zur Verfügung gestellt werden

Qualität Annotations-Guidelines



Nichtübereinstimmung bei manueller Annotation

Inter-Annotator Agreement in manuellen Annotationsexperimenten (strikte Übereinstimmung): SNOMED 37%, UMLS: 36% (Krippendorff's Alpha)

Nichtübereinstimmung bei manueller Annotation

Inter-Annotator Agreement in manuellen Annotationsexperimenten (strikte Übereinstimmung): SNOMED 37%, UMLS: 36% (Krippendorff's Alpha)

ASSESS CT

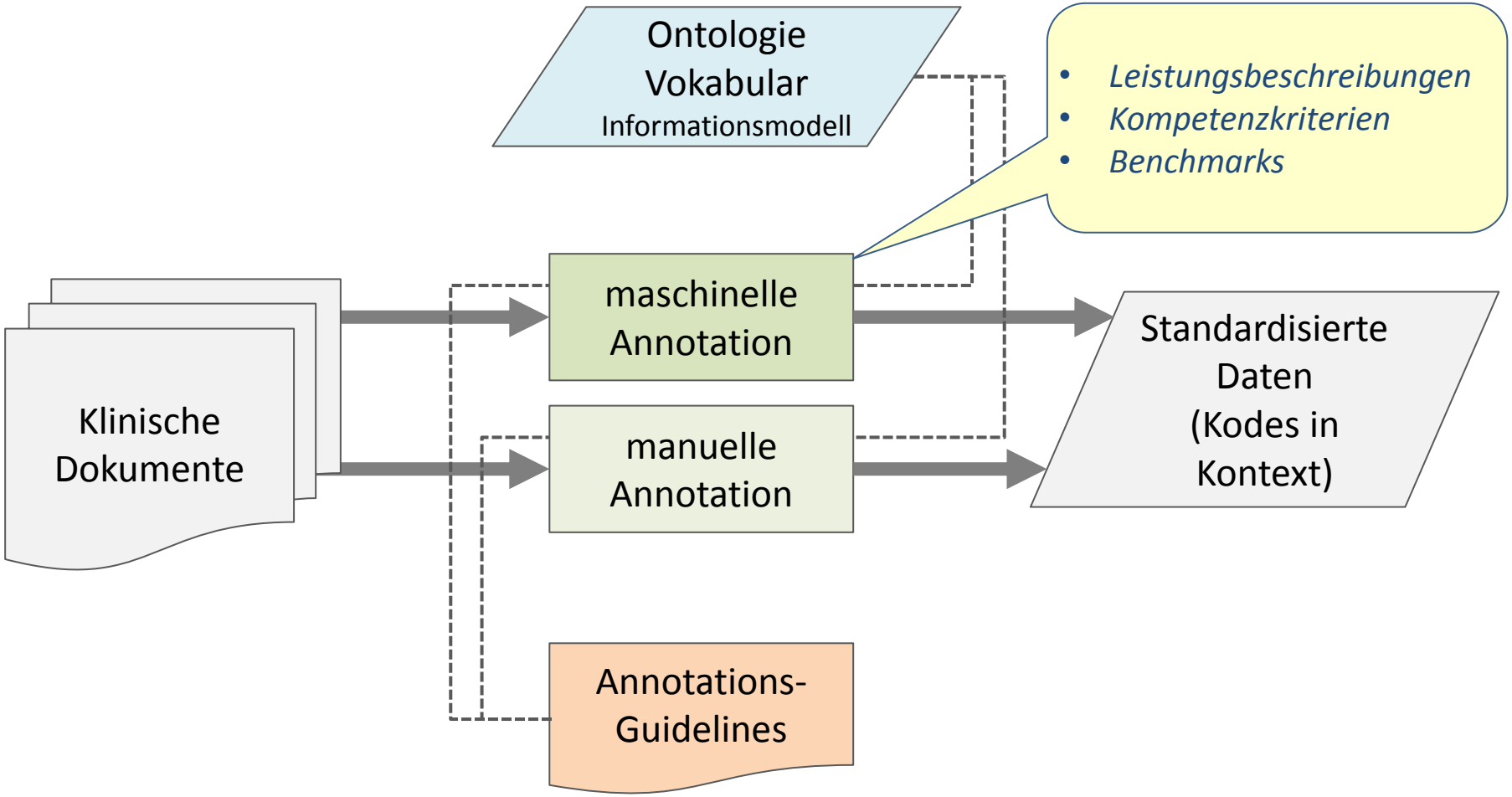
Tokens	Annotator #1	Annotator #2	Gold standard
'Lymphoma"	'*Malignant lymphoma (disorder)*'	'*Malignant lymphoma - category (morphologic abnormality)*'	'*Malignant lymphoma (disorder)*'

Tokens	Annotator #1	Annotator #2	Gold standard
"Former smoker"	'*In the past (qualifier value)*'	'*History of (contextual qualifier) (qualifier value)*'	'*Ex-smoker (finding)*'
	'*Smoker (finding)*'	'*Smoker (finding)*'	

Tokens	Annotator #1	Annotator #2	Gold standard
"Former smoker"	'*In the past (qualifier value)*'	'*History of (contextual qualifier) (qualifier value)*'	'*Ex-smoker (finding)*'
	'*Smoker (finding)*'	'*Smoker (finding)*'	

Miñarro-Giménez JA, Martínez-Costa C, Schulz S. Qualitative assessment of annotations using SNOMED CT. ODLS 2016, Halle/Saale

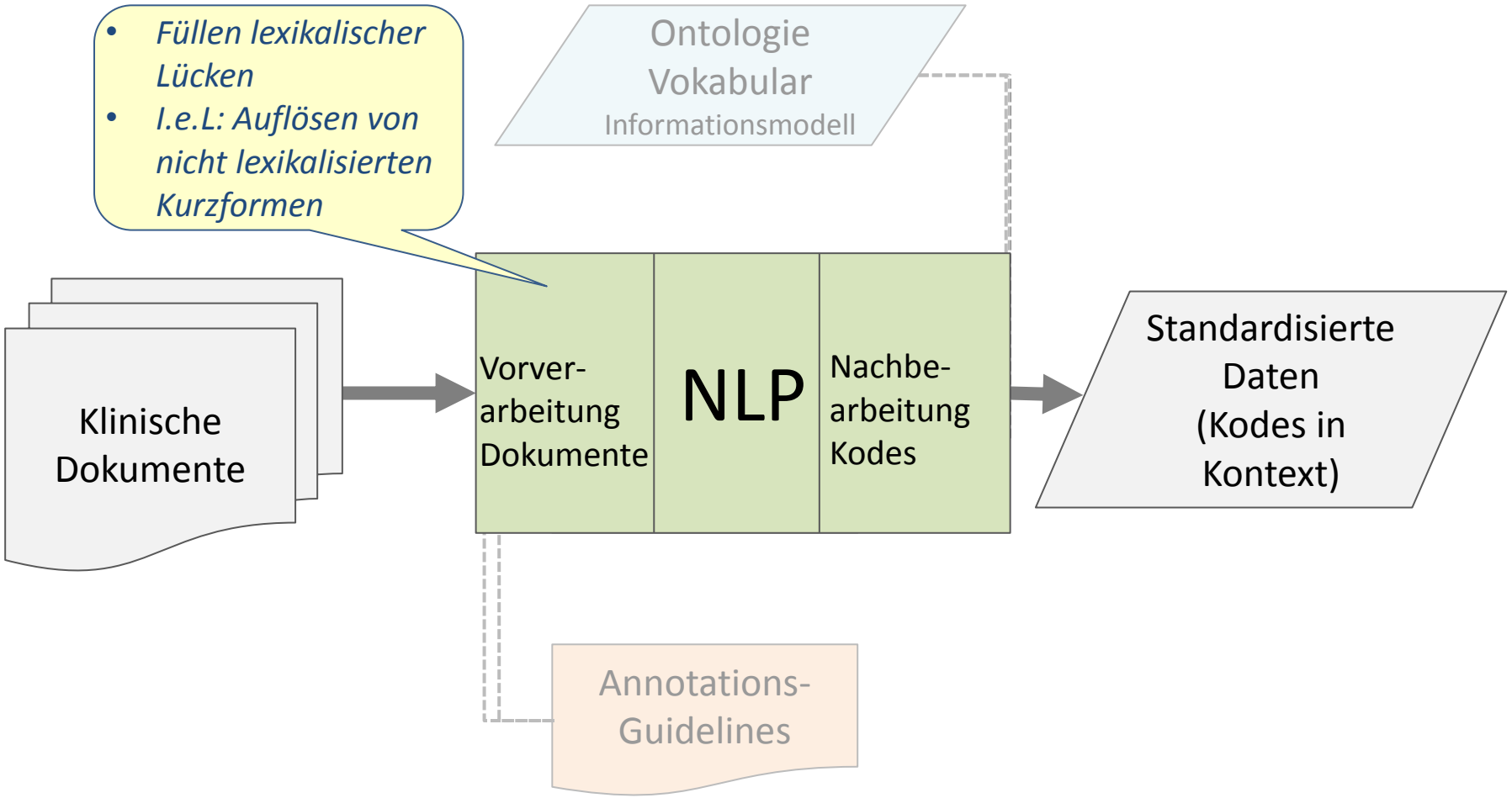
Qualität NLP



NLP- Software – wichtige Qualitätskriterien

- Fuzzy matching
  - Schreibfehler, Flexionen, Derivationen, Komposita
    "Eutyorx", "Gastritiden", "Prozacunverträglichkeit"
- Kontexterkennung
  - Negation: "kein Anhalt für Rezidiv"
  - Zeit: "NTx 3/2007"
  - Sicherheit: "Appendizitisverdacht"
  - Dokumentenabschnitte: Familienanamnese, Labor
- Koordinationen:
  - "Fraktur von Elle und Speiche". "Krea und Harnstoff erhöht"
- Disambiguierung
  - "DM": "Diabetes mellitus" vs. "Durchmesser"
- Auflösung nichtlekikalisierter Kurzformen:
  - "sek. knot.
- Anaphernauflösung

Qualität NLP - Dokumentenvorverarbeitung

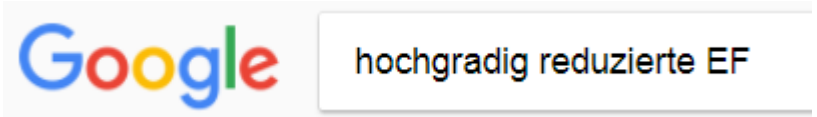


Dokumentenvorverarbeitung

- Beispiel: Auflösung von Abkürzungen und Akronymen
  - N-gram-Modelle "dilat. Kardiomyopathie, hochgr. red. EF"
  - Neuronale Netze ?
  - Web mining

N-gram-Modell aus 30.000 Arztbriefen

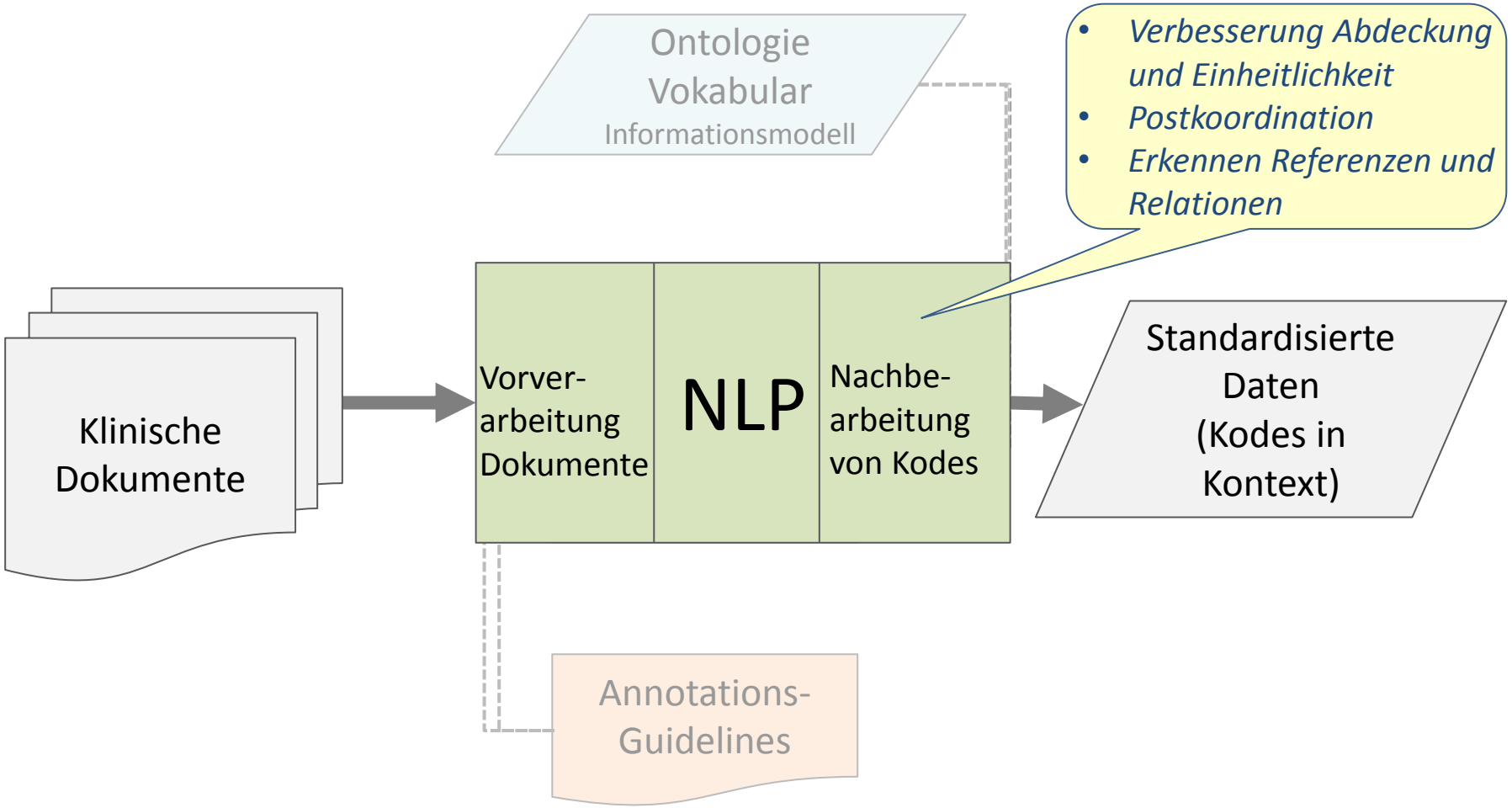
1035	dilat. Kardiomyopathie
1442	dilatative Kardiomyopathie
7	hochgr. red. EF
4	hochgradig reduzierte EF



Ejektionsfraktion – Wikipedia
https://de.wikipedia.org/wiki/Ejektionsfraktion ▾ Translate this page
Die Ejektionsfraktion (**EF**) oder Auswurffraktion (auch Austreibungsfraktion) ist ein Maß für die ... 30 %, **hochgradig** eingeschränkt ... Eine **reduzierte** Ejektionsfraktion wird als objektivierbarer Parameter

Oleynik M, Kreuzthaler M, Schulz S. Unsupervised Abbreviation Expansion in Clinical Narratives. Stud Health Technol Inform. 2017;245:539-543.

Qualität NLP – Kode-Nachbearbeitung



Beispiel Postkoordination SNOMED CT

Präkoordination

"Verbrennung 2. Grades eines einzelnen Fingers"

```
211908006 |Deep partial thickness burn of a single finger (disorder)|
                                    ≡
<<< 29673001 |Second degree burn of single finger, not thumb (disorder)| :
{ 116676008 |Associated morphology| = 262588000 |Deep partial thickness burn
(morphologic abnormality)|,363698007 |Finding site| = 56213003 |Skin of
finger (body structure)| }
```

Postkoordination

"Verbrennung 2. Grades der Rückseite des rechten Zeigefingers"

```
<<< 29673001 |Second degree burn of single finger, not thumb (disorder)| :
{ 116676008 |Associated morphology| = 262588000 |Deep partial thickness burn
(morphologic abnormality)|,363698007 |Finding site| = 37314006 | Skin
structure of dorsal surface of index finger (body structure) |, 272741003
|Laterality| = 24028007 |Right (qualifier value)| }
```

Beispiel Nachbearbeitung: Code Refinement (z.B. Auflösung anaphorischer Referenzen)

Textfragment	Direkte Codes (SNOMED CT)	Inferierte Codes (SNOMED CT)
Resektat nach Whipple: Ein noch nicht eröffnetes Resektat, bestehend aus einem distalen Magen …	65801008 |Excision (procedure)| 69695003 |Stomach structure (body structure)|	53442002 |Gastrectomy (procedure)|
Die Schleimhaut ist insgesamt livide. Auf lamellierenden Schnitten weißliches, teilweise nodulär konfiguriertes Gewebe.	414781009 |Mucous membrane structure (body structure)| 85756007 |Body tissue structure (body structure)|	78653002 |Gastric mucous membrane structure (body structure)|
2 cm aboral des Pylorus zeigt die Dünndarmwandung eine sanduhrartige Stenose	38848004 |Duodenal structure (body structure)| 415582006 |Stenosis (morphologic abnormality)|	73120006 |Stenosis of duodenum (disorder)|

Möglichkeiten und Grenzen

- Ziel: interoperable semantische Repräsentation hoher Qualität
- Klinische Texte: manueller Goldstandard problematisch, bzgl. der "richtigen" Kodierung mit großen Terminologiesystemen
- SNOMED CT: Konzeptuelle Abdeckung gut, besonders bei Nutzung von Postkoordination
- Ausreichendes lexikalisches Matching erfordert Investition in Interface-Terminologien
  → Crowdsourcing, Use-Case getrieben
- Mehrdeutige Akronyme und nichtlexikalisierte Abkürzungen:
  →Lernen von großen klinischen Korpora vielversprechend
- Nachbearbeitung / Interpretation von Annotationssequenzen:
  → Forschungsbedarf (Überführung von Sequenzen in Graphen)
- Nutzung bestehender Informations-Templates (z.B. HL7 FIHR)

Stefan Schulz
Medizinische Universität Graz

purl.org/steschu

MEDIZIN INFORMATIK INITIATIVE

Workshop „Datenqualität“
TMF, Berlin, 03.05.2018

Natural Language Processing und Ontologie-Mapping Möglichkeiten und Grenzen bezüglich der Datenqualität

Fragen?

Kontakt: stefan.schulz@medunigraz.at

Acknowledgements:
CBmed GmbH
KAGes GmbH
SAP AG
FFG Austria