Reference terminologies vs. Interface terminologies
Common problems with specialised domain terminologies

- I need to automatically annotate textual content
  - that uses an idiosyncratic sublanguage
  - which abounds of abbreviations and errors (syntax, spelling)
  - which uses highly ambiguous terms
  - the meaning of which depends on local contexts
  - is characterised by constantly new terms
  I need to find close-to-user terms for structured entry

- Current situation:
  - Numerous quasi-standard terminologies
  - None of them cover all my concepts
  - Many terms I use are not covered (although concepts are available)
## Popularity of Terms in Pubmed

<table>
<thead>
<tr>
<th>Preferred term (SNOMED CT)</th>
<th>Count</th>
<th>Synonym</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary malignant neoplasm of lung</td>
<td>0</td>
<td>Lung cancer</td>
<td>120682</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bronchial carcinoma</td>
<td>3452</td>
</tr>
<tr>
<td>Cerebrovascular accident</td>
<td>3819</td>
<td>Stroke</td>
<td>191559</td>
</tr>
<tr>
<td>Block dissection of cervical lymph nodes</td>
<td>1</td>
<td>Neck dissection</td>
<td>7512</td>
</tr>
<tr>
<td>Electrocardiographic procedure</td>
<td>1</td>
<td>Electrocardiogram</td>
<td>33670</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ECG</td>
<td>55120</td>
</tr>
<tr>
<td>Backache</td>
<td>3489</td>
<td>Back pain</td>
<td>38132</td>
</tr>
<tr>
<td>Capillary blood specimens</td>
<td>32</td>
<td>Capillary blood samples</td>
<td>574</td>
</tr>
</tbody>
</table>
Two Aspects of Terminologies

- Normative
  - Codes + Labels (names) denote well-defined entities in a realm of discourse
  - "Explanatory" labels, e.g. "Primary malignant neoplasm of lung (disorder)."
  - Scope notes and definitions explain meaning of concepts
  - Meaning formalised by logic descriptions → formal ontology

- Descriptive
  - Describes language as being used: lexicon "Lung cancer"
Reference terms – Interface terms

- Reference terms: Normative
  - Primarily language-independent: described by textual and/or formal (ontological) definitions
  - Labels univocally identify a concept, often not terms of choice in communication/documentation

- Interface terms: Descriptive
  - Collection of human language terms/expressions used in (informal) written and oral communication
  - Grounding by connection with reference terminologies
  - Inherent ambiguity of interface terms, often not perceived
  - Need to deal with short forms

- Many domain terminologies contain interface terms, but far from being exhaustive
Case study: German interface terminology for SNOMED CT

- Context: Information extraction from clinical narratives → CBmed IICCAB* (see poster)
- Limited resources, incremental approach
- Top-down:
  - Modularization of original terms (split into noun phrases):
  - Translating / finding synonyms of derived, highly repetitive phrases:
    - "Magnetic resonance imaging" in 627 SNOMED terms
    - "Second degree burn" in 166 SNOMED terms
  - Acquisition of translations and synonyms by decreasing frequency (machine translation, automated synonym acquisition)
  - Manual revision
- Bottom-up:
  - Addition of terms from n-gram frequency lists from reference corpora

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

- **Translatable SCT descriptions (EN)**
- **All SCT descriptions (EN)**
- **Raw full terms (DE)**
- **Non-Translatable SCT descriptions**
- **Term reassembling heuristics**
- **Phrase generation rules**
- **Rules**
- **Chunker**
- **n-grams (EN)**
- **Google Translate**
- **untranslated tokens**
- **n-gram translations**
- **New Token translations**
- **Reference corpus (DE)**
- **Clinical corpus (DE)**
- **Human curation**
  - correct most frequent mis-translations
  - remove wrong translations
  - check POS tags
  - normalise adjectives
  - add synonyms
- **Curated ngram translations (DE)**
- **Human Validation**
  - dependent on use cases
  - e.g. input for official translation
  - e.g. starting point for crowdsourcing process for interface term generation
  - lexicon for NLP approaches
Case study: German interface terminology for SNOMED CT

- Core vocabulary:
  - Constantly checked and enhanced by domain experts (medical students)
  - Priorisation by Use cases

- Guidelines
  - Avoidance of ambiguous entries: inclusion of composed terms (e.g. "delivery" → "drug delivery", "preterm delivery")
  - Acronyms only in context: not "CT" but "CT guided"
  - Inflection, compositions rules requires special markers (German) and rules for reassembly of terms

- Current state:
  - ca. 2 Million Interface-Terms
  - Automatically generated from core vocabulary with 92,500 n-grams, out of 85,400 English n-grams
  - Benchmark: parallel corpus extracted from Medline: term coverage 33.1% for German vs. 55.4% for English
<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Frequency</th>
<th>Co-occurrence 1</th>
<th>Co-occurrence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>vaginal</td>
<td>1</td>
<td>1478 vaginales</td>
<td>J</td>
</tr>
<tr>
<td>fluoroscopic guidance</td>
<td>2</td>
<td>1477 Durchleuchtungskontrolle</td>
<td>NN</td>
</tr>
<tr>
<td>disc</td>
<td>1</td>
<td>1476 Scheibe</td>
<td>NN</td>
</tr>
<tr>
<td>lower limb</td>
<td>2</td>
<td>1473 unteres</td>
<td>JJ</td>
</tr>
<tr>
<td>brain</td>
<td>1</td>
<td>1468 Gehirn</td>
<td>NN</td>
</tr>
<tr>
<td>preparation</td>
<td>1</td>
<td>1464 Zubereitung</td>
<td>NN</td>
</tr>
<tr>
<td>method</td>
<td>1</td>
<td>1463 Verfahren</td>
<td>NN</td>
</tr>
<tr>
<td>of bone</td>
<td>2</td>
<td>1462 des Knochens</td>
<td>NN</td>
</tr>
<tr>
<td>Red</td>
<td>1</td>
<td>1455 rotes</td>
<td>JJ</td>
</tr>
<tr>
<td>Monitoring</td>
<td>1</td>
<td>1453 Überwachung</td>
<td>NN</td>
</tr>
<tr>
<td>Computed</td>
<td>1</td>
<td>1453 berechnetes</td>
<td>JJ</td>
</tr>
<tr>
<td>phalanx</td>
<td>1</td>
<td>1449 Phalanx</td>
<td>NN</td>
</tr>
<tr>
<td>subsp.</td>
<td>1</td>
<td>1449</td>
<td></td>
</tr>
<tr>
<td>anastomosis</td>
<td>1</td>
<td>1447 Anastomose</td>
<td>NN</td>
</tr>
<tr>
<td>vessel</td>
<td>1</td>
<td>1446 Blutgefäß</td>
<td>NN</td>
</tr>
<tr>
<td>Computed tomography</td>
<td>2</td>
<td>1443 Computertomographie</td>
<td>NN</td>
</tr>
<tr>
<td>uterus</td>
<td>1</td>
<td>1436 Uterus</td>
<td>NN</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>1432 Schwierigkeit</td>
<td>NN</td>
</tr>
<tr>
<td>elbow</td>
<td>1</td>
<td>1429 Ellbogen</td>
<td>NN</td>
</tr>
<tr>
<td>high</td>
<td>1</td>
<td>1429 hohes</td>
<td>JJ</td>
</tr>
<tr>
<td>food</td>
<td>1</td>
<td>1423 Lebensmittel</td>
<td>NN</td>
</tr>
<tr>
<td>Observation</td>
<td>1</td>
<td>1423 Beobachtung</td>
<td>NN</td>
</tr>
<tr>
<td>using fluoroscopic</td>
<td>2</td>
<td>1422</td>
<td></td>
</tr>
<tr>
<td>unable</td>
<td>1</td>
<td>1421 unfähiges</td>
<td>JJ</td>
</tr>
<tr>
<td>Peripheral</td>
<td>1</td>
<td>1419 peripheres</td>
<td>JJ</td>
</tr>
<tr>
<td>unable to</td>
<td>2</td>
<td>1418 unfähig zu</td>
<td></td>
</tr>
<tr>
<td>Vascular</td>
<td>1</td>
<td>1417 vaskuläres</td>
<td>JJ</td>
</tr>
<tr>
<td>using fluoroscopic guidance</td>
<td>3</td>
<td>1416 mit Durchleuchtungskontrolle</td>
<td></td>
</tr>
<tr>
<td>Benign neoplasm</td>
<td>2</td>
<td>1415 gutartiges</td>
<td>JJ</td>
</tr>
<tr>
<td>ID</td>
<td>Code</td>
<td>Term</td>
<td>Translation</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>-----------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>20170315_240011_002</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneubildung</td>
</tr>
<tr>
<td>20170315_240011_003</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neubildung des Hirns</td>
</tr>
<tr>
<td>20170315_240011_004</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneubildung</td>
</tr>
<tr>
<td>20170315_240011_005</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Gehirns</td>
</tr>
<tr>
<td>20170315_240011_006</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneoplasie</td>
</tr>
<tr>
<td>20170315_240011_007</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Hirns</td>
</tr>
<tr>
<td>20170315_240011_008</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneoplasie</td>
</tr>
<tr>
<td>20170315_240011_009</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasie des Gehirns</td>
</tr>
<tr>
<td>20170315_240011_010</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Gehirnneoplasma</td>
</tr>
<tr>
<td>20170315_240011_011</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Neoplasma des Gehirns</td>
</tr>
<tr>
<td>20170315_240011_012</td>
<td>126952004</td>
<td>Neoplasm of brain</td>
<td>Hirnneoplasma</td>
</tr>
<tr>
<td>20170315_241010_001</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neubildung des Großhirns</td>
</tr>
<tr>
<td>20170315_241010_002</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neoplasie des Großhirns</td>
</tr>
<tr>
<td>20170315_241010_003</td>
<td>126953009</td>
<td>Neoplasm of cerebrum</td>
<td>Neoplasma des Großhirns</td>
</tr>
<tr>
<td>20170315_242015_001</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neubildung des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_002</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neubildung des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_242015_003</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_004</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasie des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_242015_005</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasma des Frontallappens</td>
</tr>
<tr>
<td>20170315_242015_006</td>
<td>126954003</td>
<td>Neoplasm of frontal lobe</td>
<td>Neoplasma des Lobus frontalis</td>
</tr>
<tr>
<td>20170315_243013_001</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neubildung des Temporalappens</td>
</tr>
<tr>
<td>20170315_243013_002</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neubildung des Lobus temporalis</td>
</tr>
<tr>
<td>20170315_243013_003</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasie des Temporalappens</td>
</tr>
<tr>
<td>20170315_243013_004</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasie des Lobus temporalis</td>
</tr>
<tr>
<td>20170315_243013_005</td>
<td>126955002</td>
<td>Neoplasm of temporal lobe</td>
<td>Neoplasma des Temporalappens</td>
</tr>
</tbody>
</table>
Crowdsourcing for terminology development

- Functionality: entry of new terms, commenting and validating existing terms
- Possible central data element: Interface Term – External Code
  "DM" - 81827009 |Diameter (qualifier value)
- Possible Attributes:
  - Creator, creation type, date, (sub)domain, user group
    Max Muster, manual, 20170803, Dermatology Graz, Doctos
  - Example annotation, e.g.
    "ein 3 cm im DM haltender Tumor"
  - Validation/ commenting by other users
    John Doe, 20180912, ★★★★★★
    "Example incomprehensible – additional examples needed"
Short forms – Document preprocessing

- Ambiguous short forms not in dictionary
- Difficulty of maintenance
  - Abundance of concurring readings
  - High productivity
- Instead:
  - Main assumption: short forms and expansions occur in the same corpus
  - Automatically create N-gram model from specific reference corpus
  - Replace short forms by most plausible expansions
  - The same for other out-of-lexicon words, e.g. misspellings
  - If assumption fails: try Web mining
Example: Resolution of short forms

- "dilat. Kardiomyopathie, hochgr. red. EF"
- Word-n-gram model (30,000 discharge summaries)
  
<table>
<thead>
<tr>
<th></th>
<th>1035</th>
<th>1442</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dilat. Kardiomyopathie</td>
<td>dilatative Kardiomyopathie</td>
</tr>
<tr>
<td></td>
<td>hochgr. red. EF</td>
<td>hochgradig reduzierte EF</td>
</tr>
</tbody>
</table>

- Web mining

  Ejektionsfraktion – Wikipedia
  
  Die **Ejektionsfraktion (EF)** oder Auswurffraktion (auch Austreibungsfraktion) ist ein Maß für die ... 30 %, hochgradig eingeschränkt .... Eine **reduzierte** Ejektionsfraktion wird als objektivierbarer Parameter
Example: Resolution of short forms

- "Pat. mit rez. HWI und VUR"
- Wort-N-gram Modell

381 Pat. mit
120 Patient mit
  2 rez.
707 rezenten
468 rezidivierende

- No clear expansions of "HWI" and "VUR"
- Web mining

[Google search for "HWI und VUR"]
Example: Resolution of short forms

- **Interpretation:**
  - **Frequency**
  - **Acronym-definition patterns**
  - **Regular expressions created from short forms**
Recommendations

- Need for interface terms not satisfied by domain terminologies
- Acquisition of interface terms
  - Top-down (from reference terminologies)
  - Bottom up (from corpora)
- Use collaborative approach (crowdsourcing)
- Avoid ambiguous terms in lexicon
- Use n-gram models (alternatively deep learning ?) for out-of-lexicon terms, extracted from related corpora
  - Benefitting from typical local contexts in "similar" documents
    - Resolution of short forms
    - Correction of typos
Questions?

Contact: stefan.schulz@medunigraz.at
Additional Slides
"Is SNOMED CT well suited as an European reference terminology?"

- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract
- Assessment:
  - concept coverage
  - term coverage
"Is SNOMED CT well suited as an European reference terminology ?"

- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract

- Assessment:
  - concept coverage
  - term coverage

H2020: Assessing SNOMED CT for Large Scale eHealth Deployments in the EU
"Is SNOMED CT well suited as an European reference terminology?"

- Manual annotation of a corpus of clinical texts with SNOMED CT vs. UMLS-Extract
- Assessment:
  - concept coverage
  - term coverage
- Differences SNOMED CT
  - Swedish: one term per concept
  - English: on average 2.3 terms per concept (Preferred terms, synonyms)
Primary malignant neoplasm of lung (disorder)
SCTID: 93880001

Primary malignant neoplasm of lung
Primary malignant neoplasm of lung (disorder)
Lung cancer

Label (Fully Specified Name)
Code (Concept ID)
Interface terms (Synonyme)
Text definition
SNOMED CT als formal 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

- **Taxonomy**
  - 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)

- **Logik Axioms**
  - Finding site → Lung structure
  - Associated morphology → Malignant neoplasm, primary
Interface-Terms ➔ Interface-Terminology

http://browser.ihtsdotools.org

Interface terms (Synonyms):
- Primary malignant neoplasm of lung (disorder)
- Primary malignant neoplasm of lung
- Primary malignant neoplasm of lung (disorder)
- Lung cancer

SCTID: 93880001
"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"