**Überblick: Inter-Annotator-Agreement**

* Kategorisierungen von 2898 Keywords verglichen
* Cohen’s Kappa Coefficient: 0.55 (= moderate Übereinstimmung) – berechnet exakte Übereinstimmung pro Keyword, keine partiellen Übereinstimmungen berücksichtigt
* auf Kategorisierungslevel (1-3 Kategorisierungen pro Keyword, je einzeln überprüft):
  + Matches: 2182 (62%)
  + Mismatches: 1334 (38%)
* auf Ebene von Keywords:
  + Matches (= alle Kategorisierungen stimmen überein: 1746 Keywords (60,2 %)
  + Partial Matches (= einzelne Kategorisierungen stimmen überein): 415 Keywords (14,3 %)
  + Mismatches (= gar keine Übereinstimmung): 737 Keywords (25,4 %)

**Individuelle Tendenzen in der Kategorisierung**

**Ein Bild, das Text, Screenshot, Schrift, Diagramm enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Schrift, Diagramm enthält.

Automatisch generierte Beschreibung**

**Problemfälle / Beobachtungen**

**Topic als sekundäre Kategorie?**

* Topic nur dann vergeben, wenn keine andere Kategorisierung möglich?
* unklare Grenzen zwischen „method/activity“ and „topic“
  + z.B. *historical gis, discovery, academic publication*
* unklare Grenzen zwischen „discipline“ und „topic“ (und „method/activity“)
  + z.B. *social computing, islamic history, spatial humanities, data management, medieval brain anatomy*, *narratology, diplomatie, compling*

**Zusammenführen von „too specific“, „too broad/unspecific“ und “unclear” zu “not categorizable” sinnvoll?**

* Hintergrund: letztlich alle irrelevant

**einzelne Tags als Ausschluss anderer Kategorien verstehen?**

* „not categorizable“ and „multiple tags” 🡪 hier dürfen keine weiteren Tags vergeben werden

**Generalisierungen vs. Entitäten/Klasse**

* *immigrants* 🡪 topic vs. entity (person)
* *african languages, historical language* 🡪 topic vs. language
* *church, prison*, *libraries* 🡪 topic vs. entity (place)
* *parser, pos-tagger*, *publishing tools* 🡪 topic vs. resource/tool
* *long-term projects on cultural heritage* 🡪 topic vs. entity (project)
* *pandemic, audio conference, conferences* 🡪 topic vs. entity (event)
* *organisation* 🡪 topic vs. entity (institution/organisation/brand)
* individuelle Tendenzen:
  + Nina: „entity“ nur im Sinne benannter, eindeutiger identifizierbarer Entitäten, ansonsten „topic“
  + Martin: aktuell „entity“ etc. auch für generalisierte Gruppen/Orte etc. verwendet

**Unklare Spezifizitätsgrenzen**

* Wann ist etwas „too broad/unspecific“ und wann (noch) nicht?
* z.B. *other social sciences, other disciplines*
* Nina: “too broad/unspecific”, Martin: “discipline”

**Zusammenfall verschiedener Informationen in einem Keyword**

* z.B. *english literature, 18th century french literature, dutch literature*
* object-type (“literature”)
* language (“english”, “french”, “dutch”)
* entity (place) (“english”, “french”, “dutch”)
* topic
* multiple tags
* z.B. *historical geography, bulgarian history*
  + Nina: nicht aufgesplittet vs. Martin: aufgesplittet
  + historical geography 🡪 discipline + time period?
  + bulgarian history 🡪 discipline + place?
* z.B. *history matters* 🡪 discipline + topic?

**Fachtermini benötigen Hintergrundwissen für Einordnung**

* z.B. *code switching, code mixing* 🡪 topic NOT method/activity
* z.B. *squeezes* 🡪 object-type

**Adjektive als Herausforderungen**

* z.B. *raw, (teil-)automatisch generiert, hand corrected, computational linguistic, audio-visual*

**Metakommentare als Herausforderungen**

* z.B. *for interns, (teil-)automatisch generiert*

**Object-type als Herausforderung**

* wann ist etwas ein „object“ vs. wann ist es kein „object“ mehr und z.B. schon ein „topic“?
* *books, newspapers, vases* 🡪 eindeutig
* schwieriger: *hotels, collection, cultural heritage, linked data, query language, digital heritage, digital infrastructure, platform, word-counts, workflows, words, kunst*
* Versuch einer Abgrenzung von object-type und topic (der auch bei Martins Annotierungen zur Anwendung kam) in Einklang mit den Vorgaben aus dem Kategorienschema

|  |  |
| --- | --- |
| *Object-type* | *Topic* |
| alles, was man "sich gegenüberhalten kann" = in gewisser Weise angreifbar/physisch umreißbar | vor allem Ideen, gedankliche oder institutionelle ("parliament") Konstrukte, Phänomene ("weather", "migration") oder Sammelbegriffe ("food", "art"), disziplinäre Fachbegriffe |
| konkreter Forschungsgegenstand | abstrakter Forschungsgegenstand |
| vor allem für quantitative Analyse bestimmt | vor allem für qualitative Analyse bestimmt |

**Ambiguität als Herausforderung**

* z.B. *replication* 🡪 Tätigkeit vs. Objekt

**Nächste Schritte**

* 2 von Martin noch unkategorisierte Fälle
* Massimilianos ARCHE-Kategorisierung einbauen
* Schreibfehler und ähnliches korrigieren bzw. Varianten zusammenführen?
* Umgang mit unterschiedlichen Sprachen
  + zusammenführen? (s. evtl. auch zusätzliche Liste von Martin)
  + nur Englisch und Deutsch behalten?
  + andere fremdsprachliche Fälle identifizieren und herausfiltern?
* Konzentration auf konkrete Gruppen, um diese zu vereinheitlichen, zusammenzuführen und relevante Vokabularien zu identifizieren?