Keyword-Kategorisierung

**Vorgehen bei Kategorisierung**

* 3782 Keywords aus diversen Quellen (dha-zotero, dariah-campus, acdh-ch-website, arche, …)
* LegalKraus entfernt, nur in ARCHE vorkommende Tags aktuell außen vor gelassen
* unabhängige Kategorisierung übriger 2900 Keywords durch Martin und Nina
* jedes Keyword konnte dabei 1-3 Kategorien zugeordnet werden
* folgende Kategorien wurden vergeben:
* **discipline** 🡪 Disziplin (z.B. Geschichte, Digital Humanities)
* **entity (event)**
* **entity (institution/organisation/brand)**
* **entity (person)**
* **entity (place)**
* **entity (project)**
* **entity (product)** 🡪 Film, literarisches Werk etc.
* **format/standard** 🡪 (Daten-)Formate und/oder -Standards (z.B. XML/TEI, CIDOC-CRM)
* **language** 🡪 Sprache (z.B. German, English)
* **method/activity** 🡪 Methode und/oder Aktivität (z.B. sediment analysis, gis, co-creation, photographing)
* **multiple tags** 🡪 mehrere Tags in einem Keyword (z.B. history & archaeology; russian and ukrainian; berliner tageblatt, kerr, wollf) – Problem der ursprünglichen Kategorisierung; wird letztlich ausgebessert, aber zuerst gesammelt, um Muster zu erkennen
* **object-type** 🡪 Art/Format einer Ressource/eines Untersuchungsobjekts (z.B. newspaper, data, software, Buch, pottery)
* **resource/tool** 🡪 (meist digitale) Ressourcen oder Tools (z.B. gams, transkribus, txtlab); hierunter können auch soziale Medien (z.B. Facebook, Twitter) zählen
* **time period** 🡪 Zeitpunkt/-raum (z.B. 18th century, Frühe Neuzeit) bzw. jegliche Art von zeitlicher Spezifizierung (z.B. historical, ancient)
* **topic** 🡪 Thema (z.B. food, art, saints, citizen innovation, code mixing, habsburg-ottoman relations)
* **not categorizable** 🡪 fasst folgende Ausschlusskriterien zusammen: **too broad/unspecific** (= Keyword ist zu offen/ungenau, z.B. other, Sonstiges, all), **too specific** und **unclear** (Bedeutung und/oder Zuordnung des Keywords unklar)
* im Anschluss Berechnung von Inter-Annotator-Agreement (verschiedene Maße) und Diskussion von Problemfällen

**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 %)

**Inter-Annotator-Agreement auf Kategorien-Ebene**

(prozentueller Anteil von Matches an allen vergebenen Instanzen einer Kategorie)

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

Automatisch generierte Beschreibung

**Individuelle Tendenzen in der Kategorisierung**

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

Automatisch generierte Beschreibung**

**Ein 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*

**Ambiguität als Herausforderung**

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

**Überlegungen zu nächsten Schritten**

* 2 von Martin noch unkategorisierte Fälle
* Massimilianos ARCHE-Kategorisierung hinzufügen
* Vereinheitlichung der Kategorisierung anhand optimierter Kategorisierungsregeln
* Reduktion:
  + Schreibfehler und ähnliches korrigieren bzw. Varianten zusammenführen?
  + unterschiedliche Sprachen zusammenführen?
  + nur Englisch und Deutsch behalten und andere fremdsprachliche Fälle identifizieren und herausfiltern?
* Konzentration auf konkrete Gruppen, um diese zu vereinheitlichen, zusammenzuführen und relevante Vokabularien zu identifizieren?
* Endziel: wie gelangen wir zu einem Vokabular / einer Richtlinie, damit Studierende Zotero-Einträge systematisch taggen können?
  + u.a. welche Kategorien sind für uns relevant, d.h. welche sollten Annotator\*innen vergeben?