

.dhmuc – Workshop:

**Wörterbücher als Datenschatz: Digitale Wörterbuchdaten
und ihre (zukünftigen) Verwendungsmöglichkeiten**

**4. / 5. Oktober (Dienstag/Mittwoch) 2022
in der BAdW, Sitzungssaal im 2. Stock
(online: us06web.zoom.us/j/87433087754)**

Die Zeit der großen, über viele Jahrzehnte und sogar Jahrhunderte erstellten Wörterbuchprojekte (z.B. Grimmsches Wörterbuch), scheint sich – auch aus förderpolitischen Gründen – einem Ende zu nähern. Die Zeit ihrer Wirkung ist damit aber nicht vorbei, sondern beginnt gerade erst, auch weil viele dieser Wörterbücher mittlerweile retrodigitalisiert wurden und damit in Form digitaler Daten vorliegen. Die so gewonnenen Wörterbuchdaten zeichnen sich durch einige für die datentechnische Verwertung sehr erfreuliche Eigenschaften aus:

- es handelt sich um wissenschaftlich-geprüfte, „**handgeschöpfte**“ **Daten**, d.h. die Daten sind von extrem hoher Datenqualität.
- die Daten decken ihren Gegenstandsbereich **relativ umfassend** ab.
- insbesondere, wenn man auch die Forschungsintermediärdaten, z.B. digitalisierte Zettelkästen hinzunimmt, handelt es sich um **große Datenmengen**.

Im Workshop möchten wir uns darüber austauschen, welche weiteren Verwendungsmöglichkeiten es für digitale Wörterbuchdaten jenseits des bloßen Wechsels der Ausgabeform als Online-Wörterbuch gibt. Denkbare Verwendungsmöglichkeiten sind dabei:

- Training und **Verbesserung von Sprachmodellen** oder generell automatischen sprachverarbeitenden KI-Systemen mit Wörterbuchdaten.
- Wörterbuchdaten als **Wissenspool für Expertensysteme**.
- Anwendungen, die sich aus der **Verknüpfung mit anderen Daten** (anderen Wörterbuchdaten, Korpora, Materialien ganz anderer Art) ergeben.
- Wörterbuchdaten als **Teil umfassenderer (Sprach-)informationssysteme**.

Neben diesen (künftigen) Anwendungsmöglichkeiten selbst, interessieren uns aber auch die Fragen nach den Voraussetzungen ihrer technischen Realisierung, etwa wie müssen Wörterbuchdaten aufbereitet werden, um sie z.B. für das Training einer KI einzusetzen.

.dhmuc – Workshop: Wörterbücher als Datenschatz

Dienstag, 4. Oktober 2022 – Vision und Exploration

- 14:00 Begrüßung
- 14:15 *Eckhart Arnold (BAdW):*
Mehr als nur Nachschlagewerke: Spekulative Überlegungen dazu, was man mit Wörterbuchdaten noch anfangen kann.
- 15:00 *Christiane Bayer (LMU):*
Vom Codex Abrogans zur digitalen Lexikographie.
- 15:45 --- Kaffeepause ---
- 16:15 *Isabelle Hardt / Bettina Hofmann-Käs (Collegium Carolinum):*
Das Sudetendeutsche Wörterbuch auf dem Weg zum Forschungsinstrument
- 17:00 *Piroska Lendvai / Claudia Wick (BAdW):*
In-House Data Utilized for the Computational Semantics Task of Word Sense Disambiguation: Finetuning Latin BERT on the Thesaurus Linguae Latinae.
- 18:30 Gemeinsames Abendessen in der Stadt

Mittwoch, 5. Oktober 2022 – Technologien (Redaktionssysteme, Datenformate, APIs)

- 9:15 *Florian Zacherl (LMU):*
Extraktion von strukturierten Daten aus Wörterbuchtexten
- 10:00 *John McCrae:*
The OntoLex Lemon Lexicography Module – an ontology for linked open dictionaries
- 10:45 --- Kaffeepause ---
- 11:15 *Matthias Schöffel (BAdW):*
Untersuchung von Sprachwandel mit Hilfe von Wörterbuchdaten (Beispiel: Latein → Französisch)
- 12:00 Abschlussdiskussion