Lemmatisierung

Resch, Claudia; claudia.resch@oeaw.ac.at

In Zusammenhang mit der Erschließung von digitalen Textdaten meint Lemmatisierung die Rückführung eines vorkommenden Wortes – einer Vollform – auf seine Grundform (auch: Lemma, Nennform, Basisform oder kanonische Form), die stellvertretend für das gesamte Flexionsparadigma eines Wortes steht. So werden etwa die Wortformen *helfe*, *hilfst*, *hilft*, *helft*, *geholfen* oder *hilf* auf ein gemeinsames Lemma *helfen* zusammengeführt. Durch diesen Arbeitsschritt kann die Suche erheblich erleichtert werden: Anstatt alle Formen eines Wortes abfragen zu müssen, erhalten Benutzerinnen und Benutzer durch die Eingabe einer Grundform alle ihr zugeordneten Wortformen. Besondere Bedeutung hat die Lemmatisierung für historische Korpora mit höherer grafischer und formaler Varianz bzw. für regionale Sprachvarietäten oder Daten gesprochener Sprache. Durch die Rückführung der Non-Standard-Daten auf eine einheitliche Grundform können diese Varianten ebenfalls mit einem einzigen Suchbefehl gefunden werden.

Die Ansetzung des Lemmas erfolgt nach bestimmten Richtlinien und kann auch mit Hilfe von Tools – sogenannter 'Lemmatisierer' (*lemmatizer*) – durchgeführt werden. Diese versuchen verschiedene Wortformen mit ihrer jeweiligen Grundform zu verbinden und sind dazu mit anderen Ressourcen, etwa mit maschinenlesbaren Lexika, ausgestattet, in denen hinterlegt ist, welcher Flexionssystematik bestimmte Worte folgen. In jedem Fall muss aber für Benutzerinnen und Benutzer nachvollziehbar dokumentiert sein, nach welchen Regeln lemmatisiert worden ist.

Die Lemmatisierung ist – gemeinsam mit der Tokenisierung und der Wortartenzuordnung (*Part-of-Speech-Tagging*) – Teil der linguistischen Annotation.

Literatur:

- Perkuhn, Rainer; Keibel, Holger; Kupietz, Marc: Korpuslinguistik. Paderborn: 2012.
- Harras, Gisela; Proost, Kristel: Strategien der Lemmatisierung von Idiomen. In: Deutsche Sprache 30: 2002, S. 167–183.
- Hirschmann, Hagen: Korpuslinguistik. Eine Einführung. Mit Abbildungen und Grafiken Korpuslinguistik. Berlin: 2019, URL: `https://link.springer.com/book/10.1007%2F978-3-476-05493-7`.
- Lemnitzer, Lothar; Zinsmeister, Heike: Korpuslinguistik. Eine Einführung. Tübingen: 2010.

- Manning, Christopher D.; Raghavan, Prabhakara; Schütze, Hinrich: Introduction to information retrieval. Cambridge Univ. Press. 2008. - Google Suche. URL: https://www.google.com/search?client=safari&rls=en&q=Manning,+Christopher+D.;+Raghavan,+Prabhakara;+Sch%C3%BCtze,+Hinrich:+Introduction+to+information+retrieval.+Cambridge+Univ.+Press.+2008.&ie=UTF-8&oe=UTF-8

Software:

weblicht, Natural Language Toolkit (nltk), spacy , LemmaGen, MontyLingua, Morfette, CST's Lemmatiser

Verweise:

Textannotation, Part-of-Speech-Tagging, Weblicht, xsl-Tokenizer, NLP, NER

Themen:

Einführung, Natural Language Processing

Projekte:

Deutsches Textarchiv, Austrian Baroque Corpus (ABaC:us), travel!digital

Lexika

- Edlex: Editionslexikon

Zitiervorschlag:

Resch, Claudia. 2021. Lemmatisierung. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.115