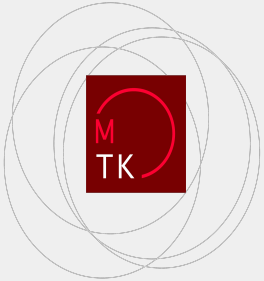
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



TEI-Praxisworkshop

Einführung in die Kodierung mittelhochdeutscher Texte

Heidelberg 13.-14. Dezember 2019

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Linguistische Annotation

Linguistische Einheiten

- **s** (s-unit) contains a sentence-like division of a text.
- **cl** (clause) represents a grammatical clause.
- **phr** (phrase) represents a grammatical phrase.
- **w** (word) represents a grammatical (not necessarily orthographic) word.

Attributen für linguistische Annotation

Beim <w> und <pc>

- **@lemma** provides a lemma (base form) for the word.
- **@lemmaRef** provides a pointer to a definition of the lemma for the word, for example in an online lexicon.
- **@pos**(part of speech) indicates the part of speech assigned to a token (i.e. information on whether it is a noun, adjective, or verb), usually according to some official reference vocabulary (e.g. for German: STTS, for English: CLAWS, for Polish: NKJP, etc.).
- **@msd** (morphosyntactic description)
- **@join** when present, it provides information on whether the token in question is adjacent to another, and if so, on which side.

Lemma und Referenzen in Wörterbücher

Verknüpfung in: Mittelhochdeutsches Handwörterbuch von Matthias Lexer

Um auf diesen Artikel zu verknüpfen, können Sie folgende URL(s) verwenden:

http://www.woerterbuchnetz.de/Lexer?lemma=minne
http://www.woerterbuchnetz.de/Lexer?bookref=1,2144,36

Attributen für linguistische Annotation

STTS (Stuttgart-Tübingen-TagSet)

https://www.sketchengine.eu/german-stts-part-of-speech-tagset/

https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/TagSets/stts-1999.pdf

Attributen für linguistische Annotation

<w lemma= “nëmen” pos= “VVIMP“ msd= “Pl.“ >Nement</w>
<w lemma= “vrouwe” pos= “NN“ msd= “Fem.Nom.Sg._“ >Frouwe</w>
<w lemma= “diser” pos= “PDAT“ msd= “Masc.Akk.Sg.“ >disen</w>
<w lemma= “kranz” pos= “NN“ msd= “Masc.Akk.Sg.St.“ >cranz</w>

[...]

<w lemma= “tragen” pos= “VVFIN“ msd= “3.Sg.Pres.Ind.“ >traget</w>

Linguistische Einheiten unter dem Wort-Ebene

- **m** (morpheme) represents a grammatical morpheme.
- **c** (character) represents a character.
- **pc** (punctuation character) contains a character or string of characters regarded as constituting a single punctuation mark.