Meta Data in LeiKo

Sarah Jablotschkin & Heike Zinsmeister
Universität Hamburg
contact: sarah.jablotschkin@uni-hamburg.de

LeiKo is a comparable corpus of German easy-to-read news texts. This freely available resource is systematically compiled and linguistically annotated for linguistic and computational linguistic research. LeiKo consists of 216 news and newspaper texts (approx. 56,600 tokens) and their meta data structured in four subcorpora according to the websites they were published on. All texts are tokenized, lemmatized, part-of-speech tagged and dependency parsed and can be queried in ANNIS (Krause/Zeldes 2016). A core corpus of 40 texts is manually corrected. Further corpus versions with additional manual annotation levels will follow.

Zenodo link: `https://zenodo.org/record/3711987`
(choose latest version on the right under “Versions”)

For each text in LeiKo, there are currently the following meta data attributes:

1 Title
2 URL
3 Source
4 Type
5 Original Text Available
6 Date
7 Authors
8 Agency
9 Genre
10 Part
11 Version

1 Title

Contains the headline of the text. In ANNIS, this attribute is called *title*.

2 URL

The URL from which the text was downloaded (might be expired). In ANNIS, this attribute is called *URL*.

3 Source

The broadcaster or the newspaper publishing the simplified texts. There are four possible values:

- ndr

- tazleicht
- sr
- nachrichtenleicht

In ANNIS, this attribute is called *source*.

4 Type

The variant of simplified German. There are two possible values:

- einfache Sprache
- Leichte Sprache

In ANNIS, this attribute is called *type*.

5 Original Text Available

Indicates whether there is an original text in Standard German that served as a basis for the simplified text. There are two possible values:

- yes
- no

In ANNIS, this attribute is called *original_text_available*.

6 Date

The publishing date of the text. Indicated in the format DD.MM.YY. In ANNIS, this attribute is called *date*.

7 Authors

The author(s) of the text (if provided). In ANNIS, you can search for the authors of each text via the attributes *author1* and *author2*.

8 Agency

The agency that created or checked the text. In ANNIS, this attribute is called *agency*.

9 Genre

Genre of the text. There are four possible values:

- column: The text was labelled “column” if it was explicitly introduced as such in the text itself.
- commentary: The text was labelled “commentary” if it was explicitly introduced as such in the text itself.

- interview: The text was labelled “interview” if it was a translation or a summary of an interview in Standard German.
- news: All other texts were labelled “news”.

In ANNIS, this attribute is called *genre*.

10 Part

Indicates to which part of the corpus the text belongs. There are two possible values:

- core: The core corpus was automatically tokenised, lemmatised, part-of-speech tagged and dependency-parsed. Additionally, the syntactic segmentations, pos and dependency annotations were manually corrected. A revised version of the corpus will contain additional manual annotation levels.
- extended: The extended corpus was automatically tokenised, lemmatised, part-of-speech tagged and dependency-parsed. Additionally, the syntactic segmentations were manually corrected.

In ANNIS, this attribute is called *part*.

11 Version

Indicates the corpus version. In ANNIS, this attribute is called *version*.