Testing the validity of Wikipedia categories for subject matter labelling of open-domain corpus data

Supplementary Material

Description of Supplement

The data underlying this study are provided in one supplementary Excel file (*supplement_02_dataset.xlsx*). The file contains the following data, stored as separate sheets:

1.	extracted Wikipedia taxonomy:	sheets *Taxonomy_Visual* and *Taxonomy_Tabular*
2.	labelled test corpus	sheet *Corpus_Labelled*
3.	annotator feedback from survey data	sheet *Survey*
4.	anonymous annotator profiles	sheet *Annotator_Profiles*

Licensing information

1.	extracted Wikipedia taxonomy	CC BY-SA 4.0	https://creativecommons.org/licenses/by-sa/4.0/deed.en
2.	labelled test corpus		
2.1	Wikipedia texts	CC BY-SA 4.0	https://creativecommons.org/licenses/by-sa/4.0/at/deed.en
2.2	BNC texts	BNC User License	http://www.natcorp.ox.ac.uk/docs/licence.pdf
2.3	labelling of test corpus	CC BY-SA 4.0	https://creativecommons.org/licenses/by-sa/4.0/deed.en
3.	annotator feedback from survey data	CC BY-SA 4.0	https://creativecommons.org/licenses/by-sa/4.0/deed.en
4.	anonymous annotator data	CC BY-SA 4.0	https://creativecommons.org/licenses/by-sa/4.0/deed.en

Note: The column *License* of the sheet *Corpus_Labelled* specifies the license for each text in the test corpus.

Note: The raw BNC texts are **not** included into this dataset to ensure compliance with §2 of the BNC User License. For a non-public copy of the BNC texts used in the test corpus, please contact the authors