Neuroscience Corpus - Dataset Description

1. source_text - Segment in English.

2. target_text - Segment in French.

3. domain - one of the three domains targeted in the project, here LS5.

4. disciplines - Metadata collected from data source; where available.

5. publication_type - Type of publication the segment was collected from, one of these values: article, conference paper, abstract, journal article abstract, report, research journal article, review abstract, thesis abstract

6. publication_source - Metadata collected from data source; the name of the journal/website/university/etc.

7. URL_source - URL address of the page the source segment is coming from.

8. URL_target - URL address of the page the target segment is coming from. Note: when EN and FR segments are collected from the same page, URLs are also the same.

9. title_source - Metadata collected from data source; the title of the publication the segment is coming from, in English. Not always available.

10. title_target - Metadata collected from data source; the title of the publication the segment is coming from, in French. Not always available.

11. keywords_source - Metadata collected from data source; keywords from the publication the segment is coming from, in English. Not always available.

12. keywords_target - Metadata collected from data source; keywords from the publication the segment is coming from, in French. Not always available.

13. author - Metadata collected from data source; author or authors of the publication the segment is coming from. Sometimes not available.

14. language_variety - Not explicitly declared metadata, but if a segment was collected from a source with a high likelihood of French language variety, this column will have this variety. Possible values: fr-CA, fr-CH, fr-FR

15. publication_license - In case publications from the same data source can have different licences, this licensing information is collected and saved in this column.

16. similarity_score - Semantic similarity score calculated between source and target segment, a score between 0 and 1.