SNF Horizonte Corpus

|  |  |  |
| --- | --- | --- |
| HTML | | |
| English  136 html pages | German  136 html pages | French  136 html pages |

|  |  |  |
| --- | --- | --- |
| PIPELINE: | | |
| Scrape html pages | Script: SNF\_html\_crawler.py |  |
| Convert html2xml | Script: convert\_html2xml.py |  |
| Tagging | Treetagger-cmd |  |
| Convert TreeTagger output (.txt) to valid corpus XML | Script: ? (perl)  Author: Martin Volk |  |
| Prepare for InterText input | Script:  build\_parallel\_corpus\_for\_  sentence\_alignment\_v2.pl  Author: Martin Volk |  |

|  |  |  |
| --- | --- | --- |
| PDF | | |
| English  NEW: 100 – 114 (2014-2018) | German  NEW: 96 – 118  MID: 81 – 95  OLD: 66 – 80 | French  NEW: 96 – 118 (2013-2018)  MID: 81 – 95 (2009-2012)  OLD: 66 – 80 (2005-2009) |
|  | | |

|  |  |  |
| --- | --- | --- |
| PIPELINE: | | |
| Scrape PDFs from web | Script: PDF\_collector.py | Notes: doesn’t automatically name the pdfs files accordingly |
| Extract text with tet | tetml --page |  |
| Simplify tetml to xml | For NEW PDFs:  convert\_tetml2xml\_new\_v2.py  Author: MP/tkew  For MID PDFs:  Conver\_tetml2xml\_mid.py | Dependent on: Content\_Parser.py  TODO: check parallel-ness of article alignment. |
| Combine split sentences |  |  |

TODO:

* Check for number of noisy articles, e.g. PDF 103 pg. 24 has noise from underlying map image.
* Sentence merging for English. (MP – DE/FR)