Issues and resources for multilingual digital research

Many digital tools and platforms were originally designed for English-language texts, and while some have made efforts to support a wider range of language, you are likely to run into language-specific issues in your research. Below are some examples of issues you might face, which are largely adapted from Quinn Dombrowski's forthcoming tutorial on 'Preparing non-English texts for computational analysis' (*Modern Languages Open*, 2020):

- **Character encoding:** Unicode (UTF-8) is now widely used across writing systems for encoding and displaying text. As a result, most text analysis tools assume the use of Unicode, and Unicode also allows you to easily move across different languages (e.g. in a bilingual parallel edition). However, earlier many different standards were used, many of which were language-specific and which you may encounter in older digital text archives. While encoding information is not included in text file properties, with a plain text editor like Atom (https://atom.io/) you can try out different encodings to find one that makes the text readable with the convert-file-encoding add-on (https://atom.io/packages/convert-file-encoding).

- **Segmentation:** For languages like Chinese and Japanese you may need to artificially insert spaces between characters to "segment" your text into words for computational text analysis.

- **Stopwords:** Stopword lists are language-specific lists of common words usually filtered out for computational text analysis. While you can often configure these yourself, many tools such as Voyant have built-in lists for multiple languages. However, these are very variable in terms of the types of words they include (e.g. the Russian list includes numbers but the Spanish list does not). As a result, you should always check which words are included first, and modify these based on your knowledge of the language and the types of texts you are analysing.

- **Lemmatizing:** Particularly for languages with many grammatical cases and complex verbal systems, it may be necessary to lemmatize your text before computational text analysis. This is also a standard procedure in corpus linguistics, where texts are linguistically annotated to assign words with their dictionary form or 'lemma' when preparing linguistic corpora. While for more general text analysis, lemmatization is not always necessary, it does for example mean that you only have to enter the 'lemma' in your stopword list rather than a word's many different variations. Natural Language Processing (NLP) lemmatizers have been created for different languages, although English is still best resourced.

- **Named Entity Recognition:** Named Entity Recognition (NER) can be useful for quickly identifying place or person names in a text, and tools like Recogito provide an option to create semi-automatic annotations using NER. It is important to remember that these are, however, language-specific and English is again best resourced (Recogito also offers French, Spanish and German). While multilingual tools are being developed, if you're working with a multilingual text you may need to redo the task with multiple NER tools to get the most accurate results (the same is true for other computational tasks like sentiment analysis).

Resources and tools for multilingual digital research:

- Multilingual DH (https://www.multilingualdh.org/): an international network for digital humanities research in languages other than English. You can join the mailing list to share language-specific issues and tools, and they have also developed a list of free NLP resources for different languages (https://github.com/multilingual-dh/nlp-resources)

- CLARIN – European Research Infrastructure for Language Resources and Technology resource families (https://www.clarin.eu/resource-families): overview and links to resources and tools for different languages (e.g. NER tools)

- Stanford word segmenter (https://nlp.stanford.edu/software/segmenter.shtml): segmentation schemes for Chinese and Arabic

- The Classical Languages Toolkit (http://cltk.org/): provides NLP support (e.g. lemmatization) for the languages of Ancient, Classical and Medieval Eurasia (primarily Latin and Greek but with other languages under development).

- The Italian NLP Tool (http://tint.fbk.eu/)

- TEI Internationalization and Localization (https://tei-c.org/tools/i18n/): translations of the TEI reference documentation into Chinese, French, German, Italian, Japanese, Korean and Spanish, with some including language-specific examples.