Stylistic analysis

For morphosyntactic parsing, we will rely on functions from the *udpipe* package (Straka et al., 2016). Language models are available for over 50 languages, and of course for French and English. *Udpipe* outperforms the well know *spaCy* package[1].

For measures of linguistic complexity, we will consider average word length, average sentence length, and amount of punctuation.

For measures of immediacy, we will compute ratios of use of active voice and present tense and rely on wordlists for temporal immediacy and spatial immediacy.

For measures of vividness, we will compute expressivity with the formula (# ADJ + # ADV) / (# N + # V), activation (the intensity of an affect) with Whissell (2009)'s Dictionary of Affect in Language, and dominance (the behavioral activation associated toward (appetitive motivation) or away (aversive motivation) from a stimulus) with Warriner et al. (2013)'s wordlist.

For measures of concreteness, we will rely on Muraki et al. (2022)'s wordlist.

Finally, for measures of emotional valence, we will consider Warriner et al. (2013)'s wordlist.

For the measures which do not simply derive from the parsing of parts-of-speech, we will consider *embedding-based cross-lingual transfer learning*. For activation, dominance, concreteness and emotional valence, available dictionaries of terms are in English. To be able to process other languages and terms, we will use our multilingual sentence-transformer to compute the embedding of each term in the various dictionaries. For each measure, we will then adopt a supervised approach to learn the mapping between the numerical embedding and the target score (support vector machines or extreme gradient boosting). For temporal and spatial immediacy, we will again use multilingual embedding to find targets in the embedded space and screen the closely related words in any language. For the detection of argumentation, we will rely on an annotated dataset

[1] https://www.bnosac.be/index.php/blog/75-a-comparison-between-spacy-and-udpipe-for-natural-language-processing-for-r-users

of short texts with or without argumentation (Stab et al., 2018) to again follow our supervised approach based on multilingual embedding, this time for a binary classification. Finally, for the characterization of evidentiality, we will consider three possible cases: no source mentioned, at least a vague source mentioned (e.g., "Some inhabitants reported…"), and at least a precise source mentioned (e.g., "the mayor of Kigali declared…"). We will build an annotated dataset by categorizing 1,000 randomly selected documents (resulting from the segmentation of the articles) and repeat our supervised approach to derive our categories of evidentiality from embeddings.

References

Muraki, E. J., Abdalla, S., Brysbaert, M., & Pexman, P. M. (2022). Concreteness ratings for 62,000 English multiword expressions. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01912-6

Stab, C., Miller, T., Schiller, B., Rai, P., & Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 3664–3674. https://doi.org/10.18653/v1/d18-1402

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of LREC 2016*, 4290–4297. http://www.lrec-conf.org/proceedings/lrec2016/pdf/873_Paper.pdf

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, *105*(2), 509–521. https://doi.org/10.2466/PR0.105.2.509-521