EUROPRO Digital Corpus Description

EUROPROwebs (European Project Websites) is a pilot corpus, which contains 30 research project websites funded under the H2020 programme, and is part of our EUROPROwebs database comprising 100 website projects. It can be considered a specialized, static corpus compiled to analyse digital academic discourse and, through its analysis, insights can be gained into international research group's evolving discursive and professional practices. EUROPROwebs corpus contains 394 072 words with a mean of 13 136 words per website.

Several criteria were followed for the selection of the 30 international funded projects and their websites.

1. The research projects should aim at producing new knowledge and not at training PhD students or professionals to undertake the research. This was done to ensure comparability within the corpus.

2. A convenience sampling method was followed, which entailed choosing research projects with at least a member from the Universidad de Zaragoza (Spain) or from a Zaragoza-based institution. This would allow us to complement our text-based analysis with contextual evidence from potential informants.

3. The date of the H2020 projects should at some point overlap with the development of our own project (2017-2021), so that the most recent digital academic practices could be studied.

4. Given their organic nature, the websites were selected and downloaded on a fixed date, which would entail that the duration of each of the projects was different at the moment of compilation. Since creating a monitor corpus (a corpus that is regularly updated) lies beyond the scope of our research project, websites had been developed to different stages when downloaded.

EUROPROtweets (European Project Twitter Accounts) pilot corpus was compiled as an extension of EUROPROwebs pilot, considering the importance of social media, and Twitter in particular. The use of Twitter is encouraged by institutional guidelines, together with other social networks, and was found to be the preferred option for dissemination purposes by H2020-funded research projects in our corpus. Subsequently, out of the 30 websites analysed, 20 were found to make use of a Twitter social account. EUROPROtweets pilot consists of 3 822 tweets containing 88 970 words, with a mean of 191 tweets and 4 449 words per research project Twitter account.

Methodological decisions for the compilation of EUROPRO Digital Corpus

EUROPROwebs pilot corpus was designed taking the following decisions.

All texts from the websites were downloaded and labelled using different codes to refer to the pages or sections of the website.

1. **DYNAMICITY** The corpus was compiled from April to May 2019. The specific start and end date of each project was retrieved in order to determine the point of development of the corresponding research project at the moment of the corpus compilation.

2. **HYPERMODALITY AND HYPERMEDIALITY**. The corpus was tagged for hyperlinks (external, internal and peripheral), for visuals (such as tables, figures, pictures, logos, etc.), for videos and audios.

3. **LAYOUT AND WEB DESIGN**. Besides downloading the texts contained in the website, screenshots of every page were taken and saved. Information was also recorded of the extent to which the text could be directly accessed from the website menu.

EUROPROtweets pilot corpus was designed taking the following decisions.

1. **DYNAMICITY**. All tweets and retweets from research project Twitter accounts were coded and downloaded at a set date, June 2019.

2. **HYPERMODALITY AND HYPERMEDIALITY**. Information regarding multimodal elements and hyperlinks in the tweets was coded. Regarding multimodality, three multimodal elements were found and, accordingly, coded: pictures, videos and GIFs.

3. **INTERACTIVITY**. Information was retrieved on the number of likes obtained by each tweet, and the number of retweets by other users. Information was also retrieved and coded on the number and range of hashtags (#) used in the Tweets by the research group and the number and range of mentions (@) to other Twitter users.