World Academy of Science, Engineering and Technology
International Journal of Humanities and Social Sciences
Vol:15, No:12, 2021

Tagging a corpus of Media Interviews with Diplomats: Challenges and Solutions

Authors : Roberta Facchinetti, Sara Corrizzato, Silvia Cavalieri

Abstract : Increasing interconnection between data digitalization and linguistic investigation has given rise to unprecedented potentialities and challenges for corpus linguists, who need to master IT tools for data analysis and text processing, as well as to develop techniques for efficient and reliable annotation in specific mark-up languages that encode documents in a format that is both human and machine-readable. In the present paper, the challenges emerging from the compilation of a linguistic corpus will be taken into consideration, focusing on the English language in particular. To do so, the case study of the InterDiplo corpus will be illustrated. The corpus, currently under development at the University of Verona (Italy), represents a novelty in terms both of the data included and of the tag set used for its annotation. The corpus covers media interviews and debates with diplomats and international operators conversing in English with journalists who do not share the same lingua-cultural background as their interviewees. To date, this appears to be the first tagged corpus of international institutional spoken discourse and will be an important database not only for linguists interested in corpus analysis but also for experts operating in international relations. In the present paper, special attention will be dedicated to the structural mark-up, parts of speech annotation, and tagging of discursive traits, that are the innovational parts of the project being the result of a thorough study to find the best solution to suit the analytical needs of the data. Several aspects will be addressed, with special attention to the tagging of the speakers' identity, the communicative events, and anthropophagic. Prominence will be given to the annotation of question/answer exchanges to investigate the interlocutors' choices and how such choices impact communication. Indeed, the automated identification of questions, in relation to the expected answers, is functional to understand how interviewers elicit information as well as how interviewees provide their answers to fulfill their respective communicative aims. A detailed description of the aforementioned elements will be given using the InterDiplo-Covid19 pilot corpus. The data yielded by our preliminary analysis of the data will highlight the viable solutions found in the construction of the corpus in terms of XML conversion, metadata definition, tagging system, and discursive-pragmatic annotation to be included via Oxygen.

Keywords : spoken corpus, diplomats' interviews, tagging system, discursive-pragmatic annotation, english linguistics
Conference Title : ICDH 2021 : International Conference on Digital Humanities
Conference Location : Tokyo, Japan
Conference Dates : December 02-03, 2021

Open Science Index, Humanities and Social Sciences Vol:15, No:12, 2021 publications.waset.org/abstracts/143495.pdf