Challenges of Digital Data for Linguistics: New Units and Hidden Processes

Manuel Alcantara-Pla*[1]

[1]Universidad Autonoma de Madrid – Spain

Abstract

Internet has provided us with an amount of data without precedents. For those who research in Linguistics, it is an unexpected gift with a huge potential. The most obvious difference of the Internet discourses -also known as Computer Mediated Communication (CMC) discourses- with the previous ones is that every interaction is registered and can be retrieved. Unlike other data, texts have not suffered a typical process of digitalization (loosing information). Digital texts can be easily stored, copied, and modified. It makes communication more complex because of phenomena such as remediation, but it also gives us a great opportunity for studying discourse on a large scale.
However this gift comes with important challenges that we have to face. I focus here on two specially relevant for data analysis: the definition of units of analysis and the influence of the platforms we get the data from. Though the cases will be taken from the discourse analysis discipline, both challenges are general to Digital Humanities.

Social media and mobile phones are two inventions that have radically changed our relation with the Internet. Users are (and are requested to be) "more active, participatory, and collaborative" (Heyd 2016, 90). Many new communication genres have been created in the last decades (blogs, wikis, social networks, and micro-blogs). All of them are multimedia, and texts compete in importance with images, videos, and sounds. The most recent platforms (such as YouTube and Instagram) explicitly relegate texts to a secondary role.

When we call "chats" to interactions made up of written texts, videos, and images, we are using this term in a very broad sense. Most of the mentioned genres are asynchronous, even those more conversational. Interaction strategies are not the same as in spoken conversations, they are written or multimodal instead of spoken, and they follow new politeness rules. Therefore they do not correspond with traditional conversations, and the definition of the units of analysis of these new genres is still work in progress (Alcántara-Plá 2014).

The work of defining these new units has to meet two basic requirements. On the one hand, being multimodal communication, texts can not be analyzed independently. Words are interconnected with images and videos, and they must be studied as such. Limits between different modes have become fuzzy, and the different disciplines that are focused on them should take it into account.

On the other hand, boundaries have also become fuzzy in another sense. Spoken conversations can be divided into smaller units, from the conversation itself to turns, utterances, and

*Speaker

lexical units. However, these units are difficult to delimit in the Internet. Digital conversations can start in a platform and continue in a different one: a tweet might end up as part of a Whatsapp interaction or embedded in a video in YouTube. Digital information is very easily remediated, and it is a frequent habit in the Internet.

Regarding the second challenge we face when using digital data for research in Humanities, the design of the platforms we get the data from should be taken into account in our analysis. This design is a key element in the context of any mediated communication, and rarely acknowledged as such. The main problem here is that most designs are not open and they can only be interpreted analyzing their user's behavior. It is not possible to analyze them as independent variables. In order to have the most reliable information, collaboration with the owners of the platforms and their designers is crucial.

Taking the discourse analysis as example, the mentioned characteristics (its units, politeness, multimodality...) are not free choices made by the users, but affordances of the platforms where the communication is taking place. The technological mediation of the digital communication and its hidden processes determines the language we use when communicating in digital contexts. Therefore affordances and restrictions of every platform should be part of the data we use for our research.

Keywords: CMC, digital data, units, analysis, Linguistics