”On the record” Transcribing and valorizing qualitative interviews with XML-TEI

Florian Cafiero, Marie Puren

"On the record"

Transcribing and valorizing qualitative interviews with XML-TEI

Florian Cafiero (1) and Marie Puren (2)

(1) CNRS / Université Paris Sorbonne - Groupe d'Etude des Méthodes de l'Analyse Sociologique de la Sorbonne, (2) CNRS - Laboratoire de recherche historique Rhône-Alpes

Qualitative interviews constitute an important research tool for disciplines such as history, sociology, ethnology or political science. Yet, despite rare initiatives (Cadorel et al., 2018), transcriptions are scarcely shared with other researchers. And their annotation is most of the time done only for a personal use, without following any sort of standard, and not meant to be shown to anyone.
In this paper, we advocate for the necessity of a more open management of these resources, and present a proposition for a XML-TEI-conformant standard, allowing for their accurate transcription and annotation. The ODD we present is aimed at facilitating systematic analyses of corpora of interview transcriptions, as well as at ensuring a better dissemination and re-usability of these resources. We rely as much as possible on existing TEI elements, but introduce a new element and a new attribute, to address the specificities of this kind of materials.

Interviews: a precious resource

Figuratively and literally, interview transcriptions are a precious resource. Producing them comes at a high cost: researchers must dedicate a lot of time and money to organize the interviews, travel, speak with the interviewees and transcribe the interviews. They should thus be used and re-used to their fullest potential. Finding a way to properly encoding them will help further qualitative or quantitative analyses by the researcher as well as by other colleagues taking interest in the source. Sharing effectively these resources would allow for comparisons between results obtained by researchers from various disciplines, at different periods and places, or to build larger corpora to obtain new results.

Fig. 1: Various disciplines use qualitative interviews

Addressing the reproducibility crisis

Human and social sciences have been targeted by many critics during the "replication crisis" (Ioannidis, 2005 ; Camerer et al., 2018) controversy. Research relying on qualitative analyses are not easily subject to the same reproducibility assessment, but should in many cases allow for "comparative re-production" (Markee, 2017): in a similar context, and following the same principles and questioning, will a new interview lead to results comparable to the ones previously made? Making one's transcriptions and annotations available would be a way to make this possible. It would also ensure that conclusions drawn from an interview are trusted, as a reviewer, colleague, or reader in general, could access the annotations underlying the researcher's analyses
We propose to create an ODD (One Document Does It All), setting out which TEI elements and associated attributes can be used and in which context, and documenting our choices to future users. An ODD also enables to add new elements and attributes. And within a given community, it is possible to agree on an available ODD customization that will ensure the interoperability, shareability and reusability of the TEI files. To create this ODD, we mostly combine the elements and attributes declared by the modules "Transcription of Speech" and "Language Corpora", and propose to add one new element and one new attribute.

Ethics and respect of privacy

A key concern in sharing qualitative interviews should be the respect of legal constraints and ethics principles. We thus propose the creation of a new element, to annotate passages that could not be freely shared. This element allows for a description of the deleted passage and the reason for its deletion. It is meant to ensure the protection of interviewees, while concealing as little relevant information as possible.
The creation of this new element relies on TEI best practices, and draws on the use of the element <damage>, employed to encode the damages done to a primary sources - for example by indicating that some text is lacking, and by supplying the lacking part with alternate text. The new element <privacy> works in the same way to mention that a passage has been deleted and why it has been deleted. As well as the <damage> element, <privacy> bears attributes such as "unit", "quantity" and "extent" to precise the length of the deletion. The reason of the deletion could then be expressed within an attribute "reason", more suitable in this context than the attribute "agent" born by <damage>. An element <desc> enclosed in the element <privacy> gives also more information on the causes of the deletion (legal reasons, ethics code, personal moral judgement). We also propose an alternate encoding strategy by enclosing a <gap> element within <privacy> to provide more information on the deletion. Moreover if the transcriber wishes to replace the deleted passage - e.g., replacing the name of a person with a pseudonym for ease of reading -, he or she may also use a <supplied> element enclosed within <privacy>.

Reflecting on one's interview practices

The <u> element is used to encode the different parts of speech given by the interviewers and the interviewees, with a "who" attribute to express who is the speaker. But qualitative interviews are not ordinary conversations: they are prepared by a researcher, implementing a strategy to get as much information as possible on a topic of interest. It is thus crucial to encode the researcher's comments on its own speech (Beaud, 1996): was the question prepared? spontaneous? what was its purpose (changing the subject/knowing more/confirming a previous statement etc.) ? This is why we propose to add a "type" attribute to the list of already existing attributes born by <u>, describing this kind of information.
In addition to simple content annotations (persons or places cited, dates evoked etc.), our model offers the possibility of sharing one's interpretations about relevant passages of the transcription.
We propose to use the <seg> element bearing an "xml:id" attribute to delimit the parts of the speech that are in need of further analyses. These analyses can then be provided via a element bearing a "target" attribute to identify which <seg> element is concerned, and a "type" attribute to express the nature of this analyses. The <interpr> element may be used in conjunction with the element, but <interpr> is more suitable to identify various parts of speeches under unique conceptual categories. Associating identified parts of speeches and specific conceptual categories is easy with TEI pointer mechanisms: for example, an "ana" attribute born by a <seg> element enables to associate this element with an <interp> element bearing an "xml:id" attribute.

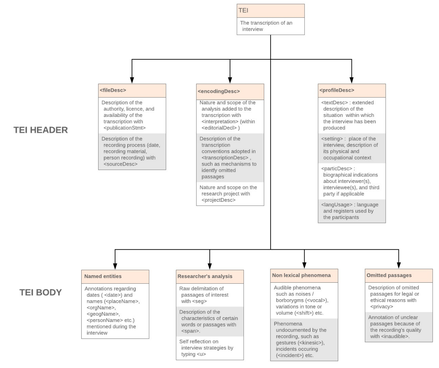

Fig. 2: ODD for quantitative interviews: excerpt of the proposed structure

- BEAUD, Stéphane. L'usage de l'entretien en sciences sociales. Plaidoyer pour l'entretien ethnographique. *Politix. Revue des sciences sociales du politique*, 1996, vol. 9, no 35, p. 226-257.
- CADOREL, Sarah, et al. beQuali: Une Plateforme d'Archives Numériques en Sciences Sociales. In : *Proceedings of the 1st International Conference on Digital Tools Uses Congress*. ACM, 2018.
- CAMERER, Colin F., DREBER, Anna, HOLZMEISTER, Felix, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2018, vol. 2, no 9, p. 637.
- IOANNIDIS, John PA. Why most published research findings are false. *PLoS medicine*, 2005, vol. 2, no 8, p. e124.
- MARKEE, Numa. Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. *Language Teaching*, 2017, vol. 50, no 3, p. 367-383.
- The TEI Guidelines, [online]. [Accessed 22 October 2019]. Available from: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html