

Defining a protocol and assessing natural language metadata for a Databank of Oral Teletandem Interactions (DOTI)

Paola Leone

University of Salento
P.zzetta A. Rizzo, 1
Lecce (Italy)

paola.leone@unisalento.it

Abstract

The current study addresses the definition of a protocol for collecting, storing data and describing (in a simple and generic way) a repository. Particularly, the transparency of a form aimed at gathering information about the pedagogical context of oral telecollaboration for language learning named Teletandem (TT; Telles, 2006) will be tested before it is spread more widely. To uncover problems in submitting information, data-input-triggers quality and reliability have been tested interviewing professors and language instructors who will be involved in a preliminary phase of Teletandem corpus implementation. General goals of the study are to enlarge the research group, to increase data and to improve efficiency in data collection.

Keywords: corpora, data collection, metadata, telecollaboration, language learning, protocol

1. Teletandem as learning context

Teletandem (TT) is a learning context, in which pairs of native and non-native speakers of different languages talk alternatively in their L1 and L2 in order to learn each other's native language (www.teletandembrasil.org). The communication between partners is multimodal, via video calls and chat (e.g. by employing Skype, MSN). Students "virtually meet" during sessions that normally last one hour. From a pedagogical perspective, Teletandem has a positive impact on students' learning experience for different reasons: i) it is a form of learning from peers (Hanushek, Kain, Markman, & Rivkin, 2003), ii) it is extensible to perform and develop different forms of plurilingualism (e.g. to foster L2 production and/or to enhance interlocutors' ability to mutually understand each other when speaking different languages). Furthermore, since communication is enabled by VoIP technology, TT is a context for practicing computer mediated interactions which are becoming more and more used for job application and for university admission. The existence of websites which give "essential Skype interview tips" is a proof of the future relevance of this technology transmitted interaction. The positive impact of Teletandem in language learning (e.g. Leone & Telles, 2016) is a good basis for expecting a gradual, but sustained increase of its practice in higher education. This trend calls for further empirical research and implies a high demand for video/audio data.

2. Background information

Bearing the above in mind, two universities, in which for

several years TT has been experienced, have undertaken a project which aims at organizing already existing Teletandem data and at defining natural language metadata. Teletandem databank is currently named Databank of Oral Teletandem Interaction (DOTI; Aranha & Leone, 2016; forthcoming) and it is build out of video recordings of Teletandem sessions thus data includes participants' moving visual images as well as voice and chat texts. Metadata allow the description of the context in which video recordings have been done and they are organized hierarchically. They stem from computer mediated interaction standardized metadata (Chanier, et al., 2014) and from pedagogical research (Mangenot, 2008); the former displays general characteristics of computer mediated interaction, the latter concerns the learning context and are currently not yet in a standardized form.

Following the general framework of "interaction space" (Chanier, et al., 2014), which includes comprehensively guidelines for describing all CMC genres in a multimodal perspective, teletandem as learning context is characterized in terms of: a) participants (i.e. n.2); b) location, meaning online location that is b.1) how interactions are transmitted (e.g. via VoIP technology, via mail etc.), b.2) where data are originally recorded (e.g. university server), b.3) the place where teletandem sessions happen (i.e. at the university or outside the institution); c) time frame which shows the beginning and the end of each session, but also the length of the telecollaboration project with specific information (days of the week, year, month). Concerning the technological environment, Teletandem is multimodal - i.e. diverse communicative modalities are used: audio-video via

VOIP technology and Internet Relay Chat (IRC)- ; communication is synchronous; it is a dyadic exchange since two people are involved; it implies oral, written, gestural and iconic modes of communication; language used can be different.

For recording characteristics related to the pedagogical implementation of Teletandem, the concepts of “learning scenario” and task have been referred to. The hierarchical organization of the learning scenario encompasses as well concepts of macrotasks (e.g. described in terms of number, typology) and microtasks.

3. Research

The general goal of the current research is to enlarge the research group involved in data collection, to increase data and to improve efficiency in data collection. Hinging on the above mentioned studies (Aranha & Leone, 2016, forthcoming), we aim at developing a protocol which offers guidelines and rules for forthcoming data collection (Aranha & Leone, forthcoming). In the protocol two main implementation levels are laid down. Level 1 consists in collecting Teletandem interactional video recordings as “raw data”, storing them and describing video file contents using natural-language definition metadata (Aranha & Leone, 2016, forthcoming). Professors and/or language instructors will carry out this activity uploading video files in a cloud storage system and submitting as well - in a digital form - information which highlights properties of the specific TT learning context - i.e. information range from the institutional profile for the course in which TT has been institutionalized (Aranha & Cavalari, 2014) to the tasks assigned for the whole activity-. In level 1, researchers, professors and language instructors must agree on which terms to be used for property description and on which definition terms have. In level 1 DOTI will be enriched. Level 2 encompasses all those actions aimed to create an interoperational databank. It consists in implementing metadata interoperability by checking the compatibility of level 1 metadata with standardized metadata (e.g. Text Encoding Initiatives and Dublin Core Metadata Initiative). In this phase of corpora development, metadata will give access to digital data and they will be targeted to applied linguistic researchers.

The current research sticks into level 1 actions and it is a test of the clarity of the form which will be used to collect information about Teletandem as learning context and then for describing Teletandem databank. Informants are colleagues who already employed Teletandem in teaching. Research questions are: Are language instructors and professors familiar with terms employed as natural language metadata for describing Teletandem as a learning context in Higher Education? Does terminology facilitate entry of information? How can metadata be improved so to be more user friendly? The methodology for collecting colleagues’ opinions is an in depth interview supported by a written questionnaire. Interviews have been recorded. Informants have to fill the

form in (i.e. the assumption was “the form is well defined if colleagues can submit proper information in the right place”) and then discuss if each question and/or concept description was/were clear (or not) and, if unclear, how it could be improved. The form contains data-input-triggers which are originated from the metadata provided for describing already collected data.

4. Results and conclusion

A first analysis of the findings shows that for some form fields (and subfields) there is coherence between required and given information (e.g. for the concept of learning scenario), whereas in other entries there was no match between purposes and information (e.g. macrotask and microtask are not easy to understand). To place interface terminologies in context, results will be discussed considering the kind of background informants show in the field of foreign language teaching. Suggestions on how natural language metadata could be improved and how the form interface layout could be implemented will be given.

References

- Aranha, S., & Cavalari, S. (2014). A trajetória do projeto Teletandem Brasil: da modalidade institucional não-integrada à institucional integrada. *The ESPecialist*, 35(2), 70-88.
- Aranha, S., & Leone, P. (2016). DOTI: Databank of Oral Teletandem Interactions. In S. Jager, & Kurek, M. (Eds.), *New directions in telecollaborative research and practice: selected papers from the second conference on telecollaboration in higher education* (pp. 327-332). Research-publishing.net.
- Aranha, S., & Leone, P. (forthcoming). State of the arts and DOTI-Databank of Oral Teletandem Interaction.
- Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Seddah, J. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 2(29), 1-30.
- Hanushek, E., Kain, J., Markman, J., & Rivkin, S. (2003). Does peer ability affect student achievement? *Applied econometrics*, 527-544.
- Leone, P., & Telles, J. (2016). The Teletandem network. In T. Lewis, & R. O'Dowd (A cura di), *Online Intercultural Exchange: Policy, Pedagogy, Practice* (p. 243-248). London: Routledge.
- Mangenot, F. (2008). La question du scénario de communication dans les interactions pédagogiques en ligne. *Jocair (Journées Communication et Apprentissage Instrumentés en Réseau)*, 13-26 .
- Telles, J. A. (2006). *Projeto Teletandem Brasil: Línguas Estrangeiras para Todos – Ensinando e Aprendendo línguas estrangeiras in-tandem via MSN Messenger*. Faculdade de Ciências e Letras de Assis, UNESP.