



HAL
open science

**Virtual exchanges as complex research environments:
facing the data management challenge. A case study of
Teletandem Brasil**

Solange Aranha, Ciara R. Wigham

► **To cite this version:**

Solange Aranha, Ciara R. Wigham. Virtual exchanges as complex research environments: facing the data management challenge. A case study of Teletandem Brasil. *Journal of Virtual Exchange*, 2020, 3, pp.13-38. 10.21827/jve.3.35748 . hal-02917206

HAL Id: hal-02917206

<https://hal.science/hal-02917206v1>

Submitted on 18 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aranha, S. & Wigham, C.R. (2020). Virtual exchanges as complex research environments: facing the data management challenge. A case study of Teletandem Brasil. *Journal of Virtual Exchange*, 3. pp. 13-38. <https://doi.org/10.21827/jve.3.35748>.

Virtual exchanges as complex research environments: facing the data management challenge.

A case study of Teletandem Brasil

Solange ARANHA and Ciara R. WIGHAM

Abstract

Although there is a move towards Open Data, with research funding bodies more frequently requiring data management plans and dissemination strategies, the data management challenges inherently linked to virtual exchange research are understudied. Data collection is often reported upon in papers addressing interaction analysis or language development, but little attention has been paid to offering critical discussion of data collection and structuration methods or practical advice to encourage data / corpora dissemination. This paper reports on two phases of the Multimodal Teletandem Corpus project (Aranha & Lopes, 2019) that structured 581 hours of video data from Portuguese-English teletandem sessions, 351 chat logs, 956 written productions exchanged between the partners (original, revised and corrected versions), 91 initial and 41 final questionnaires, and 666 learning diaries. We describe the data management problems faced that included the organization of data collected, ethical consent, management of a large quantity of data, inclusion of sociolinguistic information, expansion of learning theories and the solutions found. We then outline data management planning steps that, consequently, are being introduced for future telecollaboration instantiations.

Keywords: corpora, teletandem, telecollaboration, data management, data collection, LEarning and TEaching Corpora (LETEC)

Introduction

Telecollaboration or virtual exchange “engages groups of students in online intercultural interaction and collaboration with partner classes from other cultural contexts or geographical locations under the guidance of educators and/or expert facilitators” (Lewis & O’Dowd, 2016:3)¹. When organisers wish to offer students a pedagogical exchange but also capture their interactions for research purposes, data becomes a central issue. Whilst the data lifecycle traditionally emphasized data collection, analysis and publication of results, in recent years, there has been an increasing interest in the ‘new data lifecycle’ which focuses on data sharing, preservation, and reuse (DDI Alliance, 2013; Briney, 2015).

One challenge for research into telecollaborative exchanges is the complexity of the research environment: (i) the participants’ exchanges occur in online as well as face-to-face spaces, (ii) researchers involved are frequently from different institutions, (iii) colleagues conduct research from distant locations and may have different research cultures. Indeed, most telecollaborative exchange research is conducted by researchers on exchanges for which they themselves were one of the pedagogical coordinators. Whilst complete participation will shape the events the researcher considers as important and relevant to the research inquiry and allow deeper insights into the data, there is also the risk of losing objectivity. Another challenge is that telecollaborative exchanges often generate a wealth of data meaning that some data may not be analysed due to the time-consuming task of transcription or the sheer volume of data collected.

¹ See O’Dowd (2018) for a discussion of the terms ‘telecollaboration’ and ‘virtual exchange’.

Given the complex nature of any telecollaboration terrain, data management planning needs to become an integral part of research. Several initiatives have applied the new data lifecycle to telecollaborative exchanges to facilitate data dissemination and repurposing (Mulce, 2013; Guichon, 2017; Vimelf, 2018) and allow external researchers access to data from telecollaborations in which they were not initially involved. However, these practices are far from being common and little attention has been paid to offering critical discussion of data collection and structuration methods or practical advice to encourage data / corpora dissemination. With the aim of addressing this gap in the field, this paper describes how data from the *Teletandem Brasil: Língua Estrangeira para Todos* (Telles, 2006) project was collected at São José do Rio Preto campus then structured into a LEarning and TEaching Corpus (LETEC). Teletandem is considered an approach to virtual exchange where the focus is on language learning. Telles defines it as “a virtual, collaborative and autonomous context in which two speakers of different languages use the text, voice, and webcam image resources of VOIP technology to help each other learn their native language” (2015: n.p.). This paper takes the form of a case-study of the *Teletandem Brasil* project and focuses on some of the data management difficulties encountered and the data management planning steps that, consequently, are being introduced for future telecollaboration instantiations. It is hoped that it offers some useful reflections for colleagues who are setting up their own research protocols for telecollaboration projects.

The first section of this paper examines data management, firstly within the context of the data lifecycle and, secondly, with respect to LETEC and with reference to the Computer-Assisted Language Learning (CALL) field. We then outline the pedagogical and research contexts of the *Teletandem Brasil* project before detailing why it was decided to structure data from this project into the Multimodal Teletandem Corpus (MulTeC, Aranha & Lopes, 2019). Data management challenges in the processing of data that relate to consent terms,

corpus metadata and data storage are then discussed in the third section and strategies introduced within the research project to overcome these challenges explained. Based on lessons learnt, we finish by offering colleagues a series of questions for consideration in the data management planning stages of any telecollaborative exchange. These questions might serve as a toolkit for colleagues considering structuring and sharing data from other telecollaboration projects.

Data management

Research data

Data is often seen as the foundation upon which scientific knowledge is constructed and refers to “numerical quantities or other factual attributes derived from observation, experiment or calculation” (National Research Council, 1992, np). Given that the term ‘data’ is often used to refer to information that can be stored in a digital form, the term ‘research data’ as a collocate is often preferred to underline that, unlike other types of information, research data is “collected, observed or created for the purposes of analysis to produce original research results” (University of Edinburgh, 2018, np).

The Organisation for Economic Co-operation and Development suggest one definition of research data: “Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research and that are commonly accepted in the scientific community as necessary to validate research findings” (OECD, 2007, p.13). This definition places the emphasis on primary ‘raw’ data and research data as situated before the analysis and interpretation processes. However, the movement towards OpenData that is free for anyone to access, use, modify and share, implies that the term also needs to include ‘research-ready data’ or ‘compiled data’ (i.e. datasets that have already been processed, cleaned,

annotated, interpreted and made available to a scientific community) and the notion that ‘research data’ forms part of a data lifecycle and is not simply the focus of analysis and observation.

The data lifecycle

The prevalence of digital data has shifted the ways in which research data is understood with reference to terminology and has impacted how research data is treated within each stage of the research process. Traditionally, for research projects for which the main focus was on the publication of research results, data-centric activities were those of data collection and data analysis (Figure 1). Indeed, Briney (2015) explains that data did not play a central role in the planning of the research project or, apart from excerpts, in reporting on the data analysis in publications. Publications were the output of the research project and data was viewed as a by-product of research rather than an output, having value in itself.

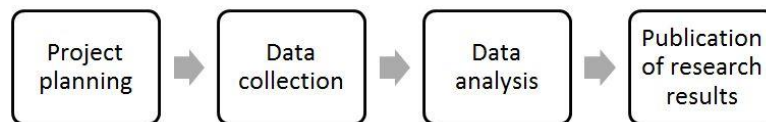


Figure 1. Traditional data lifecycle

As data has become more frequently digital, and researchers are able to do more with data in terms of data curation and preservation, it has become possible for data to have a lifespan that is longer than the research study itself. Thus, in the 1990s and early 2000s, the data lifecycle concept was promoted as a way to support data sharing, preservation, reuse and repurposing beyond the original study for which the data was collected. This led to models in which data-centric activities play a role in each step of the research project from project planning to the publication of research results and research data itself.

Humphrey (2006) proposed a research knowledge creation lifecycle model adapted to digital data (Figure 2). In the model, each chevron represents a separate stage of knowledge creation. The knowledge transfer cycle stage refers to dissemination of the research via conference presentations and journal papers. The transition points between each stage are shown by the gaps between the chevrons and represent points in the lifecycle model where the project may become victim to information and data loss and for which the study design should include data management tasks in order to reduce these potential losses. Data is crucial throughout all project stages and is one of the research outputs that may then feed into a future research project.

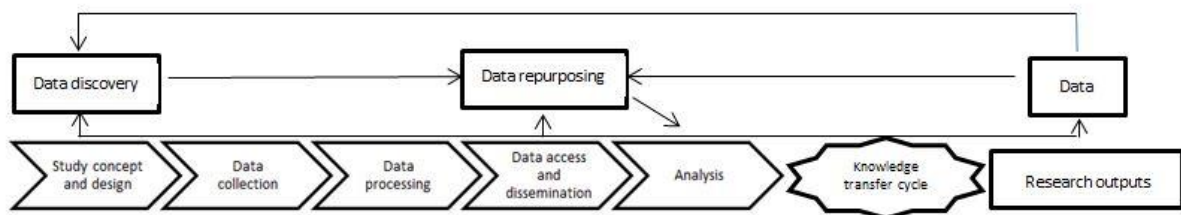


Figure 2. Humphrey's 2006 lifecycle model

The Data Documentation Initiative (DDI Alliance, 2013) proposed another conceptualisation of the data lifecycle, specifically designed for the social sciences (Figure 3). Again, data is seen as an actual product of the research, rather than a by-product, and is central to each step of the research project, from the initial planning steps to the publication of the research data and its long-term conservation. The circular nature underlines that data from one research study or project can lead into and contribute to a new project.

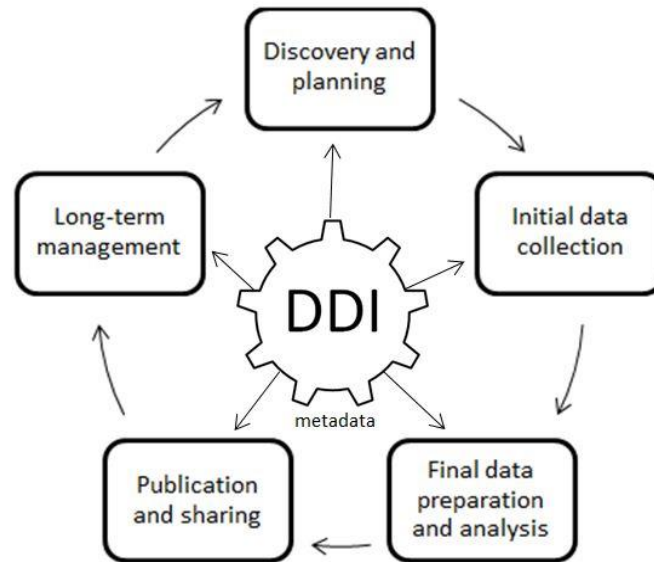


Figure 3. Data Documentation Initiative Lifecycle model

Data lifecycle models are useful to help researchers visualise the different stages of a research project and consider both the diverse aspects of data management that need to be planned for at each stage and actions that need to be accounted for throughout the project: They allow data management to be strategized and integrated into the workflow.

Data management and CALL

Data management, a term that arose in the mid-2000s, is most commonly regarded with respect to a data management plan that “helps design, put into practice, and follow up on how research data are collected, organized, used and looked after to achieve the highest quality and long-term sustainability” (Corti et al, 2014, p.25). Often required by funding agencies, it establishes, for each stage of the research cycle, the actions needed and the different collaborators’ roles and responsibilities. This is particularly important within collaborative research projects that involve different institutions, including projects around telecollaboration, as the data management actions will not always be the responsibility of the researcher who has collected the data and research management actions depending on the

researchers' academic cultures will not always be similar. Thus, it is a necessity to explicitly discuss and document the data management steps during the project planning phase.

In CALL, the LEarning and TEaching Corpora (LETEC) paradigm has been developed by Reffay et al. (2008), Chanier & Ciekanski (2010) and Chanier & Wigham (2016) as one possible staged methodology for collecting, structuring, publishing through open-access repositories and reusing learner-computer interaction data. This methodology has been applied to virtual exchange projects including the *Simuligne* project based on a global simulation for learning French as a foreign language that used email and forum exchanges and that was inspired by the *Cultura* project (Reffay et al., 2014); the German-French virtual exchange project *Infral* that focused on exchanges via blogs and an audio-conferencing platform (Abendroth-Timmer et al, 2014); and the Second Life InterCulturel telecollaborative project between trainee teachers of French as a foreign language in France and learners of French in the United States (Bayle et al., 2013).

Detailed in Chanier & Ciekanski (2010), the LETEC corpus paradigm is guided by four central principles:

- 1) *Systematic Data Collection* involves the collection of the whole data set from one instantiation of the exchange, including the learner interactions, productions and log files. Although an individual researcher may be interested in pursuing one particular research question for which the study of only one specific kind of interaction may be necessary, systematic data collection of the whole data set is a prerequisite for LETEC as it will allow other researchers in the future to reuse the data and study it from various perspectives or in the light of other research questions. Furthermore, it ensures quality for future data analysis as a researcher will have the data available to show that either a selected subset of data does not correspond to a simple disconnected

episode under analysis but rather is representative of other interactions during the online course, or on the contrary, is disparate and thus worthy of investigation.

2) *Detailed Data Description*. As the aim of LETEC is to encourage data dissemination and reuse, the description of both the learning scenario for the virtual exchange and the research protocol that was followed to collect the data are both essential in order that researchers not involved in the initial data collection may understand the context from which it stems. Data description is often referred to as 'metadata'. It involves the explanation of how data has been collected, edited and organised; sociolinguistic information about the participants; explanations of the learning scenario, tasks and outcomes and the roles of teachers/tutors during the exchange; descriptions of the platforms used during the exchange; descriptions of the research instruments used to collect data and anonymization procedures, and information about the researchers involved in data collection and organisation and their roles. For data stemming from virtual exchanges, this information is vital in order that data can be both understood by researchers not involved in the initial project but also, for the researchers involved in the virtual exchange themselves so that they have the contextual information concerning how the learning scenario fits with wider learning programmes and curricula at the partner institution(s) and differences in the instantiation for the different institutions involved in the exchange.

3) *Data Formatting and Conversion*. Research data collected during virtual exchanges may exist in many different forms: textual data, numerical data, geospatial data, images, audio-video recordings, machine-generated data (log files) etc. Often, the format is specific to a software programme that allows the data to be read and interpreted. Converting data to standard and interchangeable formats ensures the longer-term usability of the data and should facilitate the compatibility of data

between different programmes used in distinctive analysis stages or in a variety of analysis approaches. When choosing formats for digital data, Corti et al (2014) recommend that the “choice should be planned early in the research cycle to ensure that the format suits all purposes that may be necessary” (p57). This involves considering the formats that are most suitable for data creation, analysis, long-term sustainability and sharing as well as opting for open versus proprietary formats. Moving beyond file formats, researchers often use standards for the representation of texts in digital form such as Extensible Markup Language (XML) or Text Encoding Initiative (TEI) to align and structure sets of computer-mediated communication data that include different data formats (cf. TEI CMC workgroup).

- 4) *Data Release and Distribution*. Data sharing can occur at many levels. Researchers need to consider, firstly, how to organise data so that analyses and outputs from a previous study can be used as a starting point for a new study. This involves ensuring that data from different research projects are structured in a similar manner to ease “sharing with your future self” (Briney, 2015, p.145). Secondly, whether data can be easily accessed and understood between researchers from the range of institutions involved in the virtual exchange or shared with researchers in new collaborations? Lastly, public data sharing where the whole LETEC and its related analysis can be freely accessed and this access is guaranteed as permanent. The focus on reproducibility and complementary analyses are central to the increased push towards data sharing. Within the area of virtual exchange, this could encourage the comparison of different learning scenarios across different languages and different platforms. Of course, this LETEC principle must allow for the management of sensitive data through data enlightenment, guided consent and data anonymisation procedures, and be guided by ethics and consent protocols.

Clearly, the corpus paradigm principles fit closely with data lifecycle steps: Systematic data collection aligns with the need to plan data management and data acquisition; detailed data description relates to documenting - a prerequisite for data sharing; data formatting and conversion is related to long-term storage, and data preservation, release and distribution concern publication, data sharing and data re-use.

We now turn to a case-study to detail how data from the *Teletandem Brasil: Língua Estrangeira para Todos* (Telles, 2006) project was collected before being structured into a LETEC.

Pedagogical context

Teletandem Practice: Foreign Language for all

Teletandem Brasil is a Brazilian telecollaborative project for learning foreign languages (Telles, 2006; www.teletandembrasil.org). The project has been implemented at three campi of São Paulo State University and in partner universities including Universidade del Salento (Italy), Georgetown University (USA), University of Georgia at Athens (USA), University of Sheffield (United Kingdom) and Arizona State University (USA). Over a twelve-year period, Teletandem Brasil has allowed more than one thousand students to experience telecollaborative learning.

Learning design

At the São José do Rio Preto campus, two teletandem modalities are offered: the integrated and the semi-integrated. Students meet weekly from the teletandem laboratory on campus and from a language lab in the foreign university, usually during classtime.

The institutional integrated modality (Aranha & Cavalari, 2014) has been implemented since 2011 and is followed by roughly 100 students every year. This modality implies that the telecollaborative project is incorporated in regular language courses and implemented by pedagogical tasks developed for students (Aranha & Leone, 2017; Aranha & Cavalari, 2014; Cavalari & Aranha, 2016). The authors understand pedagogical tasks as the ones carried out both in class and in the context of telecollaboration with pedagogical purposes linked to the overall objectives of one specific course. The learning design, typically for a eight-week course, is composed of two macro tasks: (i) participating in the synchronous oral sessions (TOSs, Teletandem Oral Sessions) following the principles detailed in Section 3.1, and (ii) engaging in mediation sessions, which are face-to-face group discussions about problems encountered during TOSs, achievements related to different competences, self-evaluation about the student's own learning process, and other issues raised by the participation in the project . To help prepare and accompany these tasks, other micro tasks are completed by the students. These include: (i) answering questionnaires - a pre-course questionnaire to establish learning goals and self-evaluate proficiency and a post-course questionnaire to assess the benefits of the practice and the extent to which goals were met; (ii) attending a tutorial that gives students an overview of the project; (iii) writing reflective diaries after each TOS; (iv) writing texts in the language one is learning; (v) correcting written productions from one's partner (Figure 4).

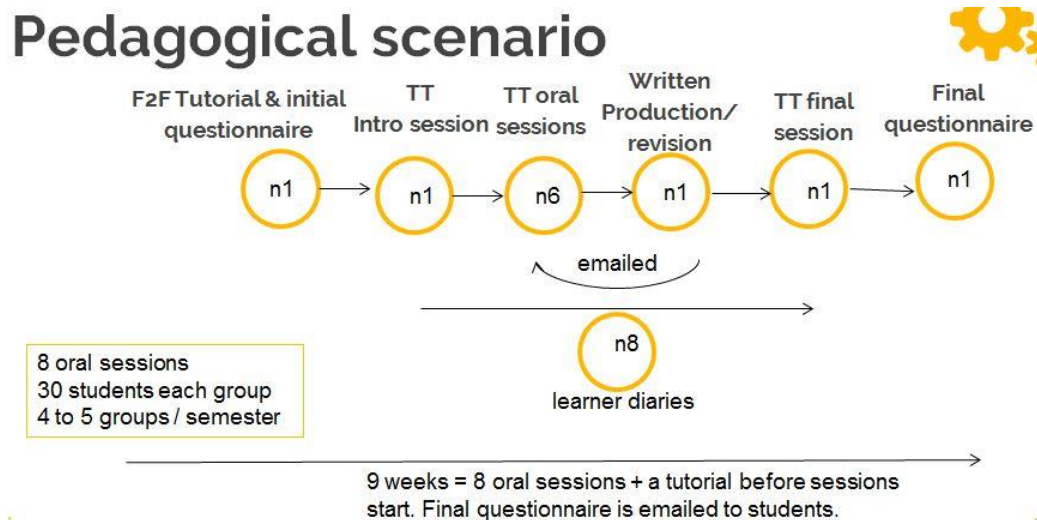


Figure 4. The flow and tasks of one group project (n= frequency).

Besides participating in the TOS and completing the tasks each TOS encapsulates, students also join the mediation sessions, held every fortnight on a face-to-face basis.

For students who participate in the semi-integrated modality, during which one institutional partner integrates the teletandem project into regular language classes whilst the other group is composed of volunteer students who participate in their free time, similar tasks are completed. However, as Aranha and Leone point out (2017, p.179), the two macro tasks (mediation sessions and the TOSs) are the only compulsory tasks in the teletandem pedagogical scenarios and the micro tasks that feed into these macro tasks may vary from one instantiation to another. Around eighty students participate in this modality every year at São José do Rio Preto campus.

Research context

The Teletandem Brasil project (Telles, 2006) has, since its creation, been the focus of research studies in the field of Applied Linguistics (see <http://www.teletandembrasil.org/publications.html>). Besides being a pedagogical project, whose aim is to develop linguistic and cultural skills, the research staff also study diverse

aspects of this learning environment. The amount of data generated by learners within the Teletandem Brasil context has been huge but until lately had never been compiled: Data that was the focus of each research study was collected and treated individually, and used solely to fulfil the purposes and needs of specific research questions. Indeed, as each new Master's or PhD student integrated the project, or permanent researchers explored a different research angle, new data was collected. Previously, no standardized framework for the collection of visual, spoken and written data was established and multimodal data analysis procedures were not agreed upon within the overarching project. Therefore, one may find a variety of procedures for data collection, processing and analysis in the different academic publications that relate to the project. Long-term data planning was not considered prior to each research project. This practice underlines that preparing data for reuse was perceived by these young researchers as time-consuming and, sometimes, unviable. Permanent researchers were then pushed to consider a more systematic data collection approach for teletandem instantiations that took place between 2012 and 2015 in order to make data available, via initially a databank (Aranha, Luvizari-Murad & Moreno, 2015) for the on-site researchers involved in the project.

Data collection

Data from 16 different groups (11 integrated and 5 semi-integrated) were systematically collected between 2012 and 2015. Figure 5 describes the type of data that comprises the MulTeC corpus.

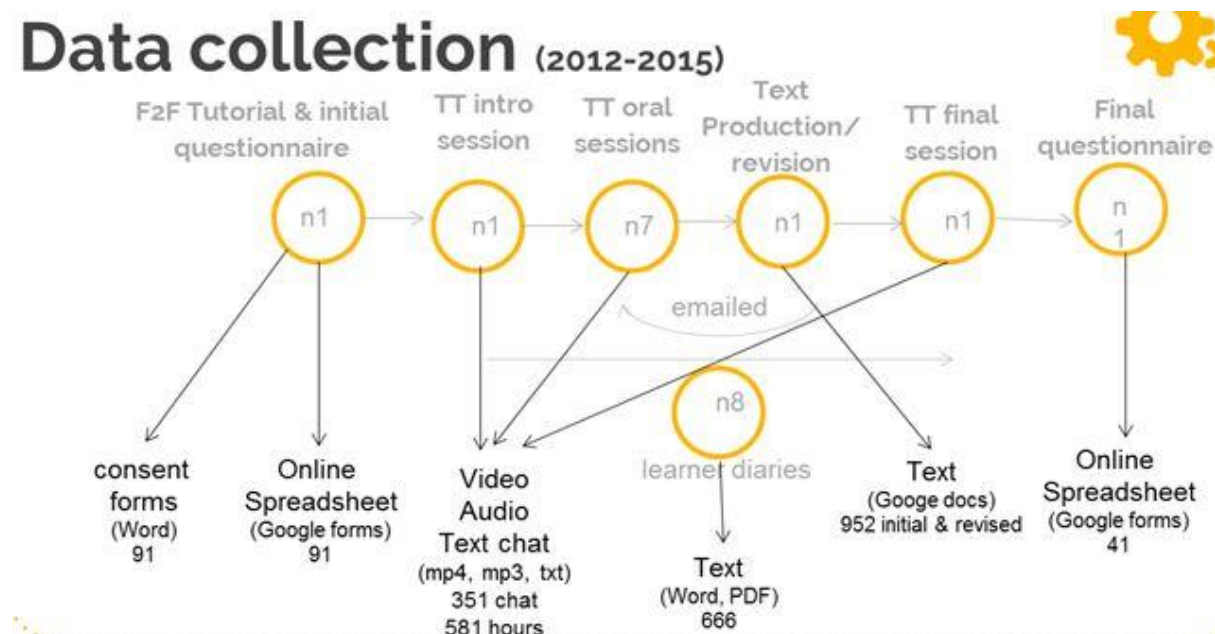


Figure 5. Types of oral and written data at MulTeC

Video recordings were made of the TOSs using Evaer (www.evaer.com). Each session was recorded on a specific machine in the teletandem lab and later transferred to a hard disk. Video files were labelled according to the semester/group/machine the individual used. Written material - reflective journals and written productions in three versions (first draft, commented text and final text)- were uploaded first to the pedagogical platforms (TELEDUC from 2012-2014 and later to Moodle 2015 before later being shared via secure cloud storage). Questionnaires were initially administered using paper-based versions but nowadays online questionnaires are used. This centralised data collection was then structured into the MulTeC databank for use by Master's and Doctorate students as well as permanent researchers.

Data Processing: The MulTeC Corpus

Given the positive response from the research team to the teletandem databank (Aranha, Luzivari-Murad & Moreno, 2015), in 2016 the permanent researchers involved in the project decided to take the data management process one step further and transform the databank into

a corpus by following LEarning and TEaching Corpora principles (Chanier & Wigham, 2016). The objectives behind this transformation were (i) to encourage collaboration among other researchers interested in investigating telecollaborative learning practice who could not come to the teletandem lab, (ii) incite research studies that are cumulative, contrastive and longitudinal, (iii) promote internationalization of data and (iv) be able to publish the data alongside research studies.

The MulTeC corpus (Aranha & Lopes, 2019), built from the teletandem databank, structures the data of 282 teletandem participants from both the integrated teletandem and semi-integrated modalities. Representing 145 GB of data, it comprises 581 hours 19 minutes of video data from TOSs, 666 learning diaries, 351 chat conversations used during TOS, 956 texts (both in Portuguese, written by L2 students and commented upon by L1 native speakers and in English, written by L2 students and commented by a proficient English speaker), 91 initial questionnaires and 41 final ones.

The path to build the MulTeC corpus was not a smooth one; however, it taught us many lessons that will be used to collect, compile and organize other corpora. We now illustrate three areas of data management for which our procedures changed. These changes were introduced to facilitate the future sharing of the corpus with external, as well as on-site, researchers.

Data Management: Lessons learnt and future data collection facilitation

This section considers, firstly, three important aspects for further data collection in this context: (i) consent terms; (ii) metadata; (iii) data preservation and storage. Secondly, we show how changes have been made to project and data management planning to overcome and manage these issues for current and future teletandem instantiations.

Consent terms

As data was to be collected for a corpus that was, in turn, to be shared with third-party researchers, consent was a delicate yet important issue.

The MulTeC database comprised data from UGA (University of Georgia at Athens) and UNESP/IBILCE (São Paulo State University at São José do Rio Preto). When collection started in 2011, both universities created locally their consent forms. UGA had their consent terms approved by their Institutional Review Board whilst UNESP had theirs approved by the local ethics committee. The result was that, depending on the institution in which they were enrolled, students signed different documents and consented to different terms.

Consequently, we faced several issues when trying to track, for each teletandem pair, whether both students had agreed for their data to be included in MulTeC. These included:

- a) The problem of tracking students one by one;
- b) The difficulty that if one student refused to sign the consent form, the use of the teletandem partnership's data was impaired even if the student from the other institution had agreed to the consent terms. As all data was recorded, this required researchers to find and delete captured data for which consent had not been given;
- c) The students' interpretations of the statements and their wording created odd consent terms, e.g. a student allowing the use of his/her video recordings for research purposes but not his/her image.

To facilitate the integration of data into the MulTeC corpus, adjustments were made. Written consent procedures are now discussed by both partners and students from both institutions are asked to agree to the consent terms as approved by the UNESP Ethics Committee. If desired, partner universities may distribute local consent forms that adhere to their institution rules.

Nevertheless, to be part of data collection for integration into the corpus, both sides have to sign the Brazilian research agreement. This has helped to streamline the consent process. Before asking participants to agree to share their data, we decided to offer tutorials during which permanent researchers (i) explain the procedures and the importance of collecting data for research purposes, (ii) answer questions, (iii) reassure participants regarding confidentiality and anonymization and (iv) inform them that they may decide to leave the project at any time and the procedure to do so. Teletandem oral sessions are not recorded unless both participants agree to the same consent terms. Thus, if only one student gives consent, the teletandem pedagogical partnership may still occur, but will not be subjected to data collection.

Metadata

The TI (Teletandem Identity) coding scheme to catalogue and label the data files from the participants was initially created for the MulTec database (see Aranha & Lopes, 2019). It consists of information concerning the student's institution, course, gender and Skype number. In compiling the MulTeC database, one problem faced by the team was to establish the gender of each participant, as this was not initially included in the sociolinguistic part of the questionnaire and could only be determined by watching the video. The female/ male assignment could only be based on observation and not on participant's orientation/identity. Besides, the research team felt that other sociolinguistic information would also be relevant, particularly to share data with external researchers who did not know the students first-hand and who did not have the possibility to contact them to ask for additional information. Nowadays, having in mind that the TI is a 'unique identification (ID) so that later it can be easily listed and described' (Chanier & Wigham, 2016, p.230), participants inform the following metadata in the first questionnaire:




- Gender (male, female, other);
- Age;
- First language: (not necessarily interaction language); second or a third language (previous research shows that they sometimes interact in three languages);
- Proficiency in their foreign languages. This information was self-determined after the students consulted the CEFR descriptors;
- School background. Whether they come from private / public schools may determine target language level. This metadata is relevant to try to track the same students over a longitudinal period;
- Experience in target language learning before college (language schools).

This information is collected via an online form (Appendix A) .

Besides the inclusion of sociolinguistic metadata, allowing longitudinal studies and specific target group research, corpus metadata related to the pedagogical design is included so that external researchers who will work on the data in the future, and who were not involved in the pedagogical project, can better understand the learning conditions in which the exchange took place. The concept of learning scenarios describes what happens in the real learning situation; it is the description of what actually occurs in a given learning context and includes specific characteristics of each group. A document was thus designed for each teletandem instantiation to record pertinent information regarding the scenario design (pedagogical scenario) and its actual instantiation (learning scenario).

The document is organized in five tabs. The first one, named scenario design (Figure 6), includes specific information about the two groups: modality of teletandem (integrated, semi-integrated), institutions (home and partner institution), groups involved in both universities,

professors in charge (in case of integrated modality), mediators (staff that facilitates the sessions in the lab), dates of mediation sessions and oral sessions, length of the project, days of oral sessions, specificities about timezone, number of sessions, discourse type (free conversation, task realization, specific theme discussion), discourse typology (intercomprehension or alternate monolingualism), Voip technology, number of texts expected to be exchanged, number of diaries, and dates of the tutorial and of the questionnaires.

MODALITY		
INSTITUTIONS		
CLASSES		
PROFESSORS		
MEDIATORS		
MEDIATION		
PERIOD		
DAY		
Time		
MARCH		
APRIL		
TOSs #		
PLACE		
DISCOURSE TYPE	() Free conversation	() Free conversation
	() Task realization	() Task realization
	() Specific theme discussion	() Specific theme discussion
TPOLOGY	() alternate monolingualism intercomprehension	()

+ ☰
Scenario design ▾
Attendance ▾
TOSs ▾
Mediation ▾
Follow up ▾


Figure 6. Scenario design

The second tab (Figure 7) presents an attendance list to be completed by the staff during each session. This document facilitates the anonymisation process: researchers know if any partnerships changed pairs due to absences on a specific day.

PAIR	E-MAIL	STUDENT NAME	PHONE #	SKYPE USER	IT	TOSs Dates			
1									
2									
3									
4									
5									


Figure 7. Attendance list

The third tab lists any incidents which occur during the session. The first column describes the plan for that specific oral session, e.g. *“Oral Session I - Getting to know each other. No previous pairing. No text to share or review”*. The second column includes any occurrence during the session, as in a third session of a group: *“Most interactions started five minutes late. Natássia took even longer to start her interaction (technical problems in ASU). Letícia's partner complained she couldn't hear her very well, so Brittany had to change computer. Manuela's case was even worse cause she couldn't contact her partner before 17:00 p.m. She complained her partner is always late.”* In the process of anonymising data, all names are substituted by the TI (teletandem identity).



teletandem brasil
línguas estrangeiras para todos


Teletandem Oral Sessions (TOSs)



TOS	Date	WORK PLAN	UNESP OCCURENCES	XXXXX OCCURENCES
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				


Figure 8. Session records

The fourth tab is intended to be completed after each face-to-face mediation session, describing which aspects were proposed and how they were developed and discussed by students (Figure 9).



teletandem brasil
línguas estrangeiras para todos

MEDIATION



UNESP MEDIATION SESSIONS			
SESSION	DATE	DESCRIPTION	OBSERVATIONS
1			
2			
3			
4			
5			

XXXXX MEDIATION SESSIONS			
SESSION	DATE	DESCRIPTION	OBSERVATIONS
1			
2			
3			
4			
5			

Figure 9. Mediation session record

The fifth tab presents a follow-up table that is completed after each project is over. It intends to provide a quantitative overview of each task (Figure 10).

The inclusion of the learning scenario concept and the document based on it introduced a system of session record sheets to track participation. These are completed by professors or lab assistants and include any observations about the session (i.e. what changed from the learning design to the actual pedagogical instantiation). In future corpora built from teletandem exchanges, this information will be integrated from the outset in order that external researchers can cross-reference this metadata with the data upon which they are working.

Data preservation / storing data

When data was collected for the MulTeC databank, teletandem groups used a platform from UNESP named TELEDUC (the platform has now been replaced by Moodle). The functionalities of the platform had been designed for pedagogical purposes by local staff and students used it in their blended courses. All the written files collected within the context of the institutional integrated teletandem modality in São José do Rio Preto were supposed to be uploaded to the platform in .doc format. As for the oral data, the sessions were recorded by EVAER (www.evaer.com) and left on each computer in the teletandem lab for later storage, organised by lab assistants, but which potentially could have exposed the data to third-parties, e.g. other students using the teletandem lab should students have forgotten to log off their individual sessions. To overcome this, it is important that teletandem sessions take place in labs where students use unique identifiers and passwords to log on to the institutional computing system.

When the staff decided to collect data for an unpretentious databank, a coding system was created (see Aranha, Luvizari-Murad & Moreno, 2015). At the end of every session, each Brazilian student was responsible for saving each video file in which their SOT had been recorded on the local hard drive of each computer. However, the responsibility for naming and uploading files depended on students' digital literacy skills. This resulted in the incorrect

naming of files and the loss of data. Another problem encountered was that the Brazilian lab had obsolete computers resulting in frequent system errors and an unstable Internet connection. This made lots of data unusable. In 2016, we attempted to overcome these constraints by asking the American partner University to collect the data. However, we realised that this was not a viable solution because the sheer amount of data collected was too heavy to transfer to UNESP via online tools.

After encountering these problems, nowadays, written data is stored in a collective online storage space that is shared by both universities (Google Drive). Paper-based questionnaires have been transformed into online questionnaires allowing us to curate the data and store this online. Students also write their texts and correct them online using collaborative writing tools. Raw data can thus be accessed by both sides, downloaded by the UNESP researchers and transformed out of its proprietary formats in order to be integrated into the corpus.

To facilitate the integration of data into the MulTeC corpus, a unique identification coding system to uniquely identify a resource type among the overall data set, was negotiated, agreed upon and adopted by both institutions. Staff members are responsible for collecting, naming and storing written and oral data after each session. Thanks to FAPESP (São Paulo Research Foundation) grant # 16/18705-9 - *Institutional Integrated Teletandem: building a multimodal databank for Applied Linguistics Research* - the lab staff can be maintained and further collections and successful cataloguing of the data should be guaranteed.

Indeed, lab assistants are now employed on student contracts to anonymise oral data file according to the IT instructions (Aranha & Lopes, 2019) and convert written data into .txt files for long-term data preservation. In addition to naming the files according to the unique identify coding system, metadata is also included in each file. For example, for the video recordings of the SOT sessions, this includes:

1. The file name formatted according to the unique identification coding instructions e.g.
2016_I27F12_UGA1i_SOTin1>
2. <recording place TTD Rio Preto>
3. <recording date 12 June de 2015>
4. <Transcribing date 14 October 2016>
5. <type of authorship>
6. <languages Portuguese and English>
7. <Authors I27F12 e U0M12>
8. <Duration of file 40'32''>
9. <number of words 5.132>
10. <Source Teletandem Rio Preto>
11. <Turn indication B=brasileiro E=Estadunidense>
12. <Transcriber >

Introducing these procedures has minimised the amount of data that has been lost due to human errors concerning data manipulation, has allowed both partner institutions access to data and has helped the corpus-building partner more easily process and prepare the data dissemination. In addition, the procedures will facilitate future external researcher's ease of use of the corpus, allowing them, for example, to easily choose files to study that have similar characteristics (e.g. files for which students are completing the same task and for which they took a similar time to complete) and to cross reference different data types generated by the same student within the overall data set (e.g. to link one student's SOT recordings with his/her diary entries).

Planning for data management prior to teletandem exchanges

Based on the Teletandem Brasil experience, there are several considerations that we would encourage colleagues to discuss whilst designing a research protocol around a teletandem exchange in order to facilitate data management.

To strategize data management, we have framed these considerations as open-ended questions under five sub-categories (Table 1) which closely fit with the new data lifecycle model. The first category ‘What data will the virtual exchange create?’ relates to data collection. Data processing aspects are grouped together under ‘How will the data be documented and organised?’ before steps to consider regarding data access and storage. For colleagues interested in data dissemination, we hope the questions listed under ‘What is necessary to enable data reuse’ may be used as discussion prompts to encourage the publishing of data as a research product in itself. Finally, long-term data management aspects are included. Although we have not discussed these final issues in this paper, it seems essential if virtual exchange data is going to be considered a research product in itself to plan for the long-term preservation of research data. If we consider the range of synchronous tools and platforms available for virtual exchanges that educators have at their disposal, as well as the rapidity regarding technical development of these, the need to store data in interchangeable formats is crucial: Considering long-term preservation of virtual exchange data from the outset of current projects would allow the research field, in years to come, to be able to compare data from different exchanges in order to analyse features of virtual exchange that are related to the features of a specific platform being used and those which are cross-platform.

What data will the virtual exchange create?

What data will be collected during the telecollaboration project?

Which policies will apply to telecollaboration data (legal, institutional?)

Are written consent procedures in both countries similar? Will they allow the same analyses of data?

Will partners produce separate consent forms or a generic form for the overall project?

Who is responsible for which part of data management?

Can supporting documentation be saved?

How will the data be documented and organized?

How will the data be organised?

In what formats will the different data types be saved?

What file naming schema will be used?

How will different data types be catalogued, labelled?

How will different versions of the data be catalogued, labelled?

How will the data be stored?

Will the data be held in one safe location or multiple locations? Do these locations fit with legal and institutional protocols?

How will data be shared? How can the project ensure that different partners have full access to the data?

Will data need to be backed-up at regular intervals? What are they? Who manages this?

Has data storage been costed into the project?

Who will have access to different versions of the data?

What is necessary to enable data reuse?

Has data reuse been considered in the consent forms?

How can you preserve personal data long-term for use in future?

How might you make data available to future users? How can you ensure third parties follow ethics guidelines for data use?

How can you ensure third parties will understand the data? What metadata needs to be introduced to aid this process?

How will data be managed after the project?

How can the readability of files be ensured? Will file formats need to be updated in order not to become obsolete?

Will storage hardware need updates?

If colleagues change institutions, how will data be managed?

Table 1. Teletandem data management planning

Conclusion

This paper aimed to underline the importance of data management planning steps related to the complex research terrain of virtual exchange. The methodological case study described the organization of the Multimodal Teletandem Corpus (MulTeC) for which the aim is to give access to researchers, both those involved and not involved in the pedagogical teletandem project, to conduct collaborative and/or complementary or indeed replication research studies (cf. Smith & Schulze, 2013) on the same data. LEarning and TEaching Corpora procedures were followed to structure the MulTeC corpus. However, this experience shed light on data management difficulties relating to consent terms, corpus metadata and data storage that the research team had not anticipated. Following a description of these

issues, we offered explanations of the steps introduced to overcome these difficulties in future instantiations of the project.

Our case study underlines the necessity to agree upon common consent terms between partner institutions prior to the project and offer face-to-face tutorials to participants to encourage them to contribute their data. It suggests possible sociolinguistic metadata to collect to facilitate longitudinal studies as well as allow researchers to delineate the data to address specific research questions. Regarding corpus dissemination, introducing metadata related to learning scenarios was also necessary so that colleagues not involved in the pedagogical instantiation can better understand the learning conditions under which data collection occurred. Building on Chanier & Wigham (2016), a file naming system was also introduced to facilitate data storage given the wealth of data the project has generated.

Despite the movement towards OpenData, very few virtual exchange projects foresee data as a research output in its own right despite the benefits that structuring and sharing corpora of virtual exchange data could bring to our research community. Guichon (2017) underlines several of these, including the division of time-consuming transcription and annotation practices, widening of the empirical basis and scope of subsequent studies and, from an epistemological viewpoint, analysis of the corpus from different research perspectives, theoretical standpoints, and methodological approaches. Research into virtual exchange has focused on pedagogical scenarios and interaction analysis but further work is needed regarding data collection and structuration methodology to encourage data/corpora access and dissemination. Perceiving data as a research product, and not simply a by-product, upon which researchers from an international research community can collaborate could potentially empower the virtual exchange transglobal research community by offering more consolidated efforts to cross geographical and epistemological boundaries and, subsequently, strengthen the associated body of knowledge (Dooly, 2015).

We acknowledge that the scientific ambition of sharing corpora requires considerable institutional support, especially regarding personnel both in terms of time required for the data structuration and dissemination and the technical training these processes may demand. Certainly, a mind-set shift is required whereby individual researchers no longer think of producing one-off analyses on individual learning situations but look towards long-term team research projects. Whilst this mind-set shift might not be so easy for established researchers, introducing data curation skills into PhD and post-doc programmes to better equip the future generation of virtual exchange researchers could be a first step. In Europe, measures are being introduced. For example, the European Research Infrastructure for Language Resources and Technology (CLARIN, 2019) offers services and tools to researchers, including user involvement events and mobility grants for training in corpora approaches and the ORTOLANG repository (ORTOLANG, 2012) aims to construct a network infrastructure including a language data repository and offers online spaces for researchers to deposit and collaborate on the analysis of corpus data.

Negotiating better institutional recognition for virtual exchange will also contribute to promoting this scientific ambition. For instance, the work achieved by UNICollaboration (UNICollaboration, n.d.) to promote virtual exchange at institutional and policy-making levels and its involvement in Erasmus+ Virtual Exchange. Finally, with the ‘publish or perish syndrome’ (Colpaert, 2012), course institutions need to support their researchers by recognising corpora production as an important step within research. Modifying teaching loads for researchers involved in such collaborative projects or acknowledging corpus production within professional evaluation practices would help encourage researchers to spend time building corpora. For example, in France, the national committee for higher education and research evaluation now recognises corpus publication as equivalent to a rank A research article within the field of Linguistics (HCERES, 2018). It is also necessary to

consider the creation of standards for bibliographic entries for corpora so that they can be recognised in citation indexes as scientific publications.

Given that data management plans and data dissemination are more frequently required by public funding agencies who encourage international collaborations, we hope that the toolkit offered in this case study may help colleagues interested in collecting data from virtual exchanges explicitly strategize their data management planning and dissemination steps.

REFERENCES

- Abendroth-Timmer, D., Bechtel, M., Chanier, T. & Ciekanski, M. (2014). *Corpus d'apprentissage INFRAL*. Banque de corpus CoMeRe. Ortolang.fr: Nancy.
<https://hdl.handle.net/11403/comere/cmr-infral>.
- Aranha, S., Lopes, Q. B. (2019). Moving from an internal databank to a sharable multimodal corpus: the MulTeC case. *The ESPECIALIST*, 40(1). <https://revistas.pucsp.br/esp/article/view/41127>.
- Aranha, S., Leone, P. (2017). The development of DOTI (Data of oral teletandem interaction). In D. Fiser & M. Beibwenger (Eds), *Investigating computer-mediated communication: corpus-based approaches to language in the digital world*. Ljubljana: University Press, Faculty of Arts. pp.172-190. <https://research-publishing.net/publication/chapters/978-1-908416-41-4/525.pdf>.
- Aranha, S., Luvizari-Murad, L., & Moreno, A. C. (2015). A criação de um banco de dados para pesquisas sobre aprendizagem via teletandem institucional integrado (TTDii). *Revista (Con)textos Linguísticos*, 9(12), 274-293.
<https://periodicos.ufes.br/contextoslinguisticos/article/view/9633>.
- Aranha, S., & Cavalari, S. M. S. (2014). A trajetória do projeto Teletandem Brasil: da modalidade Institucional Não Integrada à Institucional Integrada. *The ESPECIALIST*, 35(2), 183-201. <https://revistas.pucsp.br/esp/article/view/21467>.
- Bayle, A., Youngs, B., & Foucher, A.-L. (2013). *Learning and Teaching Corpus (LETEC) of (Second Life InterCultural) SLIC*. Mulce.org : Clermont Université. [oai: mulce.org:mce_slic_letec_all]
- Briney, K. (2015). *Data Management for Researchers*. Exeter: Pelagic Publishing.

- Cavalari, S. M. S., & Aranha, S. (2016). Teletandem: integrating e-learning into the foreign language classroom. *Acta Scientiarum: Language and culture*, 38(4), 327-336.
<http://periodicos.uem.br/ojs/index.php/ActaSciLangCult/article/view/28139>
- Chanier, T. & Ciekanski, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. *Apprentissage des Langues et Systèmes d'Information et de Communication*, 13.
- Chanier, T. & Wigham, C.R. (2016). A scientific methodology for researching CALL interaction data : Multimodal LEarning and TEaching Corpora, In Hamel, M-J. & Caws, C. (Eds.). *Learner Computer Interactions : New insights on CALL theories and applications*. Amsterdam : John Benjamins. <https://halshs.archives-ouvertes.fr/LRL/edutice-01332625v1>
- CLARIN (2019). European Research Infrastructure for Language Resources and Technology.
<https://www.clarin.eu/>
- Colpaert, J. (2012). The “Publish and Perish” syndrome, *Computer Assisted Language Learning*, 25(5), 383-391.
- Corti, L, Van den Eynden, V., Bishop, L. & Woollard, M. (2014). *Managing and Sharing Research Data. A guide to good practice*. London: Sage publications.
- DDI Alliance (2013). *Data Documentation Initiative*. <http://www.ddialliance.org>
- Dooly, M. (2015). It takes research to build a community. Ongoing challenges for scholars in digitally-supported communicative language teaching. *CALICO*, 32(1), pp.172-194.
- Guichon, N. (2017). Sharing a multimodal corpus to study web-mediated language teaching. *Language Learning and Technology*, 21(1). pp.56-75
- HCERES (2018). *Guide des Produits de la Recherche et Activités de Recherche : Linguistique*.
<https://www.hceres.fr/sites/default/files/media/downloads/Guide%20des%20produits%20de%20la%20recherche%20et%20des%20activit%C3%A9s%20de%20recherche%20-%20Sous-domaine%20SHS%204%20-%20Discipline%20-%20linguistique.pdf>
- Humphrey, C. (2006). *e-Science and the Life Cycle of Research*, University of Alberta.
<http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>
- Lewis, T. & O’Dowd, R. (2016). Introduction to Online Intercultural Exchange and this volume. In O’Dowd, R. & Lewis, T. (Eds.) *Online Intercultural Exchange: Policy, Pedagogy, Practice*. New York and London: Routledge. pp.3-20.

- Mulce (2013). *MULTimodal Corpus Exchange repository for LEarning and TEaching Corpora*. <http://mulce-doc.univ-bpclermont.fr/>
- National Research Council (1992). *Setting priorities for space research: Opportunities and imperatives*. Washington, DC: The National Academies Press.
- O'Dowd, R. (2018). From telecollaboration to virtual exchange: state-of-the-art and the role of UNICollaboration in moving forward. *Journal of Virtual Exchange*, 1, pp.1-23. Research-publishing.net.
- OECD (2007). *Principles and Guidelines for Access to Research Data from Public Funding*. <http://www.oecd.org/sti/inno/38500813.pdf>
- ORTOLANG (2020). *Ortolang Repository*. <https://www.ortolang.fr/>
- Reffay, C., Chanier, T., Noras, M. & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche. *Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation*, 15. [oai:edutice.archives-ouvertes.fr:edutice-00159733].
- Reffay, C. Chanier, T. Lamy, M.-N. & Betbeder, M.-L. (2014) Corpus d'apprentissage Interactions Simuligne (Simulation en ligne en apprentissage des langues). In Chanier T. (ed.) Banque de corpus CoMeRe. Ortolang.fr : Nancy. <https://hdl.handle.net/11403/comere/cmr-simuligne>.
- Smith, B., and Schulze, M. (2013). Thirty years of the Calico Journal – replicate, replicate, replicate. *CALICO*, 30 (1), i–iv.
- 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. <http://www.teletandembrasil.org/>
- UNICollaboration (n.d.) Cross-disciplinary professional organisation for telecollaboration and virtual exchange in Higher Education. <https://www.unicollaboration.org>
- University of Edinburgh. (2018). *How to manage research data: Defining research data*. <https://www.ed.ac.uk/information-services/research-support/research-data-service>.
- ViMELF (2018). *Corpus of Video-Mediated English as a Lingua Franca Conversations*. Birkenfeld: Trier University of Applied Sciences. Version 1.0. The CASE project <http://umwelt-campus.de/case>.

ACKNOWLEDGEMENTS

We are grateful to our peer reviewers for the constructive feedback provided.

APPENDIX A

Teletandem: initial questionnaire

The main objective of this questionnaire is to help you to establish your goals for this teletandem partnership.

*Obrigatório (Required field)

e-mail *

1) Full Name *

2) Phone number

3) Course *

4) Declared Gender (M= male; F= female; Other -- please, specify) *

5) How old are you? *

6) What is your birth date? *

Data

7) What is your first language? *

8) What is the foreign language you are learning in teletandem? *

9) Do you speak a third language? (If yes, please, specify. If no, please, write "Not applicable") *

10) What is your hometown? *

11) Where did you have most part of your high school studies? *

12) Have you had any previous experience with teletandem? For how long? *

13) In case you answered "yes", please, tell us your opinion about it. How did/didn't teletandem help you? *

14) Click on the link below, and read the self-assessment:

[grid.https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?docum](https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?docum)

entId=090000168045bb52 . Now answer the following question: what is your proficiency level in the foreign language (the language you are learning in teletandem)? *

After checking the self-assessment grid above, write down what your level is in the different skills.

15) Based on your self-assessment, establish a learning goal for your participation in teletandem. Think about these questions: (i) what do you need to know in order to go to the next level in the self-assessment grid? (ii) how can you learn what you need (or want) with the help of your teletandem partner? *

Try to write a realistic, specific goal, considering that you will practice teletandem for about 6 weeks.

16) Have you ever studied this foreign language anywhere else before college? *

17) If you answered "yes" in the previous question, please, tell us where did you study and for how long.