revista.pucsp.br/esp | ISSN: 2318-7115

The Especialist

Volume 40 | Número 1 | Ano 2019

Moving from an internal databank to a sharable multimodal corpus: the MulTeC case

A trajetória da elaboração de um banco de dados para um corpus multimodal compartilhável: o caso do MulTeC.

Solange, ARANHA (UNESP)[1]
Queila, LOPES (UFAC)[2]

ABSTRACT

The amount of data produced during teletandem practice can be considered extensive. Since 2011, teletandem researchers at UNESP (Sao Paolo State University) at São José do Rio Preto have been worried about compiling and organizing the data from groups of Brazilian and American students who interact via internet tools to learn each other´s language. This paper aims at describing how the data generated from 2012-2015 and collected according to the procedures described by Aranha, Luvizari-Murad and Moreno (2015) were treated and transformed into Multimodal Teletandem Corpus (MulTeC).

Keywords: *Telecollaboration, teletandem, pedagogical corpus, MulTeC*

RESUMO

A quantidade de dados produzidos durante a prática de teletandem pode ser considerada extensa. Desde 2011, os pesquisadores em teletandem da UNESP (Universidade Estadual Paulista) de São José do Rio Preto tem compilado e organizado os dados dos grupos de brasileiros e estadunidenses que interagem via ferramentas de internet para aprender a língua um do outro. Esse trabalho objetiva descrever como os dados gerados de 2012 a 2015, coletados conforme descrito por Aranha, Luvizari-Murad e Moreno (2015) foram tratados e transformados no MulTeC – Multimodal Teletandem Corpus.

Palavras-Chave: Telecolaboração, Teletandem, Corpus pedagógico, MulTeC

1. Introduction

[1] Universidade Estadual Paulista Júlio de Mesquita Filho, São Jose do Rio Preto, São Paulo, Brasil. Departamento de Letras Modernas; ORCID: http://orcid.org/0000-0002-8092-1875 ; solange.aranha@unesp.br. Os resultados dessa pesquisa são frutos de um projeto FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo)
[2] Universidade Federal do Acre, Rio Branco, Acre, Brasil. Centro de Educação, Letras e Artes; ORCID: http://orcid.org/0000-0003-0161-9975; queilalopes@gmail.com.

Teletandem (TELLES, 2006) is a telecollaborative autonomous language learning context developed in some Unesp campi (in Assis, São José do Rio Preto and Araraquara – São Paulo, Brazil). Telles (2006) based his project on in tandem learning (BRAMMERTS, 1996) and teletandem follows the same basic tandem principles: autonomy, language separation and reciprocity. In teletandem (TTD), learners of the language his/her TTD partner is proficient at have the opportunity to achieve goals previously set to learn the foreign language with his/her partner, using the communication resources available by the telematic networks. Partners must respect these principles during all teletandem practice and every participant is exposed to them during a tutorial before they enroll in a teletandem group. In the initial project (Telles, 2006), the partners decided when and where they would have their oral sessions, how long they would last, which tasks they would develop and how they would evaluate their progress. Since 2011, São José do Rio Preto campus has been developing a modality of teletandem practice coined by Aranha and Cavalari (2014) as institutional integrated Teletandem (iiTTD), which implies that the modality incorporates tasks to be developed along the TTD practice, is previously organized by teachers in charge of classes, feed and are fed by mediation procedures, occur in a lab with students enrolled in a specific language class and the results are graded. In the semi-integrated modality, this integration to the syllabus occur in just one of the partner universities. According to Aranha and Leone (2017), teletandem practice comprises teletandem oral sessions (TOS), during which the partners interact orally, and mediation sessions, where participants may discuss about their experience, share questions and feelings and exchange impressions with the group they are enrolled in. These are mandatory for any teletandem project to occur. According to Cavalari e Aranha (2016, 329) in iiTTD the procedures are also expected in the integrated modality: i) preparing students to TTD through a tutorial; ii) blending teletandem oral sessions and foreign language (FL) syllabus, integrating TTD tasks; iii) integrating assessment (by the learner, by the partner, and by the FL teacher).

Both modalities imply that a great amount of data is produced by each and every participant, both oral and written. The first is the video conference between two language learners who interact orally for an hour weekly during eight weeks. Each video recording, named TOS (teletandem oral session) lasts approximately 50 minutes during which 25 are in Portuguese and 25 are in English. The latter comprises questionnaires, chats, learning diaries, texts written in the foreign language, corrected by the foreign partner and rewritten by the learner. Tutorials given prior to the beginning of each group is also data to be considered.

In 2011, with the beginning of the institutional integrated teletandem modality at São José do Rio Preto, the teletandem researchers from that campus decided to organize and collect these data produced by Brazilian and American learners from one specific university that had their IRB (Institutional Review Board) approved. UNESP also had a consent term.

 DOI: 10.23925/2318-7115.2019v40i1a8

Aranha, Luvizari-Murad and Moreno (2015) described how the collection and organization procedures took place. The particular purpose of recording, filing and organizing data was to save researchers´ time spent during collection management for their thesis and dissertations. From 2011 to 2016, some thesis and dissertations used this databank in their researchers and authors helped, somehow, the collection because most of them were members of the Teletandem Group at UNESP/São José do Rio Preto. Roughly speaking, data were collected and organized in a databank, but, if we wanted the databank to be accessed by other researchers and be shared among other scholars, it should be transformed into a corpus, which means that it should include metadata, codes creation documents, anonymization, and data conversion into standard formats.

This paper describes the path through which the data organized by Aranha, Luvizari-Murad and Moreno (2015) became available in MulTeC (Multimodal Teletandem Corpus), which was built based on i) the corpus conception (Berber Sardinha, 2004; Sinclair, 2004); and ii) the staged methodology for structuring Learner Computer Interactions (LCI) data presented by Chanier and Wigham (2016). The result is a wide researchable corpus that will be available for researchers around the world who are interested in investigating telecollaborative language learning processes and any other subject that can be interrogated by means of the metadata.

2. Aranha, Luvizari-Murad and Moreno (2015): iiTTD databank

Aranha, Luvizari-Murad and Moreno (2015) presents how the data produced by teletandem partners during their interactions was compiled with the help of undergrad assistants and graduate students interested in researching TTD context. The set of data collected and organized was named institutional integrated teletandem databank. Roughly, it consisted of the TOSs, texts produced and exchanged with the partner, diaries, chats and questionnaires produced during each interaction from 2011 to 2015, and consent forms. Data was organized in folders according to the year in which they were produced, then according to the semester. Inside the folder of each semester, there were folders separated according to the kind of data: TOS, texts, chats, questionnaires. Besides, inside each one, the data were also organized into other, separated by learner. It means that a learner who used the Skype username “unespriopreto01” had all TOSs from that class in the folder “user 01” (Usuário 01) and so on.

It is important to mention that data were named using a code created according to the order of occurrence of the task, then the Skype user number, kind of task, and day of the week the TTD class occurred. For instance, the learner who used Skype “unespriopreto02” in a TTD class that occurred every week had his/her 7th session named as 7GRAV2QUI[3] . The Brazilian participants recorded the oral

[3] More details about the file codes used to name the files are described in Aranha, Luvizari-Murad and Moreno (2015).

 DOI: 10.23925/2318-7115.2019v40i1a8

session (using Evaer ®) and saved it using the codes after each TOS. Each learner was responsible for saving his/her work after each TOS.

3. Building a corpus

According to Sinclair (2004) '[a] corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.' So, to build a corpus is a decision that requires the establishment of criteria to collect and organize the data so that it will be usable for researchers.

Huang and Yao Yao (2015) state that corpora are used

> […] in language sciences, including linguistics, computational linguistics (CL), language education psycholinguistics, sociolinguistics, and translation studies, as well as other studies where a collection of texts or other language uses are crucial such as anthropology, communication studies, literary studies or political science. (2015, p.949)

Their statement helps us visualize how many researchers and fields can be benefited from corpora. However, building a corpus is a time-consuming work that must be well planned so that it will be feasible, as we can understand from Reppen (2010) statement: 'having a clearly articulated question is an essential first step in corpus construction since this will guide the design of the corpus.' (Reppen, 2010, p. 31). The researcher who decides to organize a corpus must, then, determine a priori what the purpose of the corpus is so that it can help him/her to establish ways of organization.

According to Chanier and Wigham (2016), systematic data collection, detailed data description, data conversion and data release/distribution are issues to work on the corpus paradigm. Their statement was the starting point in our work of scrutinizing the bank organized by Aranha, Luvizari-Murad and Moreno (2015) in order to transform it into a researchable corpus. Besides, the databank was used internally, by teletandem researchers (master and doctorate students) and the need to share these valuable data and have more researchers working from different perspectives made us develop the corpus.

The first thing was related to the compilation of the data. Chanier and Wigham (2006) state the whole data set in a language use should be collected even if the compiler has a strict question guiding his/her compilation, so that it will 'allow other researchers to reuse the corpus' (CHANIER, WIGHAM, 2016, p. 220). In order to make it feasible, the researcher compiling and organizing a corpus must describe the corpus in details, explaining how it was compiled, edited and organized. These details include giving information to the user of the corpus about the corpus participants and anonymizing all data. A second issue mentioned by Chanier and Wigham (2016) was about data conversion. A corpus must be read in different kinds of computers and systems, consequently, the texts that constitute the corpus must be open formats (plain text).

 DOI: 10.23925/2318-7115.2019v40i1a8

For Chanier and Wigham (2016, p. 221) '[…] a language corpus and its related analysis can only become part of the scientific research cycle if it can be freely accessed and when this access is guaranteed as permanent.' It implies that all data in the corpus should be available for downloading by the researchers interested in using them in their works. For the authors, data collection must be planned in advance, having issues as ethics, kinds of interactions to collect and data formats. The authors state that the compiler must care about three points: (i) identification of the resources, (ii) anonymisation and (iii) format conversion.

Each resource produced must contain a 'unique identification (ID) so that later it can be easily listed and described' (Chanier & Wigham, 2016, p. 230). In this phase, it is important to decide for meaningful codes. As example, the authors suggest to create a participant ID with 'the name of the student group which s/he belongs, the course name and course session name.' (Chanier & Wigham, 2016, p. 230). Once the IDs are created, the researcher can anonymise the data.

Anonymisation is a work whose results will make the data usable by the researcher, but will also preserve the face of the participants, i.e., the participant cannot be easily identified, however, the data can be interpreted. So, if in the data produced, the participants talk about some member of her/his family name, pets, phone number or address, the information must be modified in a way that the researcher can analyze it but not identify the interactant. Conversion means that all data are machine-readable. It refers to convert text created in extension .doc (Microsoft Office Word ®) into plain text.

In the next section we describe how these steps were followed in building MulTeC.

4. Building MulTeC

4.1 Pre-treatment phase

Multiple data, according to Guichon (2017, p. 65), "can only qualify as a corpus once they have undergone a certain number of operations", which means the data must be structured, labeled and even converted to readable formats, and the complementary documentation must be included; therefore, the context of the data production can be understood. These procedures are also related to what Guichon (2017) calls 'multimodal quality'. It includes to pay special attention to sound and image quality of the sessions.

We deleted files that could not be used because of the bad quality of audio or video and the ones with no consent form signed by both learners. In some cases, just one learner of the pair had signed the consent. For those, we sent emails asking for the document. Many of them returned positively and others did not. These had their files removed from the databank. Thus, we had two different amounts of data: the first with all data saved in iiTTD databank and the second with the data in the MulTeC, as illustrated on Table 1.

 DOI: 10.23925/2318-7115.2019v40i1a8

Based on Sardinha (2004), Chanier & Wigham (2016) Wigham & Chanier (2013) and Guichon (2017), the first step to the corpus design was to create meaningful codes to be used in the organization. As we decided to include each learner identification in the file names, the learner code was the first we needed to create. Subsequently, it could be used in the name of the files as well as in data anonymisation. After creating the code, the next steps were (i) reorganizing the folders, (ii) anonymising, (iii) renaming the archives, (iv) creating the headers, and (v) adding documents necessary to understand the context in which the data were generated.

4.1 Creating MulTeC codes

The creation of a meaningful code to name each data source (learning diaries, questionnaires, TOSs, chats, texts) as well as each learner led to the creation of what we have called IT (Teletandem Identity – 'Identificação no Teletandem' in Portuguese). This will give specifications about each participant and also allow longitudinal studies, once some participants may have enrolled in teletandem activities more than once. To our knowledge, LETEC do not show such amount of data with these specificities so far.

4.1.1 IT code

IT is a code assigned to each participant in each teletandem group s/he participated. It is formed by information about institution, course, gender[4] and Skype number[5] used. Therefore, in order to create the IT, we numbered all courses Unesp – Rio Preto offers, using the list available in the university homepage[6], in a total of thirty-two (32) majors and post-majors.

Aranha, Luvizari-Murad & Moreno's databank (2015) is constituted by data produced during a partnership with an American university whose participants are Portuguese learners enrolled in a Portuguese as a foreign language course, from various courses.

Figure 1 is an example of IT. **I7F3** indicates a Unesp student (**I**), majoring in Chemistry[7] (**7**), who was identified as female (**F**) and was using the Skype ® number **3** during that teletandem session.

4.1.2 Tasks codes

After deciding about the constitution of the learner's ITs, we established the codes for each task.

D – Diary;

C – Chat;

[4] When the data were compiled there was no opportunity for the learner to declare her/his sex. So, this metadata in the corpus refers to the one we could identify in the videos. (TOSs)
[5] Skype usernames at teletandem Rio Preto have a number to differentiate them: unespriopreto01, unespriopreto02, unespriopreto03 and so on.
[6] http://www.ibilce.unesp.br/
[7] The list created with Unesp courses is attached.

 DOI: 10.23925/2318-7115.2019v40i1a8

SOT – Sessão Oral de Teletandem (Teletandem Oral Session);

SOTi – Sessão Oral de Teletandem inicial (initial Teletandem Oral Session);

SOTin – Sessão Oral de Teletandem intermediária (intermediary Teletandem Oral Session);

SOTf- Sessão Oral de Teletadem final (final Teletandem Oral Session);

TOI – Texto Original em Inglês (Original Text in English);

TRevI – Texto Revisado em Inglês (Revised text in English[8]);

TReI – Texto Reescrito em Ingles (Rewritten Text in English);

TOP – Texto Original em Português (Original Text in Portuguese);

TRevP- Texto Revisado em Português (Revised text in Portuguese);

TReP – Texto Reescrito em Português (Rewritten text in Portuguese);

QI – Questionário Inicial (Initial questionnaire);

QF – Questionário Final (Final questionnaire);

D, C, SOTin, TOP, TRevP, TReP, TOI, TRevI, and TReI are codes that can be followed by a number indicating the order of occurrence during the period of sessions. For instance, the learning diary of the first teletandem oral session (TOSi) has as code D1, the second diary, written about the teletandem oral session intermediary 1, has as code D2 and so on.

4.1.3 Teletandem classes codes

There was a need to easily differentiate the data produced in different groups, so we created a code for each one, as follows: the name of the institution where the partner is affiliated to, a number indicating the order of the occurrence and the letters "i" or "si" designating respectively integrated and semi-integrated modality.

Having all codes been created, we began the anonymisation.

4.2 Anonymisation

Anonymisation is an important process in any research once the identity of the participants must be protected according to what they agreed when individually signed the consent form (CHANIER, WIGHAM, 2016). It is an ethical issue. Guichon (2017) gives an example of the importance of anonimization based on ISMAEL, a corpus organized by a group of researchers "[…] the data involve teachers who were learning to teach online and could be clumsy; thus online clips displaying them in awkward situations can potentially harm their future employability." (GUICHON, 2017, p. 61)

In order to avoid any kind of damage in the participants image and save their face, the researchers involved in building a corpus must be aware of any information "in the produced data that could lead to

[8] The learners produce texts to be revised by their respective partners. After being revised, the text with comments or/and suggestions, was saved as "revised".

the identification of a participant or skew a researcher's analysis of the data" (CHANIER, WIGHAM, 2016, p. 223).

The process of anonymising MulTeC involved the reading of each text in order to identify any word that could lead to one of the learner's identification, such as proper names, addresses, telephone numbers and hometowns.

If identified, the information was replaced by the IT and the first letters of the words that represent such information. For instance, when a learner mentions her/his boyfriend's name in any written text, we use the learner's IT and the letter "n" ("namorado" that means "boyfriend").

Another important procedure during this step was to check the authorship of the documents in the saved files in the tool comments in "revised texts". As most learners used their own computer to produce the texts, they were saved with their names as author, All authorship signs were removed and substituted by the ÏT.

4.3 Naming the files

The naming of each of the files used the codes created to each learner, class and tasks preceded by the year of occurrence. For instance, a learning diary produced by learner I9M13 for the first teletandem oral session with an American student in semi-integrated teletandem modality in 2012 is named as follows:

2012_I9M13_UGA[9]1si_D1

On table 2, the codes for naming each document is described.

It is important to mention that the files were renamed. Teletandem databank files had been previously named (as already mentioned), but the designation established for MulTeC has been more helpful for research purposes, since it shows some important details about file specifics.

During this process, we faced two different problems in the naming process: i) learner substitution by an assistant or a volunteer because he/she was absent that day; and ii) sessions whose recording was divided into two or more parts. To solve the first problem, if the learner was substituted in one or two SOTs, we decided to identify this substitution already in the file name, adding "s" to the file name. For instance, in 2012_I9F15_UGA1i_SOT2_s – indicates that this file is teletandem oral session 2 and the partner, I9F15 of the class UGA1i in 2012, was absent that day and was substituted by. For the second problem, we added a number just after the task code. Thus, if the first TOS (SOTi) of the learner I9F15 of the class UGA1si of 2012 was saved into two parts, they were named as 2015_I9F15_UGA1si_SOTi_1 and 2015_I9F15_UGA1si_SOTi_2.

[9] UGA stands for University of Georgia at Athens, the American partner university.

Once the renaming process finished, we created documents to help researchers understand each piece of data context production. Building a corpus also consists of including documents (complementary data) which were not produced by participants but were elaborated or included in the corpus to help researchers. The documents are available in the corpus so that the user can understand the data production context. Such documents inclusion is an essential procedure in corpus organization because they allow researchers to decide upon any subcorpus to study, which can be based on metadata such as linguistic level or gender, and can apply different perspectives, as syntactic, phonological or cultural approaches. Still, the researcher is able to compare number of hours of oral sessions, amount of other files. According to Guichon (2017) and Reffay, Betbeder and Chanier (2013) this contextual information facilitates the access to the data and explains the production of the data, which makes the corpus structured and coherent. MulTeC complementary documents comprise the Powerpoint ® tutorial file, attendance and pairing lists, spreadsheets with participant's sociolinguistics information.

4.4 Adding documents

Teletandem databank (Aranha, Luvizari-Murad & Moreno, 2015) provided information about the teletandem context, i.e., attendance and pairing lists; however, it did not have detailed information about the learners, the amount of data and group specifics. In order to get such information, we informally interviewed teletandem coordinators at Unesp-RP and included what they could remember about the groups and the learners. By having access to the teletandem coordinators personal computers we could find files about each teletandem class and then create documents to support researchers. The interviews and the survey in the coordinator's computers answered important questions to the corpus building: What courses were the learners enrolled in? When did each teletandem group start and finish? Which TTD modalities are there in TTD databank? What are the differences in data collection among groups?

After analyzing the interviews and the files, two different kinds of documents for each TTD group were created. Those documents provide the researcher with:

i) sociolinguistics information of the learner – Learner's information

ii) amount of data produced in each task by each learner in each group – Data survey and general data survey.

The documents will be detailed below.

4.4.1 Learners' information

Learners' information document is an Excel® spreadsheet where one can get details about first language, age, target-language proficiency according to CEFR (Common European Framework), gender and major of each learner. Age was calculated according to the birthdate, gender by watching the first

 DOI: 10.23925/2318-7115.2019v40i1a8

minutes of each TOS and language proficiency is the one alleged by the learner in the initial questionnaire, where the framework is presented and each learner is asked to self-assess according to its levels.

Information about students' major is also only from Unesp students achieved from the "Pairing List[10]" document and by watching the first minutes of initial session. According to Rampazzo and Aranha (2018), exchanging information about majors is part of a typical rhetorical move. The authors studied 15 sessions and there was rhetorical recurrence of this information in a move they named "exchanging information"(p. 461).

In addition, information about native and second language is displayed in this document with the indication L1, L2 and L3. A third language column (Figure 2) in the spreadsheet was necessary due to some European Portuguese and Mexican leaners whose L1 was not the one they were using to tutor their partners.

4.4.2 Teletandem class data survey

The information about the amount of each task produced by the learners is displayed in the spreadsheet named "Levantamento dos dados" (Data survey). In this spreadsheet, number '1' indicates that the file is in the corpus and '0' indicates it is not for most data files, except for TOSs files, which have a different register. If the TOS is saved in the corpus the time length is registered in the spreadsheet.

The spreadsheet has two tabs. In the first one (Figure 2), it is possible to know how many files of each teletandem task each learner produced in that teletandem encounter. On the second tab (Figure 3), the word counting[11] for each learner can be found. That allows external researcher to decide upon a subcorpus based, for example, on comparing two pairs with similar amount of data, or with complete different amounts to try to understand why oral sessions lasted differently in the same context.

The first tab has information about learners' authorization to data use. In the column TCLE (Consent Forms) it is informed which kind of authorization the learner donated, if only data or data and image. Number 1 indicates that the file is in the corpus and "x" indicates it is not. The columns designated to TOSs have the time length of each TOS. If the TOS was recorded in two or three parts, the number will represent the length of the whole TOS. After each kind of task, there is a column named as "total" where the amount of documents for each task and for each learner is found and at the bottom line it is the total of all data produced. Each group has one spreadsheet, totaling 16. A general data survey spreadsheet is also available, with all data in MulTeC.

[10] Name of the document produced by the teletandem mediators to each group with the pairs, which includes students' major, day of the week and time of the teletandem class. In semi-integrated teletandem class, instead of the name of a course, we can find the word "volunteers".

[11] Only TOSs did not have their words counted because the transcriptions were not finished by the time we organized the data.

 DOI: 10.23925/2318-7115.2019v40i1a8

4.4.3 General teletandem data survey

If the researcher interested in MulTeC data needs general information about the amount of hours of TOS, by modality or by teletandem groups, he or she may use information in this spreadsheet. This spreadsheet has three (3) tabs (see Figure 4).

One with all 16 groups information, which includes semi-integrated and integrated modalities; a second tab with the information about siTTD and the third one with the information only from iiTTD.

Figure 5 shows the total of participants as opposed to the amount of data of each one in the corpus (440/282), the amount of initial questionnaires answered in each group 282/91) and how long each SOT lasted (total of spoken data: approximately 24 hours) . At the end, the total numbers can be found.

4.4.4 Teletandem context document

Teletandem context document is a brief description of the teletandem tasks during one specific group. In this document, the researcher can get information about teletandem context, teletandem group formation and tasks aims and procedures.

4.5 Folders organization

MulTeC has the folders organized according to the teletandem modality and then by the year. Figure 6 represents an example of the structure of the folders organization.

Learner's folders contain the data produced by each anonymized pair. For example, considering an UGA integrated teletandem class, the first text produced by UGA student and revised by Unesp student, has as author UGA student, so the name of the file is 2015_UGA1i_U0M3_TRevP1.

Conclusion

MulTeC organization can contribute significantly with the researchers interested in working with telecollaborative multimodal learning environments and with oral and written aspects of foreign language learning. Instead of previously establishing a specific research question to design the corpus – as proposed by Chanier and Wigham (2016), we defined a research purpose of compiling, anonymizing, and filing data collected by Aranha, Luvizari-Murad and Moreno (2015) and elaborating documents to specify characteristics of teletandem context. Nevertheless, this process resulted into a researchable corpus that will allow internationalization and an open data source once we share the corpus with colleagues around the world.

Building MulTeC indicated the necessity to standardize the documents, to broaden theoretical background to include the concepts of pedagogical and learning scenarios (Foucher, 2010), to consider all the steps of corpus design for further collection to minimize the problems faced during this process.

 DOI: 10.23925/2318-7115.2019v40i1a8

All in all, as it is now, MulTeC is a researchable multimodal corpus composed of telecollaborative interactions among foreign language learners participating in different groups of teletandem.

Acknowledgments

Part of the research of this article was made possible by funding from FAPESP grant (#2016/18705-9) and Federal University of Acre (UFAC). The authors would like to thank Dr. Ciara Wigham for her invaluable support during the transformation of a raw databank into a researchable corpus. We would also like to thank the teletandem group of researchers from UNESP/São José do Rio Preto for their contributions and support, especially the careful and thoughtful readings of Dr. Cavalari and Ms. Rampazzo.

Ethical statement

All participants in this research signed a consent form volunteering the use of their data produced during their participation in teletandem. The consent forms were signed in both universities involved in teletandem partnership.

References

ARANHA, S.; LUVIZARI-MURAD, L. H.; MORENO, A. C. A Criação De Um Banco De Dados Para Pesquisas Sobre Aprendizagem Via Teletandem Institucional Integrado (TTDii). Revista (Con)textos Linguísticos, v. 9, n. 12, p. 274–293, 2015.

ATKINS, S.; CLEAR, J.; OSTLER, N. 1991. Corpus design criteria. Journal of Literary and Linguistic Computing 7.

BRAMMERTS, Helmut. 1996. Tandem language learning via the internet and the International E-Mail Tandem Network. In: Little, David & Helmut Brammerts (ed.). A guide to language learning in tandem via the Internet. CLCS Occasional Paper.

CAVALARI, S. M. S; ARANHA, S. 2016. Teletandem: integrating e-learning into the foreign language classroom Acta Scientiarum: Language and Culture, Maringá, v.38, n. 4, p. 327-336.

CHANIER, T. et al. 2014. The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. Journal for Language Technology and Computational Linguistics, v. 29, n. 2, p. 1–30, 2014.

CHANIER, T.; WIGHAM, C. R. 2016. A scientific methodology for researching CALL interaction data. Language-learner computer interactions: Theory, methodology and CALL aplications, p. 215–240, Available at: <https://benjamins.com/catalog/lsse.2.10cha>.

GUICHON, N. 2017. Sharing a multimodal corpus to study web-cam-mediated language teaching. Language Learning & Techonlogy, v. 21, n. 1, pp. 56-75.

RAMPAZZO, L.; ARANHA, S. A sessão oral de teletandem inicial: a estrutura retórica do gênero. DELTA: Documentação e Estudos em Linguística Teórica e Aplicada, 34(1), 2018, p. 449-473. Available at: << http://www.scielo.br/pdf/delta/v34n1/1678-460X-delta-34-01-449.pdf>>. Acess 27 ago. 2018.

REPPEN, R. 2010. Building a corpus: What are the basics? In A. O'Keefe & M. McCarthy (Eds.) The Routledge Handbook of Corpus Linguistics. pp. 31-38. London: Routledge.

SALOMÃO, Ana Cristina Biondo. 2006. Pequeno dicionário de Tandem. Teletandem News, ano 1, n. 02.
SARDINHA, T. B.; 2004. Linguística de Corpus. Barueri, SP: Manole.
Reppen, R. 2010. Building a corpus: What are the basics? In A. O'Keefe & M. McCarthy (Eds.) The Routledge Handbook of Corpus Linguistics. pp. 31-38. London: Routledge.
SINCLAIR, John. 2004. Corpus and Text – Basic principles. Available at: https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm Access on 6 jun. 2017.
TELLES, João Antônio. 2006. TELETANDEM BRASIL –Línguas Estrangeiras para Todos. Projeto de pesquisa. Faculdade de Ciências e Letras de Assis (UNESP)
ZAKIR, M. A. 2015. Cultura e(m) telecolaboração: uma análise de parcerias de teletandem institucional. 2015. 232 f. Tese (Doutorado em Estudos Linguísticos). Universidade Estadual Paulista "Júlio de Mesquita Filho", campus de São José do Rio Preto. São José do Rio Preto.
HUANG, Chu-Ren & YAO, Yao. 2015. Corpus Linguistics. International Encyclopedia of the Social & Behavioral Sciences. 949-953. 10.1016/B978-0-08-097086-8.52004-2.

Solange Aranha *is Professor at the Modern Languages Department, State University of Sao Paulo campus of São José do Rio Preto, São Paulo, Brazil. She holds a PhD degree in Portuguese Language and Linguistics. Her main field of interest includes foreign language teaching, ESP, and genres. Her current research is on telecollaboration, foreign language teaching and learning, and genres. E-mail: solange.aranha@unesp.br*

Queila Barbosa Lopes *is Professor at the Education, Language/Literature and Arts Department, Federal University of Acre, Acre, Brazil. She holds a PhD degree in Applied Linguistics and her main field of interest includes foreign language teaching and educational technology. Her current research is on telecollaboration and foreign language learning in digital environments. E-mail: queilalopes@gmail.com*

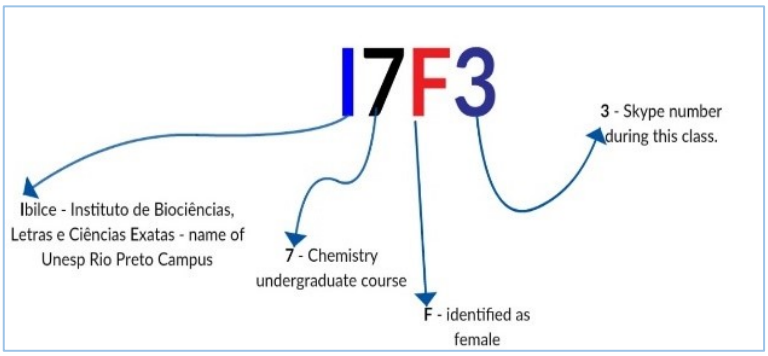
Table 1 – Core Data in Multiple Data and Shareable Corpus according to Guichon (2017)

Kind of files	***Multiple Data (Teletandem databank)***	***Shareable Corpus (MulTeC)***
TOS time	655'43"22""	581'19"07""
Learning diaries	849	666
Questionnaires	180	132
Texts	1444	956
Chats	477	351

Source: The authors

Figure 1 – Example of Unesp student's IT

 DOI: 10.23925/2318-7115.2019v40i1a8



Source: The authors

Table 2. Composition of a task name file in MulTeC

2012	I9MI3	UGA1si	D1
Year (2012, 2013,2014, …)	**IT** (created according to the standard established)	**Teletandem Class and its order of occurence** (UGA1i, UGA2i, UGA3si,...)	**Task and its order of occurrence** (**C**-chat; **D**-Diaries; **TOI** – Original text in English; **TRevI**- Revised text in English, **TReI** – Rewritten text in English; **TOP** – Original text in Portuguese; and so on)

Source: The authors

Figure 2. Exemplifies part of a spreadsheet for one group learner's information.

 DOI: 10.23925/2318-7115.2019v40i1a8

		Usuário Skype	IT	Gênero ident.	Curso	Idade	Proficiência Autoavaliada na Língua em estudo	L1	L2	L3
2012_UGA_term1_TUESDAY	UGA1i	unespriopreto03	I9F3	F	Letras	25	B2	Port	Ing	X
		teletandem13	U0M13	M	X	X	X	X	X	X
		unespriopreto04	I9M4	M	Letras	23	B2	Port	Ing	X
		teletandem14	U0F14	F	X	X	X	X	X	X
		unespriopreto06	I9F6	F	Letras	20	B2	Port	Ing	X
		teletandem16	U0F16	F	X	X	X	X	X	X
		unespriopreto07	I9F7	F	Letras	21	B1	Port	Ing	X
		teletandem17	U0F17	F	X	X	X	X	X	X
		unespriopreto09	I9F9	F	Letras	21	B1	Port	Ing	X
		teletandem19	U0F19	F	X	X	X	X	X	X
		unespriopreto012	I9F12	F	Letras	22	C1	PortP	Port	I
		teletandem22	U0F22	F	X	X	X	X	X	X
		unespriopreto013	I9F13	F	Letras	40	X	Port	Ing	X
		teletandem23	U0M23	M	X	X	X	X	X	X
		unespriopreto014	I9F14	F	Letras	X	B2	Port	Ing	X
		teletandem24	U0F24	F	X	X	X	X	X	X
		unespriopreto015	I9M15	M	Letras	21	X	Port	Ing	X
		teletandem25	U0F25	F	X	X	X	X	X	X

2012_UGA1i

Source: MulTeC

Figure 3. Teletandem class survey spreadsheet - Plan 1

Turma	IT	Usuário Skype	TCLE	QI	SOTs								
			D (Dados) DI (Dados e imagens)		SOTi	SOTin1	SOTin2	SOTin3	SOTin4	SOTin5	SOTin6	SOTf	Totais por par
2014_UGA_1sem	I9F1	unespriopreto01	DI	x	0:46:07	0:44:00	0:47:09	0:37:31	x	x	x	0:41:38	03:36:25
	U0F11	teletandem11	DI	x									
	I9M2	unespriopreto02	DI	1	0:47:02	0:13:37	0:49:25	0:41:53	0:42:32	x	x	0:47:07	04:01:36
	U0M12	teletandem12	DI	x									
	I9M4	unesprioreto04	DI	1	0:37:52	0:38:17	x	x	x	x	x	x	01:16:09
	U0M14	teletandem14	DI	x									
	I9F9	unespriorieto09	DI	x	x	x	x	x	x	x	x	x	00:00:00
	U0F19	teletandem19	DI	x									
	I9F12	unespriopreto012	DI	x	x	x	x	x	x	x	x	x	00:00:00
	U0M22	teletandem22	DI	x									
	I9F13	unespriopreto013	DI	x	0:46:49	0:41:39	0:52:37	0:41:49	0:44:54	x	x	x	03:47:48
	U0F23	teletandem23	DI	x									
	I9F15	unespriopreto015	DI	x	0:00:00	0:32:37	0:33:07	0:40:57	0:48:27	x	x	0:41:32	03:16:40
	U0F25	teletandem25	DI	x									
	I9F17	unespriopreto017	DI	x	0:47:37	0:40:36	0:49:39	0:38:37	0:42:24	x	x	0:23:48	04:02:41
	U0F27	teletandem27	DI	x									
	I9M19	unespriopreto019	DI	x	0:42:56	0:43:58	0:41:53	0:37:04	0:41:58	x	x	0:44:10	04:11:59
	U0F29	teletandem29	DI	x									
			Totais	2	4:28:23	4:14:44	4:33:50	3:57:51	3:40:15	0:00:00	0:00:00	3:18:15	24:13:18

Dados produzidos por par | Contagem das palavras

Source: MulTeC

Figure 4. Teletandem class survey spreadsheet – Tab 1

DOI: 10.23925/2318-7115.2019v40i1a8

	IT	QI	QF	Total	Diários						Totais por aprendiz	Chats						Totais por par
Turma					1º	2º	3º	4º	5º	6º		1º	2º	3º	4º	5º	6º	
2014_UGA2si	I1F1	x	x	0	x	x	x	x	x	0	0	x	x	x	x	x	x	0
	U0M13	x	x	0	x	x	x	x	x	0	0							
	I10F4	154	x	154	186	212	162	149	160	153	1022	x	x	x	x	x	x	0
	U0M19	x	x	0	x	x	x	x	x	0	0							
	I9F7	301	x	301	248	201	207	190	165	0	1011	x	50	x	x	x	x	50
	U0F12	x	x	0	x	x	x	x	x	0	0							
	I9F11	x	x	0	150	103	109	65	82	149	658	x	x	x	x	x	x	0
	U0F20	x	x	0	x	x	x	x	x	0	0							
	I9M12	100	x	100	106	87	115	88	132	84	612	x	x	x	x	x	x	0
	U0M16	x	x	0	x	x	x	x	x	0	0							
	I3F15	134	x	134	46	67	90	33	x	0	236	x	x	x	x	x	x	0
	U0F11	x	x	0	x	x	x	x	x	0	0							
	I9F17	123	x	123	114	60	74	59	49	46	402	x	131	271	241	294	122	1059
	U0M12	x	x	0	x	x	x	x	x	0	0							
	I8F19	x	x	0	99	170	213	144	146	103	875	x	50	x	x	x	x	50
	U0M18	x	x	0	x	x	x	x	x	0	0							
Totais				812	949	900	970	728	734	535	4816	0	231	271	241	294	122	1159

Dados produzidos por par | Contagem das palavras

Source: MulTeC

Figure 5. General teletandem data spreadsheet

Semestre	Turma	Qtde de interagentes		QI	SOTs									Diários				
		Participantes	Dados no banco		SOTi	SOTin1	SOTin2	SOTin3	SOTin4	SOTin5	SOTin6	SOTf	Total geral	1º	2º	3º	4º	5º
1º	2012_UGA1i	32	18	7	05:46:58	4:20:39	5:30:08	5:19:28	7:18:32	6:14:14	5:13:41	1:27:10	41:10:50	8	7	7	7	6
	2012_UGA2i	**36**	**30**	**15**	**8:48:06**	**10:12:18**	**9:17:33**	**9:20:01**	**11:14:53**	**10:51:45**	**9:42:55**	**5:02:11**	74:29:42	**13**	**14**	**13**	**13**	**13**
2º	2012_UGA3i	32	20	10	7:15:57	6:39:35	6:42:04	4:54:26	6:58:42	5:09:13	x	04:06:23	41:46:20	9	9	8	9	8
	2012_UGA4si	**24**	**12**	**1**	**4:00:07**	**2:33:38**	**2:57:35**	**2:45:52**	**3:33:53**	**3:32:58**	**2:11:10**	**1:08:14**	22:43:27	**1**	**0**	**0**	**0**	**0**
1º	2013_UGA1i	26	24	11	9:56:36	8:18:14	9:11:45	8:02:52	6:47:15	7:10:34	x	4:17:16	53:44:32	10	12	11	12	10
	2013_UGA2i	**38**	**32**	**7**	**15:49:57**	**9:17:50**	**12:31:56**	**8:31:01**	**13:13:09**	**10:40:47**	**8:13:09**	**6:20:16**	84:38:05	**9**	**9**	**7**	**6**	**6**
2º	2013_UGA3si	22	12	1	4:24:12	2:53:23	4:46:39	4:48:50	3:50:43	3:32:24	0:00:00	0:00:00	24:16:11	3	3	2	2	1
	2013_UGA4si	**26**	**6**	**1**	**1:17:23**	**2:14:23**	**1:34:34**	**2:21:45**	**1:32:58**	**1:40:59**	**0:00:00**	**0:00:00**	10:42:02	**0**	**0**	**0**	**0**	**0**
	2013_UGA5i	24	10	1	2:37:50	3:04:42	3:07:55	3:01:37	4:07:37	3:15:38	2:14:41	3:00:02	24:30:02	4	4	4	4	4
1º	**2014_UGA1i**	**22**	**20**	**2**	**5:13:19**	**4:59:40**	**5:13:42**	**4:43:09**	**3:40:15**	**0:00:00**	**0:00:00**	**4:02:15**	27:52:20	**5**	**4**	**4**	**4**	**3**
	2014_UGA2i	16	10	0	2:04:58	2:45:03	2:36:41	2:48:57	2:15:33	2:28:31	x	00:00:00	14:59:43	2	2	2	2	2
2º	**2014_UGA3i**	**22**	**20**	**10**	**6:53:09**	**4:55:29**	**7:00:04**	**5:40:45**	**6:08:10**	**x**	**x**	**6:40:00**	37:17:37	**10**	**10**	**10**	**10**	**10**
	2014_UGA4si	12	12	2	3:43:17	4:50:41	3:50:58	4:31:54	4:29:11	x	x	3:51:18	25:17:19	4	3	4	3	3
1º	**2015_UGA1i**	**18**	**6**	**2**	**0:36:40**	**2:22:33**	**2:08:51**	**2:06:42**	**1:43:21**	**x**	**x**	**0:38:46**	9:36:53	**3**	**2**	**3**	**1**	**2**
	2015_UGA2si	16	16	5	1:36:58	1:14:53	2:45:53	2:15:50	0:51:12	0:00:00	x	3:11:00	11:55:46	7	7	7	7	6
2º																		
	2015_UGA3i	**34**	**34**	**16**	**5:02:33**	**10:27:04**	**9:58:41**	**8:23:33**	**9:34:16**	**7:00:25**	**x**	**11:46:57**	62:13:29	**17**	**17**	**17**	**17**	**16**
TOTAIS		400	282	91	85:08:00	81:10:05	89:14:59	79:36:42	87:19:40	61:37:28	27:35:36	55:31:48	567:14:18	105	103	99	97	90

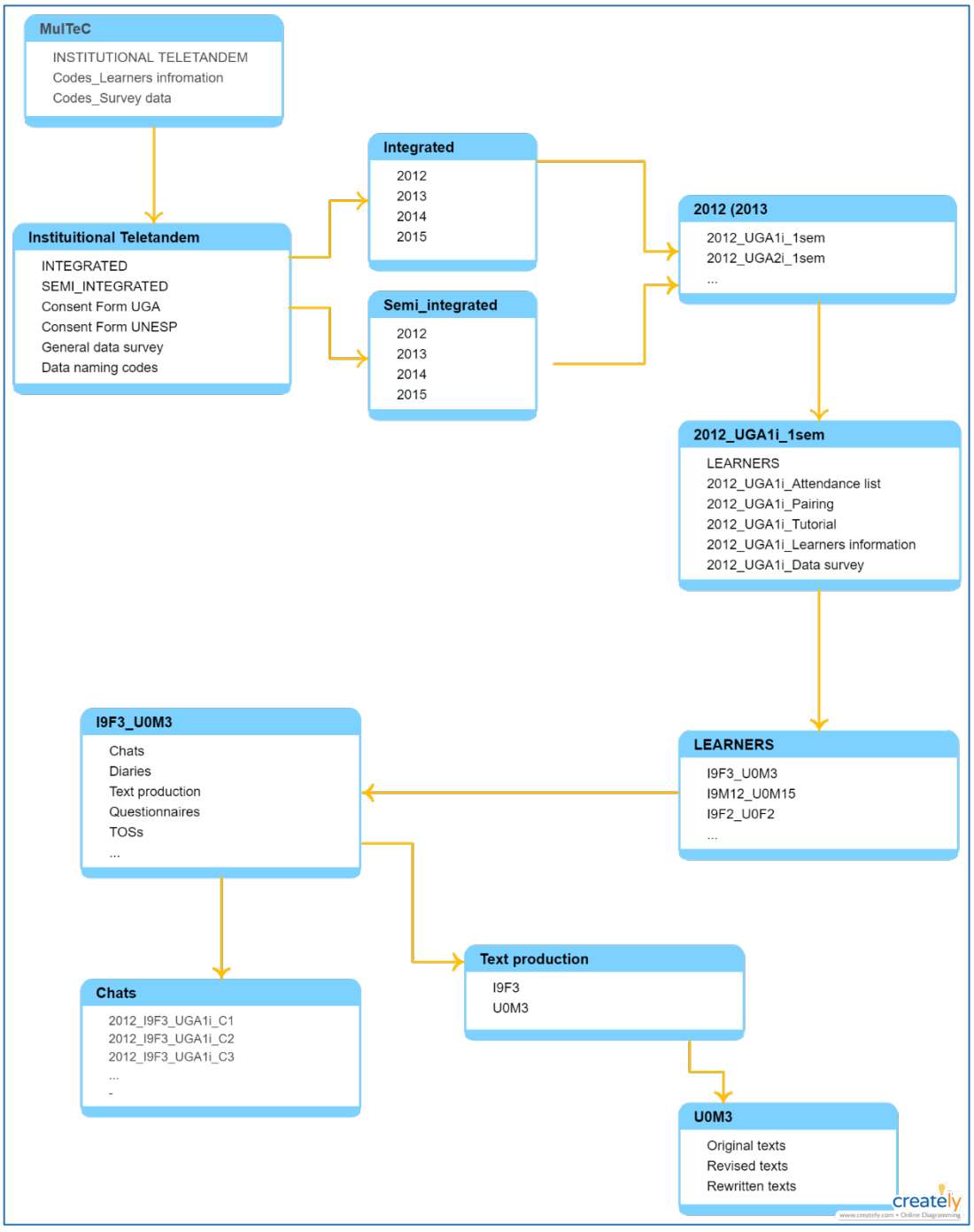
Geral | Semi-integrado | Integrado

Source: MulTeC

DOI: 10.23925/2318-7115.2019v40i1a8

Figure 6. Folders organization in MulTeC

DOI: 10.23925/2318-7115.2019v40i1a8



Source: The authors

APPENDIX

DOI: 10.23925/2318-7115.2019v40i1a8

Table 1. Courses at Ibilce Campus – Unesp in Rio Preto – Following the order presented on the university homepage.

Courses	Level	#
Ciências Biológicas	Undergraduate	1
Bacharelado em Ciência da Computação		2
Engenharia de Alimentos		3
Física		4
Matemática (Bacharelado)		5
Matemática (Licenciatura)		6
Química		7
Bacharelado em Letras - Tradutor		8
Licenciatura em Letras diuron		9
Licenciatura em Letras noturno		10
Pedagogia		11
Biofísica Molecular	Master	12
Biologia Animal		13
Ciência da Computação		14
Engenharia e Ciência de Alimentos		15
Ensino e Processos Formativos		16
Estudos Linguísticos		17
Genética		18
Letras		19
Matemática		20
Microbiologia		21
Química		22
Matemática em Rede Nacional - PROFMAT	Professional Master	23
Biofísica molecular	Doctorate	24
Biologia Animal		25
Engenharia e Ciência de Alimentos		26
Estudos Linguísticos		27
Genética		28
Letras		29
Matemática		30
Microbiologia		31
Química		32

Source: The authors

 DOI: 10.23925/2318-7115.2019v40i1a8