



UNIVERSIDADE ESTADUAL PAULISTA  
“JÚLIO DE MESQUITA FILHO”  
Câmpus de São José do Rio Preto

Queila Barbosa Lopes

**MulTeC:**

A construção de um *corpus* multimodal em teletandem

São José do Rio Preto  
2019

Queila Barbosa Lopes

**MulTeC:**

A construção de um *corpus* multimodal em teletandem

Tese apresentada como parte dos requisitos para obtenção do título de Doutora em Estudos Linguísticos, junto ao Programa de Pós-Graduação em Estudos Linguísticos, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadora: CAPES

Orientadora: Prof<sup>a</sup>. Dr<sup>a</sup>. Solange Aranha

São José do Rio Preto  
2019

L864m      Lopes, Queila Barbosa  
MulTeC : A construção de um corpus multimodal em  
teletandem / Queila Barbosa Lopes. -- São José do Rio  
Preto, 2019  
161 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista  
(Unesp), Instituto de Biociências Letras e Ciências Exatas,  
São José do Rio Preto  
Orientadora: Solange Aranha

1. Linguística Aplicada. 2. Ensino à distância. 3.  
Linguagem e línguas Estudo e ensino. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do  
Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados  
fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Queila Barbosa Lopes

**MulTeC:**

*A construção de um corpus multimodal em teletandem*

Tese apresentada como parte dos requisitos para obtenção do título de Doutora em Estudos Linguísticos, junto ao Programa de Pós-Graduação em Estudos Linguísticos, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Financiadoras: CAPES

Comissão Examinadora

Prof<sup>a</sup>. Dr<sup>a</sup>. Solange Aranha  
UNESP – Câmpus de São José do Rio Preto  
Orientador

Prof. Dr. Antônio Paulo Berber Sardinha  
PUC – São Paulo

Prof. Dr. Júlio César Rosa de Araújo  
UFC – Fortaleza

Prof<sup>a</sup>. Dr<sup>a</sup>. Claudia Zavaglia  
UNESP – Câmpus de São José do Rio Preto

Prof<sup>a</sup>. Dr<sup>a</sup>. Suzi Spatti Cavalari  
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto  
11 de março de 2019

À Solange Aranha que confiou-me o desafio de organizar um *corpus*  
com os dados do Banco TTDii.  
Minha eterna gratidão!

## AGRADECIMENTOS

A Deus, minha força e fortaleza, sem as quais a depressão teria me impedido de concluir essa etapa da minha vida com algum êxito.

À UFAC, por meu afastamento com ônus para realização de meus estudos de doutoramento, sem o qual esta pesquisa não teria sido realizada.

À minha orientadora, Prof.<sup>a</sup> Dr.<sup>a</sup> Solange Aranha, pelas inúmeras leituras cautelosas e sempre bastante criteriosas de meus textos, bem como pelos momentos em que ouviu minhas dificuldades para além de pesquisadora, como aquelas que enfrentei no papel de esposa e mãe.

À Prof.<sup>a</sup> Dr.<sup>a</sup> Ciara Wigham, cuja bondade, competência e generosidade não podem ser descritas em palavras. Agradeço por tantas vezes, mesmo antes de me conhecer, ter respondido meus e-mails, realizado sessões de Skype, ter disponibilidade em me encontrar em Liubliuana (Eslovênia) e ter respondido meus questionamentos quanto ao trabalho de organização do MulTeC enquanto o estava elaborando, sempre procurando esclarecer todas as minhas dúvidas.

Ao Prof. Dr. Rodrigo Esteves de Lima-Lopes cuja contribuição no debate SELIN de meu trabalho suscitou reflexões e contribuiu bastante para a organização do MulTeC.

Às Prof.<sup>as</sup> Dr.<sup>as</sup> Suzi Spatti Cavalari e Claudia Zavaglia que pacientemente apontaram as inúmeras falhas do texto apresentado na qualificação. Sua leitura cuidadosa e a discussão no dia da qualificação foram muito importantes para a continuidade do trabalho.

À Prof.<sup>a</sup> Dr.<sup>a</sup> Fernanda Galli pela orientação na qualificação especial. Suas palavras de encorajamento e sua competência conduziram-me a conclusão do trabalho quando imaginei que não conseguiria.

À Prof.<sup>a</sup> Dr.<sup>a</sup> Fabiana Komesu que, mesmo envolta em tantas atribuições, foi durante esses 4 anos uma presença amiga, um sorriso encorajador e uma capacidade de me convidar a realidade que foram especiais para mim.

À Ana Carolina Freschi, pela amizade, pelas revisões em minhas traduções Francês – Português e pela aula instrumental de Francês, além das conversas sempre tão sinérgicas e carinhosas.

À Ludmila Andreu-Funo pela leveza no compartilhar tantos anseios comigo, pelas leituras carinhosas e apontamentos questionadores ao meu trabalho.

À Deise Marinoto, pela tradução, feita tão prontamente, do artigo de Anne-Laure Foucher com profissionalismo e gentileza.

Aos meus filhos, Agatha Lima Lopes e Eduardo Lima Lopes, e esposo Francisco S. de Lima Silva, que estiveram apoiando minhas decisões, permanecendo carinhosamente ao meu lado, aceitando tantas mudanças, algumas restrições e o enfrentamento de algumas dificuldades em decorrência desse processo de doutoramento tão longe de nosso lar.

À minha mãe, Francisca S. Barbosa cuja paciência só aumenta e a capacidade de amar se expande com o passar dos dias. Agradeço por cuidar dos meus filhos, Agatha e Eduardo, por alguns meses para que pudesse organizar o retorno deles para nossa terra com mais tranquilidade.

Às minhas irmãs Clícia B. Lopes, Jardane B. Lopes e Betânia B. Lopes, pelo apoio incondicional sempre. Quanto amor se pode enviar virtualmente e fortalecer uma pessoa. Minha gratidão!

À minha querida amiga Lori Daiane Nunes Pereira, e seus dois filhos: João Asaph Fanini Costa Pereira e Samuel Fanini Costa Pereira, cuja luz, doçura e presenças cheias de amor foram fundamentais nos últimos dias de escrita da tese quando não pude estar mais perto dos meus filhos e esposo.

À minha querida amiga Laura Rampazzo que sempre esteve muito próxima a mim, intercedendo, enviando mensagens diárias de apoio quando estive nos piores dias da depressão, me ouvindo pacientemente sempre que precisei, lendo e relendo os meus textos, procurando me fazer acreditar em minha capacidade.

Às amigas do grupo de pesquisa teletandem de São José do Rio Preto, pelo compartilhamento de sorrisos, lágrimas e discussões sempre tão frutuosas.

Às professoras Ana Claudia Paleta e Daniela Águas que com suas aulas de natação e yoga, respectivamente, foram indispensáveis para o bem-estar necessário para concluir este trabalho.

Aos queridos psicólogos Lucas Cucato Bini e Daniella Fernandes Almeida que, com sua escuta qualificada e sempre tão respeitosa às minhas limitações e lágrimas, foram luz e reparação quando a depressão parecia que vencer minha capacidade de seguir.

Ao Prof. Dr. Júlio César Araújo cujo exemplo sempre me inspirou, seja pela escrita acadêmica, seja pelas palestras e cursos, a ser uma pessoa e uma pesquisadora melhor. Ter sua amizade desde 1997 é motivo de gratidão imensa, além de ser oportunidade de aprendizagem constante tanto no sentido acadêmico quanto espiritual e pessoal. Sei que ainda tenho muito a aprender.

À minha amiga Raquel Ishii que se fez presente durante esses quatro anos longe de meu lar, apesar de suas inúmeras atribuições. Cada e-mail afetuoso, cada presente recebido, cada gesto de amor inesperado nutriu meu coração e foi um refrigério para minha alma.

À família de Paulo e Michelle Arêas que acolheu a mim e a minha família com tanta generosidade e amor.

A tantos amigos feitos durante a pós-graduação. Vocês me constituem e me formam, porque, apesar de precisarmos nos posicionar e decidir como sujeitos, o fazemos a partir do que vivemos juntos como comunidade. Minha eterna gratidão!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - código de financiamento 001, à qual agradeço.

## RESUMO

O teletandem (TTD), contexto de aprendizagem de línguas telecolaborativo e autônomo (TELLES, 2006), desenvolvido na Unesp há mais de 12 anos, vem promovendo oportunidade de aprendizagem de línguas estrangeiras embasada na autonomia, separação de línguas e reciprocidade, princípios basilares da aprendizagem *in tandem*, como discutido por Brammerts (1996). Tal contexto de aprendizagem se tornou possível pela utilização das redes telemáticas que viabilizam a aprendizagem de línguas entre aprendizes que residem em países geograficamente distantes. No TTD, os aprendizes realizam sessões orais de teletandem (SOTs) e participam de sessões de mediação. Segundo Aranha e Leone (2016; 2017) essas são macrotarefas a partir das quais ocorrem as microtarefas: produção de diários de aprendizagem, troca de textos entre os participantes, respostas aos questionários inicial e final e uso da ferramenta chat durante a SOT. Esse contexto propicia a geração de uma quantidade significativa de dados multimodais (áudio, vídeo, texto escrito) que podem ser utilizados para pesquisas diversas. O intuito desta pesquisa é a construção de um *corpus* com os dados produzidos por 112 pares no período de 2012 a 2015. Para a composição do *corpus*, foram utilizados os dados coletados e armazenados conforme descrito por Aranha, Luvizari-Murad e Moreno (2015) e para a elaboração do *corpus* utilizei os seguintes procedimentos: a) levantamento das características de cada grupo de teletandem; b) criação de código para identificação dos parceiros; c) criação de códigos para a nomeação dos dados; d) padronização da anonimização dos dados; e) conversão dos dados para formatos compatíveis com diferentes sistemas operacionais; e f) renomeação dos dados a partir de Chanier e Wigham (2016). O MulTeC (Multimodal Teletandem *Corpus*) é composto por um total de 91 questionários iniciais, 41 questionários finais, 666 diários de aprendizagem, 355 chats, 921 textos produzidos na troca de texto por brasileiros e corrigidos por estadunidenses e 581 horas de gravações em vídeos, com transcrições em andamento. Após a finalização do *corpus*, percebeu-se a necessidade de ampliação de dados a serem coletados, incluindo a coleta da reunião tutorial e das sessões de mediação em mídia de vídeo ou áudio.

**Palavras-chave:** Teletandem. *Corpus*. Tarefas.

## ABSTRACT

Teletandem (TTD), a telecollaborative and autonomous language learning context (TELLES, 2006), which has been carried out at Unesp for more than 12 years, has been promoting language learning based on autonomy, separation of languages and reciprocity, basic principles of tandem learning, as discussed by Brammerts (1996). Such learning context has become viable using telematic networks that enable language learning between learners residing in geographically distant countries. In teletandem, learners hold teletandem oral sessions (TOSs) and participate in mediation sessions. According to Aranha and Leone (2016, 2017) these are macrotasks from which microtasks occur: production of learning diaries, text exchange among TTD participants, responses to the initial and final questionnaires, and use of the chat tool during teletandem oral sessions. The production of these tasks generates a significant amount of data that can be used for various researches. The purpose of this research is constructing a multimodal *corpus* in teletandem with the data produced by the participants in the period from 2012 to 2015. The data which constitute the *corpus* were produced by 112 pairs, in a partnership with an American university in the period of 2012 to 2015, whose collection was described by Aranha, Luvizari-Murad and Moreno (2015). The elaboration of the *corpus* from the collection and storage described in Aranha, Luvizari-Murad and Moreno (2015) adopted the following procedures: a) survey of the characteristics of each teletandem group; b) creation of code to identify the partners; c) creation of codes for data naming; d) standardization of data anonymization; e) converting the data to compatible formats with different operating systems; and f) renaming the data according to what is proposed by Chanier and Wigham (2016). After the organization of the *corpus*, MulTeC (Multimodal Teletandem *Corpus*) has a total of 91 initial questionnaires, 41 final questionnaires, 666 learning diaries, 355 chats, 921 texts produced in the text exchange between the partners and 581 hours of TOSs recordings, with transcriptions ongoing. After the completion of the *corpus*, it was perceived the need to broaden the collection of the data, including tutorial meeting and mediation sessions in audio or video media collection.

**Keywords:** Teletandem. *Corpus*. Tasks.

## LISTA DE ILUSTRAÇÕES

<u>Figura 1 – As modalidades de teletandem adaptado de Zakir (2015)</u> .....	23
<u>Figura 2 – Organização do cenário de aprendizagem no TTD por Aranha e Leone (2017)</u> .....	29
<u>Figura 3 – Versão dinâmica dos níveis de intervenção didática em Foucher (2010)</u> .....	33
<u>Figura 4 – Composição da macrotarefa por Grosbois (2009)</u> .....	35
<u>Figura 5 – Tarefas realizadas para mediação por objetivos</u> .....	36
<u>Figura 6 – Relação das tarefas na constituição do TTD</u> .....	38
<u>Figura 7 – Novo ciclo de vida dos dados</u> .....	47
<u>Figura 8 – Mapa dos dados</u> .....	49
<u>Figura 9 – Quadro comum europeu apresentado no questionário inicial do TTD</u> .....	70
<u>Figura 10 – Exemplo de anonimização nos textos revisados</u> .....	79
<u>Figura 11 – Cabeçalho do pareamento de turma TTD de 2013</u> .....	85
<u>Figura 12 – Cronologia no TTD</u> .....	89
<u>Figura 13 – Documento Anonimização dos Dados</u> .....	90
<u>Figura 14 – Documento ‘Código de Levantamento de Dados’</u> .....	91
<u>Figura 15 – Código Informações Interagentes</u> .....	92
<u>Figura 16 – Indicação de participante recorrente</u> .....	94
<u>Figura 17 – Aba ‘Dados produzidos por par’ – Planilha Levantamento de Dados</u> .....	96
<u>Figura 18 – Abertura de pasta de dados produzidos por par</u> .....	99
<u>Figura 19 – Aba ‘Contagem de palavras’ na planilha ‘Levantamento de Dados’</u> .....	101
<u>Figura 20 – Planilha ‘Levantamento de Dados Geral’</u> .....	105
<u>Figura 21 – Planilha ‘Levantamento de Dados Geral’</u> .....	107
<u>Figura 22 – Texto produzido por participante</u> .....	112
<u>Figura 23 – Representação da organização das pastas no MulTeC</u> .....	114
<u>Figura 24 – Cenário pedagógico – Aba 1</u> .....	117
<u>Figura 25 – Cenário pedagógico – Aba 3</u> .....	119
<u>Figura 26 – Cenário de aprendizagem – Aba 2</u> .....	121
<u>Figura 27 – Cenário de aprendizagem – Aba 3</u> .....	121
<u>Figura 28 – Cenário de aprendizagem – Aba 4</u> .....	122
<u>Figura 29 – Protocolo de coleta apresentado ao grupo TTD de São José do Rio Preto</u> .....	124
<u>Figura 30 –Relação entre os procedimentos de coleta, desafios e ciclo de dados</u> .....	127

## LISTA DE QUADROS

<u>Quadro 1 – Relação entre as perguntas feitas na análise da organização do Banco TTDii e suas respectivas descobertas .....</u>	58
<u>Quadro 2 – Relação entre procedimentos e objetivos específicos .....</u>	63
<u>Quadro 3 - Síntese dos questionamentos, procedimentos e fontes .....</u>	71
<u>Quadro 4 – Protocolo de criação do IT.....</u>	76
<u>Quadro 5 – Padronização da nomenclatura dos arquivos .....</u>	83
<u>Quadro 6 – Exemplo de nomeação de dados do participante I8F7 (UGA3i – 2015).....</u>	84
<u>Quadro 7 – Sistema de Transcrição para o MulTeC.....</u>	102

## LISTA DE ABREVIATURAS E SIGLAS

TTD – Teletandem

TTDii – Teletandem institucional integrado

TTDisi – Teletandem institucional semi-integrado

SOT – Sessão Oral de Teletandem

SOTi – Sessão Oral de Teletandem inicial

SOTin – Sessão Oral de Teletandem intermediária

SOTf – Sessão Oral de Teletandem final

DOTI – Databank of Oral Teletandem Interactions

IS – Interaction Space

LC – Linguística de *Corpus*

IoT – Internet of Things

TCLEs – Termos de Consentimento Livre e Esclarecido

TICEs – Tecnologias de Informação e Comunicação na Escola

CETIC – Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação

## SUMÁRIO

<b>INTRODUÇÃO</b> .....	14
<b>CAPÍTULO 1: EMBASAMENTO TEÓRICO</b> .....	19
1.1 TELETANDEM: TELECOLABORAÇÃO PARA APRENDIZAGEM DE LÍNGUAS ESTRANGEIRAS.....	19
1.1.1 AS MODALIDADES DE (TELE)TANDEM .....	21
1.1.1.1 <i>Teletandem institucional: integrado, não-integrado e semi-integrado</i> .....	24
1.2 O TELETANDEM E A PRODUÇÃO DE DADOS: CONCEITOS IMPORTANTES.....	27
1.2.1 CENÁRIOS DE APRENDIZAGEM/PEDAGÓGICO E A RELAÇÃO ENTRE MACROTAREFAS, TAREFAS E MICROTAREFAS .....	32
1.2.1.1 RELAÇÃO ENTRE MACROTAREFA, TAREFAS E CENÁRIO DE APRENDIZAGEM NO TTD.....	36
1.3 A LINGUÍSTICA DE <i>CORPUS</i> .....	39
1.3.1 COMPREENSÃO DA LINGUÍSTICA DE <i>CORPUS</i> E TIPOLOGIA DE <i>CORPUS</i> .....	40
1.3.2 <i>CORPUS</i> MULTIMODAL: A NOVA GERAÇÃO DOS CORPORA .....	41
1.4 O GERENCIAMENTO DE DADOS PARA CONSTRUÇÃO DE <i>CORPUS</i> DE PESQUISA.....	46
1.4.1 O PLANO DE GERENCIAMENTO DOS DADOS .....	50
1.4.2 OS METADADOS NO DOTI (DATABANK OF ORAL TELETANDEM INTERACTIONS) .....	53
<b>CAPÍTULO 2: PERCURSOS METODOLÓGICOS</b> .....	55
2.1 ABORDAGEM DA PESQUISA.....	55
2.2 CONTEXTUALIZANDO A PESQUISA .....	57
2.2 AS 3 ETAPAS PARA A CONSTRUÇÃO DO <i>CORPUS</i> .....	61
2.2.1 UM OBJETIVO: VÁRIOS PROCESSOS .....	62
2.2.1.1 <i>O início do processo de organização: coleta de informações sociolinguísticas</i> .....	63
2.2.1.1.1 O TCLEs .....	66
2.2.1.1.2 Proficiência dos participantes .....	69
2.2.1.1.3 Documentos por turma TTD no Banco TTDii .....	72
2.2.1.2 <i>Procedimentos para o tratamento dos dados: anonimização, conversão e nomeação</i> .....	75
2.2.1.2.1 Definição do código de identificação do participante .....	76
2.2.1.2.2 Anonimização dos dados .....	77
2.2.1.2.3 Conversão dos dados .....	81
2.2.1.2.4 A padronização da nomeação dos arquivos .....	82
2.2.1.3 <i>Organização dos arquivos no MulTeC: etapa 3 da organização do MulTeC</i> .....	85
2.2.1.3.2 Documentos com códigos utilizados na constituição do <i>corpus</i> .....	86
2.2.2 <i>Documentos contextuais</i> .....	88
2.2.2.1 Contexto TTD .....	88
2.2.2.2 Documentos com Códigos utilizados na nomeação e anonimização .....	89

2.2.2.3 Códigos das planilhas .....	90
2.2.3 PLANILHAS.....	92
2.2.3.1 Planilha ‘Informações participantes’ .....	93
2.2.3.2 Planilha ‘Levantamento de Dados’ .....	95
2.2.3.2.1 Levantamento de dados – aba 1.....	95
2.2.3.2.2 Contagem das palavras – aba 2.....	100
2.2.3.3 Planilha ‘Levantamento Geral’ .....	104
2.2.4 O CICLO DOS DADOS NO MULTEC .....	106
2.2.5 CABEÇALHOS.....	108
2.2.6 CONVERSÃO DOS DADOS .....	110
2.3 TIPOLOGIA DO MULTEC.....	112
<b>CAPÍTULO 3: RESULTADOS E DISCUSSÃO.....</b>	<b>113</b>
3.1 RESULTADOS: O MULTEC E A PROPOSTA DE COLETA PARA O TTD.....	113
3.1.1 A ORGANIZAÇÃO DAS PASTAS .....	113
3.5 DOCUMENTOS PROPOSTOS: CENÁRIOS .....	115
3.6 A PROPOSTA DOS CENÁRIOS .....	116
3.6.1 CENÁRIO PEDAGÓGICO .....	116
3.6.2 CENÁRIO DE APRENDIZAGEM.....	120
3.7 A PROPOSTA DE COLETA AO TTD .....	123
<b>CONSIDERAÇÕES FINAIS.....</b>	<b>129</b>
OBJETIVOS E RESULTADOS ALCANÇADOS .....	129
LIMITAÇÕES DA PESQUISA.....	130
ENCAMINHAMENTOS FUTUROS .....	130
<b>REFERÊNCIAS .....</b>	<b>133</b>
<b>Anexo 1 – Lista de presença .....</b>	<b>137</b>
<b>Anexo 3 – Tutorial.....</b>	<b>140</b>
<b>Anexo 4 - Tabela de Cursos Unesp.....</b>	<b>144</b>
<b>Anexo 5 - Contexto TTD .....</b>	<b>145</b>
<b>Anexo 6 – Nomeação dos dados.....</b>	<b>155</b>
<b>Anexo 7 – Plano de gerenciamento de dados - MulTeC .....</b>	<b>159</b>
<b>Anexo 8 – Organização do MulTeC.....</b>	<b>161</b>

## INTRODUÇÃO

“O trabalho colaborativo, a socialização em comunidades de pessoas que se reúnem por afinidade e por grupos de interesses comuns, são a tônica dos ambientes digitais” (COSCARELLI, 2016, p. 13)

Em tempos nos quais se discute a chegada da web 3.0, ou web semântica, que sinaliza para uma navegação muito mais significativa e produtiva, bem como a internet das coisas (IoT – *Internet of Things*), a utilização da internet parece cada dia mais fazer parte da vida da maioria dos seres humanos, como presença quase inevitável em seu cotidiano. No Brasil, segundo pesquisa realizada pelo Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br<sup>1</sup>), em 2016, 93% dos domicílios brasileiros tinham celulares e, em 79% deles, havia possibilidade de acesso à internet por quaisquer de seus moradores a qualquer momento. Tal cenário possibilita que pessoas geograficamente distanciadas possam identificar aqueles que têm os mesmos propósitos e, então criar grupos que nascem do compartilhamento desses objetivos, realizar trabalhos juntos, partilhando e gerando conhecimento colaborativamente via ambientes digitais. Esses ambientes, aos quais se refere Coscarelli (2016) na epígrafe, além de facilitarem formação de comunidades por compartilhamento de propósitos, ainda oportunizaram a emergência de um movimento na Ciência conhecido como *Open Science* (Ciência Aberta, em tradução livre).

Segundo Albagli (2015) *Open Science* se “[...] torna um termo abrangente que vai além do livre acesso a publicações científicas e inclui outros tópicos como dados científicos abertos, ferramentas científicas abertas [...]”<sup>2</sup> (ALBAGLI, 2015, p. 14-15). No mesmo sentido, Delfanti e Pitrelli (2015, p. 59) afirmam que o termo “é um conceito amplo que engloba diferentes práticas e ferramentas relacionadas ao uso colaborativo de tecnologias digitais e ferramentas de propriedade intelectual alternativas”<sup>3</sup>. Desse modo, ciência aberta se faz em todas as fases da pesquisa científica incluindo a produção e organização de dados para pesquisa de modo que sejam compartilháveis, assim como a publicação de resultados em mídias de livre acesso. *Open*

<sup>1</sup> <http://data.cetic.br/cetic/>

<sup>2</sup> No original: “Open Science becomes an umbrella term that goes beyond free access to scientific publications and includes other topics such as open scientific data, open scientific tools” (Tradução feita pela autora desta tese)

<sup>3</sup> No original: “is a very broad concept that encompasses several different practices and tools linked to the use of collaborative digital technologies and alternative intellectual property tools”. (Tradução minha)

*Science* é, portanto, um paradigma de ciência no qual os pesquisadores realizam seus trabalhos pensando no compartilhamento não apenas do produto final, por meio das publicações de seus resultados, mas viabilizando também o compartilhamento dos dados oriundos do processo da pesquisa, das anotações feitas e ferramentas utilizadas.

A rede mundial de computadores ainda tem oportunizado contexto de intercâmbio virtual<sup>4</sup> para aprendizagem de línguas a partir do uso das ferramentas de comunicação síncrona disponíveis na rede mundial de computadores, como por exemplo a aprendizagem telecolaborativa via teletandem.

Tanto *Open Science*, como a criação de grupos de interesse de qualquer natureza em ambientes digitais e o intercâmbio virtual, apesar de serem de níveis diferentes, podem ser tomados como exemplos de compartilhamento à distância que foram tornados mais céleres e amplos a partir da *web 2.0*. Segundo Primo (2007, p. 2), a internet dessa geração “caracteriza-se por potencializar as formas de publicação, compartilhamento e organização de informações, além de ampliar os espaços para a interação entre os participantes do processo”.

A formação de comunidades geograficamente distantes, agregadas virtual e sincronicamente só se tornou possível por meio da rede mundial de computadores, pois a comunicação por telegramas é uma possibilidade bem anterior a ela, mas seu caráter assíncrono a diferencia do uso da web. Essas comunidades são movidas pelo compartilhamento de propósitos com trocas mais fluídas e em menos tempo que as tecnologias anteriores como o telefone e o telegrama. Segundo Silva (2012, p. 85), o TTD é uma comunidade que “se estabelece a partir de afinidades de interesses, objetivos em comuns, partilha de um mesmo conhecimento e de uma prática”. Essa comunidade existe pelo desejo que pares interagentes têm de que ocorra aprendizagem de uma língua estrangeira. Assim, esse propósito une seus participantes em um contexto digital que desconhece fronteiras geográficas e torna factível o intercâmbio cultural tão amplo e facilitado como inimaginável até algumas décadas atrás.

No contexto TTD, participantes proficientes em línguas diferentes se dispõem a serem tutores um do outro nas respectivas línguas-alvo, de modo que, por exemplo, o participante proficiente em português auxilia a aprendizagem do parceiro, proficiente em inglês – língua-alvo do primeiro. Os participantes devem respeitar os princípios basilares da aprendizagem *in*

---

<sup>4</sup> Gostaria de destacar que essa terminologia – intercâmbio virtual – é utilizada por O’Dowd (2018) para nomear o “engagement of groups of learners in *on-line* intercultural interactions and collaboration projects with partners from other cultural contexts or geographical locations as an integrated part of their educational programmes (p. 1)”

*tandem* inicialmente propostos por Brammerts (1996), de autonomia e reciprocidade, mantidos por Telles (2006), como é possível perceber no excerto em que afirma que a decisão sobre “o como” e “o que” irão aprender deve ser feita “em comum acordo por ambos os participantes, em consonância com um dos dois princípios da noção de tandem – a “autonomia dos aprendizes” (o outro princípio é a “reciprocidade”)” (TELLES, 2006, p. 10). No entanto, há ainda um terceiro princípio, a separação de línguas. Desse modo, no TTD, os princípios são três, a saber: autonomia, reciprocidade e separação de línguas. Seguindo esses princípios, os participantes devem: i) sob o princípio da autonomia, estabelecer seus objetivos de aprendizagem e o modo como desejam aprender; assim como ii) respeitar as necessidades de aprendizagem do parceiro, sendo-lhe recíproco (reciprocidade) e iii) dividir igualmente o tempo dedicado às línguas que objetivam aprender durante a parceria (separação de línguas).

Além de ser projeto de aprendizagem de línguas, o TTD também se caracteriza como projeto de pesquisa, envolvendo desde 2006, ano de início de suas atividades, até o presente momento, diversos pesquisadores que utilizaram (e ainda utilizam) os dados produzidos pelos participantes para investigações científicas. Cada pesquisador interessado em utilizar os dados produzidos no TTD fazia sua coleta a partir de suas perguntas de pesquisa, organizando os dados de seu estudo conforme protocolos estabelecidos individualmente, o que implicava em uma multiplicidade de formas de sistematização de material documentário e uma consequente impossibilidade de sistematização de um banco de dados mais abrangente, capaz de favorecer futuros estudos<sup>5</sup>.

A partir de 2011, conforme Aranha, Luvizari-Murad e Moreno (2015) o grupo de pesquisadores em TTD da Unesp de São José do Rio Preto começou a realizar coleta dos dados produzidos pelos participantes para futuras pesquisas, a serem utilizados por pesquisadores do TTD. Assim, as atividades realizadas entre os participantes, a saber: as interações via vídeo, os chats, os textos produzidos para revisão pelo parceiro, os diários de aprendizagem (produzidos para ajudar na reflexão sobre a aprendizagem via TTD) e os questionários iniciais e finais, foram sendo coletados e organizados conforme protocolo estabelecido pelo grupo de pesquisa e detalhado por Aranha, Luvizari-Murad e Moreno (2015), criando o Banco TTDii - Banco de Dados para pesquisas sobre aprendizagem via Teletandem institucional integrado<sup>6</sup> (TTDii).

---

<sup>5</sup> Informações mais detalhadas sobre pesquisas realizadas no TTD podem ser obtidas <https://teletandemriopreto.wixsite.com/ibilce/pesquisas>

<sup>6</sup> O teletandem institucional integrado é uma modalidade possível no TTD. Discorro sobre este trabalho mais detalhadamente sobre as modalidades de TTD no capítulo de fundamentação teórica.

Em síntese, uma quantidade significativa de dados foi salva e armazenada, contudo, os dados não estavam de acordo com o previsto para que pudessem ser compartilháveis, ou seja, havia necessidade de anonimização, padronização da nomeação e elaboração de documentos contextuais. A dimensão do problema engendrado pela ausência de procedimentos relevantes para tornar os dados compartilháveis, evidencia-se especialmente, quando consideramos que o compartilhamento possibilitaria que pesquisadores das mais diversas áreas pudessem acessá-los, contribuindo ainda mais para o desenvolvimento da ciência.

Desse modo, a presente pesquisa emerge da necessidade de potencializar o alcance do relevante trabalho realizado pelo grupo, tornando o Banco TTDii em um *corpus*. Assumindo essa missão e partindo da hipótese de que seria possível construir um *corpus* a partir dos dados coletados, estabeleci as seguintes perguntas de pesquisa:

- Quais procedimentos precisam ser adotados para que o banco de dados TTDii possa se constituir um *corpus* cujos dados possam ser compartilháveis com pesquisadores que não participaram da coleta?
- Quais informações precisam ser acrescentadas?
- Quais documentos precisam ser criados para auxiliar o uso dos dados em pesquisas?

Ao responder essas perguntas alcanço o objetivo geral desta pesquisa, a saber:

- Construir um corpus multimodal em teletandem com as produções coletadas no contexto TTD da Unesp de São José do Rio Preto, a partir do levantamento, anonimização e organização dos dados produzidos no período de 2012 a 2015.

Para alcançar esse objetivo, defini como objetivos específicos os seguintes:

- Analisar a organização dos dados produzidos e coletados no período de 2012 a 2015, conforme descrito em Aranha, Luvizari-Murad e Moreno (2015);
- Propor padrão de anonimização e nomeação dos arquivos (BRINEY, 2015; CHANIER; WIGHAM, 2016; EYNDEN et al., 2011);
- Organizar e analisar os dados a partir dos pressupostos teóricos de organização de dados de pesquisa (BRINEY, 2015; CHANIER; WIGHAM, 2016; EYNDEN, 2013; EYNDEN et al., 2011), elaborando e acrescentando documentos para a construção de um corpus de pesquisa.

Esta pesquisa então apresenta, portanto, um *corpus* a ser compartilhado. Esta tese, que discute a construção desse *corpus*, tem as informações distribuídas em três capítulos. No primeiro capítulo, apresento o embasamento teórico utilizado para a realização do trabalho,

mais especificamente, partindo do conceito de telecolaboração e do contexto TTD de aprendizagem de línguas, discutindo, em seguida, o conceito de tarefas assim como a constituição de *corpus* a partir dos pressupostos da Linguística de *Corpus* (LC) e dos protocolos internacionais para a construção de *corpus* pesquisa.

No segundo capítulo, discorro sobre a metodologia utilizada na organização do *Multimodal Teletandem Corpus* (MulTeC), explicitando os procedimentos e padrões adotados a partir das leituras de trabalhos que organização de *corpus* como os de Chanier e Wigham (2016), Briney (2015) e Eynden *et al.* (2011), bem como os resultados obtidos a partir da organização dos dados.

No capítulo três, apresento os resultados e uma discussão a partir do que experienciei durante a organização do MulTeC, no qual apresento ainda a proposta para coleta e organização dos dados produzidos e coletados no contexto de aprendizagem de línguas TTD. Encerro esta tese com as considerações finais com a retomada dos objetivos de pesquisa, discussão sobre as limitações da pesquisa e encaminhamentos futuros.

O percurso discursivo se inicia com os embasamentos teóricos que viabilizaram um olhar sobre os dados produzidos, seu contexto de produção e estratégias de gerenciamento de dados que otimizem o tempo dedicado a coleta e organização.

## CAPÍTULO 1: EMBASAMENTO TEÓRICO

O capítulo que se segue está dividido em quatro seções. A primeira apresenta ao leitor uma discussão sobre o teletandem como projeto telecolaborativo para aprendizagem de línguas, entendendo telecolaboração nos pressupostos de Dooly e O’Dowd (2018), O’Dowd (2016) e do que é proposto por Telles (2006), a partir da aprendizagem *in tandem* discutida por Brammerts (1996). Ainda nessa seção apresento, principalmente a partir de Aranha e Cavalari (2014), Salomão (2006, 2008, 2010) e Zakir (2015), as modalidades de teletandem cuja compreensão é relevante para este trabalho.

Na seção 2 do capítulo, faço a apresentação do teletandem como um *locus* profícuo de produção de dados para pesquisas, apresentando o DOTI (Databank of Oral Teletandem Interaction), proposto por Aranha e Leone (2016;2017), além dos conceitos utilizados por essas autoras para caracterizar o TTD: espaço de interação, discutido por Chanier *et al* (2014) e os cenários de aprendizagem e pedagógico, bem como a divisão de tarefas em categorias como discutido por Foucher (2010) e Guichon (2008).

A terceira seção do capítulo discorrerá sobre a Linguística de *Corpus* a partir de Tagnin (2018), Berber Sardinha (2004, 2002, 2000) e Sinclair (2004), constituído de duas subseções: a primeira com a apresentação dos trabalhos sobre os *corpora* multimodais de Knight (2011), Reffay, Betbeder e Chanier (2013), Chanier e Wigham (2016) e Guichon (2017), e a segunda apresenta embasamentos sobre o gerenciamento de dados para construção de *corpus* de pesquisa. Nessa última, apresento, embasada nos trabalhos de Briney (2015), Eynden (2013) e Eynden *et al* (2011), procedimentos a serem obedecidos na construção de um *corpus* cujo objetivo final é o compartilhamento dos dados com outros pesquisadores.

### 1.1 TELETANDEM: TELECOLABORAÇÃO PARA APRENDIZAGEM DE LÍNGUAS ESTRANGEIRAS

Segundo Dooly e O’Dowd (2018), “telecolaboração envolve engajar aprendizes geograficamente distribuídos em algum tipo de interação para um propósito verdadeiramente comunicativo (idealmente para co-construir conhecimento de algum tipo).” (DOOLY;

O'DOWD, 2018, p. 19)<sup>7</sup>. Os autores afirmam que, embora a telecolaboração<sup>8</sup> não seja recente na área da educação, havendo registros de telecolaboração no final dos anos 1800 e início dos anos 1900, quando algumas escolas incentivavam a troca de cartas entre turmas de estudantes geograficamente distantes, com o advento das tecnologias de comunicação e informação torna-se mais fácil propor e realizar projetos telecolaborativos. Um exemplo de projeto telecolaborativo é o Teletandem, idealizado por Telles (2006) no âmbito da Unesp, inicialmente nos campi de Assis e São José do Rio Preto.<sup>9</sup>

Telles (2006) propôs o “Projeto Teletandem Brasil: Línguas Estrangeiras para Todos” a partir da noção de aprendizagem *in tandem* de Brammerts (1996), que a define como “uma forma de aprendizagem aberta, em que duas pessoas com diferentes idiomas nativos trabalham juntos em pares” (BRAMMERTS, 1996, p.10) para aprender uma língua estrangeira e auxiliar seu par na aprendizagem da língua estrangeira dele e também discutir questões culturais e trocar conhecimento sobre assunto de interesse. No mesmo texto, Brammerts apresenta os princípios basilares da aprendizagem *in tandem*: a autonomia e a reciprocidade, citando a separação de línguas como uma consequência da reciprocidade, pois ao serem recíprocos, os participantes dedicam tempo igual para a aprendizagem um do outro como afirma no excerto:

“[...] aprendizagem bem-sucedida *in tandem* está baseada na dependência recíproca e apoio mútuo dos parceiros; [...] Eles devem não apenas dedicar a **mesma quantidade de tempo para cada língua**: eles também devem investir a mesma quantidade de energia na preparação, no interesse que mostram pelo sucesso da aprendizagem do parceiro [...] (BRAMMERTS, 1996, p. 11)<sup>10</sup> (Meu destaque)

Vassallo e Telles (2006) apresentam o princípio da separação de línguas para o teletandem como apresentei na introdução deste trabalho. Para eles, esse é um dos três princípios, segundo o qual os participantes devem pautar sua participação, além da autonomia e da reciprocidade, princípios já estabelecidos por Brammerts (1996).

<sup>7</sup> Minha tradução para “Telecollaboration involves engaging geographically distributed learners in some sort of interactions for a truly communicative purpose (ideally to co-construct knowledge of some sort).”

<sup>8</sup> A Moderna Gramática Portuguesa, de Bechara (2004), apresenta, no capítulo intitulado “Formação de Palavras do Ponto de Vista Construcional”, menção ao “**tele-**” como um **prefixo** ou elemento grego, cujo significado é “distância, afastamento, controle feito à distância”, exemplificando com telégrafo, telepatia, teleguiado.

<sup>9</sup> Em Araraquara o TTD teve início em 2012. O leitor poderá obter mais informações sobre o TTD no campus de Araraquara na página <http://teletandem.wixsite.com/fclar/historico>

<sup>10</sup> Minha tradução para: “[...] successful learning in tandem is based on the reciprocal dependence and mutual support of the partners; [...] They should not only dedicate the same amount of time to each language: they should also invest the same amount of energy in preparation, in the interest they show in the learning success of their partner [...]” (BRAMMERTS, 1996, p. 11)

Brammerts (1996) afirma que, devido ao princípio da autonomia, os participantes são os responsáveis por sua aprendizagem. No que se refere ao princípio da reciprocidade, o autor destaca a importância da dedicação mútua entre os parceiros. Tal assertiva ainda ressalta a relevância do comprometimento por parte dos participantes com a aprendizagem um do outro.

Vale lembrar que a aprendizagem colaborativa de línguas *in tandem* não se constitui uma novidade. Desde a década de 60, parceiros na Europa se dispunham a realizar encontros presenciais nos quais um auxiliava o outro na aprendizagem da língua-alvo, conforme Brammerts (1996). O teletandem proposto por Telles (2006) é presença virtual e mantém os princípios do *tandem*.

Desse modo, seguindo os princípios da aprendizagem *in tandem*, o TTD, que se caracteriza como “um tandem virtual que utiliza as ferramentas on-line de escrita, leitura, áudio e vídeo” (TELLES e VASSALO, 2006, p. 189), foi proposto por Telles, tendo em 2006 o início de sua implementação. Conforme Vassalo (2010, p. 17) “o desenvolvimento do tandem pode ser integrado às atividades curriculares de uma instituição, ou pode ser realizado de forma completamente independente”, caracterizando diferentes modalidades que apresento na próxima subseção.

### 1.1.1 AS MODALIDADES DE (TELE)TANDEM

No livro em que Brammerts (1996) discute os princípios do *tandem*, Calvert (1996) menciona a integração da aprendizagem *in tandem* via e-mail, se referindo à integração do *tandem* nos cursos<sup>11</sup>, uma modalidade possível de *tandem*. Contudo, é em 2002 que Brammerts (*apud* SALOMÃO, 2006) apresenta as características das diferentes modalidades de *tandem*, sempre mantendo os princípios basilares apresentados em Brammerts (1996). Brammerts (2002 *apud* SALOMÃO, 2006) se refere a diferenças de modalidades por:

- i) Contexto físico: *tandem* presencial e *tandem* a distância, esse último variando conforme a ferramenta tecnológica utilizada para a comunicação: *tandem* telefônico e *e-tandem*, via e-mail ou chat;
- ii) Contexto: *tandem* institucional – os participantes são vinculados a uma instituição de ensino; semi-institucional – institucional para apenas 1 dos participantes; e não-institucional – nenhum dos participantes está vinculado a uma instituição. O contexto institucional, segundo Salomão (2006) pode variar entre:

---

<sup>11</sup> “In order for e-mail tandem learning to be integrated into courses, account must be taken of three key factors: skills, facilities, and accreditation.” (CALVERT, 1996, p. 35)

- a. Integrado – vinculado a uma disciplina de LE, objeto de avaliação por um professor;
  - b. Complementar – “Reconhecido pela instituição para parte integrante do curso” (BRAMMERTS *et al*, 2002, p. 86)
  - c. Opcional – Reconhecido pela instituição, mas não integra um curso.
  - d. Não-integrado – A instituição disponibiliza os recursos para realização do tandem, mas não o reconhece oficialmente.
- iii) Participantes: *tandem* didático – entre alunos de cursos de línguas; *tandem* leigo – entre pessoas que não participam de cursos de línguas; *tandem* formativo – os participantes são professores de língua em formação; e *tandem* profissional – participantes que atuam em mesma área profissional.

Assim, considerando a vinculação do *tandem* à uma instituição, as modalidades podem ser institucional, semi-institucional e não-institucional. Enquanto *tandem* institucional, pode haver as variações: integrado, opcional, complementar e não-integrado. Essas modalidades de *tandem* têm sido transpostas à realidade do teletandem (SALOMÃO, 2006; ARANHA e CAVALARI, 2014; ZAKIR, 2015; dentre outros).

Desse modo, o *tandem* integrado em Brammerts (2002) corresponde ao teletandem integrado, por exemplo, levando em consideração que ambos contextos de aprendizagem estão embasados na aprendizagem (tele<sup>12</sup>)colaborativa em que os princípios de autonomia, reciprocidade e separação de línguas, devem ser respeitados.

Do mesmo modo, como mostro na figura 1, no TTD são dois níveis diferentes de vinculação que caracterizam as modalidades: o primeiro nível está relacionado ao âmbito institucional, o segundo é referente a vinculação do contexto TTD a um(a) curso/disciplina. Desde a implementação, o TTD é institucional, promovido pela Unesp, nos campi de Assis, São José do Rio Preto e posteriormente em Araraquara. No âmbito da vinculação do TTD com uma disciplina, há 5 diferentes modos de realização do teletandem institucional: integrado – cuja implementação foi descrita por Aranha e Cavalari (2014), semi-integrado, complementar, opcional e não-integrado.

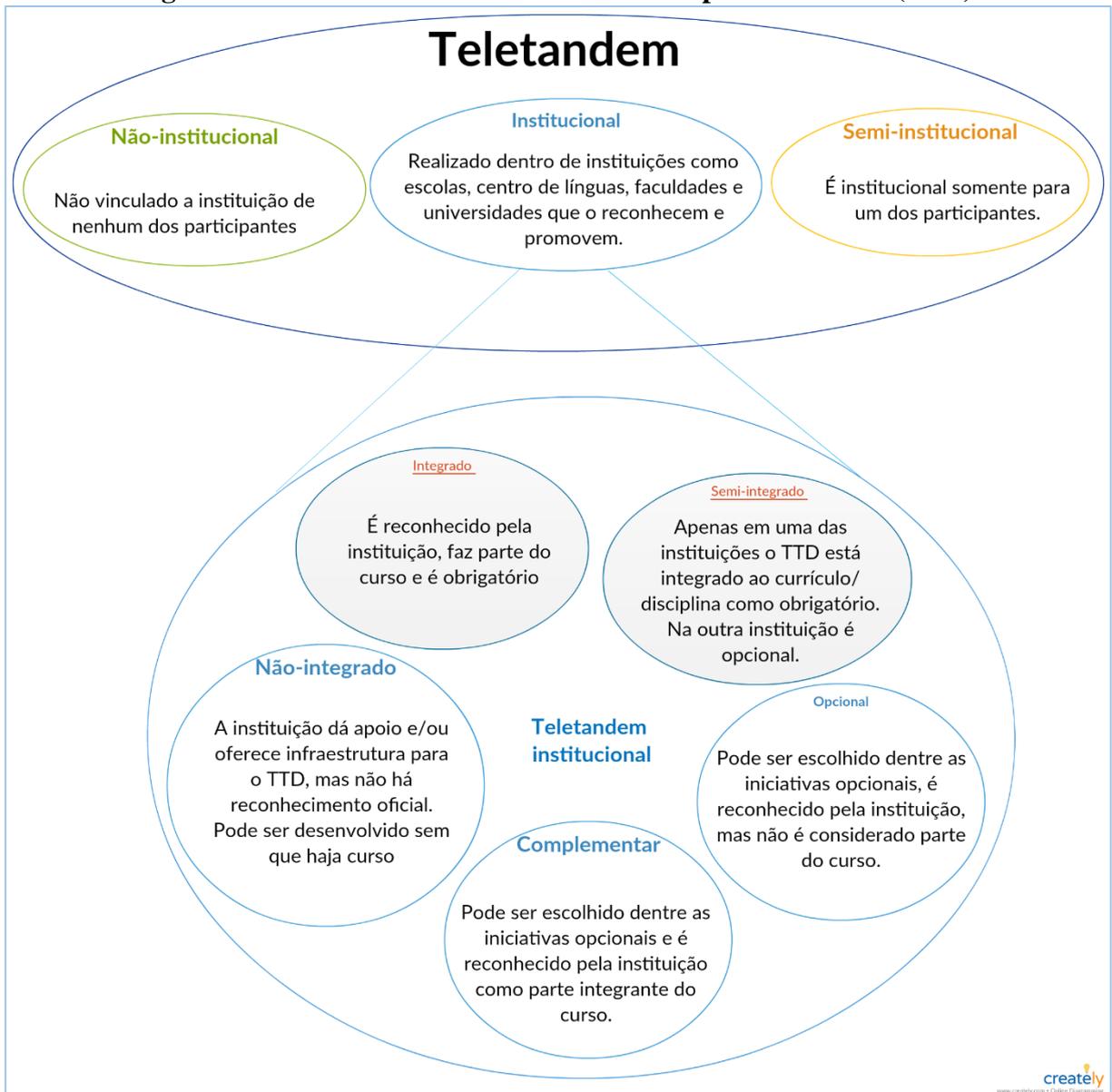
Na modalidade de teletandem institucional integrado (TTDii), o contexto teletandem está integrado à disciplina de língua estrangeira, sendo objeto de avaliação para aferição de nota na disciplina, pelo professor regente, todas as atividades realizadas pelos participantes, ocorrendo durante o horário da disciplina à qual o TTD está integrado. Sobre as demais

---

<sup>12</sup> No teletandem.

modalidades de TTD, ainda há poucos trabalhos que discutem suas características diferenciais. A figura a seguir resume a caracterização das modalidades de teletandem e foi elaborada a partir do trabalho de Zakir (2015).

**Figura 1 – As modalidades de teletandem adaptado de Zakir (2015)**



Fonte: Elaborado pela autora a partir de Zakir (2015, p. 35)

Na figura anterior, priorizei a diferenciação das modalidades nos dois níveis: no primeiro nível, estão as modalidades possíveis no que se refere ao vínculo institucional dos participantes do TTD, enquanto que as modalidades cuja caracterização está vinculada à integração do TTD a uma disciplina estão em um segundo nível, todas sendo institucionais. As

modalidades de institucional em destaque, integrado e semi-integrado, são aquelas nas quais os dados utilizados nesta pesquisa foram produzidos.

No período de 2006 a 2010, na Unesp em São José do Rio Preto, conforme Aranha e Cavalari (2014), a atividade de teletandem ocorria na modalidade institucional não-integrado, que se caracteriza pelo reconhecimento da instituição educacional que o promove e viabiliza, como afirmam Salomão (2006), Zakir (2015) e Aranha e Cavalari (2014). Não havia até 2011 integração do TTD em uma disciplina de língua estrangeira. Nessa modalidade de teletandem, as parcerias eram consideradas independentes, como se referem Salomão (2008, 2010) e Cândido (2010) ao optar pelo termo “parcerias independentes” em seus trabalhos. Essas autoras descrevem tais parcerias como aquelas que ocorreram de modo que os participantes, que formaram cada dupla, tinham total autonomia para definir quando e por quanto tempo realizariam os encontros virtuais. As “parcerias independentes” contavam com o apoio de professores de línguas, também pesquisadores (alunos de pós-graduação interessados em investigar o TTD), orientando-os no sentido de auxiliá-los a refletir sobre a atividade de teletandem. Tal modalidade foi a única a ocorrer no contexto da Unesp de São Jose do Rio Preto até 2011, a partir de quando inicia a oferta de teletandem institucional integrado (TTDii).

Cavalari e Aranha (2016) descrevem como sendo características distintivas do teletandem institucional integrado: a) o uso da reunião ‘tutorial’ para dar orientações gerais quanto a questões teóricas e práticas da modalidade; b) integração das atividades TTD no currículo de disciplina de língua estrangeira; e c) avaliação integradora, uma avaliação pelo participante com quem forma parceria, de si mesmo, e pelo professor regente da disciplina de LE à qual o TTD está integrado.

#### *1.1.1.1 Teletandem institucional: integrado, não-integrado e semi-integrado*

Apresento, nesta subseção, as relações que caracterizam as modalidades integrado, não-integrado e semi-integrado. Essas modalidades têm como característica convergente o fato de que são apoiadas e promovidas por uma instituição que fornece infraestrutura para sua realização e as reconhece como atividade institucional. O fator que as difere reside na integração ou não do TTD a uma disciplina de LE nas duas instituições que constituem a parceira como discutem Aranha e Cavalari (2014).

Na modalidade de TTDii, “a participação no TTD deixa de ser voluntária por parte do participante e passa a ser parte integrante de suas atribuições acadêmicas/pedagógicas dentro de uma disciplina” (ARANHA; CAVALARI, 2014, p. 185). Segundo as autoras, os elementos que distinguem o TTDii das modalidades em que o TTD não está integrado são:

- As Sessões Orais de Teletandem (SOTs) (na ocasião nomeadas “interações”<sup>13</sup>) ocorrem durante o horário da disciplina à qual o TTD está integrado;
- Há troca de textos produzidos em LE para correção entre os parceiros;
- Temas relacionados aos programas das disciplinas são propostos para as interações;
- Os aprendizes participam de uma reunião “tutorial” (presencial ou através de videoconferência), durante a qual são introduzidos ao contexto TTD, seus princípios basilares, cronograma da parceria a qual farão parte, além das tarefas planejadas;
- Os participantes respondem a questionários cujo intuito é auxiliar no estabelecimento das metas individuais de aprendizagem, bem como suscitar a reflexão quanto ao processo de aprendizagem no contexto TTD;
- Há a produção de diários<sup>14</sup> de aprendizagem após cada SOT, no intuito de provocar a reflexão sobre seu percurso de aprendizagem no TTD;
- Há aferição de nota pelo professor regente da disciplina de LE no tocante à participação nas tarefas do TTD. Essa avaliação fica a critério do professor.

Nas palavras das autoras, foi a modalidade TTDii “que permitiu a criação de um banco de dados que contempla todos os passos do processo em cada uma das parcerias”(ARANHA; CAVALARI, 2014, p. 197). De acordo com as autoras, os Termos de Consentimento Livre e Esclarecido (TCLE) eram assinados ao término do semestre, caso o participante concordasse em doar os dados produzidos durante o período para pesquisas. Os documentos de textos eram arquivados pelos participantes na plataforma Teleduc® enquanto que os arquivos em vídeo e/ou áudio eram gravados, utilizando ferramenta de captura de tela<sup>15</sup>, e posteriormente salvos na área de trabalho da máquina utilizada pelo participante para as interações (SOTs), levando em consideração que cada participante usava a mesma máquina durante todo o período de interações do semestre.

---

<sup>13</sup> Desde 2017 esses eventos comunicativos no TTD da Unesp de São José do Rio Preto foram nomeados como Sessão Oral de Teletandem, a partir de uma sugestão da Prof<sup>ª</sup>. Dr<sup>ª</sup> Rosinda de Castro Guerra Ramos - no intuito de discernir entre as diversas possibilidades de interação e aquela que ocorre no teletandem via vídeo conferência - por ocasião do debate do trabalho de mestrado de Rampazzo (2017) no VII Seminário de Estudos Linguísticos da Unesp em 2016. Trabalhos desenvolvidos por pesquisadores das outras unidades Unesp que se dedicam ao teletandem podem ainda usar o termo ‘interações’ para o mesmo evento comunicativo durante o qual parceiros interagem via ferramentas de comunicação síncrona, seguindo os princípios de: autonomia, reciprocidade e separação de línguas.

<sup>14</sup> Caracteriza-se como um texto reflexivo que deve ser escrito tão logo a sessão oral de teletandem termine. O participante é convidado a registrar seu processo reflexivo sobre a aprendizagem durante aquela sessão bem como possíveis questionamentos que possam ter sido suscitados durante o encontro virtual.

<sup>15</sup> O aplicativo Evaer® tem sido utilizado como ferramenta de captura de tela para os vídeos/áudios. Mais informações sobre o aplicativo podem ser encontradas em [www.evaer.com](http://www.evaer.com).

No TTD<sub>Disi</sub> (Teletandem institucional semi-integrado), as características de: i) integração do TTD à disciplina de LE, ii) avaliação e iii) uso de tutorial podem estar presentes em apenas um dos lados da parceria. Nas palavras de Zakir (2015, p. 39) o TTD<sub>Disi</sub> acontece “quando apenas uma das universidades parceiras adota a prática de teletandem como parte do currículo”, logo em uma universidade ele está integrado a uma disciplina de LE e em outra não está, no entanto, as duas universidades estabelecem e realizam a parceria, constituindo assim uma turma de TTD<sub>Disi</sub>. Sobre essa modalidade ainda há necessidade de mais trabalhos publicados com uma descrição detalhada de como essa atividade ocorre nos campi Unesp que realizam TTD. Contudo,

Embora ainda haja poucos trabalhos que têm como objeto de estudo o contexto no qual a instituição estrangeira promove o teletandem institucional integrado e a instituição brasileira, o não-integrado, pode-se dizer que nessa configuração há um acompanhamento mais sistematizado das sessões de teletandem e das atividades relacionadas a essa prática. (ZAKIR, 2015, p. 35)

Na afirmação, Zakir (2015) aponta ao leitor a diferença no acompanhamento das atividades do TTD quando as turmas realizam TTD<sub>Disi</sub> em comparação com o institucional não-integrado. A autora afirma haver mais sistematização na instituição onde o TTD está integrado ao currículo de uma disciplina de LE, mesmo que apenas em uma das instituições, do que na instituição em que os participantes são voluntários.

Como afirmei anteriormente, o contexto no qual desenvolvi esta pesquisa envolve tanto teletandem institucional integrado como semi-integrado. Nessas modalidades, aos participantes do TTD tem se oportunizado contexto para aprendizagem de línguas por meio de: Sessão Oral de Teletandem (SOT), chat, diários de aprendizagem, textos produzidos para revisão pelo parceiro, questionários inicial e final. Como será possível perceber até o final deste trabalho a produção desses dados pode ocorrer independente da modalidade, estando sua produção mais diretamente relacionada às decisões tomadas pelos mediadores das turmas envolvidas do que pela modalidade *per se*. Esses dados têm sido utilizados para realização de pesquisas pelo Grupo TTD da Unesp, mas dada a quantidade e multimodalidade que caracteriza o conjunto dos dados, a construção de um *corpus* a ser compartilhado com um número maior de pesquisadores se torna um processo relevante.

## 1.2 O TELETANDEM E A PRODUÇÃO DE DADOS: CONCEITOS IMPORTANTES

Dado que as SOTs, trocas de textos, chats, produção dos diários e respostas aos questionários têm ocorrido mediados por computadores no teletandem, está construído o ambiente propício para a coleta de dados, haja vista a utilização de máquinas e ferramentas que viabilizam a gravação dos dados produzidos e seu posterior arquivamento. Contudo, não apenas o uso das máquinas e suas possibilidades, mas principalmente a oferta de TTDii, como discutem Aranha, Luvizari-Murad e Moreno (2015) mostrou-se o *kairós*<sup>16</sup> para o início de uma coleta sistematizada que culminou na criação do Banco de Dados TTDii, sobre o qual discorreremos mais detalhadamente em nosso capítulo metodológico.

Dentre os dados coletados, as SOTs constituem o DOTI (*Databank of Oral Teletandem Interactions*) (ARANHA, LEONE; 2016, 2017). A apresentação do DOTI neste trabalho é necessária haja vista que ele é constituído por sessões orais de teletandem que integram o *corpus* cuja construção é objeto desta pesquisa. Para a organização desse *corpus* multimodal utilizo conceitos que ampararam a caracterização do TTD pelas autoras para a organização do DOTI.

O DOTI é fruto de uma parceria entre pesquisadoras da Unesp – São José do Rio Preto (Brasil) e da Universidade de Salento (Itália) que compreenderam que, a partir dos dados coletados no Brasil e na Itália, a constituição de um banco com as SOTs das duas universidades poderia contribuir sobremaneira para “linguistas (aplicados), professores e especialistas em computadores que queiram desenvolver *softwares* baseados na Comunicação Mediada por Computador para aprendizagem de línguas”<sup>17</sup> (ARANHA e LEONE, 2016, p.332). Segundo Aranha e Leone (2016), os objetivos do DOTI são: “1) valorizar o teletandem como contexto situado de aprendizagem; 2) fundamentar pesquisa realizada na área; e 3) oferecer a outros pesquisadores a possibilidade de acessar dados para confirmar ou refutar pesquisas publicadas” (ARANHA; LEONE, 2016, p. 327).

A maior parte das SOTs que formam o DOTI foram produzidas pelos participantes da Unesp do campus de São José do Rio Preto e seus parceiros, o que, nesse caso, significa que ele contém gravações das sessões realizadas entre falantes de português e inglês (essas integram o MulTeC), bem como as gravações das sessões entre aprendizes de italiano e inglês produzidos na Universidade de Salento. Aranha e Leone (2016; 2017) embasaram-se nos conceitos de tarefas e cenários de aprendizagem, procurando explicar o que ocorre no TTD, levando em

---

<sup>16</sup> Momento oportuno, certo.

<sup>17</sup> [...] (applied) linguists, professors, and computer experts who want to develop software based on CMC for language learning” (Tradução minha)

consideração sua complexidade, para, a partir desses conceitos proporem a organização do DOTI.

As autoras partem do conceito de espaço de interação, utilizado por Chanier *et al.*, (2014) que definem espaço de interação (Interaction Space – IS) como “um conceito abstrato, localizado no **tempo** (com uma data de início e término com tempo absoluto, portanto, um período de tempo) no qual interações entre **um conjunto de participantes** ocorrem em uma **localização on-line**”. (CHANIER *et al.*, 2014, p.5, grifo dos autores). Tempo pré-determinado, grupo de participantes e localização *on-line* constituem-se características do contexto TTD, o que justifica a utilização do conceito de IS (*Interaction Space* – Espaço de interação) que, para Aranha e Leone (2016), auxilia a caracterizar esse *locus* de aprendizagem de línguas, no qual há um período pré-determinado de interações durante o qual um grupo de participantes encontra-se *on-line*, a partir da ciberinfraestrutura disponível, para aprender um a língua do outro. Gostaria de ressaltar que, neste trabalho, ‘turma TTD’ designa um conjunto de participantes (estrangeiro/brasileiro) que é resultado do pareamento<sup>18</sup> entre duas turmas de universidades parceiras. Esses participantes se reúnem presencialmente (os brasileiros na universidade brasileira e os estrangeiros em suas instituições), com periodicidade determinada no planejamento de cada turma, para interagir via ferramenta de comunicação síncrona com o parceiro com o intuito de aprender a língua na qual o par é proficiente. Cada turma conta com o auxílio de, pelo menos, uma mediadora<sup>19</sup> em cada instituição e tem um período determinado, desde o planejamento, de início e término das atividades.

Outro conceito que as autoras (ARANHA; LEONE, 2016) tomam como relevante para a caracterização do TTD, portanto para a explicação do DOTI, é o conceito de cenário de aprendizagem em contextos de aprendizagem *on-line*, discutido por Mangenot (2008) e Foucher (2010). Segundo esses autores o cenário de aprendizagem pode ser considerado como aquele que é constatado na realização do que foi planejado para a turma de participantes, pois diversas situações podem interferir e ocasionar alguma alteração no planejamento inicial elaborado pelo professor/mediador. Mangenot (2008) e Foucher (2010) entendem que esses cenários se constroem embasados em tarefas que classificam como podendo ser macrotarefas, tarefas e microtarefas. A relação entre essas tarefas (macro, micro e tarefas) serão discutidas detalhadamente adiante. Nessa subseção, considero relevante informar apenas que esses conceitos foram base para Aranha e Leone (2016) na organização do DOTI. As autoras

---

<sup>18</sup> Estabelecimento de pares fixos por período para realização do TTD.

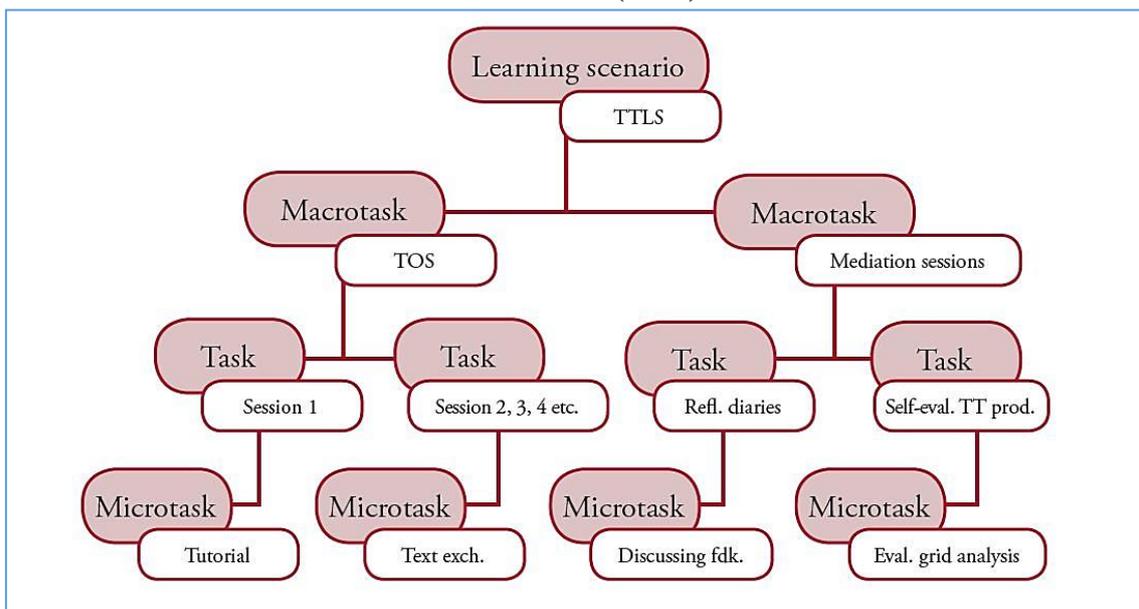
<sup>19</sup> Em São José do Rio Preto todas as ações de mediação são realizadas por mulheres. Desse modo, opto pelo uso do substantivo no gênero feminino neste trabalho ao me referir às responsáveis pela mediação das turmas.

propõem um organograma, neste trabalho apresentado na figura 2, que representa a organização do cenário de aprendizagem no TTD, utilizando os conceitos de macrotarefas, tarefas e microtarefas, também discutidos por Foucher (2010). Em Foucher (2010) a definição de macrotarefa, tarefa e microtarefa<sup>20</sup> é feita a partir da concepção francófona desses termos, no qual há o uso do termo

macrotarefa que retoma uma tarefa global sem foco linguístico preciso, o que parece, portanto, corresponder a tarefa sem foco, tal como definido por Ellis: “Um conjunto realista de ações que leva a uma produção de linguagem não limitada ao universo escolar” (Demaizière & Narcy- Combes, 2005: p.50) “Um projeto de aprendizagem global em que os alunos têm que lidar com informações escritas ou orais na L2 para construir um objeto de sentido escrito ou oral” (Guichon, 2006: p. 54) (FOUCHER, 2010, p. 80)

A partir de Foucher (2010), Aranha e Leone (2017) caracterizam o teletandem como sendo o cenário de aprendizagem no qual há duas macrotarefas que têm suas tarefas e microtarefas diretamente relacionadas a macrotarefa que lhe corresponde. As autoras organizam esse cenário de aprendizagem por meio da figura 2.

**Figura 2 – Organização do cenário de aprendizagem no TTD por Aranha e Leone (2017)**



Fonte: Aranha e Leone (2017, p. 180)

<sup>20</sup> [...] macro-tâche qui renvoie une tâche globale sans visée linguistique précise, ce qui semble donc correspondre à la unfocused task telle qu'elle est définie par Ellis: “un ensemble d’actions réalistes conduisant à une production langagière non limitée à l’univers scolaire” (Demaizière & Narcy- Combes, 2005 : p.50) “un projet d’apprentissage global au cours duquel les apprenants sont amenés à traiter l’information écrite ou orale en L2 pour un construire un objet de sens écrit ou oral » (Guichon, 2006 : p. 54) [...] (Traduzido por Deise Marinoto)

Desse modo, segundo Aranha e Leone (2017), há duas macrotarefas no TTD (as SOTs e as sessões de mediação) e a relação das microtarefas são tipológicas. As microtarefas pertinentes à interação entre os pares estão relacionadas à macrotarefa “sessão de teletandem” (neste trabalho SOT –Sessão Oral de Teletandem). As microtarefas são aquelas nas quais há o objetivo de refletir (como na produção dos diários de aprendizagem e no visionamento<sup>21</sup> da própria gravação da SOT) e/ou receber orientação sobre a aprendizagem no TTD (como no tutorial e nas sessões de mediação). Sob a visão de Aranha e Leone (2016;217) estão relacionadas com a macrotarefa “sessão de mediação” as tarefas os diários reflexivos e autoavaliação, e a macrotarefa ‘sessão oral de teletandem’ tem as sessões orais como tarefas e tutorial e troca de textos como microtarefas, conforme representam na figura 2.

Embora não seja foco deste trabalho a discussão sobre a mediação, e ainda que seja um tema complexo e não se tenha a pretensão de esgotá-lo, há necessidade de explicar ao leitor o conceito de mediação no contexto de produção de tarefas do teletandem de modo que se tenha uma noção básica do que representa a mediação entre os participantes do teletandem, haja vista que os dados foram produzidos nesse contexto.

A mediação, assim como a interação entre os pares, no TTD tem sido objeto de estudo para pesquisadores que realizam suas pesquisas no Projeto TTD. Desde o trabalho de Salomão (2008), que sob a ótica do professor-mediador em um estudo de caso, investigou as estratégias e procedimentos utilizados no processo de mediação no teletandem como a dissertação de Souza (2012) até trabalhos mais recentes, como os de Andreu-Funo (2015) e Freschi e Lopes (2016), pesquisadores do TTD têm atentado para a relevância do papel do mediador no TTD<sup>22</sup>.

Salomão (2008) discute o papel do professor-mediador como aquele que auxilia o participante a realizar suas decisões quanto ao processo de aprendizagem no TTD. Segundo ela, a figura do mediador promove o processo reflexivo relevante para a aprendizagem, ao afirmar que:

Suas estratégias de gerenciamento, ao incluir elementos reflexivos e oferecimento de andaimes, abrangendo uma postura não diretiva, proporcionamento de alternativas em relação a questões procedimentais e (auto-)exploração da prática por meio de

---

<sup>21</sup> Essa prática adotada na Universidade de Salento até o presente momento não tem ocorrido no Brasil. No visionamento, os participantes vêm uma SOT junto com a turma e discutem questões referentes à cultura, língua e atitudes dos parceiros, como afirma Leone (2014, p. 48) Durante oficinas os estudantes são encorajados a investigar sua produção analisando clips que eles selecionaram das gravações das sessões de teletandem (Nossa tradução para: “During workshops students are encouraged to investigate their production analysing clips they have selected from Teletandem sessions’ video-recordings.”)

<sup>22</sup> Cabe-nos informar ao leitor que há muitos outros trabalhos sobre mediação no TTD que podem ser acessadas no site [www.teletandembrasil.org](http://www.teletandembrasil.org) e que não discutiremos aqui por não ser esse o objeto de nossa pesquisa, razão pela qual não discutiremos os dois últimos trabalhos citados.

questionamentos que levassem o interagente a refletir sobre suas decisões, ações e procedimentos em relação ao processo de ensino e aprendizagem colaborativo, trouxeram ao contexto forte carga de autonomia e desenvolvimento. A figura do mediador pareceu também responsável por trazer objetivos pedagógicos para que a relação não se restringisse a um simples bate-papo entre os interagentes. (SALOMÃO, 2008, p. 302)

Nas palavras de Salomão, o leitor é convidado a perceber que o mediador não perpetua a antiga visão de um professor que deve decidir sozinho os caminhos a serem percorridos pelo aprendiz até o alcance dos seus objetivos de aprendizagem. O professor-mediador, a quem me refiro nesta tese como mediador, desempenha papel de problematizador, aquele que ao invés de simplesmente indicar um caminho, apresenta questionamentos e suscita (auto)reflexões quanto às possíveis decisões a serem tomadas no processo, deixando a decisão ser tomada pelo aprendiz, auxiliando-o a desenvolver autonomia. Na mesma esteira, Souza (2012) afirma que o papel do mediador deve confluir de modo que auxilie o participante e “construa com ele uma reflexão sobre a interação que o ajude a identificar possíveis dificuldades, estratégias de aprendizagem e pontos que possam ser melhor explorados pelos interagentes.” (SOUZA, 2012, p.38).

A partir da compreensão do papel do mediador, o leitor poderá compreender que por meio da escrita dos diários de aprendizagem, participação nas sessões de mediação, resposta aos questionários inicial e final, bem como na participação na reunião tutorial, o participante obtém orientação e ainda lhe é proporcionado espaço reflexivo para aprender no TTD da melhor maneira possível para ele. Desse modo, embora considere a ‘interação entre os aprendizes’ como cerne das atividades do TTD, há que se considerar que há duas tarefas como principais, na visão de Aranha e Leone (ARANHA; LEONE, 2016, 2017), duas macrotarefas: a sessão de mediação e as sessões orais de teletandem.

Na subseção a seguir procuro detalhar a relação entre as macrotarefas e microtarefas a partir de Foucher (2010) e Guichon (2008), um dos embasamentos teóricos de Aranha e Leone (2016, 2017) para a compreensão da caracterização das tarefas no TTD em perspectiva um tanto diversa da proposta por Aranha e Leone (2016;2017), porquanto tais conceitos embasam este trabalho também.

### 1.2.1 CENÁRIOS DE APRENDIZAGEM/PEDAGÓGICO E A RELAÇÃO ENTRE MACROTAREFAS, TAREFAS E MICROTAREFAS

A compreensão de cenários e tarefas pela qual opto nesta tese, a partir do trabalho de Aranha e Leone (2016, 2017) é discutida no trabalho de Foucher (2010), no qual a autora discorre sobre a relação entre a Didática das línguas-culturas e as Tecnologias de Informação e Comunicação na Escola (TICEs), entendendo que, no uso das TICEs, é relevante distinguir três tipos de cenário: o de navegação, o pedagógico e o de comunicação. Considerando o contexto TTD onde os dados (vídeos/áudios com as SOTs, arquivos de texto com os chats, textos produzidos para a troca, questionários e diários de aprendizagem) foram coletados, pondero, em consonância com Aranha e Leone (2016, 2017), que a discussão de Foucher (2010) contribui significativamente para a compreensão e caracterização do contexto e como ele ocorre. Para a autora, os cenários (de navegação, pedagógico e de comunicação) estão estritamente relacionados ao cenário de aprendizagem<sup>23</sup>, pois de algum modo influenciam sua completude, bem como é nele que são desenvolvidas as tarefas, durante a realização das quais há a possibilidade da aprendizagem objetivada.

O cenário de navegação em Foucher (2010) consiste numa previsão das possibilidades de percurso do aprendiz no dispositivo ou ambiente, sendo “mais ou menos linear ou limitado conforme as opções didáticas tomadas”<sup>24</sup> (p. 82). Esse cenário, segundo a autora, está estritamente ligado ao cenário de comunicação que ela define, citando Nissen, como a “junção das possibilidades de interação que o aprendiz tem à sua disposição e que lhes são claramente indicadas no quadro de sua formação (parcialmente) *on-line*”<sup>25</sup> (Nissen *apud* FOUCHER, 2010, p. 85). Portanto, ambos estão relacionados ao percurso possível no ambiente *on-line* para a aprendizagem, mas a autora se detém mais na definição do cenário pedagógico, que para este trabalho também é profícuo.

Para o cenário pedagógico, Foucher, citando Mangenot (2006 *apud* FOUCHER, 2010), apresenta principalmente duas concepções possíveis. A primeira, de cenário pedagógico como planejamento. Nessa concepção, o cenário é aquele que estrutura as sequências de tarefas selecionadas para o período de ensino e aprendizagem. A segunda definição seria de cenário pedagógico como simulação do real em que há encenação do papel de aprendiz numa situação

---

<sup>23</sup> Aquele no qual as atividades planejadas ocorrem.

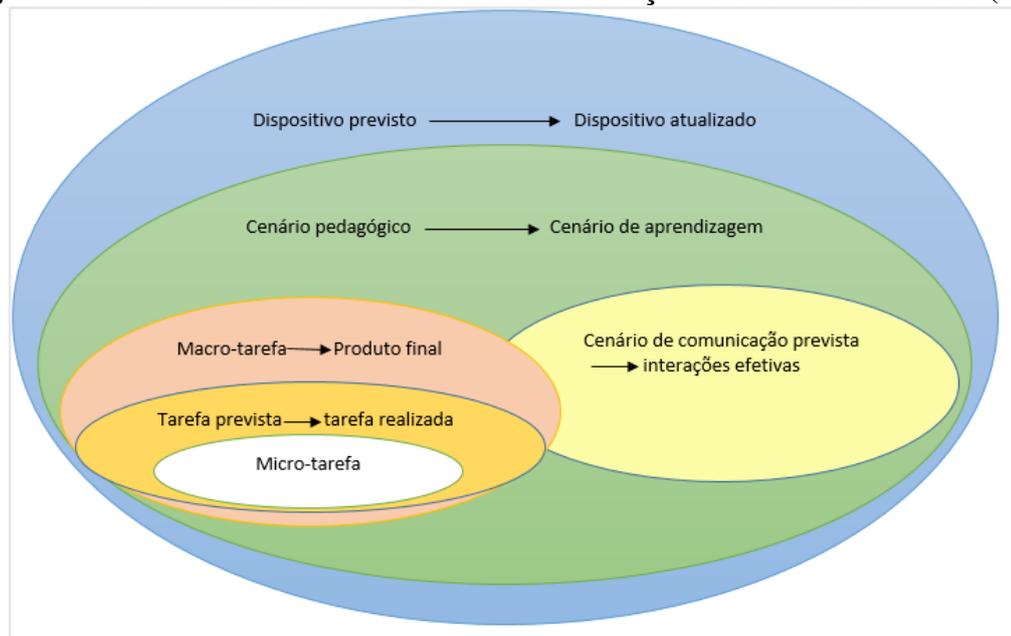
<sup>24</sup> Tradução de Deise Marinoto para “*être plus ou moins linéaire ou contraint selon les options didactiques prises*”

<sup>25</sup> Tradução de Deise Marinoto para “*l’ensemble des possibilités d’interaction qu’a l’apprenant à sa disposition et qui lui sont clairement indiquées dans le cadre de sa formation (partiellement) en ligne*”.

imaginada de aprendizagem ou de encenação de papel fictício em situação igualmente imaginada.

Logo, os três cenários (de navegação, pedagógico e de comunicação) discutidos por Foucher estão relacionados ao planejamento, organização da aprendizagem *on-line* e seus percursos, enquanto que o cenário de aprendizagem, segundo a autora, é o cenário constatado, aquele que vai se tecendo no decorrer do desenvolvimento do cenário pedagógico. Na figura 3, a autora apresenta uma versão dinâmica dos níveis de intervenção didática em cada cenário, as setas representando a relação temporal entre o previsto e a realidade constatada.

**Figura 3 – Versão dinâmica dos níveis de intervenção didática em Foucher (2010)**



**Fonte: Foucher (2010), tradução de Deise Marinoto**

Na figura, Foucher desenha um quadro da relação entre os cenários e a realização de tarefas, no qual o cenário pedagógico é a situação composta por uma ou mais tarefas (macrotarefa, tarefa e microtarefa) que foram planejadas para que ocorra a aprendizagem. A opção pelo uso do termo tarefas, em detrimento de exercício ou atividades está embasada na discussão feita por Foucher da noção de exercício e atividade, esta última composta por tarefas.

Para Foucher, no contexto de ensino aprendizagem de uma L2 (segunda língua), o exercício, que está relacionado a um componente ou dificuldade da língua-alvo, é selecionado pelo professor, ‘respondido’ pelo estudante e é objeto de correção. Por outro lado, segundo a autora, o termo atividade pode ser usado para designar “atividades de aprendizagem”, “atividades de linguagem” e “atividades pedagógicas”, podendo causar confusão pelos vários níveis de apreensão do termo.

A autora pondera principalmente sobre a atividade de aprendizagem, haja vista seu interesse pedagógico, como uma situação composta por tarefas, que prefere designar como “cenário pedagógico”. Para Foucher (2010), a atividade de linguagem extrapola a aprendizagem, sendo contingente, ocorrendo em situações do cotidiano que não necessariamente envolvem o aprender. Por fim, a pesquisadora considera que tanto tarefa quanto atividade estão claramente relacionadas, citando o conceito adotado por Ellis de que “‘Tarefas’ são atividades que exigem o uso da linguagem focada no significado. Em contraste, ‘exercícios’ são atividades que denominamos para o uso da linguagem focada na forma<sup>26</sup>” (ELLIS, 2003, p. 3).

Outras definições discutidas em Foucher (2010) são as de macrotarefas, tarefas e microtarefas. Segundo a autora, elas constituem o cenário pedagógico assim como o de aprendizagem. Como é possível observar na figura 3 e ainda como explica a autora, (FOUCHER, 2010, p. 80-81) “a macrotarefa engloba as tarefas que se constituem das microtarefas”, similarmente ao que afirma Guichon (2006) ao distinguir micro e macro tarefas

[...] É por isso que nos propomos a distinguir as "microtarefas" que consistem em trabalhar certos aspectos específicos da língua e a "macrotarefa" que corresponde a um projeto global de aprendizagem no qual os alunos têm de lidar com informação escrita ou oral em L2 para construir um objeto de significado escrito ou oral.(GUICHON, 2006, p. 54)<sup>27</sup>

As afirmações desses autores indicam que as tarefas e suas microtarefas possibilitam a realização da macrotarefa, contexto no qual ocorre a aprendizagem, nas palavras de Guichon – o projeto global de aprendizagem.

No intuito de procurar esclarecer essa relação, a autora apresenta o esquema de Grosbois (2009 *apud* FOUCHER, 2010 p. 81) apresentado aqui na Figura 4.

---

<sup>26</sup> Minha tradução para “‘Tasks are activities that call for primarily meaning-focused language use. In contrast, ‘exercises’ are activities that call for primarily form-focused language use.” (ELLIS, 2003, p.3)

<sup>27</sup> Tradução de Ana Carolina Freschi para: nous proposons de distinguer les "micro-tâches" qui consistent à faire travailler certains aspects précis de la langue et la "macro-tâche" qui correspond à un projet d'apprentissage global au cours duquel les apprenants sont amenés à traiter de l'information écrite ou orale en L2 pour construire un objet de sens écrit ou oral.

**Figura 4 – Composição da Macrotarefa por Grosbois (2009)**



**Fonte: Foucher (2010) Traduzido por Ana Carolina Freschi<sup>28</sup>**

A figura acima apresenta a relação de constituição entre os conceitos. Como é possível ver na figura, ao “criar um suporte de mídia evocando um fato cultural”, a macrotarefa se ampara em quatro tarefas que se realizam através de três microtarefas, pois “microtarefas e tarefas são claramente articuladas, as primeiras permitindo alcançar os objetivos mais amplos das segundas que se alimentam da macrotarefa<sup>29</sup>” (FOUCHER, 2010, p.81).

A relação entre macrotarefas, tarefas e microtarefas discutida por Foucher (2010) em contexto de aprendizagem telecolaborativo auxilia a compreensão do que ocorre quando da produção de dados como os que utilizamos para a construção do *corpus* de pesquisa que apresentamos nesta tese. Essa relação, bem como os conceitos de cenário de aprendizagem, são minha opção para descrever o contexto telecolaborativo de aprendizagem de línguas que é o teletandem. Opto por divergir de Aranha e Leone (2016; 2017) principalmente por entender que o cenário de aprendizagem é conceito que descreve melhor cada uma das turmas de TTD e não o TTD como um todo, pois, considero esse cenário, em consonância com Foucher (2010) como aquele que é constatado no decorrer do desenvolvimento do que foi planejado no cenário pedagógico. Desse modo, cada turma TTD constitui um cenário de aprendizagem com suas

<sup>28</sup> Ana Carolina Freschi é uma das pesquisadoras do grupo de pesquisa em teletandem da Unesp – Rio Preto, tendo concluído seu mestrado em 2017. Em seu trabalho, Freschi analisou as categorias de *feedback* corretivo recorrentes nas sessões orais de teletandem.

<sup>29</sup> Tradução de Ana Carolina Freschi para “[...] la tâche, micro-tâches et tâches sont clairement articulées, les premières permettant d’atteindre les objectifs plus larges des secondes qui elles-mêmes nourrissent la macro-tâche.”

especificidades e nuances diferenciais. Nesse contexto, as microtarefas focam os objetivos linguísticos que são podem ser definidos pelos professores-mediadores e pelos participantes. Discorro melhor a seguir.

#### 1.2.1.1 RELAÇÃO ENTRE MACROTAREFA, TAREFAS E CENÁRIO DE APRENDIZAGEM NO TTD

Considerando que a macrotarefa se alimenta das tarefas e das microtarefas como discuti anteriormente, a partir de Foucher (2010), entendo que, no TTD a macrotarefa seja o teletandem, como o espaço de interação no qual os aprendizes se comunicam na língua-alvo um do outro e tomam decisões sobre sua aprendizagem. Essa macrotarefa é alimentada por tarefas de duas naturezas: mediação (auxílio para a aprendizagem) e interação (*locus* de aprendizagem telecolaborativa), dessa compreensão decorreu a composição da proposta de cenário de aprendizagem, sobre a qual discorro posteriormente neste trabalho, bem como orientou a organização dos dados em pastas e subpastas no HD externo que contém o MulTeC.

Dada a complexidade da relação entre essas tarefas realizadas para mediação no TTD, parece-me auxiliar na compreensão, a sua divisão a partir dos objetivos para os quais cada tarefa foi inserida no contexto TTD, como podemos ver na figura 5.

**Figura 5 – Tarefas realizadas para mediação por objetivos**



**Fonte: A autora**

A sessão de mediação (assim como as aulas que podem ser parte do TTD quando ocorre na modalidade integrada) pode servir como espaço de reflexão, mas também de

orientação, motivo pelo qual, na figura acima, está na intersecção entre os dois objetivos fins da mediação, representando que durante ela pode haver tanto orientação quanto reflexão, como nos apontam os pressupostos teóricos que discuti anteriormente (ANDREU-FUNO, 2015; SALOMÃO, 2008; SOUZA, 2012). Nessas tarefas, estão envolvidos mediadoras e participantes, havendo inevitavelmente uma relação hierárquica que pode intervir nas relações entre eles. Gostaríamos de destacar que essa divisão por objetivos é apenas um auxílio para a compreensão, mas em qualquer dessas tarefas da natureza de mediação, reflexão e orientação podem ocorrer ainda no caso dos diários de aprendizagem quando produzidos no Google Docs®, nos quais a mediadora pode inserir comentários tanto problematizando, suscitando reflexões, quanto fornecendo orientações no que se refere ao que o participante discorreu.

As tarefas que considero como diretamente relacionadas à aprendizagem telecolaborativa têm como envolvidos os participantes, dispensando a participação da mediadora e caracterizando uma relação entre pares, sem que haja, a princípio, uma relação hierárquica entre eles. Entre essas tarefas listo no TTD: a Sessão Oral de Teletandem (SOT) e a troca de textos cujos objetivos são: interagir com o parceiro para aprender a língua/cultura alvo, assim como auxiliar na aprendizagem do outro. Vale ressaltar que durante a SOT os parceiros podem lançar mão do chat como ferramenta auxiliar na interação síncrona, logo, constitui-se um dado a ser coletado, cujo acesso pelo pesquisador tornará mais ampla a compreensão de como ocorreu a aprendizagem no contexto. Estudos sobre chat no contexto TTD estão sendo iniciados (FERRO, em andamento), mas considerando que o chat ocorre durante a SOT, não como uma tarefa, haja vista que está disponível como ferramenta auxiliar para os participantes, mas o chat é um dado produzido na SOT e, portanto, importante para a compreensão da SOT caso tenha sido utilizado pelos parceiros, o que não ocorreu em todas as SOTs coletadas.

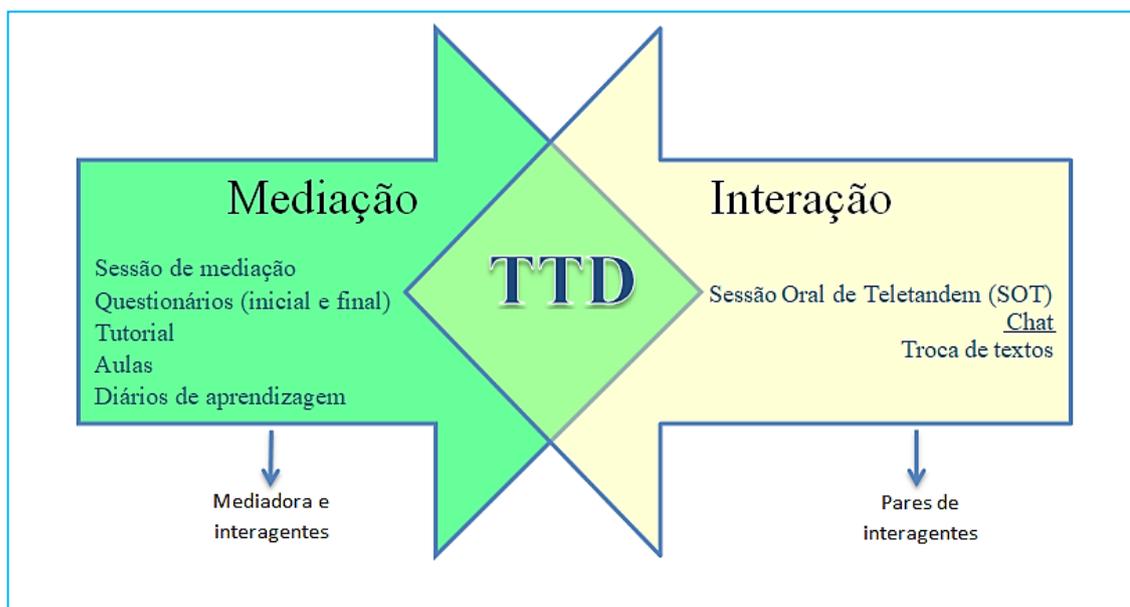
Desse modo, neste trabalho, a relação entre macrotarefa e tarefa no TTD fica estabelecida da seguinte maneira: i) o TTD é a macrotarefa, o ambiente no qual os participantes buscam desenvolver suas habilidades linguísticas e culturais; ii) as tarefas que se alimentam dessa macrotarefa podem ser caracterizadas considerando suas naturezas, seja de promover reflexão e/ou de orientar a aprendizagem no TTD e de interagir para aprendizagem de línguas estrangeiras; iii) as microtarefas no TTD não podem ser descritas ou listadas por estarem circunscritas no âmbito da autonomia do aprendiz, um dos princípios basilares do TTD, de modo que ele possa optar entre dedicar-se mais a aprimorar sua pronúncia, ampliar vocabulário em alguma área específica ou aprender mais sobre a cultura do país da língua-alvo, etc. Nas

microtarefas ainda incluo aqueles objetivos estabelecidos quando o mediador opta por uma tarefa em detrimento de outra.

Para Guichon (2006, p. 54) a distinção entre microtarefas e macrotarefa é que as “microtarefas consistem em fazer certos aspectos específicos do trabalho de linguagem e a ‘macrotarefa’ que corresponde a um projeto de aprendizagem global no qual os alunos têm que lidar com informação escrita ou oral em L2 para construir um objeto de significado escrito ou oral.” O autor ainda afirma que microtarefas estão voltadas para a forma (objetivos linguísticos) enquanto que a macrotarefa está voltada para a construção do significado (interação). A partir desses pressuposto, justifico que macrotarefa seja o contexto TTD no qual os parceiros trabalham telecolaborativamente para a construção do significado, realizando tarefas – de natureza da mediação ou da interação - podendo, no âmbito de sua autonomia, um dos princípios do TTD, decidir que questões formais (ou não) da língua em estudo deseja aprender, ou seja, definindo quais microtarefas irá desenvolver.

Embora haja compreensão de que seja viável a realização de TTD sem mediação, defendo que o contexto TTD se efetiva de modo mais significativo quando da junção das tarefas das duas naturezas: mediação e interação como procuramos representar na figura 6.

**Figura 6 – Relação das tarefas na constituição do TTD**



Fonte: Elaborado pela a autora para o MulTeC

Cabe destacar que essas são tarefas possíveis de ocorrer no contexto de aprendizagem, ao qual esta pesquisa se dedica, e ocorrência delas dependerá em primeira instância da

modalidade de TTD e em seguida do que fora planejado para ocorrer em cada turma. Uma turma de TTDii poderá<sup>30</sup> apresentar todas as tarefas listadas na figura acima desde que as mediadoras tenham planejado assim para a turma, enquanto que em uma turma não-integrada poderá não haver troca de textos, por exemplo. O chat está sublinhado na figura por ser um dado coletado complementar a SOT, mas que apresenta indícios de não se constituir uma tarefa, havendo necessidade de estudos aprofundados para podermos chegar à identificação das características do chat no TTD.

Essas tarefas são organizadas pelas professoras coordenadoras do TTD nas instituições que estabelecem a parceria, tendo uma sequência que pode ser dividida em três fases: 1) a fase do planejamento, que antecede a primeira SOT<sup>31</sup>; 2) a fase de interação entre os pares; e 3) a fase de conclusão do período com aplicação do questionário final.

Se faz necessário destacar que há tarefas que ocorrem apenas entre as mediadoras (professoras-coordenadoras), cuja coleta de dados não tem sido realizada com frequência até então, tendo ocorrido uma coleta por ocasião da pesquisa de mestrado de Araújo (2012) que objetivou sob a ótica da Teoria da Atividade pesquisar a formação de parcerias de TTD. Araújo (2012) utilizou como um dos dados de pesquisa os e-mails trocados entre os professores-coordenadores das duas instituições que estabeleciam parceria. Essa troca de e-mails costuma ocorrer em maior quantidade no período do planejamento das SOTs, com o intuito de decidir (normalmente a partir do calendário das instituições) quantas SOTs poderão ocorrer no semestre, se haverá ou não troca de textos ou ainda se os textos serão produzidos colaborativamente, por exemplo.

Para além desses conceitos que descrevem o contexto TTD no qual os dados foram produzidos, é necessário apresentar o conceito de *corpus* de pesquisa e quais procedimentos podem ser considerados adequados na busca de garantir sua construção de modo eficiente e evitando desperdícios tanto de investimentos quanto de tempo.

### 1.3 A LINGUÍSTICA DE *CORPUS*

Começo esta seção, que discute brevemente a Linguística de *Corpus*, com uma afirmação de Crystal: “Eu não acho que eu poderia tê-lo escrito há 5 anos em razão da falta de estudos acadêmicos para fornecer algum conteúdo e da dificuldade geral em obter grande

---

<sup>30</sup> Vale ressaltar que o uso do chat é atualmente complementar a SOT, podendo não ser utilizado em todas elas.

<sup>31</sup> Desde o trabalho de Rampazzo (2017) está nomeada como SOTi – Sessão Oral de Teletandem inicial.

amostra de dados” (CRYSTAL, 2001, p. viii)<sup>32</sup>. Situada no prefácio de sua obra “Língua e a Internet”, escrita há mais de uma década, essa afirmação de Crystal reforça e aponta para a relevância da coleta de dados e a gama de possibilidades que a internet trouxe aos seres humanos, o que conduz à importância da Linguística de *Corpus*.

Já se tornou truismo afirmar que a internet e a ciberinfraestrutura que a viabiliza oportunizam coleta de grande quantidade de dados para pesquisa. No entanto, a sistematização desses dados em *corpora* que deem conta de prover os pesquisadores com as informações e dados necessários para suas pesquisas, configura-se como bastante relevante. Da mesma forma ocorre com a Comunicação Mediada por Computador (CMC). O que significa afirmar que, a partir da popularização dos computadores e do uso da internet, a possibilidade de organização de grande quantidade de dados é uma realidade que, amparada no trabalho da Linguística de *Corpus*, viabiliza a construção de *corpora* a serem utilizados por pesquisadores das mais diversas áreas.

Com base nisso, apresento (não exaustivamente, dado que não é esse o propósito deste trabalho), a Linguística de *Corpus*, destacando sua importância nesse processo de construção de *corpus* bem como o gerenciamento de dados como modo de tornar factível o uso ou re-uso dos dados coletados por pesquisadores, mesmo que não tenham participado do processo da coleta.

### 1.3.1 COMPREENSÃO DA LINGUÍSTICA DE *CORPUS* E TIPOLOGIA DE *CORPUS*

Desde o lançamento, em 1964, do *corpus Brown*, com 1 milhão de palavras, até o presente é considerado um marco para a Linguística de *Corpus* (LC) até o presente. Vale ressaltar que atualmente há corpora constituídos de “5 bilhões de palavras como o *News on the Web* (NOW)” (TAGNIN, 2018, p. 11). Na circunscrição da pesquisa acadêmica, a constituição e o uso de corpora, a partir dos fundamentos da LC, têm contribuído em pesquisas de diversas áreas, não apenas as que estão voltadas para a investigação em Linguística e Linguística Aplicada, mas em áreas como Ciência da Informação e Biblioteconomia ou Biologia. Sobre isso são lúcidas as palavras de Tagnin a seguir:

Mesmo vislumbrando áreas não linguísticas em que a Linguística de *Corpus* possa fazer uma contribuição significativa, qualquer área em que a língua tenha um papel relevante – e qual é a área que pode dispensar a língua? – só tem a ganhar em termos

---

<sup>32</sup> “I do not think I could have written it five years ago because of the lack of scholarly studies to provide some substance, and the general difficulty of obtaining large sample of data” (Tradução minha)

de objetividade e confiabilidade com o uso da Linguística de *Corpus*, principalmente porque a tecnologia avança a passos largos nesse setor, permitindo a criação de ferramentas que permitem análises cada vez mais específicas e direcionadas. (TAGNIN, 2018, p.14)

Desse modo, a dedicação de tempo para o gerenciamento de dados a constituir um *corpus*, representa um ganho para a ciência. O compilador, no entanto, precisa atentar para diversos detalhes no gerenciamento dos dados de maneira que o *corpus* que almeja construir seja utilizável, compartilhável.

Os corpora podem ter tipologia e extensão diversas. Segundo Berber Sardinha (2004, p. 20-21), a tipologia depende de fatores como: o modo (falado ou escrito); o tempo (sincrônico, diacrônico, contemporâneo ou histórico); a seleção (de amostragem, monitor, dinâmico, estático, equilibrado); o conteúdo (especializado, regional/dialetal, multilíngue); a autoria (de aprendiz, de língua nativa); a disposição interna (paralelo e alinhado) e a finalidade (de estudo, de referência, de treinamento ou teste). Com o desenvolvimento de ferramentas de comunicação, tanto síncronas quanto assíncronas, multimodais, a produção de dados dessa natureza tem se ampliado, o que conseqüentemente contribui para que mais pesquisadores se dediquem à construção de corpora multimodais, embora ainda estejamos presenciando o amanhecer dessa prática.

No tocante à extensão de um *corpus*, Berber Sardinha (2004) refere-se a três dimensões que podem ser consideradas: a quantidade de palavras, de textos e de gêneros. Nas três dimensões, o autor afirma que quanto maior a quantidade, maior a chance de o pesquisador encontrar registros que sejam significativos para sua pesquisa.

A contagem das palavras deve ser realizada por softwares de processamento linguístico ou por ferramentas de contagem de palavras disponibilizadas por softwares de processamento de textos como o *Wordsmith Tools*® e a ferramenta *Wordlist*.

A maioria dos corpora que podemos encontrar atualmente é monomodal, tanto que na classificação de Berber Sardinha (2004) há apenas essas duas possibilidades no critério modo, para tipologia dos corpora: falado ou escrito. Porém, o desenvolvimento das tecnologias síncronas de comunicação via vídeo tem possibilitado o acréscimo de um terceiro tipo: o multimodal – um *corpus* que contenha tanto o modo falado quanto o escrito.

### 1.3.2 CORPUS MULTIMODAL: A NOVA GERAÇÃO DOS CORPORA

Segundo Knight (2011, p. 392), “a construção e o uso de corpora multimodais está ainda em sua relativa infância, com a maioria das pesquisas associadas a essa área tendo seu

início há apenas uma década”<sup>33</sup>. Embora Knight se refira a corpora apenas com vídeos<sup>34</sup>, o *corpus* objeto desta tese é considerado multimodal em decorrência do fato de que ele é constituído não apenas por textos escritos, mas por vídeos que estão em certa medida relacionados com os textos, e ainda que em alguns textos há presença de imagens<sup>35</sup>, utilizadas por alguns dos participantes do TTD em suas produções textuais. Nesta subseção, discuto alguns trabalhos que apresentam corpora multimodais, a saber: Guichon (2017), Wigham e Chanier (2013), Reffay, Betbeder e Chanier (2013) e especialmente Chanier e Wigham (2016) por apresentarem uma metodologia científica para pesquisar dados de interação mediada por computador, no qual discutem pontos relevantes na construção de um *corpus* de natureza similar ao que nos propomos nesta pesquisa.

Guichon (2017) ao discutir o compartilhamento de um *corpus* multimodal para o estudo de ensino de língua mediado por *webcam*, apresenta os princípios norteadores para a criação do ISMAEL, *corpus* constituído por dados produzidos em um projeto telecolaborativo entre professores em formação e aprendizes de francês que teve duração de seis semanas. Para o autor, os princípios a serem considerados para a construção do *corpus* ISMAEL objetivam atender aos questionamentos de pesquisa do grupo interessado na constituição desse banco de dados. No caso do ISMAEL os princípios que nortearam a coleta e a sistematização do *corpus* foram: i) a qualidade multimodal, ii) a integridade contextual, iii) a qualidade no desenvolvimento – coletando todo o período de interação – e iv) a diversidade dos participantes, incluindo tantos professores em formação quanto fosse possível.

Para Guichon (2017)

Coletar dados, estruturá-los em um *corpus*, prover informação contextual e facilitar sua acessibilidade para pesquisa são as diferentes etapas envolvidas na criação de um *corpus* compartilhável. Todas essas etapas necessitam de tempo e esforço, procedimentos estratégicos de tomada de decisão e uma constante atenção às questões éticas e às necessidades dos pesquisadores[...]”<sup>36</sup> (GUICHON, 2017, p. 66)

A assertiva de Guichon aponta para a importância de um bom planejamento em cada etapa da construção do *corpus*, demandando esforço e um tempo considerável, além da

---

<sup>33</sup> Minha tradução para “the construction and use of multimodal corpora is still in its relative infancy, with the majority of research associated with this field spanning back only a decade”. (KNIGHT, 2011, p. 392)

<sup>34</sup> Knight utilizou como fonte de análise 18 corpora constituídos de gravações em áudio/vídeo.

<sup>35</sup> Embora não seja uma solicitação para a produção textual, encontramos textos com inserção de imagens pelos participantes.

<sup>36</sup> Minha tradução para “Collecting data, structuring them into a corpus, providing contextual information, and facilitating their accessibility for research are thus the different steps involved in the creation of a shareable corpus. All these steps necessitate time and effort, efficient and strategic decision-making procedures, and a constant attention to ethical issues and to the researchers’ needs[...]” (GUICHON, 2017, p. 66)

necessidade de não se descuidar da documentação necessária bem como do estabelecimento de procedimentos de anonimização que objetivam considerar questões éticas, sem, no entanto, deixar de atender às necessidades do pesquisador que opta por construir um *corpus* de pesquisa que seja compartilhável/re-utilizável.

Reffay, Betbeder e Chanier (2013), tendo como ponto de partida o objeto científico criado no Projeto Mulce - o LETEC (Learning & TEaching *Corpus*), discutem pontos a serem considerados quando da estruturação de um *corpus* multimodal como o organizado pela equipe do projeto. Os autores afirmam que a “simples coletânea de dados de interação *on-line* de estudantes não representa um objeto científico” (REFFAY, BETBEDER, CHANIER, 2013, p. 3). É indispensável, segundo os autores, estar preocupado com aspectos humanos, éticos e técnicos que viabilizem ou que dificultem a estruturação coerente dos dados em um *corpus*. Os autores enfatizam que a estruturação e coerência do *corpus* está relacionada com a inclusão de informações contextuais no que se refere ao *locus* de produção dos dados, a saber: “1) o contexto do cenário educacional; 2) as perguntas originais de pesquisa; 3) o contexto educacional e técnico; e 4) as interações que ocorreram.”<sup>37</sup> (REFFAY, BETBEDER, CHANIER, 2013, p. 4). De acordo com os autores, o trabalho de estruturação desses dados demanda moderação e planejamento cauteloso antes mesmo do início da coleta, inclusive organizando a coleta das assinaturas dos participantes nos Termos de Consentimento Livre e Esclarecido (TCLEs).

Esses procedimentos devem ser seguidos no intuito de garantir que os dados possam ser reutilizáveis, seguindo uma tendência da ciência definida em 2003, em Berlin, como “contribuição de acesso aberto”<sup>38</sup> que convida a comunidade científica a compartilhar os dados produzidos em suas pesquisas. Apesar dos desafios impostos ao compartilhamento e reutilização de dados de pesquisa, para os autores, os benefícios para a ciência motivam a dedicação de esforços na organização de dados compartilháveis, que para o serem precisam de um tratamento adequado.

No tocante ao tratamento de dados produzidos em ambiente *on-line*, Chanier e Wigham (2016) propõem uma metodologia “usada para coletar, transformar e organizar dados de situações de aprendizagem *on-line*, a fim de torná-los compartilháveis através de repositórios de acesso aberto”<sup>39</sup> (CHANIER, WIGHAM, 2016, p. 215).

---

<sup>37</sup> Minha tradução para “(1) the context of the educational scenario, (2) the original research questions. (3) the educational and technical context and (4) the interactions which occurred.

<sup>38</sup> Minha tradução para “open-access contribution”.

<sup>39</sup> Minha tradução para “[...] used to collect, transform and organize data from *on-line* learning situations in order to make them sharable through open-access repositories”

Após uma breve discussão sobre *corpus* na ciência, os autores discutem a relação corpora e ensino de línguas, que teve seu início nos anos 90, exemplificando o uso de corpora para a elaboração de livros didáticos com estruturas linguísticas atualmente usadas. Chanier e Wigham (2016) afirmam que os corpora anteriores capturaram linguagem usada em situações formais ou informais, mas apenas por falantes nativos. No entanto, Sylviane Granger e sua equipe, na Bélgica, lançaram os corpora de aprendiz (Learner Corpora), com produções de aprendizes de inglês como segunda língua. Os pesquisadores desse grupo “desenvolveram uma estrutura para pesquisa de *corpus* de aprendiz onde dados foram coletados, estruturados e a partir de 2009, anotados da mesma forma<sup>40</sup>.” (GRANGER, 2004 *apud* CHANIER, WIGHAM, 2016, p. 219).

Após essa introdução, os autores apresentam um paradigma de *corpus*, que segundo eles, pode seguir os seguintes passos:

- Coleta sistemática de dados – nesse item, os autores afirmam que mesmo havendo um questionamento que direciona a coleta, o conjunto dos dados produzidos no contexto (interações, produções, arquivos log) devem ser coletados, como pré-requisito para que outros reusem o *corpus*.
- Descrição detalhada dos dados (metadados) – o pesquisador precisa nomear o *corpus*, dar uma lista de créditos, assim como explicar como os dados foram coletados, editados (anonimizados) e organizados, além de apresentar informações sociolinguísticas<sup>41</sup> detalhadas sobre os participantes.
- Conversão dos dados - os dados precisam estar em formatos de livre acesso<sup>42</sup>.
- Divulgação e distribuição dos dados – Precisa ser de acesso livre e garantido como permanente.

Assim, durante a construção de um *corpus*, os envolvidos no processo precisam atentar a cada um desses itens, destacando-se a relevância do detalhamento contextual. Para os autores, um documento que delineie uma situação de aprendizagem *on-line* demanda descrições detalhadas que contemplem a construção do cenário de aprendizagem para o pesquisador que não participou das atividades do projeto. Logo, no *corpus*, o pesquisador precisa encontrar documento com descrição das ferramentas utilizadas no desenvolvimento das atividades *on-*

---

<sup>40</sup> Minha tradução para “They developed a framework for learner corpus research where data were collected, structured and, from 2009 onwards, annotated in the same way.”

<sup>41</sup> Sexo, idade, curso de graduação, línguas e nacionalidade, por exemplo.

<sup>42</sup> Formato independente, que pode ser acessado em qualquer sistema operacional de computadores.

*line* bem como sua sequência de ocorrência, os papéis desenvolvidos por cada participante e as orientações fornecidas a eles.

No que se refere à organização dos dados para constituírem um *corpus*, foco desta tese, Chanier e Wigham (2016) apresentam “um modo de transformar dados não tratados em dados de pesquisa<sup>43</sup>” (CHANIER e WIGHAM, 2016, p. 229) que podem ser consideradas fases na construção do *corpus*. As fases apresentadas pelos autores foram elaboradas a partir do trabalho realizado por sua equipe na construção do LETEC.

A primeira fase consiste na criação de códigos únicos de identificação para os dados, de modo que cada dado possua identificação ímpar. Na segunda fase, deve ocorrer a anonimização dos dados “[...] através de um processo sistemático<sup>44</sup>” (CHANIER e WIGHAM, 2016, p. 230). Os autores sugerem que os códigos utilizados na anonimização sejam o mais significativo possível, assim como afirmam a necessidade de criação de tabelas com os códigos dos participantes, suas informações sociolinguísticas e o que nomeiam como ‘biografia da língua’ que seriam as informações sobre a “[...] língua estrangeira falada, nível de proficiência, quantidade de anos estudando a língua e contexto de estudo<sup>45</sup>” (CHANIER e WIGHAM, 2016, p. 230). Na anonimização, as substituições de dados sensíveis, orientam os autores, deve ser feita com informação significativa, como por exemplo, “[...] o número de telefone de um participante em uma mensagem de texto no chat poderia ser substituído por um código e rotulado para destacar que a informação original correspondia a um número de telefone<sup>46</sup>” (CHANIER e WIGHAM, 2016, p. 230).

A terceira fase estaria voltada à preservação dos dados a longo prazo, o que implica na conversão dos dados em formatos atualizados e livres de modo que o acesso por pesquisadores seja garantido ao longo do tempo.

Os autores citados nessa subseção indicam que há diversos procedimentos a serem cautelosamente observados para a construção de um *corpus* bem organizado que é fruto de um gerenciamento de dados estabelecido desde o início das atividades de pesquisa, conforme discutem Briney (2015), Eynden (2013) e Eynden *et al.* (2011).

---

<sup>43</sup> Minha tradução para “[...] one way to transform raw data into research data”.

<sup>44</sup> Minha tradução para “[...] through a systematic process”.

<sup>45</sup> Minha tradução para “[...] foreign language spoken, language level, number of years studying the language and context of study”.

<sup>46</sup> Minha tradução para “[...] a participant’s phone number in a text chat message could be replaced with a code and labelled to highlight that the original information corresponded to a phone number.”

#### 1.4 O GERENCIAMENTO DE DADOS PARA CONSTRUÇÃO DE *CORPUS* DE PESQUISA

O gerenciamento dos dados a constituírem um *corpus* demanda um planejamento que precede e orienta a coleta e que pode sofrer algumas alterações, servindo como instrumento norteador das ações em cada fase do projeto de pesquisa até a sua conclusão. Em se tratando da coleta de dados para constituir um *corpus*, o plano elaborado precisa ser consistentemente seguido nas fases de formatação, documentação, armazenamento e busca de consentimento para uso dos dados para pesquisas conforme instruem Briney (2015) e Eynden et al. (2011), por exemplo.

O planejamento para a coleta deve considerar o tipo de *corpus* a ser constituído e, por conseguinte, os tipos de textos, bem como a coleta de Termos de Consentimento Livre e Esclarecido (TCLEs). A assinatura de TCLEs deve preceder a coleta de modo que não sejam coletados dados que não poderão ser utilizados, isso levando-se em conta a capacidade finita de armazenamento das máquinas e o tempo que o compilador precisará destinar para a formatação dos dados, bem como a questão ética, pois se não há autorização, os dados não devem ser coletados. Em seu trabalho sobre gerenciamento de dados, Briney afirma que "um bom gerenciamento é a compilação de uma série de pequenas práticas de rotina que se tornam bons hábitos"<sup>47</sup> (BRINEY, 2015, p. 24) desde o planejamento da coleta até a análise dos dados, completando-se um ciclo de vida dos dados.

O ciclo de vida dos dados sofreu alteração nos últimos anos, segundo Briney (2015), ampliando suas fases e subfases a partir do avanço das tecnologias digitais e de sua capacidade de armazenamento e compartilhamento. Nesse 'novo' ciclo, além de serem instrumentos para a publicação como resultado de uma pesquisa, os dados se constituem ainda como um produto da pesquisa. Assim, com "[...] a prevalência dos dados digitais na pesquisa significa que nós podemos fazer mais com os dados da pesquisa além de perdê-los no final de um projeto"<sup>48</sup>(BRINEY, 2015, p. 20).

No ciclo anterior de vida dos dados, segundo Briney (2015) havia o planejamento do projeto de pesquisa, a coleta de dados, seguida da análise que culminava com a publicação dos resultados da análise. Os dados, nesse ciclo, são descartados após o encerramento do projeto, o que representa um desperdício de recursos tanto financeiros quanto temporais, considerando os

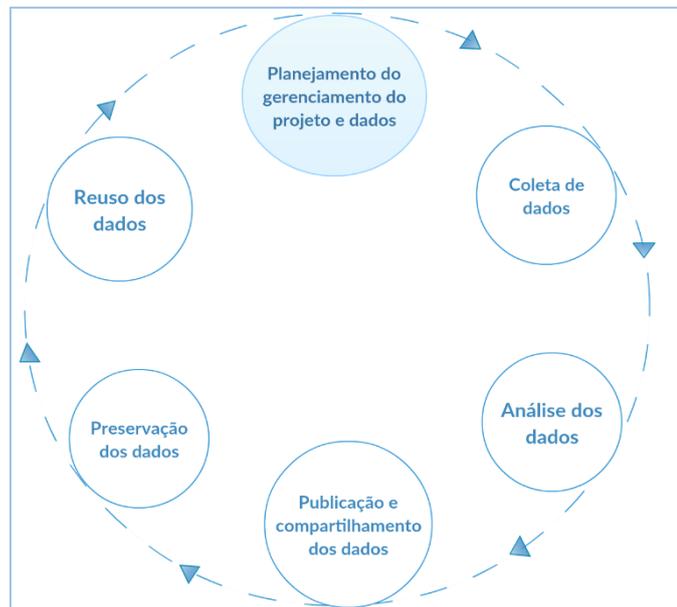
---

<sup>47</sup> Minha tradução para "Good data management is the compilation of a number of small routine practices that add up to good habits."

<sup>48</sup> Minha tradução para "[...] the prevalence of digital data in research means that we can do more with research data beyond losing it at the end of a project".

investimentos de tempo e dinheiro designados para a organização de um *corpus* para um projeto de pesquisa de qualquer que seja a área de investigação. Diante disso, o atual ciclo de vida dos dados, insere o planejamento do gerenciamento dos dados bem como o acesso aos dados por outros como podemos ver na figura 7.

**Figura 7 – O novo ciclo de vida dos dados**



**Fonte: Briney 2015, p. 21.**

Nesse ciclo, os dados - que foram anonimizados, salvos em formato de livre acesso e disponibilizados - poderão ser reutilizados por outros pesquisadores, tendo uma vida útil maior e ampliando as possibilidades de pesquisa com economia de recurso e tempo a ser investido em nova coleta, caso os dados não fossem preparados para uso posterior. Essa possibilidade se torna viável pelo cauteloso planejamento do gerenciamento dos dados que objetiva sua preservação para além do projeto de pesquisa que os gerou, constituindo-se um produto da investigação.

Cada fase desse novo ciclo demanda do pesquisador uma série de ações cuja realização procura garantir que os dados possam ser compartilháveis. Desse modo, no processo inicial de planejamento e gerenciamento do projeto de pesquisa e dos dados, o pesquisador deverá: planejar que tipos de dados serão coletados, em que local serão salvos, por quanto tempo serão mantidos, em que formato; criar os metadados e a documentação necessária para compreensão do contexto de produção dos dados. Esse planejamento inclui ainda como se dará a coleta, como serão organizados os arquivos, preparar e oportunizar a assinatura dos TCLEs (BRINEY, 2015;

EYNDEN, 2013; EYNDEN *et al.*, 2011). Criteriosa e consistentemente realizada, cada uma das etapas do ciclo poupará recursos, esforços e tempo do pesquisador.

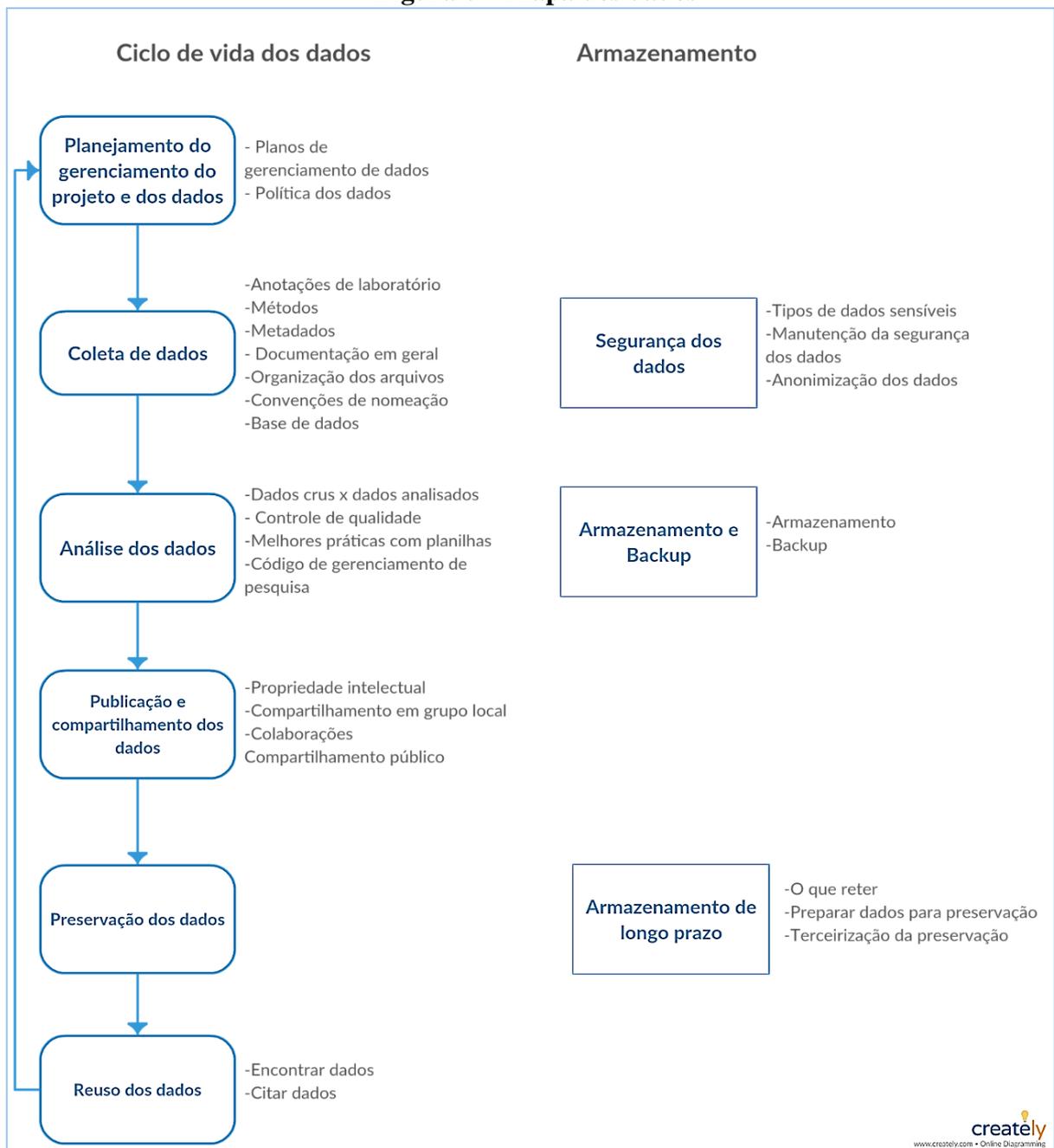
Na fase de aquisição dos dados, as ações estarão voltadas para a coleta, o que envolve a organização dos arquivos como planejado na etapa anterior, bem como a nomeação dos arquivos e documentação complementar. A nomeação convencionalizada precisa ser seguida de modo que se procure proporcionar a melhor compreensão possível do contexto no qual os dados foram produzidos.

Durante a fase de análise dos dados, como atividades a serem realizadas posso listar, embasada em Briney (2015), Eynden (2013) e Eynden *et al.* (2011): o controle de qualidade e a checagem de possíveis erros, bem como a análise dos dados a partir do embasamento teórico selecionado. A fase seguinte de publicação e compartilhamento dos dados envolverá a forma de compartilhamento (*on-line*, por exemplo) e as formas de acesso, podendo ser através de senha ou livremente pelo simples acesso ao repositório do *corpus*.

Encerrada a fase de publicação e compartilhamento, o pesquisador realiza verificação do modo de preservação dos dados. Para um período longo de preservação, os dados precisam estar em formatos que possam ser lidos não apenas por softwares mais antigos, assim como não deixar salvo em apenas um local (HD externo), para evitar possível perda de todos os dados caso ocorra dano irreparável no local de arquivamento. Seguindo-se todas as fases anteriores, os dados poderão ser reutilizados em pesquisas com mesmo propósito da anterior, ratificando ou retificando resultados, bem como com objetivo de investigação diverso do inicial inclusive pelo pesquisador que gerou o *corpus*.

Briney sintetizou em um fluxograma, que ela nomeou como mapa dos dados, os procedimentos necessários em cada etapa do ciclo de vida, como demonstro a seguir na figura, originalmente apresentada por Briney (2015).

**Figura 8 – Mapa dos dados**



Fonte: Traduzido de Briney (2015, p. 28)

O mapa apresentado por Briney (2015) resume sinoticamente os procedimentos sobre os quais discorri anteriormente e que precisam ser considerados quando nos propormos a construir um *corpus* de pesquisa compartilhável. Segundo Briney (2015), as estratégias adotadas em cada etapa do ciclo devem considerar inclusive as ferramentas disponíveis para a realização da pesquisa, assim como seu percurso, adequando as estratégias de acordo com as necessidades e possibilidades. A fase que concentra a maior quantidade de procedimentos é a fase de aquisição dos dados, durante a qual o pesquisador deve estar atento ainda às questões

relacionadas à segurança dos dados quanto ao arquivamento, como, por exemplo, o tipo de dados sensíveis coletados e anonimização dos dados.

Quanto à reutilização dos dados em pesquisas posteriores, Briney (2015) exemplifica com o excerto:

Em uma noite não há muito tempo atrás, um grupo de pesquisadores estava em um bar local, conversando sobre seu trabalho sob a lua cheia. A combinação da lua e da cerveja os levou a pensar sobre um estudo sobre o sono que eles conduziram dez anos antes e perguntaram-se se os dados poderiam dizer-lhes algo sobre o efeito dos ciclos lunares sobre o sono. Os pesquisadores voltaram para os dados antigos e, felizmente, encontraram bastante dados que apontaram para a correlação entre o sono e fase lunar (Zivkovic, 2013). Usando os dados reciclados e uma nova análise, os pesquisadores publicaram seus resultados no artigo “Evidências de que o ciclo lunar influencia o sono humano” (Cajochen et al. 2013).<sup>49</sup> (*apud* BRINEY, 2015, p. 126)

No caso, os pesquisadores reutilizaram dados que coletaram e organizaram para pesquisa realizada há uma década, mas, na reutilização, os mesmos dados serviram a propósito de investigação diverso. Essa possibilidade se tornou factível porque, quando da coleta dos dados para pesquisa anterior, os pesquisadores foram cautelosos para compilarem a maior quantidade possível de informações sobre os dados coletados. O plano de gerenciamento de dados elaborado na primeira fase do ciclo de vida dos dados se constitui ferramenta útil nessa empreitada.

#### 1.4.1 O PLANO DE GERENCIAMENTO DOS DADOS

A relevância de um documento como o plano de gerenciamento de dados reside no estabelecimento das metas e levantamento das necessidades para que o pesquisador possa concluir o trabalho de realização da pesquisa, como afirma Briney (2015) ao referir-se à importância de colaboradores no gerenciamento dos dados: “Em particular, o planejamento do gerenciamento de dados desempenha um papel ainda maior, pois a decisão antecipada das estratégias de gerenciamento de dados pode agilizar o uso de dados”<sup>50</sup> (BRINEY, 2015, p. 142).

---

<sup>49</sup> “On a night not too long ago, a research group was hanging out at their local bar, chatting about their work under the full moon. Something about the combination of the moon and the beer led them to think about a sleep study they conducted ten years previously and wonder if that data could tell them anything about the effect of lunar cycles on sleep. The researchers went back to their old data and, happily, found enough data points to correlate sleep and the lunar phase (Zivkovic 2013). Taking the recycled data and a new analysis, the researchers published their results in the paper “Evidence that the lunar cycle influences human sleep” (Cajochen et al. 2013).” (Tradução nossa)

<sup>50</sup> In particular, data management planning plays an even greater role because deciding on data management strategies ahead of time can streamline data use. (BRINEY, 2015, p. 142)

Segundo Eynden (2011) e *Eynden et al* (2013) o plano de gerenciamento de dados envolve planejar quem será responsável e como serão realizadas as tarefas de coleta, organização dos dados, arquivamento e definição dos investimentos necessários para a realização da pesquisa. Embora o plano de gerenciamento de dados deva ser flexível a uma adaptação a partir do que os pesquisadores vivenciarem durante o processo da pesquisa, ao elaborar o plano de gerenciamento de dados eles precisarão planejar questões referentes a todas as fases do ciclo de vida dos dados, obtendo respostas para perguntas como:

- Que dados serão coletados? Haverá necessidade de coletar quantos dados?
- Como os dados serão coletados?
- Que documentação e metadados acompanharão os dados?
- Como será gerenciada a questão ética?
- Como os dados serão arquivados? Como será feito o *backup* dos dados?
- Como serão gerenciados o acesso e a segurança dos dados?
- Que dados serão salvos a longo prazo? Como esses dados serão preservados?
- Como os dados serão compartilhados?
- Haverá restrições quanto ao compartilhamento?
- Quem será responsável pelo gerenciamento dos dados?
- Que recursos serão necessários para o desenvolvimento do plano?

Essas questões contemplam cada fase do ciclo de vida dos dados: a coleta, a análise dos dados, publicação e compartilhamento, preservação dos dados e reuso dos dados. Assim, antes mesmo do início da coleta, os pesquisadores estabelecerão um percurso a seguir, otimizando seu tempo e seus recursos.

Caso os pesquisadores estejam atentos ao ciclo de vida dos dados e seguindo o plano de gerenciamento de pesquisa, a reutilização dos dados pode ocorrer com pesquisadores que não utilizaram os dados anteriormente, mas que podem encontrar nos dados reciclados respostas às suas inquietações investigativas atuais.

A partir dessas considerações, do ciclo e do mapa dos dados (BRINEY, 2015) e do trabalho de Berber Sardinha (2004), que discutiu procedimentos para a constituição de *corpus* de pesquisa, e ainda levando em consideração o contexto de produção dos dados deste trabalho, o TTD, estabeleci os procedimentos de construção do MulTeC. Cada um desses procedimentos será detalhado neste trabalho no capítulo metodológico. No procedimento de organização dos arquivos, destacamos a criação dos cabeçalhos nos arquivos de texto.

Segundo Berber Sardinha (2004, p. 73), “cabeçalhos são uma parte do arquivo de cada texto do *corpus* que contém informações sobre o texto, tais como origem, a data de coleta, o grupo de pesquisa responsável, o tamanho do texto, sistema de transcrição, detalhes do *copyright*, a autoria, os participantes”. Para Berber Sardinha (2004), a codificação do cabeçalho pode ser “por meio de etiquetas SGML [...]” ou por “códigos do tipo Cocoa (*Count and Concordance on Atlas*)” (BERBER SARDINHA, 2004, p. 74-75). A codificação por SGML (*Standard Generalized Markup Language*) se configura como uma linguagem mais complexa, “usada para formatar (*encode*) o texto todo (incluindo as divisões internas do texto, as pontuações, os caracteres especiais) e não apenas as informações do cabeçalho” (BERBER SARDINHA, 2004, p. 74). Levando em consideração a quantidade de textos e o fato de que tal codificação pode representar uma dificuldade no processamento, haja vista que cada arquivo ficaria maior, optamos pela codificação do tipo Cocoa, com etiquetamento simples como apresenta Berber Sardinha (2004, p. 75) no “exemplo de cabeçalho no formato Cocoa de uma publicação dos sonetos de Shakespeare:

```
<author William Shakespeare>
<title Shake-speares Sonnets>
<stitle WS.Sonnets>
<placepub London>
<printer G. Eld>
<publisher T. T.>
<datepub 1609>
<datecomp ca. 1590-1609>”
```

O cabeçalho apresenta subsídios gerais de modo a resumir informações sobre o dado, seu tipo e origem, por exemplo. Logo, durante a constituição de um *corpus*, quando estiver a definir o padrão do cabeçalho, o pesquisador precisa considerar, que a opção que fizer deverá ser aquela que melhor auxiliará o usuário do *corpus* a compreender as condições gerais de produção do dado, de “modo sistematizado e consistente, para que não se tenha de recorrer à memória do pesquisador” (BERBER SARDINHA, 2004, p.76).

Cabe ressaltar que a construção de corpora, como discutido, demanda ainda metadados detalhados que contemplem os critérios que caracterizam um *corpus*, como o que foi realizado por Aranha e Leone (2016; 2017) para a criação do DOTI (Databank of Oral Teletandem Interactions) sobre o qual discorro a seguir.

#### 1.4.2 OS METADADOS NO DOTI (DATABANK OF ORAL TELETANDEM INTERACTIONS)

A descrição dos metadados no DOTI segue a lógica conceitual estabelecida por Aranha e Leone (2016; 2017) da divisão do contexto de produção dos dados em macrotarefas (Sessão Oral de Teletandem e sessão de mediação) e suas microtarefas. Como o DOTI é constituído das sessões orais, segundo as autoras, os metadados referem-se a essa macrotarefa – a Sessão Oral de Teletandem. Em seus trabalhos, Aranha e Leone (2016; 2017) descrevem os metadados do DOTI. No texto de 2017, as autoras dividem os metadados em dois grandes grupos: um voltado às questões pedagógicas, o cenário de aprendizagem, e o outro com as questões concernentes ao espaço de interação – ambos conceitos utilizados pelas autoras para amparar a compreensão do leitor do que consiste o TTD, como discutimos acima.

Quanto ao espaço de interação, Aranha e Leone (2017) apresentam como metadados as informações sobre os participantes, as instituições envolvidas nas parcerias<sup>51</sup> e o período em que ocorreram as interações (semestre, ano, número de SOTs, duração de cada SOT). Um item também descrito quanto ao espaço de interação é a língua utilizada nas SOTs, que precisa ser informada, e o ambiente tecnológico que medeia o desenvolvimento das SOTs. Segundo Aranha e Leone (2017), no TTD as SOTs são multimodais e síncronas.

O cenário de aprendizagem apresenta as informações pedagógicas “relacionadas ao percurso formativo no TTD<sup>52</sup>” (ARANHA, LEONE, 2017, p. 185). Logo, são apresentados como metadados do cenário de aprendizagem as informações referentes à modalidade de TTD realizada, período da parceria, nomes dos professores e mediadores, descrição das tarefas (conversação livre, direcionamento das conversas por tema, por exemplo). Segundo as autoras, foi criado para o DOTI um documento com informações censitárias dos participantes, como por exemplo, sexo, proficiência autoavaliada e curso de graduação ao qual o participante está vinculado quando participa do TTD.

Encerro o capítulo lembrando os conceitos que embasaram as escolhas metodológicas que pautaram a organização do MulTeC. Para realizar essa tarefa considerei os conceitos de cenário de aprendizagem, a partir do qual optei pela organização das pastas e subpastas do MulTeC, o contexto de produção dos dados – TTD – como espaço de aprendizagem telecolaborativo, ainda a concepção de tarefas e no TTD a relação entre essas tarefas para o alcance da macrotarefa, como o leitor poderá compreender melhor no decorrer deste trabalho. Em todo o processo de organização, estando atenta ao plano de gerenciamento

---

<sup>51</sup> No caso do DOTI: Unesp/UGA e Unisalento/universidades britânicas e estadunidenses.

<sup>52</sup> Nossa tradução para: [...] pedagogical issues related to the formative path based on Teletandem”.

de dados elaborado para atingir o objetivo geral da pesquisa. A seguir apresento os percursos metodológicos pelos quais optei para construir o MulTeC.

## CAPÍTULO 2: PERCURSOS METODOLÓGICOS

Neste capítulo, descrevo os procedimentos metodológicos adotados na construção do MulTeC – Multimodal Teletandem *Corpus*, os quais foram estabelecidos a partir da concepção de *corpus* apresentada por Sardinha (2009) e Sinclair (2004), cujas proposições apontam as características primordiais de um *corpus*. Além deles, estou embasada na metodologia apresentada por Chanier e Wigham (2016) para construção de *corpus*, pois apresentam critérios para organização de dados produzidos em interações de aprendizes mediadas por computador - contexto similar ao TTD - bem como nos trabalhos de Briney (2015), Eynden (2013) e Eynden *et al.* (2011) que discutem gerenciamento de dados de pesquisa.

De acordo com Briney (2015) e Eynden *et al.* (2011), há procedimentos a serem considerados cautelosamente quando existe a proposição de disponibilizar dados para uso a longo e médio prazo e/ou por outros pesquisadores, ou seja, quando se propõe a construir um *corpus* de pesquisa sob a ótica da *Open Science*. Isso significa que mesmo que os dados que constituem o MulTeC não cheguem a ser disponibilizados para livre acesso, a construção do *corpus* precisa seguir os critérios que viabilizem tal acesso, seguindo tendência internacional de compartilhamento dos dados para fortalecimento da Ciência.

O capítulo está dividido em quatro seções: a abordagem metodológica, a contextualização da pesquisa e os procedimentos adotados, na terceira seção detalho os procedimentos para a construção do MulTeC em três categorias (pré-tratamento, tratamento e organização), considerando os objetivos específicos da pesquisa que serão apresentados na contextualização. Vale ressaltar ainda que, neste trabalho, os procedimentos metodológicos foram se tecendo a partir da análise da organização dos dados do Banco TTDii. Início com a discussão da abordagem de pesquisa adotada.

### 2.1 ABORDAGEM DA PESQUISA

Esta pesquisa se insere na abordagem mista em que procedimentos qualitativos e quantitativos se complementam para o alcance dos objetivos propostos. Segundo (DÖRNEY, 2009, p. 45) “a principal atração pela pesquisa com métodos mistos tem sido pelo fato de que

ao usar tanto as abordagens QUAL como QUAN<sup>53</sup>, os pesquisadores podem trazer o melhor dos dois paradigmas, assim combinando forças da pesquisa quantitativa e qualitativa”<sup>54</sup>.

Sobre pesquisa qualitativa Bogdan e Bilken (1996, p. 47-51), elencam algumas características que resumem esta abordagem.

Para esses autores:

- a) Na pesquisa qualitativa, a fonte direta de dados é um ambiente natural. O investigador qualitativo considera o contexto de produção dos dados, entendendo que a observação dos locais reais das ocorrências auxiliará na compreensão dos fatos investigados;
- b) A investigação qualitativa descreve os fatos estudados ao tentar analisar toda a complexidade que os envolve;
- c) A importância vital se encontra no significado, de modo que o pesquisador esteja atento às diferentes perspectivas para apreendê-las adequadamente, registrando da melhor forma possível o fenômeno que o interessa, seja através de vídeos e entrevistas no intuito de registrar rigorosamente tanto “quanto possível o modo como as pessoas interpretam os significados” (BOGDAN; BIKLEN, 1994, p.51).

Essas características denotam principalmente a relevância do contexto na produção dos dados utilizados para/na pesquisa, levando em consideração a complexidade da realidade que os viabiliza. O pesquisador que realiza pesquisa qualitativa objetiva garantir que os dados analisados sejam reais e, conseqüentemente, possam apontar indícios de respostas às questões investigadas. Em função disso, ele está interessado na compreensão do recorte da realidade que selecionou. Assim, há que se considerar o contexto de produção dos dados e procurar reconstruir o contexto com o auxílio de documentos descritivos, por exemplo.

Ao diferenciar pesquisa qualitativa de pesquisa quantitativa, Günther (2006) caracteriza a primeira como ato de construção subjetivo, buscando a compreensão como “princípio do conhecimento” (GÜNTHER 2006, p. 202), através de textos que devem ser analisados com instrumentos e procedimentos específicos, contemplando a singularidade e complexidade do objeto de estudo. No entanto, ele enfatiza que é perfeitamente viável a utilização de ambas abordagens. O que definirá o uso de uma, de outra ou de ambas são as

---

<sup>53</sup> Ao usar as abreviaturas QUAL e QUAN o autor se refere respectivamente a qualitativa e quantitativa.

<sup>54</sup> Tradução minha para: The main attraction of mixed methods research has been the fact that by using both QUAL and QUAN approaches researchers can bring out the best of both paradigms, thereby combining quantitative and qualitative research strengths. (DORNIEY, 2009, p. 45)

questões “prática, empírica e técnica” (GÜNTHER, 2006, p. 207) a depender da pergunta de pesquisa que move o investigador e das possibilidades e ferramentas que puder utilizar. Outros autores também tratam da questão da junção de ambas. Para Strauss e Dörney (1998 *apud* DÖRNEY, 2009, p. 43), “a questão não é usar uma ou outra forma, mas como podem trabalhar juntas para promover o desenvolvimento da teoria.”

A opção pela junção de ambas (pesquisa quantitativa e qualitativa), nesta tese, se dá pelo fato de que o trabalho demandou uma série de levantamentos quantificados, pois era necessário ter ciência de quantos dados tínhamos por tipo de dados produzidos e número de participantes, no que diz respeito aos dados coletados no Banco TTDii. Assim, para a realização deste trabalho precisei realizar entrevistas, coletar documentos, realizando uma pesquisa exploratória e documental no intuito de alcançar o objetivo de construir o MulTeC. Alcançar esse objetivo desencadeou vários processos que descrevo adiante ainda neste capítulo, mas antes situo o leitor no contexto da pesquisa.

## 2.2 CONTEXTUALIZANDO A PESQUISA

Um *corpus*, retomando o conceito de Sinclair (2004, p.19), “é um conjunto de textos em formato eletrônico [...] que representa uma língua ou variedade da língua como fonte de dados para pesquisas linguísticas”. De modo que os dados de pesquisa estejam organizados para esse objetivo, o pesquisador que se propõe a organizar o *corpus* precisa selecionar o contexto de produção e o tipo de dado a ser coletado. Como afirmam Chanier e Wigham (2016), Sinclair (2000, 2004), Berber Sardinha (2009), Briney (2015) e Eynden *et al.* (2011), a pergunta de pesquisa é que orientará o processo da coleta de dados, pois a partir dela é que os textos serão selecionados e o modo como serão coletados será estabelecido, a pergunta de pesquisa que orientou este trabalho foi: Como é possível organizar os dados do Banco TTDii de modo que sejam compartilháveis?<sup>55</sup>

Em contextos ideais, o detalhamento do processo de coleta, embasado por perguntas de pesquisa, é parte importante para a constituição de um *corpus*. Entretanto, o que ocorreu com o MulTeC foi um caminho diverso. A maioria dos dados foi coletada em período anterior à proposição desta tese.

---

<sup>55</sup> Segundo Guichon (2017) dados compartilháveis são os que foram tratados, transcritos, estruturados e contextualizados. A contextualização dos dados demanda informações complementares como, por exemplo, idade e sexo dos participantes, proficiência na língua em estudo, natureza das tarefas, documentos usados, instruções dadas e duração de parcerias.

Os dados, provenientes da realização de parcerias teletandem efetivadas entre 2011 e 2015, foram preliminarmente organizados em HD externo, no qual os arquivos foram divididos em pastas conforme o ano e o semestre de ocorrência; seguindo uma organização que separava os dados por tipo de dado produzido em cada turma de teletandem (turma TTD). Ao HD externo só as pesquisadoras e monitoras do projeto Teletandem de São José do Rio Preto tinham acesso.

No que se refere a quantidade de arquivos constantes do Banco TTDii, não havia informações, assim como não havia como saber o número de participantes geral ou quantas turmas tiveram seus dados salvos. A ausência dessas informações me conduziu à necessidade de um primeiro levantamento dos dados antes da elaboração de um plano de gerenciamento dos dados coletados de modo que soubesse a quantidade de cada tipo de dado e sua qualidade para então começar a construção do *corpus*. A partir dos embasamentos teóricos deste trabalho, realizei uma análise da organização dos dados no Banco TTDii, procurando responder às primeiras perguntas necessárias para começar a organização do *corpus*. Apresento-as e a relação delas com as descobertas feitas no quadro a seguir.

**Quadro 1 – Relação entre as perguntas feitas na análise da organização do banco TTDii e suas respectivas descobertas**

<i>Perguntas</i>	<i>Descobertas</i>
<ul style="list-style-type: none"> <li>• <i>Qual a quantidade de arquivos de textos?</i></li> </ul>	<ul style="list-style-type: none"> <li>- Arquivos de texto corrompidos;</li> <li>- Arquivos com nomeação fora do padrão estabelecido pelo grupo de pesquisa em TTD de Rio Preto;</li> <li>- Arquivos com identificação dos participantes.</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Quantas SOTs há no Banco TTDii?</i></li> </ul>	<ul style="list-style-type: none"> <li>- Arquivos de áudio/vídeo corrompidos;</li> <li>- Arquivos salvos como sendo SOT, mas contendo apenas gravações teste.</li> </ul>
<ul style="list-style-type: none"> <li>• <i>Que metadados são fornecidos no Banco TTDii?</i></li> </ul>	<ul style="list-style-type: none"> <li>- Presença de turmas de TTDisi e TTDii;</li> <li>- Ausência de alguns TCLEs.</li> <li>- Das 16 turmas com dados no banco, apenas de 2 havia lista de pareamento.</li> <li>- Não havia lista de presença com descrição das tarefas ou registro de ocorrências.</li> </ul>

**Fonte:** Elaborado pela autora

Cada uma dessas perguntas feitas no intuito de descobrir detalhes sobre os dados do Banco TTDii revelou minúcias que não tinham sido previstas, como por exemplo ao procurar identificar metadados referentes as turmas com dados no banco, encontrei listas de pareamento que indicavam que em um dos lados (os brasileiros) os participantes eram voluntários, o que é

uma das características do teletandem institucional semi-integrado. Ao abrir os arquivos das SOTs, com objetivo de verificar a duração de cada uma delas, identifiquei aquelas sem qualidade para uso em pesquisas, além das que foram salvas como SOT, mas que continham apenas teste de gravação com duração de 50 minutos. Enquanto a pergunta que se referiu a quantidade de arquivos de textos no banco levou a descoberta de arquivos corrompidos, com identificação dos participantes (nome, e-mail, *link* para perfil em redes sociais), assim como havia ausência de alguns TCLEs.

Além dessas descobertas, a partir desse levantamento foi possível saber a quantidade de dados com os quais seria realizado o trabalho de organização. Na tabela 2 apresento o quantitativo dos dados que encontrei no Banco TTDii, por tipo e a quantidade de dados que constituem o MulTeC (*Multimodal Teletandem Corpus*). Após o tratamento houve uma redução dos dados que foi realizada seguindo os critérios de qualidade dos dados e autorização para uso dos dados em pesquisas.

**Tabela 2 – Dados múltiplos de acordo com Guichon (2017) – Quantitativos do Banco TTDii**

<i>Tipos de arquivos</i>	<i>Dados múltiplos (Banco de dados TTDii)</i>	
<i>SOTs<sup>56</sup> (arquivos de video/áudio)</i>	655'43''22'''	
<i>Diários de aprendizagem</i>	849	
<i>Questionários</i>	180	
<i>Textos</i>	1444	
<i>Chats</i>	477	
<b><i>Total de turmas TTD</i></b>	<b>11 (TTDii)</b>	<b>5 (TTDisi)</b>

Fonte: Elaborado pela autora

Aranha, Luvizari-Murad e Moreno (2015) tiveram o cuidado de indicar que na organização do referido banco havia pastas nomeadas como “Interações sem aproveitamento” nas quais deveriam ser salvos os dados sem TCLE ou com qualidade abaixo do aceitável. Esse primeiro levantamento resultou na constatação de que havia inconsistências a serem resolvidas antes de iniciar o trabalho de organização do *corpus*. Este trabalho teve como objetivo geral:

- Construir um *corpus* multimodal em teletandem com as produções coletadas no contexto TTD da Unesp de São José do Rio Preto, a partir do levantamento, anonimização e organização dos dados produzidos no período de 2012 a 2015.

<sup>56</sup> Em horas, minutos e segundos.

Para alcançar esse objetivo, os dados provenientes das parcerias de TTD entre alunos da Unesp – São José do Rio Preto e da *University of Georgia* estabeleci, embasada em Briney (2015), Chanier e Wigham (2016), Eynden (2013) e Eynde *et al.* (2011) os procedimentos que procurariam transformar os dados múltiplos, segundo Guichon (2017) em dados compartilháveis. Esses procedimentos foram:

- Levantamento dos TCLEs;
- Seleção de dados com qualidade necessária para o uso em pesquisas;
- Coleta de documentos contextuais por turma TTD;
- Criação de códigos para anonimização;
- Estabelecimento de padrão de nomeação e anonimização, seguido da realização desses processos;
- Conversão dos arquivos para formatos livres;
- Organização dos dados em pastas por turma e parceria TTD;
- Elaboração de documentos complementares.

O estabelecimento desses procedimentos procura atender as exigências internacionais para o trabalho de organização de *corpus* de pesquisa. Sobre cada um dos procedimentos detalharei neste capítulo.

Ao realizar esses procedimentos consigo alcançar os objetivos específicos estabelecidos para este trabalho, a saber:

- Analisar a organização dos dados produzidos e coletados no período de 2012 a 2015, conforme descrito em Aranha, Luvizari-Murad e Moreno (2015);
- Propor padrão de anonimização e nomeação dos arquivos (BRINEY, 2015; CHANIER; WIGHAM, 2016; EYNDEN *et al.*, 2011);
- Organizar e analisar os dados a partir dos pressupostos teóricos de organização de dados de pesquisa (BRINEY, 2015; CHANIER; WIGHAM, 2016; EYNDEN, 2013; EYNDEN *et al.*, 2011), elaborando e acrescentando documentos para a construção de um *corpus* de pesquisa.

Alcançados esses objetivos haveria encontrado a resposta para minha pergunta de pesquisa e estaria confirmada a hipótese inicial de que era possível constituir um *corpus* com os dados que integram o Banco TTDii, excluindo os dados produzidos em 2011 que não foram selecionados para constituir o MulTeC por não haver ocorrido aplicação do questionário inicial – documento que auxilia na coleta de informações, como descrevo neste capítulo.

Na organização do MulTeC segui etapas criteriosas, estabelecidas a partir do pressuposto por Briney (2015), Eynden (2013) e Eynden *et al* (2011), que descreverei no decorrer deste capítulo. Na seção seguinte, detalho três etapas para a construção do *corpus* de modo que seja compartilhável.

## 2.2 AS 3 ETAPAS PARA A CONSTRUÇÃO DO *CORPUS*

As etapas para a construção do *corpus* objeto desta pesquisa estão embasadas, principalmente nos trabalhos de Chanier e Wigham (2016), Briney (2015), Eynden (2013), Eynden *et al.* (2011) bem como na discussão de Berber Sardinha (2004) quanto à constituição de *corpus* de pesquisa. Dessa discussão, retomo principalmente o ciclo de vida dos dados (Planejamento, coleta, análise, publicação/compartilhamento, preservação e reuso dos dados) e seus procedimentos, a partir do qual apresento os procedimentos adotados para a sistematização do MulTeC, considerando suas especificidades.

Cada etapa do “novo ciclo de vida dos dados” pode ser realizada com procedimentos diversos, conforme o mapa de dados de Briney (2015) que apresento no capítulo teórico. Esse mapa inclui tanto os procedimentos do ciclo de vida dos dados como aqueles relacionados ao arquivamento, parte relevante do trabalho de construção de um *corpus* de pesquisa. Na organização do MulTeC as 3 etapas foram nomeadas de pré-tratamento, tratamento e organização a partir da característica de cada procedimento realizado em cada etapa. Assim, os procedimentos que antecedem o tratamento dos dados, como, por exemplo, a coleta de documentos como o Powerpoint® da reunião tutorial de cada uma das turmas e as listas de pareamento, estão categorizados como procedimentos da etapa de pré-tratamento e assim sucessivamente, como será detalhado ainda neste capítulo.

A realização desses procedimentos para esta pesquisa precisou enfrentar alguns desafios. O primeiro desafio foi estabelecer um padrão de nomeação que fosse o mais significativo possível para o pesquisador interessado nos dados sem ter caracteres em demasiado. O segundo desafio, que surgiu em decorrência do primeiro, foi a criação de um código identificador para os participantes. O código do participante estaria inserido na nomeação de todos os arquivos, logo, tornou-se indispensável decidir sobre o código para então propor um padrão de nomeação.

Foi a partir do padrão de nomeação, estabelecido para este trabalho, que os processos de organização das pastas, anonimização e construção do *corpus* puderam ser padronizados e

realizados, como discuto ainda neste capítulo. Assim, estabelecido o padrão de nomeação - que indica características do participante como sexo e curso, bem como o ano de produção e o tipo de dado - foi possível começar a definir os procedimentos necessários para o alcance do objetivo desta pesquisa de cunho qualitativo e quantitativo, desenvolvendo vários processos para o alcance de cada objetivo específico como explicarei na próxima seção.

### 2.2.1 UM OBJETIVO: VÁRIOS PROCESSOS

Cada uma das três etapas para a organização dos dados em um *corpus* corresponde a um dos objetivos específicos estabelecidos para a realização desta pesquisa. Assim, para o primeiro objetivo específico, a saber: analisar a organização dos dados produzidos e coletados como descrito por Aranha, Luvizari-Murad e Moreno (2015), foram estabelecidos procedimentos necessários para realizar o ‘pré-tratamento’, a saber: i) exclusão dos dados cujo TCLE não foi assinado/coletado; ii) controle de qualidade dos dados, excluindo-se dados sem possibilidade de uso seja pela qualidade ou por estarem corrompidos; iii) levantamento de documentos produzidos para cada turma TTD, mas não incluídos no Banco TTDii (listas de pareamento, apresentações Powerpoint ® da reunião tutorial).

O segundo objetivo específico<sup>57</sup> demandou: i) a criação de convenção de códigos para nomeação dos dados e arquivos, a partir de Chanier e Wigham (2016); ii) o estabelecimento de padrão de anonimização específico, embasados em Briney (2015), Chanier e Wigham (2016); iii) a conversão dos dados em formatos de livre acesso conforme orientam Briney (2015) Eynden (2013), Eynden *et al.* (2011); e iv) a (re)nomeação dos arquivos no intuito de facilitar a pesquisa junto ao *corpus*.

O terceiro objetivo específico gerou a necessidade de estabelecimento de um padrão de organização das pastas e subpastas um pouco diverso do apresentado no Banco TTDii. Esse procedimento se justifica por este trabalho estar embasado, para a organização, no conceito de cenário de aprendizagem utilizados também por Aranha e Leone (2016, 2017) para a constituição do DOTI e discutido no trabalho de Foucher (2010). Ainda no escopo desse objetivo, foi necessária a elaboração de documentos com informações contextuais sobre a proveniência dos dados (ASSANTE *et al.*, 2016).

Resumo os procedimentos que estabeleci para cada objetivo específico no Quadro 2 a seguir:

---

<sup>57</sup> Padronizar a nomeação e anonimização dos dados.

**Quadro 2 – Relação entre procedimentos e objetivos específicos**

<i>Objetivo específico 1</i>	<i>Objetivo específico 2</i>	<i>Objetivo específico 3</i>
<i>Investigar e analisar os dados não-tratados</i>	<i>Padronizar a nomeação e anonimização</i>	<i>Organizar e analisar os dados compartilháveis</i>
<b>Pré- tratamento</b>	<b>Tratamento</b>	<b>Organização das pastas e subpastas do MulTeC</b>
<ul style="list-style-type: none"> <li>• Seleção dos dados a partir da qualidade;</li> <li>• Levantamento dos TCLEs;</li> <li>• Inclusão de tutoriais e listas de presença e pareamento.</li> </ul>	<ul style="list-style-type: none"> <li>• Criação de códigos;</li> <li>• Padronização da nomeação e anonimização;</li> <li>• Anonimização dos dados;</li> <li>• Conversão dos dados;</li> <li>• (Re)Nomeação dos arquivos.</li> </ul>	<ul style="list-style-type: none"> <li>• Organização considerando a modalidade de TTD, anos e turmas.</li> </ul> <p><b>Elaboração de documentos explicativos e contextuais</b></p> <ul style="list-style-type: none"> <li>• Explicação dos códigos utilizados em cada planilha;</li> <li>• Apresentação dos dados quantitativos;</li> <li>• Apresentação do contexto TTD;</li> <li>• Criação dos metadados.</li> </ul>

Fonte: A autora

Cada um dos procedimentos sumariamente apresentados acima será retomado e explicado subsequentemente, iniciando com o primeiro objetivo específico.

#### 2.2.1.1 O início do processo de organização: coleta de informações sociolinguísticas

No intuito de alcançar o objetivo de investigar e analisar os dados que constituem o Banco TTDii procedi da seguinte forma: i) assisti os 5 minutos iniciais e os 3 minutos finais<sup>58</sup>

<sup>58</sup> Rampazzo (2017) indicou que nos minutos iniciais da SOTi os participantes se referem ao curso ao qual estão matriculados.

de cada um dos 903 arquivos de SOTs, objetivando identificar substituições, cursos de alguns participantes, assim como o sexo, haja vista que não houve, durante a coleta, um levantamento<sup>59</sup> junto aos alunos quanto ao seu perfil sociocultural de modo que pudessem ser coletadas informações sociolinguísticas relevantes para a construção de um *corpus*, como a declaração de identidade de gênero pelo participante e a idade. As informações sociolinguísticas (sexo, curso, língua materna e língua estrangeira) que pude constatar foram utilizadas na criação de códigos que foram úteis no processo de anonimização e na nomeação dos dados e registrada na planilha “Levantamento Banco de Dados com IT”, que não integra o MulTeC. Esse levantamento de informações sociolinguísticas foi realizado através das SOTs, pois ao assisti-las foi possível definir um sexo e o curso, e através da leitura de cada um dos questionários iniciais, através da qual pude coletar informações como a proficiência na língua em estudo. Vale ressaltar que essa proficiência é autodeclarada pelo participante no questionário inicial.

No questionário inicial os participantes respondem sobre sua proficiência na língua que objetivam aprender e estabelecem seus objetivos de aprendizagem, informações quanto a idade, sexo e tempo de estudo não constam nos questionários do período de 2012 a 2015. Assim, realizei buscas nos documentos no intuito de coletar essas informações a partir de três tipos de arquivos: o questionário inicial (documento em Word®), o TCLE (digitalizado em .pdf) e a SOTi (arquivo em .mp4), como afirmo anteriormente.

O uso da SOTi para identificação de algumas dessas informações se deu embasado no trabalho de Rampazzo (2017) que, no intuito de investigar os movimentos retóricos<sup>60</sup> dos 15 minutos iniciais da sessão, identificou a presença do movimento “troca de informações” que seria um momento no qual os parceiros, trocam informações relativas à vida acadêmica e pessoal<sup>61</sup> no qual o participante brasileiro pergunta qual curso o estadunidense “faz” e o estadunidense responde que está cursando Inglês e Literatura. Vale informar ao leitor que esse procedimento para identificação do curso do participante foi necessário apenas nas turmas de teletandem institucional semi-integrado (TTDisi), pois as respectivas listas de pareamento dessas turmas não continham essa informação.

---

<sup>59</sup> Entendemos que o próprio participante deveria identificar seu gênero, de modo que pudéssemos ir além do binarismo do masculino-feminino cuja identificação foi feita através da observação da imagem do participante no vídeo.

<sup>60</sup> Embasada em Swales (1990) e Aranha (2004), Rampazzo (2017, p. 43) afirma que os movimentos retóricos são “partes argumentativas” e “podem ser vistos em termos de conteúdo encontrado ou como desempenhando diferentes funções que levarão ao cumprimento dos propósitos comunicativos.”

<sup>61</sup> Nas transcrições do MulTeC as indicações de troca de turno estão como B=brasileiro e E=estadunidense, indicando a origem da instituição à qual o participante está vinculado.

Assim, para a identificação do curso, L1 e L2, precisamos ouvir os 15 minutos iniciais de todas as SOTis das turmas de TTD<sub>isi</sub>, totalizando 22 SOTis, totalizando a oitava de 330 minutos, nas 5 turmas de teletandem semi-integrado que tiveram coleta de dados no período de 2012 a 2015.

Durante esse processo de levantamento, percebi que, no contexto TTD, há participantes que estão matriculados em um curso na universidade estadunidense, mas têm origem mexicana, por exemplo, o que, em alguns casos, significa que sua língua materna é espanhol, falam inglês proficientemente e estão participando do TTD como aprendizes de língua portuguesa. O que nos levou a registrar então L1, L2 e L3, sendo a L1 a língua na qual é nativo, L2 a língua na qual é proficiente e L3 a língua-alvo para a turma TTD na qual está inserido. Nos casos em que o participante ainda não fala outra língua além da nativa, há informação quanto à L1 e L2 apenas. Vale ressaltar que só foi possível realizar o levantamento quanto aos participantes brasileiros em decorrência do fato de que havia informações sobre brasileiros de 11 turmas do banco, faltando apenas sobre 5 turmas, enquanto que no que se refere aos estadunidenses teríamos que investigar todas as SOTis coletadas na modalidade de TTD<sub>ii</sub>, além das TTD<sub>isi</sub>, o que não foi possível pois não havia tempo hábil no espaço temporal deste trabalho.

Outra informação retirada da SOTi foi a identificação de sexo. Como não houve oportunidade para que os participantes identificassem o próprio gênero, a identificação foi realizada a partir das imagens do vídeo, considerando-se as características físicas de cada participante. Entendemos que idealmente a identificação de gênero deve ser declarada pelo participante, no entanto, não havia como solicitarmos essa auto-identificação considerando a quantidade de participantes cujos dados foram coletados e a insuficiência de dados que viabilizassem o contato com eles. A identificação do gênero e do curso estão inseridas na constituição do IT apresentado anteriormente neste trabalho e constituem informações constantes na planilha “Informações Participantes” sobre a qual discorreremos a seguir.

A idade dos participantes à época da participação em cada uma das turmas TTD integrantes do Banco TTD<sub>ii</sub> não fora coletada durante a organização detalhada por Aranha, Luvizari-Murad e Moreno (2015), porque não era o objetivo do grupo de pesquisa construir um *corpus*. Conforme citado na metodologia, trata-se de informação relevante para tal e como não havia sido realizada sua coleta, procurei extraí-las dos dados que constavam do Banco TTD<sub>ii</sub>. Assim, para realizar o levantamento da idade dos participantes utilizei o TCLE que solicitava a data de nascimento como uma das informações pessoais. A maioria dos participantes que assinaram o TCLE informaram a data de nascimento. Desse modo, para realizar o cálculo da

idade de cada participante considerei a data de nascimento e o ano/semestre no qual estava participando da turma utilizando os 219 TCLEs assinados pelos participantes brasileiros. No *Consent Form*, documento de autorização assinado pelos estudantes da Universidade de Georgia<sup>62</sup> não havia a informação de data de nascimento, ficando inviabilizado o levantamento de idade dos estadunidenses. A ausência de alguns TCLEs demandou a necessidade de solicitar assinatura retroativa, como descrevo na sequência.

#### 2.2.1.1.1 O TCLEs

O levantamento dos TCLEs foi realizado com auxílio da planilha “Levantamento Dados Banco com IT”. Elaborei essa planilha para registrar, por participante, se havia TCLE assinado e se a autorização de uso dos dados incluía a autorização da imagem. Para tal usei como códigos: X, D e DI – “X” para não autorizado; “D” para autorização de uso dos dados e “DI” quando o participante autorizou o uso dos dados e da sua imagem. Não houve registro de participante que autorizou apenas o uso da imagem em pesquisas. Urge informar ao leitor que, para inserir as informações na planilha de levantamento, precisei abrir cada um dos TCLEs digitalizados para identificar se a autorização concedida era de imagens e dados ou apenas dados.

Aproveitei a abertura de cada um dos TCLEs para realizar o levantamento sobre a idade dos participantes<sup>63</sup> e na mesma planilha reuni informações quanto ao ano e semestre de ocorrência de cada turma, usuário Skype utilizado e acrescentei uma coluna para inserção do código do participante que foi criado posteriormente. Assim, a planilha em Excel® para o levantamento dos TCLEs foi criada com 9 colunas, de modo que em um documento apenas houvesse o registro de: i) ano/semestre de realização da turma, ii) nome da turma, iii) nome dos participantes, iv) email, v) usuário Skype, vi) código de participante (IT-identidade Teletandem), vii) data de nascimento, viii) idade e ix) autorização de dados. Essa planilha foi ferramenta importante para a elaboração de documentos complementares com as informações dos participantes, que apresento no capítulo no qual discorro sobre os resultados.

A partir do levantamento dos TCLEs cheguei ao cômputo de 61 participantes sem os respectivos TCLEs salvos no banco, embora houvesse dados produzidos por eles salvos no banco. Não foram encontrados registros que informassem sobre a razão da ausência do documento que poderia ser justificada em função da não digitalização desses documentos ou

---

<sup>62</sup> Anexo a esta tese há um exemplo de TCLE e de *Consent Form*.

<sup>63</sup> Esse cálculo pode ser realizado apenas para todos os participantes brasileiros. No *Consent Form* dos estadunidenses não havia informação de data de nascimento que viabilizasse o cálculo.

da simples negativa do participante na doação dos dados para pesquisa. Como o número era expressivo e excluir todos os dados desses 61 participantes representava perder muitos dados do Banco TTDii, realizei uma busca no laboratório TTD da Unesp –Rio Preto onde ficam arquivados, em armário com chaves, os documentos assinados pelos alunos, que costumam ser então digitalizados e salvos.

Feita a vistoria no laboratório, constatei que os TCLEs dos 61 participantes não estavam arquivados, nem foram encontrados, razão pela qual decidi entrar em contato com essas pessoas via e-mail ou via rede social, auxiliada pelo grupo de pesquisa TTD de São José do Rio Preto, de modo que fosse possível conseguir assinatura retroativa delas. Dos 61 e-mails enviados recebi 28 respostas, dos quais 5 responderam negativamente e 23 participantes autorizaram. Os 23 participantes que aceitaram, assinaram retroativamente o TCLE, enviando-o digitalizado via e-mail ou entregando o documento assinado que, posteriormente, foi digitalizado. Os TCLEs assinados retroativamente foram salvos em HD externo utilizado para a construção do *corpus*.

A planilha em Excel® elaborada inicialmente para o levantamento dos TCLEs e tipo de autorização fornecida pelos participantes, serviu ainda de apoio para a organização dos dados durante todo o processo de constituição do *corpus*, pois informações referentes ao IT ou ao usuário Skype® foram confirmadas ou retificadas a partir da análise de cada um dos dados produzidos pelos participantes e coletados no banco. Porém, essa planilha não integra o MulTeC, por conter informações sigilosas como os nomes completos dos participantes e seus respectivos endereços de e-mail. A partir desse levantamento, foi possível chegar a uma contabilização das turmas e da quantidade de participantes, chegando ao número de 16 turmas teletandem, com 147 participantes brasileiros, 168 participantes estadunidenses que formaram no período 208 pares. Desses participantes, 282 assinaram TCLE disponibilizando imagem e/ou dados para pesquisas.

Os dados que constituem o MulTeC são, portanto, os arquivos com as SOTs, os chats, os diários, os questionários inicial e final, e os textos produzidos para a troca entre parceiros, dados constantes do Banco TTDii, excluindo aqueles que não foram autorizados ou cuja qualidade impossibilitava o uso em pesquisas.

Após essa etapa cheguei aos seguintes números:

- 581h (cerca de) de SOTs;
- 666 diários de aprendizagem;
- 351 chats;
- 956 textos produzidos na troca entre os parceiros;

- 91 questionários iniciais;
- 41 questionários finais.

Esses foram, então, os dados nos quais precisaria trabalhar no tratamento, preparando-os para constituir um *corpus* de pesquisa. Contudo para realizar o tratamento dos dados havia necessidade de compreensão do seu contexto de produção, havendo assim igual necessidade de resgatar documentos com registro elaborado pelas mediadoras do período de ocorrência das turmas, tais como: listas de presença e pareamento, além do Powerpoint® tutorial.

Não encontrei parte considerável desses documentos no Banco TTDii, como algumas listas de pareamento, conforme afirmei anteriormente neste trabalho, pois havia ainda a ausência dos arquivos utilizados nas reuniões tutoriais e listas de pareamento de várias turmas.

Reitero, em consonância com Rampazzo (2017), que (i) o tutorial, bem como (ii) a sessão de mediação auxiliam na compreensão do contexto de produção dos dados e ambos deveriam ser gravados, caso haja interesse do grupo de pesquisa TTD de São José do Rio Preto em constituir *corpus* com os dados produzidos no TTD.

Assim, o Powerpoint® tutorial precisa ser inserido no *corpus* levando em consideração que nele o pesquisador poderá inferir<sup>64</sup> ou depreender quais orientações foram dadas sobre o contexto de aprendizagem TTD, bem como as tarefas previstas para aquela turma (para cada turma ocorre uma reunião tutorial com uma apresentação preparada exclusivamente para aquele grupo e, durante a reunião tutorial, há costumeiramente a apresentação do calendário de atividades a ser cumprido pelos participantes, no qual estão descritas tanto as tarefas, quanto o prazo para a execução de cada uma delas).

De modo semelhante a sessão de mediação também precisa ser inserida no *corpus*, porque nela também ocorrem orientações, discussões sobre o contexto de aprendizagem teletandem e sugestões de resoluções de possíveis conflitos entre os parceiros, logo, essa sessão, também deveria ser considerada dado a ser coletado. Contudo, as sessões de mediação não tive como recuperar, pois não foi arquivado qualquer registro delas.

Os documentos acrescentados foram padronizados<sup>65</sup>, facilitando encontrar informações relevantes para o pesquisador usuário do MulTeC como a modalidade de ocorrência, o período de realização da parceria, as mediadoras responsáveis e a quantidade de participantes. As versões originais desses documentos não seguiam um padrão, excetuando-se o Powerpoint® utilizado no tutorial, pois não havia o intuito de construir um *corpus* com os

---

<sup>64</sup> A gravação em áudio ou vídeo da reunião ‘tutorial’ seria útil também.

<sup>65</sup> Essa padronização se refere às listas de presença e pareamento, pois a apresentação em PPT do tutorial estava padronizada.

dados coletados. Nos demais, em especial nas listas de pareamento constatei a falta de padronização que acarretou a ausência de informações, como, por exemplo, em uma lista havia indicação das turmas, nas respectivas instituições, às quais pertenciam os participantes que constituiriam a turma TTD, no mesmo documento referente a outra turma realizada no mesmo semestre não havia o registro. Como exemplo adicional, cito a lista de pareamento com apenas uma listagem sequencial dos nomes dos participantes – brasileiros e estadunidenses – mas sem haver indicação de quais seriam os participantes que formariam cada uma das duplas. Essa informação, ou seja, quem foi designado para formar par com quem no pareamento, é relevante em razão de que, ao se deparar com um conjunto de SOTs de um mesmo par, caso o pesquisador perceba no vídeo que há pessoas diferentes em alguns dos vídeos, poderá checar a listagem de pareamento de modo que seja possível inferir se houve substituição efetiva ou temporária e quem foi designado para formarem pareceria para o período. Como discuti anteriormente neste trabalho, a substituição intervém no fluxo da interação em uma SOT, como ocorre em qualquer outro ambiente em que um par está realizando um diálogo e um dos participantes precisa ser substituído por outra pessoa que não estava acompanhando o desenrolar da interação.

Dadas as características do teletandem para a aprendizagem de línguas, portanto, a dinâmica das SOTs se altera cada vez que há uma substituição, pois o substituto não tem condições de dar continuidade ao que o par estava realizando nas sessões anteriores, de modo que ao substituir, há necessidade de nova apresentação ao parceiro daquele dia e negociação do que farão durante a SOT sem haver compromisso para um próximo encontro. Dessa maneira, saber que houve substituição temporária ou efetiva faz diferença para o pesquisador. Além disso, as substituições temporárias e efetivas podem constituir outras áreas de interesse dos pesquisadores.

#### 2.2.1.1.2 Proficiência dos participantes

Uma informação também relevante no contexto de aprendizagem de línguas telecolaborativo é a proficiência na língua-alvo que no TTD é autoavaliada pelo participante antes da realização da SOTi. O instrumento a partir do qual foi possível coletar esse dado foi o questionário inicial que apresenta o Quadro Comum Europeu (*Common European Framework of Reference for Languages - CEFR*) –Figura 9 - para auxiliar o participante na autoavaliação de sua proficiência. O intuito da apresentação desse quadro é auxiliar o estudante a refletir sobre seus objetivos a partir da proficiência que almeja alcançar.

**Figura 9 – Quadro comum europeu apresentado no questionário inicial do TTD<sup>66</sup>**

Level	Listening	Reading	Speaking	Writing
<b>C2</b>	Has no difficulty to understand any kind of spoken language delivered at native speaker speed	Can understand a wide range of long and complex texts	Has a good command of idiomatic expressions	Can write clear, smoothly flowing, complex texts in a logical structure.
<b>C1</b>	Can understand enough to follow complex topics, though he/she may need to confirm details	Can understand in detail lengthy complex texts, provided he/she can reread difficult sections	Can express himself/herself fluently and spontaneously	Can express herself/himself with clarity and precision
<b>B2</b>	Idiomatic usage influences the ability to understand	Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms	Can interact with a degree of fluency and spontaneity that makes regular interaction	Can express news and views effectively in writing
<b>B1</b>	Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure	Can read straightforward factual texts on subjects related to his/her field and interest	Can exploit a wide range of simple language to deal with most situations	Can write personal letters and notes asking for a conveying simple information of immediate relevance
<b>A2</b>	Can understand enough provided speech if clearly and slowly articulated	Can understand short, simple texts containing the highest frequency vocabulary	Can communicate in simple and routine tasks requiring a simple and direct exchange of information	Can write short, simple formulaic notes relating to matters in areas of immediate need
<b>A1</b>	Can follow speech which is very slow carefully articulated, with long pauses for her/him to assimilate meaning	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words, and basic phrases	Can interact in a simple way but communication is totally dependent on repetition at a slower rate of speech	Can ask for or pass on personal details in written form

Fonte: Google Drive do grupo de pesquisa em TTD da Unesp – Rio Preto

Os participantes realizam assim uma autoavaliação a partir dos descritores apresentados no quadro. A partir da proficiência autoavaliada é que o participante deve estabelecer seus objetivos de aprendizagem para aquele período de interações. Essa informação não foi coletada dos participantes que não responderam o questionário.

O levantamento dessas informações são relevantes para tornar os dados compartilháveis, de modo que o *corpus* atenda o que preconizam as boas práticas de pesquisa da *Open Science*, e ainda, como afirma Briney (2015, p. 163), de modo que os “dados devem trabalhar para você, não contra você, porque a pesquisa científica por si já é suficientemente desafiadora”<sup>67</sup>, pois dados sem o mínimo de informações quanto aos participantes da pesquisa podem se tornar inutilizáveis. O que implica em disponibilizar a maior quantidade de informações possíveis sobre o contexto de produção dos dados e as intempéries que por ventura possam ter incidido de alguma forma sobre os dados produzidos, assim, ao iniciar o trabalho de análise dos dados o pesquisador tenha condições de compreender a razão pela qual encontra singularidades nesse material. Desse modo, informações quanto a idade dos participantes, sexo e proficiência na língua em estudo se caracterizam como informações relevantes para o usuário dos dados do *corpus*. A organização do *corpus* oportuniza maior dedicação à análise dos dados,

<sup>66</sup> Optei pelo Quadro Comum Europeu como está no questionário inicial, em inglês, apesar de haver versões em Português, por estar discorrendo aqui sobre esse questionário especificamente, com o intuito de fornecer ao leitor a mesma visão que tem o participante de TTD ao receber o documento.

<sup>67</sup> “[...] data should work for you, not against you because scientific research on its own is challenging enough.” (Tradução nossa)

contribuindo para que o pesquisador possa dedicar mais tempo na análise dos dados do que na construção do *corpus*.

Movida por perguntas que foram surgindo durante esse primeiro levantamento, fui procurando as respostas nos dados, de modo que chegasse a um *corpus* de pesquisa que atendesse às recomendações internacionais no que concerne à segurança, confiabilidade e confidencialidade para trabalhos dessa natureza como apresentadas por Chanier e Wigham (2016), Briney (2015), Eynden (2013) e Eynden *et al* (2011). Assim, esse primeiro momento da organização e análise dos dados resultou no estabelecimento dos procedimentos e das etapas do trabalho, que pode ser sintetizado no quadro 3:

**Quadro 3 – Síntese dos questionamentos, procedimentos e fontes**

<b>Questionamentos</b>	<b>Procedimentos adotados</b>	<b>Fontes</b>
<b>1) Que modalidades de TTD temos no Banco TTDii?</b>	<ul style="list-style-type: none"> <li>• Identificar características das turmas.</li> </ul>	i. Listas de pareamento; ii. Tutoriais.
<b>2) A anonimização realizada no Banco TTDii possibilita disponibilização dos dados sem a fácil identificação do participante?</b>	<ul style="list-style-type: none"> <li>• Perscrutar informações que levem a quebra de confidencialidade.</li> </ul>	i. Textos produzidos para a troca entre parceiros; ii. Chats; iii. Diários de aprendizagem; iv. Questionários inicial e final.
<b>3) Há arquivos que não podem ser disponibilizados em um <i>corpus</i>?</b>	<ul style="list-style-type: none"> <li>• Identificar dados com nomeação equivocada ou corrompidos;</li> <li>• Verificar qualidade de vídeo/áudio;</li> <li>• Localizar os TCLEs assinados com autorização de uso de imagem e/ou de dados; solicitar assinatura retroativa dos TCLEs.</li> </ul>	i. Vídeos (SOTs); ii. Textos produzidos para a troca; iii. Chats; iv. Diários de aprendizagem; v. Questionários inicial e final; vi. TCLEs salvos no banco; vii. Participantes contatados via e-mail ou rede social.
<b>4) Qual a quantidade de arquivos na modalidade oral e na modalidade escrita?</b>	<ul style="list-style-type: none"> <li>• Anotar quantidade de arquivos por tarefa, registrando em planilha Excel®, inclusive a</li> </ul>	i. Vídeos (SOTs); ii. Textos produzidos para a troca; iii. Chats;

	duração de cada arquivo de áudio/vídeo.	iv. Diários de aprendizagem; v. Questionários inicial e final.
--	---	---

**Fonte: Elaborado pela autora**

A partir da conclusão dessa etapa tornou-se possível seguir com os procedimentos a fim de constituir o *corpus* de pesquisa, ou seja, iniciar a organização dos arquivos nas pastas conforme o padrão estabelecido para o MulTeC que será apresentado neste capítulo.

#### 2.2.1.1.3 Documentos por turma TTD no Banco TTDii

Segundo Briney (2015, p. 44) a “documentação é uma das mais importantes partes do gerenciamento dos dados porque eles precisam de contexto para serem compreendidos e usados. De certo modo, dados sem documentação não têm significado”<sup>68</sup>. Como o primeiro levantamento indicou que não havia documentos no Banco TTDii que explicassem o contexto de produção dos dados, foi necessário acrescentar a este trabalho inicialmente os documentos contextuais produzidos para cada turma pelas mediadoras a partir de uma coleta realizada nos computadores pessoais da coordenação do projeto TTD em São José do Rio Preto.

Esses documentos foram: i) o Powerpoint ® utilizado para a reunião tutorial da turma; ii) a lista de pareamento – na qual estão registrados os pareamentos realizados antes da primeira sessão oral de teletandem; iii) lista de presença, com os pareamentos, data de ocorrência de cada SOT, quantidade de SOTs, registro de alguma ocorrência relevante para a compreensão do dado. Vale ressaltar que a maioria das listas de presença não contém os registros de ocorrências porque não foram encontrados. Esses documentos também não estavam padronizados. Após a coleta desses dados junto à coordenação do TTD em São José do Rio Preto, padronizei as listas de pareamento e as de presença, haja vista que os Powerpoint® tutoriais estavam padronizados. A padronização facilita a leitura pelos usuários. No período no qual os dados foram produzidos, cada mediadora criava um documento com as informações que entendia serem necessárias para o pareamento, por exemplo, de modo que não encontrei padrão.

A organização dos arquivos está ancorada nos conceitos de cenário de aprendizagem e de pedagógico a partir de Foucher (2010) e Aranha e Leone (2016) e de tarefas discutida por Foucher (2010) e Guichon (2008), levando em consideração que o cenário pedagógico refere-

<sup>68</sup> Minha tradução para “Documentation is one of the most important parts of managing data because data needs context in order to be understood and used. In some sense, data without documentation has no meaning.”

se ao planejamento e o cenário de aprendizagem ao que se realiza no período planejado para a aprendizagem. Assim, cada turma TTD tem seu período de planejamento, no qual os professores coordenadores do TTD, nas duas instituições parceiras, estabelecem o período de interações e a quantidade de SOTs, se haverá textos a serem trocados e que outras tarefas devem ser realizadas pelos participantes.

Durante esse planejamento, os professores trocam e-mails, podem realizar web-conferências, produzem documentos que registram o que decidiram sobre a realização daquela turma TTD especificamente, inclusive estabelecendo o cronograma de tarefas com prazos para envio dos textos a serem revisados pelos parceiros, por exemplo. Esses professores-mediadores estão, nesse período, projetando o cenário pedagógico nos moldes de Foucher (2010).

Quando as turmas começam a realizar o que foi proposto como cenário pedagógico, há documentos produzidos pelas mediadoras que registram o cenário de aprendizagem – ou seja aquele cenário que se materializa no intuito de oportunizar aprendizagem, incluindo as alterações feitas ao que fora planejado de modo que, possíveis dificuldades que emergem pelo caminho não representem obstáculos. Como ‘documentos de turma TTD’ produzidos para a turma nesse período temos no *corpus* as listas de presença, que existiam sem padronização e foram padronizadas para o *corpus*.

Portanto, o que está nomeado neste trabalho como ‘documento de turma TTD’ se refere aqueles que foram produzidos para cada uma das turmas TTD no processo de organização e realização de uma turma. No Banco TTDii, vale lembrar, foram coletados apenas os dados produzidos pelos participantes, não tendo sido coletado o que foi produzido pelas mediadoras. A partir do pressuposto teórico de que o *corpus* deve conter documentos elaborados durante a produção dos dados como afirmam Chanier e Wigham (2016) Briney (2015), e Eynden (2011), tais documentos são relevantes para a organização do MulTeC, portanto, devem ser acessíveis ao pesquisador que optar por utilizar esses dados em sua pesquisa.

O Powerpoint® utilizado durante as reuniões tutoriais são estruturados basicamente da mesma forma: i) apresenta em linhas gerais o que é teletandem; ii) seus princípios básicos; iii) como esses princípios básicos se materializam durante o período de interação, a saber: através da negociação, colaboração, dedicação igualitária de tempo às línguas em estudo, atenção aos aspectos tanto linguístico quanto culturais e comprometimento com a própria aprendizagem e com a aprendizagem do parceiro; iv) cronograma para aquela turma TTD; v) informação sobre a coleta de dados, considerando ser o TTD também um projeto de pesquisa. No cronograma, os participantes são informados quanto às datas e aos horários de realização

de cada SOT e os prazos para o envio de textos, quando houve planejamento de realização dessa tarefa para a turma em questão.

Todos esses pontos são discutidos durante a reunião tutorial, havendo oportunidade para esclarecer possíveis dúvidas quanto à parceria e o contexto TTD de aprendizagem de línguas. No MulTeC, os arquivos em Powerpoint® nos quais estavam as apresentações nomeadas como ‘tutorial’ foram arquivados na pasta de sua respectiva turma TTD, após sua conversão para a formatação .pdf, em um total de 14 arquivos em pdf, pois das 16 turmas TTD, dois tutoriais, cada um preparado para as turmas 2014\_UGA4si e 2015\_UGA3i não foram encontrados nos arquivos pessoais das coordenadoras do TTD em São José do Rio Preto.

A lista de pareamento, como o tutorial, é elaborada antes da SOTi. A produção dessa lista em momento anterior a SOTi ocorre em cenários em que o pareamento foi planejado como sendo prévio, a depender do cenário pedagógico, no qual se planeja as tarefas a serem realizadas. Essa decisão de haver ou não pareamento prévio está sob responsabilidade das mediadoras, que entram em consenso a respeito durante o planejamento da turma. Assim, caso as mediadoras optem por realizar pareamento prévio precisam criar um documento com informações como: nome dos participantes e e-mail, para então informar, via e-mail aos parceiros com quem formarão par no período.

Considerando o contexto de produção dos dados e a necessidade de compreensão desse contexto para a utilização deles em pesquisas, foi necessária a inclusão desse documento no *corpus*. Encontrei diferentes tipos de lista de pareamento que variaram entre ter ou não o pareamento, o período de realização da turma, usuários Skype dos participantes e modalidade de teletandem na qual a turma seria desenvolvida. Nenhuma das listas encontradas nos arquivos da coordenação do TTD de São José do Rio Preto constava todas essas informações. Em algumas das listas de pareamento, não havia a indicação clara da ligação entre os pares interagentes, havia apenas uma lista de nomes de participantes brasileiros e de estadunidenses escritos subsequentemente, assim, era impossível depreender quais seriam os pares, ou como as parcerias das duplas de interagentes se constituíram (por exemplo: ao acaso, por indicação dos docentes e baseada em algum critério etc.).

A padronização demandou coleta de informações em outros documentos, como o tutorial, por exemplo, para completar os dados que são necessários para a compreensão do contexto. Desse modo, conclui que o documento ‘lista de pareamento’ deveria apresentar informações como: o nome das mediadoras, o dia da semana, horário de realização das SOTs, modalidade de TTD, turmas que constituem a turma TTD, nome da turma TTD (a partir da

nomeação proposta no MulTeC, por exemplo 2013\_UGA2i<sup>69</sup>), calendário com os prazos para a realização das tarefas propostas para a turma e pareamento com a utilização dos ITs e cursos dos participantes, razão pela qual os ITs precisavam ser criados antes do início da realização de qualquer outra atividade, na construção do MulTeC de modo que a anonimização fosse realizada na lista de pareamento que originalmente continha os nomes e e-mails dos participantes<sup>70</sup>.

Quanto às listas de presença vale ressaltar que também foram padronizadas, para facilitar sua leitura e preenchimento, e, então, inseridas no MulTeC a partir das informações que conseguimos resgatar no Banco TTDii e nos arquivos pessoais de uma das coordenadoras do TTD na Unesp de São José do Rio Preto.

A fase seguinte do trabalho demandou o estabelecimento dos procedimentos de tratamentos dos dados sobre os quais discorro a seguir.

### *2.2.1.2 Procedimentos para o tratamento dos dados: anonimização, conversão e nomeação*

Os procedimentos que serão detalhados nesta seção foram designados para o tratamento dos dados por objetivarem prepará-los para serem compartilháveis, para o uso em pesquisas, procurando anonimizá-los e viabilizar a leitura em sistemas operacionais diversos, assim, arquivos em formatos independentes, como os arquivos na extensão .txt (texto sem formatação), que facilitem a leitura por softwares de processamento de textos.

Ademais, formatos independentes ainda contribuem para a longevidade dos dados, pois os *softwares* ou plataformas, especialmente os pagos, costumam ser atualizados com certa periodicidade. Ocorre que essas atualizações podem ocasionar a impossibilidade de leitura de arquivos em um formato dependente, como por exemplo ocorre com os pacotes da Microsoft<sup>®71</sup>, e dados em formatos dependentes seriam perdidos. Desse modo, discorro a seguir sobre o padrão de anonimização que estabeleci para o MulTeC e o processo de conversão dos arquivos em formatos independentes.

Nessa etapa, estabeleci os códigos a serem utilizados na nomeação e padronização de nomeação dos arquivos, bem como defini o padrão de cabeçalho a ser utilizado nos arquivos e

---

<sup>69</sup> Segunda turma TTD em parceira com a UGA no ano de 2013.

<sup>70</sup> Em anexo, o leitor pode encontrar uma lista de pareamento padronizada para o MulTeC.

<sup>71</sup> Em abril de 2018 foi noticiado que “versões anteriores ao Office 2007 não poderão editar arquivos de Offices mais recentes”, o que inclui documentos criados no Word®, Powerpoint® e Excel®, por exemplo, conforme disponível em: <https://www.techtudo.com.br/noticias/2018/03/microsoft-vai-encerrar-pacote-de-compatibilidade-do-word-excel-e-powerpoint.ghtml>

o modo de organização desses arquivos nas pastas e subpastas no *corpus*, cada um desses procedimentos está detalhado a seguir.

#### 2.2.1.2.1 Definição do código de identificação do participante

A criação do código, que designei como IT (Identidade Teletandem), se deu no intuito de auxiliar na anonimização dos dados, procurando preservar a identidade do aprendiz e facilitar a construção de uma convenção de nomeação dos arquivos que fosse a mais significativa possível, seguindo as orientações de Chanier e Wigham (2016), Briney (2015), Eynden (2013) e Eynden *et al.* (2011), fornecendo na nomeação informações sobre o tipo de dado e o ano de produção, por exemplo.

Considero que o IT criado para o MulTeC é fruto da análise dos dados e do contexto de produção, bem como a projeção de compartilhamento dos dados. Assim, ele indica ao pesquisador a instituição de ensino superior à qual o aluno está vinculado, o curso da instituição ao qual o participante está vinculado, o sexo<sup>72</sup> e o usuário Skype utilizado, seguindo a orientação de Chanier e Wigham (2016, p. 223) de que “é preferível criar códigos significativos que facilitem a investigação posterior”. O quadro a seguir apresenta o protocolo de criação do IT.

**Quadro 4 – Protocolo de criação do IT**

Item	Código
Instituição	<b>I</b> (Ibilce = Campus Unesp em São José do Rio Preto) <b>U</b> (UGA = University of Georgia)
Curso ao qual está vinculado	Conforme tabela <sup>73</sup> criada a partir do site da Unesp – Rio Preto; Para os estrangeiros não haverá essa identificação considerando não haver documentos no banco que nos forneçam esta informação. O espaço preenchido com um ‘0’
Sexo <sup>74</sup>	<b>M</b> (Masculino)

<sup>72</sup> Essa designação se deu pelo que foi possível perceber nas imagens de vídeo, pois não foi feito levantamento quanto à identidade de gêneros dos participantes, haja vista não ter sido uma preocupação até o momento no TTD.

<sup>73</sup> Tabela anexa (anexo 4).

<sup>74</sup> Sexo identificado no vídeo. Até o início deste trabalho não houve oportunidade para os próprios participantes declararem seu gênero, o que deve ocorrer a partir deste trabalho.

	<b>F (Feminino)</b>
Usuário Skype	Um número diferencia um participante do outro nos diferentes contextos. Na Unesp o usuário é unespriopreto <b>01</b> , unespriopreto <b>02</b> ,... Na UGA: teletandem <b>01</b> , teletandem <b>02</b> ,...

**Fonte: Elaborado pela autora**

Assim, uma aluna do curso de doutorado em Estudos Linguísticos da Unesp – Rio Preto, no campus Unesp nomeado Ibilce<sup>75</sup>, que se inscreveu para interagir no TTD e que utilizou o usuário Skype unespriopreto012 terá o seguinte IT: **I27F12**. **I** por ser aluna do Ibilce, **27** por estar cursando doutorado no programa de estudos linguísticos, **F** por ser do sexo feminino e **12** por usar, em uma turma de TTD, do ano e semestre em questão, o usuário Skype® unespriopreto012. O participante identificado no vídeo como sendo masculino, estudante da UGA, por exemplo, terá seu IT designado como **U0M22**<sup>76</sup>.

O código IT está presente na nomeação de cada um dos dados produzidos pelo participante, mas também das subpastas que contêm esses dados. Considerando que em cada subpasta há dados produzidos por uma parceria, a pasta que os contém está nomeada, tomando, por exemplo, os dados acima, como: **I27F12\_U0M22**. A criação desse IT, facilita tanto a identificação prévia dos participantes que produziram os dados, bem como o uso das caixas de pesquisas para encontrar documentos em computadores pessoais, otimizando o tempo de procura dos arquivos produzidos por um mesmo participante em determinado período, por participante feminino/masculino ou de um determinado curso.

O IT, criado para o *corpus*, se constitui elemento relevante na anonimização dos dados e padronização da nomeação dos arquivos supracitados era necessário estabelecê-lo primeiro antes da definição dos padrões de anonimização que detalho na sequência.

#### 2.2.1.2.2 Anonimização dos dados

A anonimização dos dados é eticamente indispensável para a utilização dos dados em pesquisas. O processo, segundo Chanier e Wigham (2016) significa modificar alguma

<sup>75</sup> Instituto de Biociências, Letras e Ciências Exatas.

<sup>76</sup> O numeral “0” indica que a informação quanto ao curso do participante não pode ser resgatada nos dados já coletados. Essa não era uma informação que as mediadoras brasileiras solicitavam dos parceiros estadunidenses.

informação que possibilite identificar o participante, o que implica em substituir não apenas seu nome, mas também o endereço de sua residência ou telefone, como também afirmam Eynden *et al.* (2011) e Briney (2015). Assim, anonimizar implica em tornar dados sensíveis em não-sensíveis, substituindo informações que possam ajudar na identificação do participante, como pode ocorrer em um ambiente autônomo e telecolaborativo de aprendizagem em que os parceiros em alguns momentos compartilham fatos sobre suas vidas e pessoas próximas a eles. O processo de anonimização dos dados foi realizado nos textos em Word®: diários de aprendizagem, questionários, chats e textos produzidos para a troca entre os parceiros, substituindo nomes próprios, e-mails, números de telefones, endereços ou nomes de animais de estimação de modo que se torne mais complicada a identificação dos participantes. O mesmo padrão de anonimização está norteando as transcrições que estão sendo realizadas por monitoras do projeto FAPESP sob coordenação da Prof.<sup>a</sup> Dr.<sup>a</sup> Solange Aranha.

Conforme Chanier e Wigham (2016), a anonimização precisa ser sistemática. O objetivo da anonimização é a confidencialidade da identidade do participante, o que demanda procedimentos que incluem a substituição de nomes próprios, endereços (inclusive os e-mails) e números de telefone. No contexto telecolaborativo de aprendizagem de línguas que é o TTD, essas informações podem ser repassadas ao parceiro em uma SOT com objetivos diversos, mesmo que não estejam diretamente relacionadas às tarefas previstas para ocorrerem no contexto de uma determinada turma TTD. Desse modo, havia necessidade de anonimização tanto na nomeação dos dados quanto nos arquivos de modo que as informações sigilosas fossem substituídas.

No Banco TTDii houve a padronização da nomeação dos arquivos e das pastas omitindo os nomes próprios dos participantes. No entanto, ao abrir os documentos de texto como chats, diários e textos produzidos para a troca encontrei muita informação sensível, como por exemplo, endereço de e-mail, nome de usuário Skype e nome completo de participante.

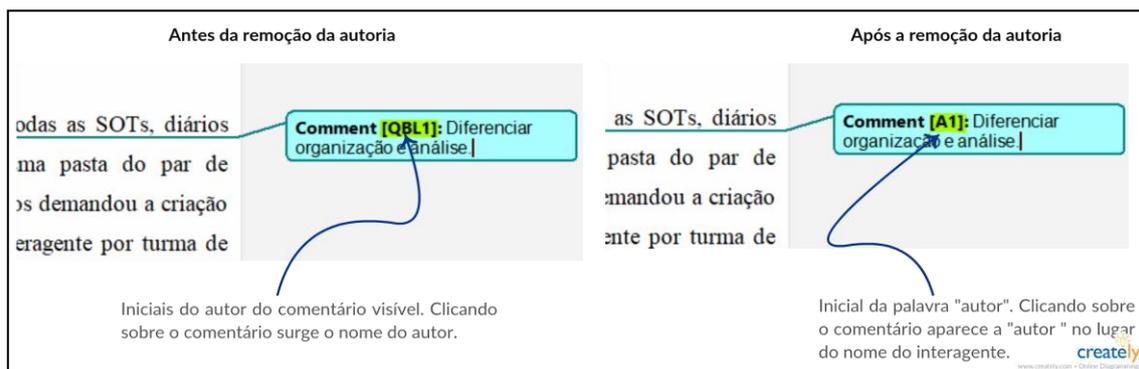
Para a anonimização dos documentos precisei estar atenta aos tipos de informação diferentes, utilizando procedimento de acordo com o tipo de dado a ser omitido, a saber: nomes próprios, endereços e quaisquer referências que pudessem levar à identificação dos participantes; substituição dos nomes nos comentários no textos revisados pelo parceiro, substituição de indicação de turno com Skype pessoal em alguns chats, “assinaturas” nos diários e/ou nos textos produzidos para a troca, alteração da informação de autoria nos arquivos de texto. Nos dados utilizados encontrei tanto nomes próprios como e-mails, links para página de redes sociais dos participantes, nomes de integrantes da família e endereços. Todos esses dados sensíveis foram substituídos utilizando o IT dos participantes e uma sigla referente ao tipo de

informação dada, como por exemplo, no caso de um participante citar o nome da mãe, substitui pelo IT acrescido da letra ‘m’ (inicial de mãe), por exemplo I8F3m. A anonimização foi realizada antes da conversão dos arquivos em documentos de acesso aberto.

Havia arquivos salvos em formato de imagem (.jpeg). Nesses casos a anonimização demandou a impressão do documento, substituição dos indícios de autoria, passando por nova digitalização, estando, a partir de então, anonimizados. Como afirmei a partir dos dados encontrados no Banco TTDii, a anonimização se restringiu a nomeação dos arquivos dos dados.

O cuidado com a anonimização dos dados no MulTeC basicamente se constituiu de substituição de nomes próprios pelos ITs de cada participante e retirada da autoria nos documentos, inclusive nos comentários inseridos nos textos revisados pelo parceiro. Esse processo de anonimização, portanto, foi realizado tanto nos textos produzidos pelos participantes, substituindo nomes próprios conforme o padrão estabelecido, bem como nos comentários inseridos quando da revisão pelo parceiro de TTD, haja vista que ao comentar havia indicação de autoria nos comentários. A figura exemplifica como estava antes da anonimização o comentário feito por um parceiro e como ficou a autoria do comentário após a anonimização.

**Figura 10 - Exemplo de anonimização nos textos revisados**



**Fonte: Elaborado pela autora**

Como mostrado na figura 10 o comentário não foi alterado, apenas foi retirada sua autoria, assim mantendo o dado sem identificar o participante que o produziu. Utilizei a ferramenta “inspecionar documentos” do Word® para apagar os indícios de autoria dos textos. Essa mesma ferramenta retira a autoria dos comentários nos textos revisados pela(o) parceira(o) com resultado demonstrado na figura acima.

Além da autoria dos comentários e dos arquivos, salvos automaticamente pelo software da Microsoft®, ainda há em alguns textos, pequenos recados, acrescentados pelo(a)

parceiro(a) revisor(a) e que, por vezes, está assinado. Nesses casos, a assinatura do participante foi substituída por seu IT.

Houve textos salvos em imagem (.jpeg) pelo participante. Nesses dados, os nomes foram apagados utilizando o aplicativo ‘fotos’ do sistema operacional do computador que utilizei para a pesquisa, substituindo-os pelo código IT inserido na imagem onde havia o nome do(a) parceiro(a).

Nos chats, como nos textos produzidos na língua-alvo para a troca com o parceiro, precisei estar atenta à leitura de cada documento no sentido de identificar nomes dos participantes, e-mails ou links para páginas nas redes sociais substituindo-os pelo IT do participante. No que diz respeito à inserção de nomes de parentes ou pessoas próximas que, por ventura, levassem a identificação dos participantes, optei por uma codificação que inclui o IT do participante e a inicial do substantivo que designa a relação entre eles. Por exemplo, o código I27F12f indica que o participante cujo IT no *corpus* é I27F12 está se referindo ao filho (f) ou filha. Essa estratégia também foi utilizada nos textos. Todas as abreviaturas utilizadas na anonimização estão descritas no documento “Nomeação dos dados”, um dos documentos acrescentados ao MulTeC.

Uma questão que entrevi de alguma forma na anonimização foi a realização do que foi nomeado pela equipe TTD da Unesp – Rio Preto como “interação extra”, cuja ocorrência se deu no período em 8 turmas TTD (7 na modalidade TTDii e 1 na modalidade TTDisi)<sup>77</sup> no período de 2012 a 2015. Nessas turmas, o planejamento, apresentado no cronograma durante o tutorial da turma, indicava a necessidade de uma SOT fora do horário (ou mesmo do dia) estabelecido para a realização das SOTs pela turma, caracterizando o que então era designado como “interação extra”. Assim, em havendo sido planejada a realização de oito SOTs para a turma, uma delas deveria ocorrer em horário e dia a ser definido pelos parceiros. Essa ocorrência se deve ao fato de os professores/mediadores envolvidos na organização da turma terem optado por um período de interação com oito SOTs, mas encontraram como obstáculo a incompatibilidade dos calendários institucionais, por vezes em decorrência de feriados ou atividades extracurriculares nas instituições parceiras. Logo, os participantes precisaram combinar melhor dia e horário para ambos realizarem a SOT extra, podendo fazê-lo no

---

<sup>77</sup> Na modalidade de TTDii (teletandem institucional integrado) as turmas que integram a turma teletandem têm tarefas mandatórias sob as quais incide avaliação pelo professor regente. Na modalidade de TTDisi (teletandem semi-integrado) apenas em uma das turmas que constituem uma turma teletandem há avaliação pelo professor regente, por exemplo, mas os participantes também realizam as tarefas que são obrigatórias. Discuti essas diferenças mais detalhadamente no capítulo teórico da tese.

laboratório ou em suas residências. Os participantes eram responsáveis por realizar as gravações, entregando-as à mediadora da turma TTD.

Esse fato produziu necessidade de anonimização diferenciada nos chats porque, além de buscar nomes próprios e quaisquer outra informação que identificasse os parceiros, nas SOTs realizadas fora do laboratório TTD, os participantes costumavam usar suas contas pessoais de Skype e, assim, as indicações de turno no dado coletado eram idênticas aos usuários Skype pessoais e não se referiam aos usuários institucionais (como “unespriopreto012”, em exemplos supracitados). Em função disso, nesses chats, todas as indicações de troca de turno precisaram ser alteradas. A anonimização, nesses casos, então, significou substituir as trocas de turno indicadas, pelo usuário Skype utilizado pelos parceiros<sup>78</sup> naquela parceira de TTD.

Concluída a anonimização cada arquivo foi salvo em texto sem formatação (.txt), atendendo a necessidade de salvar os arquivos em formato livre, com nomeação padronizada que busca facilitar a compreensão do pesquisador quanto ao contexto de produção dos dados. A conversão dos dados foi feita concomitantemente a anonimização e inserção dos cabeçalhos. A seguir apresento os procedimentos realizados para a conversão dos dados em formatos livres.

#### 2.2.1.2.3 Conversão dos dados

A conversão dos dados em formatos independentes constitui uma atividade relevante principalmente por duas questões: a longevidade e a legibilidade dos dados. Dados salvos em formatos pagos, como o Word®, impossibilitam a leitura por computadores com sistema operacional diverso da Microsoft®, por exemplo, além de não poderem ser lidos pelo mesmo sistema operacional muitos anos depois, pois tais sistemas costumam ser atualizados. Ao salvar em formatos independentes, se possibilita que o uso dos dados não esteja submetido à atualização de determinados softwares, o que amplia a longevidade dos dados que constituem o *corpus*.

De modo que possam ser legíveis por outros sistemas, os arquivos em textos, criados originalmente em Word® (.doc; .docx) foram convertidos em texto sem formatação (.txt). Alguns dos dados produzidos pelos participantes do TTD possuem no uso das cores um instrumento de marcação entre os participantes, como no caso dos textos revisados pelo parceiro. Há, por exemplo, o uso de vermelho para indicar uso equivocado de palavra e amarelo para indicar palavra que pode substituir uma inadequada ao contexto ou gênero, bem como há

---

<sup>78</sup> Os arquivos em vídeo não foram anonimizados por questão de tempo e de recursos.

textos com inserção de imagens. Os documentos que incluíam imagens ou destaques com uso de cores diversas foram salvos, também na extensão pdf, de modo que um texto produzido para troca que contém imagens, por exemplo, está salvo em duas versões: uma em txt e outra em pdf.

O formato em pdf permite a visualização da imagem e dos destaques utilizando cores, enquanto que o txt permite o processamento do texto por softwares diversos, auxiliando o trabalho de análise do pesquisador. Os arquivos<sup>79</sup> em vídeo estavam salvos em mp4, não havendo, portanto, necessidade de conversão.

Os dados salvos como imagem foram mantidos no formato salvo no banco por entendermos que, caso fossem digitados, haveria perda de informações relevantes ao pesquisador como as escolhas de cores indicativas de tipologia de equívocos, por exemplo. Cada um desses arquivos foi nomeado seguindo o padrão que estabeleci, como discuto na próxima subseção.

#### 2.2.1.2.4 A padronização da nomeação dos arquivos

A padronização, criada para a nomeação dos arquivos, procurou seguir a orientação de Chanier e Wigham (2016), assim como os princípios listados para nomeação citados por Briney (2015): “1) nomes devem ser descritivos; 2) nomes devem ser consistentes; 3) nomes deveriam ser curtos, preferivelmente com menos de 25 caracteres; 4) usar traços ou *underlines* no lugar de espaçamentos; e 5) evitar caracteres especiais.”<sup>80</sup> (BRINEY, 2015, p. 75, tradução nossa).

No intuito de evitar espaçamentos, Briney (2015) sugere o uso de dois tipos de convenções: o *camel case* e o *pothole case*, o uso de um não excluindo necessariamente o uso do outro. Ambas podem ser utilizadas para evitar os espaçamentos que podem dificultar o processamento dos dados mesmo em computadores mais modernos (BRINEY, 2015, p.77). No caso do *camel case* usa-se a alternância de letras maiúsculas e minúsculas para facilitar a compreensão da nomeação, convenção bastante utilizada em *hashtags* na rede mundial de computadores, por exemplo: #InformacoesGerais. A convenção *pothole case*, por outro lado,

---

<sup>79</sup> A transcrição desses arquivos está em andamento.

<sup>80</sup> 1- Names should be descriptive  
 2- Names should be consistent  
 3- Names should be short, preferably less than 25 characters  
 4- Use underscores or dashes instead of spaces  
 5- Avoid special characters [...]

utiliza de traços para tornar mais compreensível a nomeação, tomando como exemplo a *hashtag* acima utilizando *pothole* ficaria: *informacoes\_gerais*.

A padronização que estabeleci para o MulTeC teve como prioridade ser a mais descritiva possível. Essa característica foi citada por Briney (2015) como um dos princípios gerais a serem seguidos quando definimos o sistema de nomeação dos dados, bem como a consistência da nomeação, a opção por nomes não muito longos e sem o uso de caracteres especiais ou espaços. Esse último princípio, no que se refere ao MulTeC que tem nomeação em português, implica na substituição de “ç” por “c”, por exemplo, bem como a omissão de acentos em palavras como “produção” (no MulTeC – producao). O padrão de nomeação utilizado no MulTeC está descrito em documento salvo em pdf, para mantermos a apresentação das tabelas, em pasta anterior as que contém os dados produzidos pelas turmas TTD. O leitor poderá encontrar esse documento com a descrição da nomeação no anexo 6 desta tese.

No MulTeC mesclamos as convenções citadas por Briney (2015) *pothole case* (com o uso de *underline* entre palavras) e *camel case* (uso de iniciais maiúsculas), que apresentamos no capítulo metodológico, priorizando a constituição de uma nomeação mais significativa possível considerando o contexto TTD.

Nesse sentido, a nomeação dos dados no MulTeC inclui o IT, o ano de produção dos dados, a parceria institucional, a modalidade, a tarefa e sua ordem de ocorrência. O quadro 4 apresenta a padronização da nomenclatura dos arquivos criada para a organização proposta neste trabalho.

**Quadro 5 - Padronização da nomenclatura dos arquivos**

<b>Item</b>	<b>Código (exemplos)</b>
Ano	<b>2015</b>
IT	<b>I27F12 ou U0M22</b>
Parceria	<b>UGA1</b>
Modalidade	<b>i</b> = Integrada <b>si</b> = Semi-integrada
Tarefa	<b>SOT</b> = Sessão Oral de Teletandem ( <b>SOTi</b> = Sessão Oral de Teletandem inicial; <b>SOTin</b> = Sessão Oral de Teletandem intermediária; <b>SOTf</b> = Sessão Oral de Teletandem final) <b>D</b> = Diário de aprendizagem <b>C</b> = Documento com os registros do chat durante as SOTs. <b>TO</b> = Texto original (TOI – texto original em Inglês; TOP – Texto Original em Português) <b>TRev</b> = Texto revisado (TRevP – Texto revisado em Português; TRevI- Texto revisado em Inglês) <b>TRe</b> = Texto re-escrito (TReP – Texto re-escrito em Português;

	<p>TReI – texto re-escrito em Inglês)</p> <p><b>TCol</b> = Texto escrito colaborativamente</p> <p><b>QF</b> = Questionário Final</p> <p><b>QI</b> = Questionário Inicial</p>
Ordem da tarefa	<b>1, 2, 3</b>

**Fonte: Elaborado pela autora**

Assim, a Sessão Oral de Teletandem intermediária 1<sup>81</sup> do usuário com IT I27F12, realizada em 2015 em uma turma de TTD integrado, parceria com a UGA, primeiro semestre fica nomeada como: 2015\_I27F12\_UGA1i\_SOTin1.

Um desafio encontrado para a nomeação dos arquivos em algumas turmas TTD do Banco TTDii foi a ordem de ocorrência das SOTs, pois ela implica em nomeação diversa (SOTi–primeira, SOTin1– segunda, SOTin2–terceira, SOTf –última). Em uma turma com a previsão de uma interação extra a ocorrer quando os parceiros decidissem, a SOT extra pode ter sido a SOTin3 para um par e para outro par a SOTin5 ou ainda a SOTf. A indicação das datas de criação dos arquivos, pela busca da data de criação dos arquivos, bem como a audição dos 7 minutos iniciais dessas SOTs foram procedimentos necessários para a nomeação dessas SOTs.

No caso em que há substitutos<sup>82</sup> nas SOTs, indicamos essa ocorrência com um “sf” ou “sm” (a depender do sexo do substituto – ‘m’ para masculino e ‘f’ para feminino) no final do nome do arquivo, por exemplo: 2015\_I27F12\_UGA1i\_SOTin1\_s. O mesmo padrão de nomeação para as pastas, subpastas e demais arquivos que constituem o MulTeC. O quadro 5 a seguir demonstra a nomeação dos arquivos de cada tipo de dado produzidos por I8F7 na turma 2015\_UGA3i.

**Quadro 6 – Exemplo de nomeação de dados do participante I8F7 (UGA3i-2015)**

TIPO DE DADO	ORDEM DE OCORRÊNCIA	NOMEAÇÃO
SOT	1 <sup>a</sup>	2015_I8F7_UGA3i_SOTi
Chat	1 <sup>a</sup>	2015_I8F7_UGA3i_C1
Diário	1 <sup>a</sup>	2015_I8F7_UGA3i_D1
Questionário Inicial	-	2015_I8F7_UGA3i_QI
Questionário Final	-	2015_I8F7_UGA3i_QF
Texto revisado em Inglês	1 <sup>a</sup>	2015_I8F7_UGA3i_TRevP1

**Fonte: Elaborado pela autora**

<sup>81</sup> Sessão oral que sucede imediatamente a SOTi.

<sup>82</sup> Quando um dos parceiros não pode comparecer por algum motivo, as mediadoras precisam encontrar quem o substitua naquela SOT para que o parceiro não fique sem a realização daquela SOT.

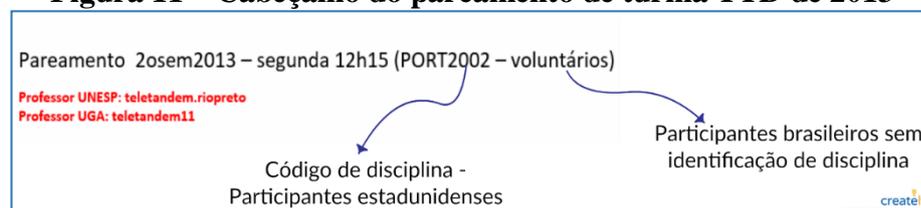
Considero a nomeação significativa por indicar tipo de dado, sexo e curso do participante, modalidade de teletandem na qual a turma se desenvolveu e a ordem de ocorrência do dado no fluxo do TTD, assim como o ano de realização da turma. Os questionários como são apenas 1 de cada tipo, não recebem sequência numérica.

### 2.2.1.3 Organização dos arquivos no MulTeC: etapa 3 da organização do MulTeC

Os procedimentos de organização dos dados no MulTeC foram: seleção dos dados de modalidades de teletandem diversas, separação dos dados produzidos por par de participantes e não por tipo de dado, como foi feito no Banco TTDii, criação de pastas por modalidades, ano, turma e parceiros. Essa organização está ancorada principalmente nas características do contexto TTD e nos conceitos de cenários de aprendizagem e pedagógico bem como a concepção de tarefas que discuti no capítulo teórico. Esses conceitos embasam a ordem das pastas e subpastas. Cada turma está em uma pasta nas quais estão contidas as pastas com os dados produzidos pelos participantes pareados no período naquela turma específica. A quantidade de pastas em cada par varia a depender dos dados salvos/produzidos. Assim, um par que produziu dados de todas as tarefas (chats, SOTs, questionários, diários de aprendizagem e textos produzidos para a troca) terá 5 subpastas, enquanto que o par que não utilizou a ferramenta chat durante as SOTs não terá a pasta nomeada ‘chat’ como subpasta de sua pasta da parceria, com um total de 4 subpastas.

O Banco TTDii surgiu em um contexto de emergência de uma nova modalidade de TTD na Unesp de São José do Rio Preto, em uma tentativa de “organizar os dados para que os atuais e futuros pesquisadores interessados nos diferentes aspectos do contexto pudessem ter acesso” (ARANHA; LUVIZARI-MURAD; MORENO, 2015, p. 284). A identificação de modalidade diversa da TTDii ocorreu pela leitura do cabeçalho das listas de pareamento. A figura 11 apresenta um desses cabeçalhos, destacando a indicação de turma TTD de semi-integrado.

**Figura 11 – Cabeçalho do pareamento de turma TTD de 2013**



Fonte: Google Drive® do grupo de pesquisas TTD da Unesp – Rio Preto

Embora não haja a identificação das turmas como sendo de brasileiros ou de estadunidenses no documento salvo, a disciplina com o código PORT2002 não integra o currículo dos cursos da UNESP- Rio Preto, assim foi possível concluir que a turma em questão foi constituída por estadunidenses matriculados na disciplina de Português como LE, enquanto que os estudantes brasileiros eram voluntários no TTD, matriculados em cursos diversos e participando do TTD no contra turno de suas atividades curriculares. Essa diferença representou uma alteração na organização dos dados na construção do MulTeC.

Nas turmas de semi-integrado, na lista de pareamento os alunos estadunidenses são matriculados na disciplina de Português como LE, mas não havia informação sobre os alunos brasileiros<sup>83</sup>, havia apenas a informação de que eram voluntários, isso porque o grupo que organizou o Banco TTDii não objetivou organizar um *corpus*. De acordo com Chanier e Wigham (CHANIER; WIGHAM, 2016, p. 220), na organização do *corpus* “Informações sociolinguísticas sobre os participantes são detalhadas<sup>84</sup>”. Essa afirmação apontou para a necessidade de levantamento de tais informações, o que preferencialmente deveria ter sido realizado quando do início da coleta de dados, na primeira fase do ciclo de vida dos dados, no planejamento do gerenciamento dos dados.

Nesse cenário, com a coleta tendo sido realizada anteriormente à realização desta pesquisa, as informações sociolinguísticas que poderíamos obter minimamente seriam: L1 e L2, idade, sexo, curso e proficiência. Cada uma dessas informações foi extraída dos TCLEs, dos minutos iniciais das SOTs, das listas de pareamento e dos questionários iniciais.

Além da identificação e organização de dados de semi-integrado e integrado em pastas diferentes, ainda se tornou necessário o acréscimo de documentos produzidos para a turma, como discuti neste trabalho, assim, foram inseridos os documentos produzidos para as turmas TTD, como apresento a seguir.

#### 2.2.1.3.2 Documentos com códigos utilizados na constituição do *corpus*

Foram estabelecidos padrões de organização, nomeação e anonimização, logo, se faz necessário para o *corpus*, que haja documentos que expliquem ao pesquisador o que cada código significa. De acordo com Eynden *et al.* (2015), a documentação dos dados implica em

---

<sup>83</sup> Não foi possível identificar os cursos de graduação aos quais os estadunidenses estavam vinculados, pois não há registro dessa informação no Banco TTDii.

<sup>84</sup> Minha tradução para “Sociolinguistic information about the participants is detailed”.

explicar ou definir os códigos, bem como as siglas e nomenclatura adotados na organização do *corpus*. Elaborei os seguintes documentos:

- Contexto TTD<sup>85</sup> – no qual o pesquisador tem uma breve explicação do contexto TTD e das tarefas que podem ocorrer a depender do que planejam os professores que optam por realizar o TTD em suas instituições. Esse documento possui ainda um exemplo de questionário inicial e final. Nesse documento há figuras elaboradas no intuito de auxiliar o leitor a compreender o contexto TTD, havendo links para auxiliar o leitor a encontrar textos explicativos sobre cada tarefa;
- Cabeçalho – nesse documento, há um exemplo da constituição do cabeçalho adotado para todos os arquivos de textos do MulTeC;
- Códigos Informações Interagentes – documento que contém a explicação dos códigos da planilha com as informações que pude recuperar de todos os participantes no que se refere a idade, sexo, línguas em estudo e língua na qual é proficiente;
- Códigos Planilha Levantamento de Dados – nesse arquivo, o pesquisador terá acesso à explicação do significado de cada código adotado na planilha “Levantamento de Dados”;
- MulTeC – documento que contém os créditos, apresentando os nomes das pessoas que auxiliaram na organização, o nome de quem criou os padrões de organização, nomenclatura e anonimização, bem como o responsável pelo MulTeC e o contato para acesso ao *corpus*;
- Nomeação dos Dados – é o documento que explica ao usuário do *corpus* como foi designada a nomeação dos dados e a explicação de cada código utilizado na nomeação;
- TCLE (Unesp) e *Consent Form* (UGA) – são documentos de consentimento das duas instituições;
- Anonimização dos Dados – nesse documento, há a explicação dos códigos utilizados para a anonimização dos dados.

A inserção desses documentos no MulTeC se constituiu em uma das últimas ações do plano de gerenciamento de dados estabelecido por mim para a organização prevista. Descrevo esses documentos a seguir.

---

<sup>85</sup> Anexo 5 desta tese.

### 2.2.2 Documentos contextuais

De modo que os dados sejam reutilizados por pesquisadores, há de se considerar o que Chanier e Wigham (2016), Briney (2015) Eynden (2013) e Eynden *et al.* (2011) declaram sobre a necessidade de documentação dos dados, ou seja, a elaboração de documentos que visem auxiliar o pesquisador na compreensão do contexto de produção dos dados disponíveis para pesquisa. Nessa categoria, produzimos documentos para: a) descrever o fluxo de realização de uma turma de TTD - explanando seus princípios resumidamente; b) explicar a convenção de anonimização e nomeação dos dados; e c) explicar os códigos utilizados nas planilhas elaboradas para o MulTeC, que serão discutidas a seguir.

Os documentos acrescidos serão explanados a seguir, de modo que o leitor compreenda a necessidade deles para o uso do *corpus* por um pesquisador que não integre o grupo de pesquisa TTD de São José do Rio Preto, pois que nos demais campi Unesp podem haver variações nos contextos.

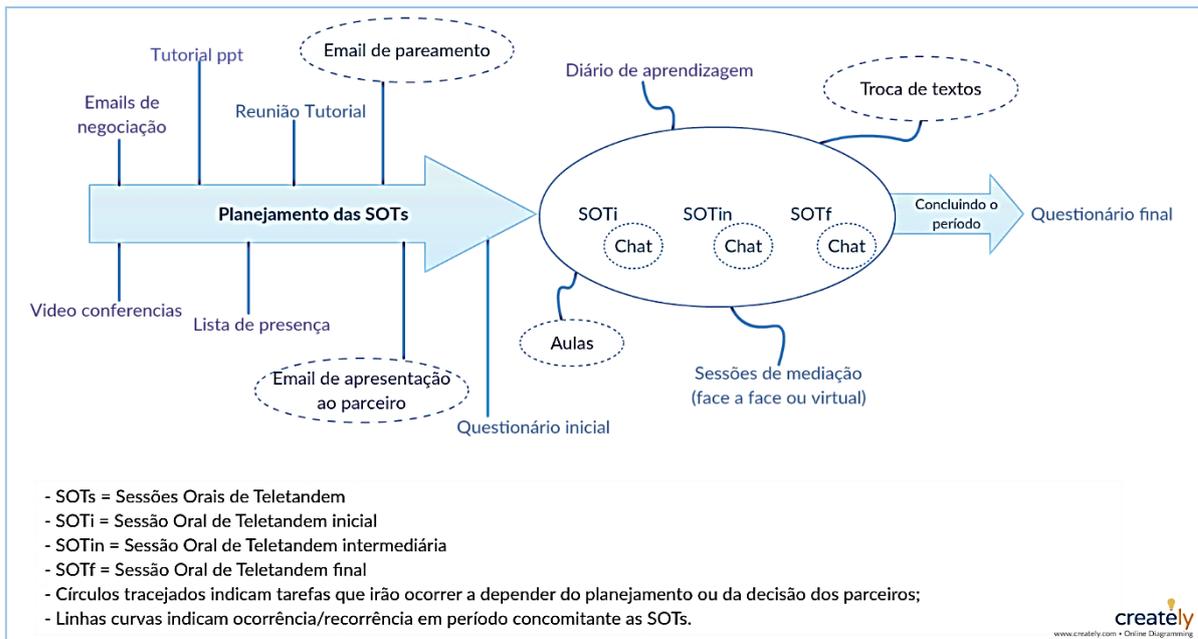
#### 2.2.2.1 Contexto TTD

O documento nomeado ‘contexto TTD’ está inserido no MulTeC com o intuito de descrever o contexto de aprendizagem TTD e suas diversas tarefas, assim como resumidamente os princípios básicos a serem seguidos pelos participantes. Nesse documento<sup>86</sup>, criei links na figura para o texto descritivo de cada tarefa, de modo que o pesquisador interessado em pesquisar apenas uma das tarefas possa, ao clicar na tarefa, ser levado diretamente ao texto que a descreve. No sentido de ampliar a compreensão de como ocorre o TTD na prática elaborei para esse documento uma figura na qual procuro representar de maneira simples a cronologia de uma turma TTD. Afirmando que seja simplesmente pelo fato de que, dada a complexidade do que ocorre em cada cenário de aprendizagem do TTD, uma única figura não teria como abarcar as possibilidades de ocorrência nas modalidades de TTD que têm sido desenvolvidas na Unesp de São José do Rio Preto. Logo, trata-se de uma tentativa de representar a ordem de ocorrência/produção de cada dado que compõe o MulTeC.

---

<sup>86</sup> Ao qual o leitor terá acesso no Anexo 5.

**Figura 12 - Cronologia no TTD**



**Fonte: Elaborado pela autora para o MulTeC**

É necessário informar que, em alguns cenários, por exemplo, não há troca de textos, enquanto em outro pode não haver e-mail de pareamento, ou seja, não haverá pareamento prévio, mas quando todas as tarefas possíveis até o momento no contexto TTD foram previstas no planejamento de uma turma TTD, a cronologia de ocorrência de cada uma das tarefas é essa que está representada na figura 12.

#### 2.2.2.2 Documentos com Códigos utilizados na nomeação e anonimização

Na categoria de documentos cujo objetivo é explicar as convenções estabelecidas estão dois documentos nomeados como 'anonimização' e 'nomeação' dos dados. Neles, o pesquisador encontra as normas de convenção de nomeação, apresentadas no capítulo metodológico desta tese, bem como a de nomeação dos dados.

O documento que trata da anonimização apresenta, inclusive, uma pequena lista das abreviaturas utilizadas na anonimização dos dados, utilizadas quando um dos participantes cita, por exemplo, o nome de um amigo próximo. Esse documento ainda explica a importância da criação do IT como fonte de informação sobre o participante com o devido cuidado de preservar a identidade dele.

**Figura 13 – Documento Anonimização dos Dados**

anonimização no MulTeC, os códigos listados no quadro abaixo:

<b>CÓDIGO IT</b> (no lugar do nome do interagente, no exemplo um estudante UGA)	⇒U0F2
<b>CÓDIGO IT_sf</b> (indicação de que houve uma substituição na SOT em questão. 'f' indica substituto feminino e 'm' masculino)	⇒U0F2_sf
<b>CÓDIGO ITn</b> (indicação de namorado citado no dado coletado)	⇒U0F2n
<b>CÓDIGO ITa</b> (indicação de amigo(a) citado(a) no dado coletado)	⇒U0F2a

**Fonte: Elaborado pela autora para o MulTeC**

Na figura acima podemos visualizar uma tabela acrescida ao texto explicativo do documento Anonimização dos Dados que tem como objetivo exemplificar o código utilizado para anonimização, a saber o código IT do participante e a primeira letra do nome a quem ele se refere durante a conversa ou no texto escrito para a troca. Há uma lista nesse mesmo documento com as substituições realizadas.

Esse procedimento foi realizado nos chats, nos diários de aprendizagem, nos textos produzidos para a troca, bem como nos questionários inicial ou final, pois, ao analisar os dados, observei que mesmo nos questionários alguns participantes citam nomes próprios que poderiam levar a identificação do autor dos dados.

### 2.2.2.3 Códigos das planilhas

No MulTeC, foi elaborado um documento com a explicação dos códigos para cada uma das planilhas criadas, considerando que cada uma delas apresenta códigos diversos, assim há um documento que explica a convenção de códigos utilizados na planilha que apresenta o levantamento de dados por turma, por exemplo. Nessa subseção, apresento os documentos explicativos de códigos utilizados nessas pastas de trabalho. Dada a especificidade de cada planilha, elaborei um documento para cada, totalizando dois documentos: um referente àquela que apresenta o levantamento de dados e outro com os códigos utilizados naquela que reúne as informações sobre os interagentes.

Esses documentos foram nomeados com referência às planilhas, por exemplo, o documento que explica os códigos utilizados da planilha LevantamentoDeDados<sup>87</sup>, foi nomeado Codigos\_LevantamentoDeDados. Na figura 14 apresento o documento que explica os códigos utilizados com o levantamento de dados por turma.

**Figura 14 - Documento Código de Levantamento de Dados por turma.**

**1** – Dado consta no banco;  
**x** – Dado não consta no banco;  
**SOTs e Chats** – São produzidos pelo par logo a célula está mesclada unindo as linhas do interagente e seu respectivo parceiro;  
**SOTs** – Nesta célula está registrada a duração de cada uma das SOTs, no seguinte formato 00:00:00 = (00-hora:00-minutos:00-segundos);  
**Totais** – Na horizontal indicam soma dos dados produzidos pelo interagente ou pelo par (no caso das SOTs e dos Chats) na vertical indicam a soma do tipo de dados produzidos pela turma no período;  
**QI** – Questionário Inicial  
**QF** – Questionário Final  
**SOT** – Sessão Oral de Teletandem

**Fonte: Elaborado pela autora para o MulTeC**

Ao ter acesso a esse documento, o pesquisador terá conhecimento das convenções adotadas para a elaboração das planilhas, facilitando a leitura das informações contidas no documento. Além dos códigos criados, há a informação do formato de preenchimento das horas, minutos e segundos, relevante para a leitura do tempo de duração de cada SOT e do total das SOTs apresentado.

<sup>87</sup> Na nomeação acentos e espaços não aparecem para facilitar a leitura por qualquer sistema.

**Figura 15 - Código Informações Interagentes**

Códigos da planilha 'Informações Interagentes'
<b>X</b> – Indica informação não encontrada;
<b>Gênero ident.</b> – Gênero identificado. Não houve levantamento sobre o gênero com o qual cada um dos interagentes se identifica. Como os dados já estavam coletados a identificação se deu pela imagem ou nome completo na lista de pareamento.
<b>L1</b> –Língua materna
<b>L2</b> – Segunda língua
<b>L3</b> – Terceira língua
<b>Obs.:</b> Na planilha “ <b>Informações interagentes_Geral</b> ” os códigos são iguais. Há apenas o acréscimo de uma coluna: Interagente recorrente. Nesta coluna o interagente que participou mais de uma vez neste período recebe um código formado por um número e a letra:
<ul style="list-style-type: none"> <li>• I (para interagente estudante da Unesp em Rio Preto)</li> <li>• U (para interagentes da UGA)</li> </ul>

**Fonte: Elaborado pela autora para o MulTeC**

O arquivo com os códigos utilizados no documento que contém as informações sobre os interagentes é o menor dos produzidos para o MulTeC, porquanto que a coleta fora realizada anteriormente e dados adicionais não puderam ser inseridos como o tempo de estudo na língua-alvo, por exemplo. A elaboração desses documentos se deu em função da necessidade de explicação dos códigos utilizados em cada planilha. Detalho as planilhas elaboradas para o MulTeC a seguir.

### 2.2.3 PLANILHAS

A inserção de planilhas no MulTeC se deu no intuito de também documentar o *corpus*, apresentando ao pesquisador um registro do levantamento feito durante o procedimento exploratório dos documentos salvos no Banco TTDii e nos documentos salvos nos arquivos pessoais de uma das coordenadoras do TTD em São José do Rio Preto. A decisão por elaborar tais planilhas foi inspirada na afirmação de Sinclair (2004) de que “o design e composição de um *corpus* deveria ser completamente documentada com informações sobre os conteúdos e

argumentos, justificando as decisões tomadas”<sup>88</sup> (p. 8), assim com Guichon (2017) que discorre sobre documentos complementares ao *corpus*. Esses pressupostos me levaram a acrescentar além dos documentos contextuais, discutidos acima, documentos com levantamento sociolinguístico dos participantes e dos dados produzidos nas 16 turmas TTD que constituem o MulTeC, cujas informações estão no documento ‘Informações Participantes’ que explico a seguir.

#### 2.2.3.1 Planilha ‘Informações participantes’

A planilha “Informações Participantes” apresenta os seguintes dados sobre os participantes: IT, usuário Skype, autorização para uso dos dados, gênero, curso, idade, autoavaliação de proficiência, L1, L2 e L3. Elaborei dois tipos de planilhas com essas informações: uma geral com todos os participantes de todas as turmas do MulTeC e uma com as informações por turma, inserida na pasta de cada turma, conforme organização apresentada anteriormente. A planilha de informações geral contém uma informação adicional que fornece ao pesquisador uma possibilidade de identificar quem foram os participantes que participaram do TTD mais de uma vez no decorrer dos quatro anos em que houve a coleta. De modo que essa identificação fosse possível, acrescentei uma coluna na planilha nomeada ‘participantes recorrentes’ na qual os identifiquei com código: I1, I2, U1, U2 e assim sucessivamente, I – indicando participante matriculado no Ibilce (Unesp de São José do Rio Preto) e U – indicando participante matriculado na UGA. Registrei na planilha como I19 o mesmo participante, embora tenha recebido IT diferente em cada turma. Utilizando a ferramenta ‘localizar’ do Excel® o pesquisador tem acesso à quantidade de vezes que esse participante realizou TTD, podendo descobrir em que turma e ano ao selecionar as células em que aparecem o código I19.

---

<sup>88</sup> Minha tradução para “The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken”.

Figura 16 – Indicação de participante recorrente

55	2	UGA 3 i	ISM1	unespriopreto01	DI	M	Tradução	x	x	Port	Ing	x	I27
56	0		U0M29	teletandem29	DI	M	x	x	x	x	x	x	
57	1		ISF6	unespriopreto06	D	F	Tradução	21	C1	Port	Ing	x	I14
58	2		U0F16	teletandem16							x	x	U2
59	-		ISM7	unespriopreto07							ng	x	I26
60	UGA		U0M17	teletandem17							x	x	U1
61	-		ISF12	unespriopreto012							ng	x	
62	1		U0F22	teletandem22							x	x	
63	sem		ISM13	unespriopreto013							ort	I	I25
64			U0F23	teletandem23							x	x	
65			ISM14	unespriopreto014							ng	x	
66			U0M24	teletandem24							x	x	
67			ISM16	unespriopreto016							ng	x	I15
68			U0M26	teletandem26							x	x	
69		ISF17	unespriopreto017							ng	x		
70		U0M27	teletandem27							x	x		
71		ISF18	unespriopreto018							ng	x	I19	
72		U0M28	teletandem28							x	x	U8	
73	2	UGA 1 i	ISM2	unespriopreto02						les	x		
74	0		U0F12	teletandem12							ort	x	
75	1		ISM3	unespriopreto03							les	x	
76	3		U0F3	teletandem3							ort	x	U3
77	-		ISF4	unespriopreto04							les	x	
78	UGA		U0F14	teletandem14							ort	x	
79	-		ISM5	unespriopreto05							les	x	I1
80	1		U0F15	teletandem15							ort	x	
81	sem		ISF6	unespriopreto06							les	x	
82			U0F16	teletandem16	DI	X	X	DI	x	Inglês	Port	x	
83			ISF7	unespriopreto07	D	26/11/1993	19	D	x	Port	Inglês	x	
84			U0M17	teletandem17	DI	X	X	DI	x	Inglês	Port	x	
85			I9M11	unespriopreto013	D	07/12/1993	19	DI	B2	Port	Inglês	x	
86			U0M21	teletandem23	DI	X	X	DI	x	Inglês	Port	x	
87		ISF13	unespriopreto014	DI	21/06/1994	18	D	B2	Port	Inglês	x		

Localizar e substituir

Localizar:

Opções >>

Localizar tudo   Localizar próxima   Fechar

Pasta	Planilha	Nome	Célula	Valor	F...
Informacoes Interagentes_Geral.xlsx	Interagentes_Geral		SMS71	119	
Informacoes Interagentes_Geral.xlsx	Interagentes_Geral		SMS149	119	
Informacoes Interagentes_Geral.xlsx	Interagentes_Geral		SMS309	119	
Informacoes Interagentes_Geral.xlsx	Interagentes_Geral		SMS321	119	

4 célula(s) encontrada(s)

Fonte: Elaborado pela autora para o MulteC

No caso apresentado na figura 16, o participante com o código I19 participou, entre 2012 e 2015, em 4 turmas TTD. A partir da ferramenta ‘localizar’ do sistema operacional de sua máquina, o pesquisador pode selecionar a célula indicada e identificar em que turma/modalidade e semestre esse participante teve oportunidade de aprender via contexto TTD. Essa coluna está presente apenas na planilha ‘Informações Participantes Geral’.

As informações apresentadas no documento foram coletadas no TCLE, questionário inicial e SOTi, como discuti anteriormente neste trabalho. Além dessa planilha, elaborei como documento complementar a planilha ‘Levantamento de Dados’.

#### 2.2.3.2 Planilha ‘Levantamento de Dados’

No tocante ao levantamento de dados, quantificando a produção coletada como descrito por Aranha, Luvizari-Murad e Moreno (2015), a elaboração de uma planilha demandou considerar o contexto de produção dos dados: o teletandem como espaço de aprendizagem de línguas telecolaborativo. Assim, havia necessidade de organizar esse levantamento por turmas, por par e individualmente. Levando em consideração que cada turma constitui-se um cenário de aprendizagem, conceito que também embasou o trabalho de Aranha e Leone (2016) e Foucher (2010), assim como concluiu Rampazzo (2017) e cada cenário tem suas características específicas, cada turma deve ter um documento no qual o pesquisador pode ver sinoticamente o quantitativo de dados produzidos no cenário. No entanto, objetivando apresentar os dados gerais do *corpus*, criei, a partir das informações das planilhas das turmas, uma planilha de levantamento geral com o cômputo dos dados produzidos por turma e modalidade, mas apresentando um número geral das produções.

A planilha ‘Levantamento dos Dados’ por turma recebeu uma nomeação que indicasse a turma, o ano, a modalidade de TTD, por exemplo, 2012\_UGA1i\_LevantamentoDeDados. Essa planilha está constituída de duas abas: uma com o levantamento da quantidade de dados produzidos e a outra contendo a quantidade de palavras produzidas por dado.

##### 2.2.3.2.1 Levantamento de dados – aba 1

A aba contendo informações no que se refere à quantidade de dados produzidos está organizada em colunas com indicação de dado salvo ou não, conforme podemos perceber na figura 17.

Figura 17 – Aba ‘Dados produzidos por par’ – Planilha “Levantamento de Dados”

Turma	IT	Usuário Skype	TCLEs	QI	SOTs							Totais por par	Diários							Totais por aprendiz
					SOTi 6/mar	SOTin1 20/mar	SOTin2 27/mar	SOTin3 3/abr	SOTin4 10/abr	SOTin5 21/abr	SOTf 28/abr		1º	2º	3º	4º	5º	6º	7º	
2013-UGA1i	I8F4	unespriopreto04	DI	1									1	1	1	1	1	1	1	7
	UOF14	teletandem14	D	x	0:50:42	00:38:53	00:48:42	00:44:47	00:46:49	00:41:08	00:31:13	5:02:14	x	x	x	x	x	x	x	0
	I8F5	unespriopreto05	DI	1									1	1	1	1	1	1	1	7
	UOF15	teletandem15	DI	x	00:52:13	00:49:00	00:47:28	00:41:42	00:46:39	00:42:02	00:29:04	5:08:08	x	x	x	x	x	x	x	0
	I8F6	unespriopreto06	D	1									1	1	1	1	1	1	1	7
	UOF16	teletandem16	DI	x	00:50:00	00:49:28	00:49:05	00:46:30	00:49:08	00:45:27	00:30:27	5:20:05	x	x	x	x	x	x	x	0
	I8F7	unespriopreto07	D	1									1	1	1	1	1	1	1	7
	UOF17	teletandem17	DI	x	00:32:42	00:38:06	00:48:08	00:39:03	00:38:08	00:20:40		3:36:47	x	x	x	x	x	x	x	0
	I8M11	unespriopreto011	D	1									1	1	x	1	1	1	1	6
	UOM21	teletandem21	DI	x	00:49:40	00:46:52	00:46:30	00:46:27	00:47:49	00:31:27		4:28:45	x	x	x	x	x	x	x	0
	I8F13	unespriopreto013	DI	1									1	1	1	1	x	1	x	5
	UOM23	teletandem23	DI	x	00:52:51	00:49:49	00:48:55	00:24:49		x	00:29:56	00:29:56	3:56:16	x	x	x	x	x	x	0
	I8F14	unespriopreto014	D	1									1	1	1	1	x	1	1	6
	UOM24	teletandem24	DI	x	00:48	x	00:23:45	00:32:21		x	00:24:26	00:10:40	2:19:12	x	x	x	x	x	x	0
	I8F15	unespriopreto015	D	1									1	1	1	1	1	1	1	6
	UOF25	teletandem25	D	x	00:49:28	00:47:15	00:44:08	00:46:40	00:39:35	00:52:47	00:35:05	5:14:58	x	x	x	x	x	x	x	0
	I8F17	unespriopreto17	D	1									1	1	1	1	1	1	1	7
	UOM27	teletandem27	DI	x	00:50:47	00:43:51	00:46:11	00:45:16	00:42:53	00:44:26	00:29:56	5:03:20	x	x	x	x	x	x	x	0
	I8F19	unespriopreto019	DI	x									1	1	1	1	1	1	x	5
	UOF20	teletandem20	DI	x	00:52:16	00:41:20	00:49:13	00:26:14		x		x	2:49:03	x	x	x	x	x	x	0
<b>Totais</b>				<b>9</b>	<b>8:08:39</b>	<b>6:44:34</b>	<b>7:32:05</b>	<b>6:33:49</b>	<b>5:11:01</b>	<b>5:32:19</b>	<b>3:16:21</b>	<b>42:58:48</b>	<b>8</b>	<b>10</b>	<b>9</b>	<b>10</b>	<b>8</b>	<b>10</b>	<b>8</b>	<b>63</b>

Fonte: Elaborado pela autora para o MulTeC

Em decorrência da quantidade de colunas na planilha, entre 35 e 44 colunas a depender do cenário de aprendizagem (FOUCHER, 2010; ARANHA e LEONE, 2016, 2017), não podemos representar em uma figura apenas toda a aba<sup>89</sup> de maneira que fosse legível. No entanto, na figura 17, o pesquisador terá acesso aos dados produzidos por par, bem como os produzidos individualmente, a depender do tipo de dado. Cada dado produzido em uma das tarefas do TTD está registrado com o numeral 1; em não tendo sido produzido, seja por ausência do participante ou porque o dado não fora encontrado no Banco TTDii, há um ‘x’.

As duas primeiras colunas informam IT e usuário do participante, a razão de manter o usuário Skype® mesmo havendo essa informação na constituição do IT se deve à necessidade de redução da possibilidade de erro, pois o trabalho de organização foi feito manualmente. A coluna seguinte informa que tipo de autorização foi dada pelos participantes. Isso porque alguns participantes doaram dados e imagem e outros apenas os dados, como expliquei anteriormente, deixando claro ao pesquisador que não autoriza a publicação de sua imagem, dado relevante ao pesquisador, no que se refere à ética da pesquisa sem haver necessidade de que ele abra outro documento (Planilha “Informações Interagentes) para ter acesso a essa informação.

As demais colunas apresentam três características diferentes. Nas colunas mais simples são apresentados os dados produzidos em textos escritos, com indicação de presença (1) ou ausência (x) do dado no *corpus* por ordem de ocorrência, assim em ‘Diários’ 1º é o diário de aprendizagem escrito sobre a SOTi, em ‘Chats’ – 1º é o chat produzido durante a SOTi e assim sucessivamente. Estão organizadas dessa maneira as colunas sobre os questionários (inicial e final), os diários de aprendizagem, os chats e os textos produzidos na troca entre os parceiros, esses últimos havendo registro quanto as suas três versões possíveis<sup>90</sup>: original, revisada e reescrita. As colunas onde há registro das SOTs indicam o dia de produção de cada uma, bem como a duração em minutos.

No intuito de contabilizar a quantidade de dados produzidos da maneira mais acurada possível, inseri uma coluna entre cada tipo de dado, acrescentando nelas fórmulas que somam a quantidade de dados produzidos por par ou individualmente a depender do tipo de autoria (dupla ou individual). Tomando por exemplo a figura 20, na qual as SOTs produzidas estão visíveis, observamos que a partir da coluna que a sucede, o pesquisador pode identificar a quantidade de horas de vídeo salvas por par, bem como, no final, a quantidade de horas produzidas por toda a turma, no exemplo, um total de 42 horas, 58 minutos e 48 segundos. Na

---

<sup>89</sup> O leitor terá acesso a uma planilha dessas anexo a este texto.

<sup>90</sup> Possíveis porque não é verdadeiro afirmar que há todas as versões em todos os textos de todos os participantes que autorizaram a coleta e uso dos dados em pesquisas.

linha em destaque, no final da planilha, com os totais, o pesquisador saberá quantos minutos de cada SOT foi produzida na turma, assim, no exemplo acima há 8 horas, 8 minutos e 39 segundos de gravação de SOTis dos pares participantes dessa turma e 3 horas, 16 minutos e 21 segundos de SOTf. O mesmo foi feito com os dados de texto, como por exemplo, na Figura 16, os diários de aprendizagem. Como foram 7 SOTs, os participantes deveriam ter produzidos 7 diários, logo há 7 colunas para o registro. A coluna que sucede o registro dos diários contém fórmula que realiza a soma dos diários produzidos por participante e na última linha a somatória dos tipos de diários (se o primeiro, o segundo,...), bem como os produzidos na turma em geral. No exemplo foram produzidos e autorizados para uso em pesquisa um total de 63 diários de aprendizagem.

Nessa aba, realizei uma linkagem, ou seja, a criação de links que levam o pesquisador ao dado produzido pelo participante, de modo que tenha acesso simplificado aos dados sem precisar sair da planilha. Ao clicar nos ITs, por exemplo, será aberta uma janela na qual o pesquisador poderá ver que tipos de dados foram produzidos pelo par.

**Figura 18 – Abertura de pasta de dados produzidos por par**

The image shows a screenshot of a Microsoft Excel spreadsheet and a Windows File Explorer window. The Excel spreadsheet is titled '2013\_UGA1i\_Levantamento dos dados - Excel' and displays a table with the following data:

	A	B	C	D	E	F	G	H	I
2		IT	Usuário Skype	TCLEs	QI				
3				D (Dados) DI (Dados e imagens)		SOTi	SOTin1	SOTin2	SOTin3
4	Turma					6/mar	20/mar	27/mar	3/abr
5		I8F4	unespriopreto04	DI	1	0:50:42	00:38:53	00:48:42	00:44:42
6		U0F14	teletandem14	D	x				
7		I8F5	unespriopreto05	DI	1	00:52:13	00:49:00	00:47:28	00:41:42
8		U0F15	teletandem15	DI	x				
9		I8F6	unespriopreto06	D	1	00:50:00	00:49:28	00:49:05	00:46:32
10	2	U0F16	teletandem16	DI	x				
11	0	I8F7	unespriopreto07	D	1	00:32:42	00:38:06	00:48:08	00:39:02
12	1	U0F17	teletandem17	DI	x				
13	3	I8M11	unespriopreto011	D	1	00:49:40	00:46:52	00:46:30	00:46:22
14	-	U0M21	teletandem21	DI	x				
15	U	I8F13	unespriopreto013	DI	1	00:52:51	00:49:49	00:48:55	00:24:42
16	G	U0M23	teletandem23	DI	x				
17	A	I8F14	unespriopreto014	D	1	00:48	x	00:23:45	00:32:22
18	1	U0M24	teletandem24	DI	x				
19	i	I8F15	unespriopreto015	D	1	00:49:28	00:47:15	00:44:08	00:46:42
20		U0F25	teletandem25	D	x				
21		I8F17	unespriopreto17	D	1	00:50:47	00:43:51	00:46:11	00:45:12
22		U0M27	teletandem27	DI	x				
23		I8F19	unespriopreto019	DI	x	00:52:16	00:41:20	00:49:13	00:26:12
24		U0F20	teletandem20	DI	x				
25		Totais			9	8:08:39	6:44:34	7:32:05	6:33:49

The File Explorer window shows the folder structure for 'I8M11\_U0M21', which includes subfolders for each 'U0F' (e.g., I8F4\_U0F, I8F5\_U0F, etc.) and 'U0M' (e.g., U0M21, U0M23, U0M24, U0M27).

Fonte: Elaborado pela autora para o MulTeC

O par da figura 18, I8M11 e U0M21, tem como dados salvos no banco os diários, os textos, os questionários e as SOTs, não havendo registro de chats. Os números ‘1’ que indicam presença do dado em questão no *corpus* estão linkados ao dado produzido e convertido em texto sem formatação .txt, excetuando-se aqueles com tabelas, imagens e/ou utilização de cores para revisão pelo parceiro. Essas informações são perdidas quando convertidas para o formato .txt, assim mesmo havendo convertido os dados em .txt para processamento, esse tipo de dado também está salvo em .pdf, de modo que o pesquisador possa ter acesso ao original do texto. O link, nesses casos, vai trazer à tela do pesquisador o dado em .pdf.

As indicações de duração de cada SOT está linkada, nessa aba ao vídeo salvo no *corpus*, sendo aberto a partir do clique sobre os números que indicam a duração da sessão. A outra aba dessa planilha contém o levantamento de quantidade de palavras produzidas por parceria.

#### 2.2.3.2.2 Contagem das palavras – aba 2

A quantidade de palavras em um *corpus* tem sido a referência para a sua classificação como grande, médio ou pequeno, como discute Berber Sardinha (BERBER SARDINHA, 2002) ao tratar da extensão de um *corpus*.

Objetivando chegar ao cálculo da quantidade de palavras que constituem o MulTeC inseri a aba ‘contagem das palavras’ com o registro da quantidade de palavras produzidas pelos participantes em cada dado. Como afirmei neste trabalho, a contagem das palavras foi realizada utilizando a ferramenta contagem de palavras do Word ® antes da conversão dos documentos para a versão texto sem formatação. Para a contagem foram excluídas os cabeçalhos e orientações para a produção textual e/ou perguntas dos questionários.

Similarmente a aba ‘Dados produzidos por par’, a aba ‘Contagem das palavras’ está organizada em colunas por dado como é possível verificar na figura 19.

Figura 19 – Aba ‘Contagem de palavras’ na planilha ‘Levantamento de Dados’

IT	QI	QF	Total	TRANSCRIÇÃO/SOTs								Diários							Totais por aprendi
				SOTi	SOTin1	SOTin2	SOTin3	SOTin4	SOTin5	SOTf	Totais por par	1º	2º	3º	4º	5º	6º	7º	
				6/mar	20/mar	27/mar	3/abr	10/abr	21/abr	28/abr									
I8F4	116	447	563								0	142	121	152	126	122	110	196	969
U0F14	x	x	0								0	x	x	x	x	x	x	x	0
I8F5	153	485	638								0	90	184	197	199	142	146	164	1122
U0F15	x	x	0								0	x	x	x	x	x	x	x	0
I8F6	130	349	479								0	218	313	215	247	167	290	251	1701
U0F16	x	x	0								0	x	x	x	x	x	x	x	0
I8F7	82	125	207								0	189	115	100	115	65	119	74	777
U0F17	x	x	0								0	x	x	x	x	x	x	x	0
I8M11	119	209	328								0	85	51	x	86	59	83	124	488
U0M21	x	x	0								0	x	x	x	x	x	x	x	0
I8F13	80	149	229								0	51	73	55	49	x	68	x	296
U0M23	x	x	0								0	x	x	x	x	x	x	x	0
I8F14	103	280	383								0	82	106	130	116	x	95	247	776
U0M24	x	x	0								0	x	x	x	x	x	x	x	0
I8F15	145	354	499								0	x	337	327	337	126	185	234	1546
U0F25	x	x	0								0	x	x	x	x	x	x	x	0
I8F17	81	149	230								0	120	150	125	121	51	84	59	710
U0M27	x	x	0								0	x	x	x	x	x	x	x	0
I8F19	x	395	395								0	x	94	181	67	81	98	x	521
U0F20	x	x	0								0	x	x	x	x	x	x	x	0
<b>Totais</b>	<b>1009</b>	<b>2942</b>	<b>3951</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>977</b>	<b>1544</b>	<b>1482</b>	<b>1463</b>	<b>813</b>	<b>1278</b>	<b>1349</b>	<b>8906</b>

Dados produzidos por par Contagem das palavras

Fonte: Elaborado pela autora para o MulTeC

O trabalho de transcrição não foi concluído no decorrer da escrita desta tese. Ele ficou sob responsabilidade de monitores do Projeto Teletandem. Inicialmente consegui coletar 2% das SOTs com transcrições concluídas. Essas transcrições foram realizadas por pesquisadores do TTD que investigaram aspectos diversos do teletandem em suas pesquisas<sup>91</sup>, no entanto, a equipe do grupo de pesquisa em TTD de São José do Rio Preto está trabalhando na adequação delas para então acrescentá-las ao MulTeC, bem como há monitoras trabalhando para ampliar a quantidade de SOTs transcritas em um esforço conjunto para contribuir com a finalização do *corpus*. As monitoras responsáveis pelas transcrições das SOTs estão mantendo o sistema de transcrição proposto para o MulTeC e aprovado em reunião pela coordenação do Projeto Teletandem, bem como o cabeçalho apresentado no *corpus*. O sistema de transcrição está apresentado no quadro a seguir.

**Quadro 7 – Sistema de Transcrição para o MulTeC**

Ocorrências	Sinais	Exemplos
Nomes na indicação de turnos	E para o participante da universidade estadunidense e B para o participante da universidade brasileira	E; B
Nomes próprios dos participantes	São utilizados os ITs (Identificação Teletandem) de cada participante	I9F3
E-mails dos participantes	É inserida nota explicativa entre colchetes simples	[email de I9F3]
Nomes próprios de pessoas relacionadas aos participantes	São acrescentadas letras ao final do IT correspondente à relação com o participante “p” para pai, “m” para mãe, “av” para avós, “n” para namorado/a, “a” para amigo/a	I9F3p
Nomes das universidades	São mantidos	UNESP; UGA
Nomes das cidades	São anonimizados. É inserida nota explicativa entre colchetes simples	[cidade onde mora] [cidade que visitou]

<sup>91</sup> Utilizaram dados do Banco TTDii os pesquisadores Rampazzo (2017), Freschi (2017), Rossi (2017), Toledo (2017) e Detomini (2018).

Numerais	Grafia por extenso	Dez; vinte
Tom interrogativo	Ponto de interrogação	?
Tom exclamativo	Ponto de exclamação	!
Interjeições	Grafia da ocorrência	ah; hum; ahn
Prolongamento de vogal	Dois pontos	Eu tava é: visitando minha família
Citação	Entre aspas duplas	Magsamen disse “When I grow up, I want to be me”
Sobreposição de vozes	Colchete simples seguido da inicial maiúscula do outro participante na mesma linha	B: então eu [E: você...
Interrupção	Colchete simples seguido da anotação “interrompe” na outra linha	B: [interrompe]
Truncamentos	Barra	e/eu co/comprei
Silabação	Hífen	a-par-ta-men-to
Pausas	Reticências	...
Ênfase	Maiúscula	Você ACREDITA?
Trecho incompreensível	Anota-se “incompreensível” entre colchetes simples	[incompreensível]
Comentários do analista ou descrição da imagem	Comentário ou descrição entre colchetes duplos	[[ ]]
Cabeçalho	Indicar nome do arquivo, a quais outros arquivos está relacionado, local e data de gravação e transcrição, tipo de autoria, línguas transcritas, autores, duração do arquivo, número de palavras (sem incluir na contagem as letras de indicação de turnos), fonte, indicação de turnos e nome do transcritor	Exemplo abaixo

Fonte: Elaborado por Laura Rampazzo e Queila Barbosa Lopes a partir de MARCUSCHI, L. A. A transcrição de conversações. In: \_\_\_\_\_. *Análise da Conversação*. 5.ed. São Paulo: Editora Ática, 2003; e GONÇALVES, S. C. L.; TENANI, L. E. Problemas teórico-metodológicos na elaboração de um sistema de transcrição de dados interacionais: o caso do projeto ALIP (Amostra Linguística do Interior Paulista). *Niterói*, n.25, 2008, p. 165-183.

Enquanto estive escrevendo esta tese, trabalhei na revisão de transcrições, no sentido de procurar garantir a consistência do padrão estabelecido para o MulTeC.

Na aba de contagem das palavras, figura 19, há a inserção do quantitativo das ocorrências de cada transcrição tão logo as monitoras encaminham-me os arquivos. Portanto, após a conclusão das transcrições a quantidade de palavras por transcrição, assim como a quantidade geral, totalizada pela fórmula inserida na coluna, ficarão acessíveis aos pesquisadores ao acessar a planilha. A linkagem para os documentos das transcrições está sendo realizada nessa aba, na coluna para cada número que indica a quantidade de palavras por SOT transcrita.

O cômputo geral das ocorrências (incluindo as ocorrências das transcrições de todas as SOTs) ainda não pode ser informado, mas a partir do trabalho que realizei para a organização e das transcrições feitas até janeiro de 2019, o cômputo das ocorrências (palavras) no MulTeC é:

- 617.904 em dados escritos;
- 186.350 em dados orais (transcritos);

A soma das ocorrências totaliza **814.063** até janeiro de 2019. A partir dessas 16 planilhas (considerado que há uma turma por planilha e são 16 turmas) surgiu a necessidade de elaborar uma planilha com os dados gerais de cada turma, mas em um só documento de modo que o pesquisador pudesse visualizar sinoticamente a quantidade de dados produzidos por turma e no geral no período. A planilha foi nomeada como “Levantamento de Dados Geral”.

#### 2.2.3.3 Planilha ‘Levantamento Geral’

Esse documento apresenta em uma aba os quantitativos de todos os dados produzidos por turma TTD no MulTeC. Nela, o pesquisador encontra informações quanto à quantidade de dados produzidos (por tipo de dado) assim como o total de dados por semestre. O documento possui três abas: Na primeira é possível verificar os quantitativos por ordem cronológica, estando na primeira linha as informações referentes a primeira turma de 2012 e assim sucessivamente. A figura 20 apresenta a primeira aba.

Figura 20 – Planilha ‘Levantamento de Dados Geral’

Semestre	Turma	Qtde de interagentes		QI	SOTs								Total geral	Diários				
		Participantes	Dados no banco		SOTi	SOTin1	SOTin2	SOTin3	SOTin4	SOTin5	SOTin6	SOTf		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	8:44:10	10:23:33	9:15:49	9:22:03	12:02:55	10:50:44	10:07:24	5:02:11	75:48:49	13	14	13	13	13
2º	2012_UGA3i	32	20	10	8:05:34	7:28:20	7:25:08	5:42:25	6:58:42	5:09:13	x	04:06:23	44:55:45	10	10	9	9	9
	2012_UGA4si	24	12	1	4:00:07	2:33:38	2:57:55	2:45:51	3:33:12	3:32:58	2:11:10	1:08:14	22:43:05	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	8:03:21	86:21:10	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	4	3	2	2	1
	2013_UGA4si	26	6	1	1:17:57	2:14:23	1:34:34	2:21:45	1:32:58	1:40:59	0:00:00	0:00:00	10:42:36	0	0	0	0	0
	2013_UGA5i	24	10	1	4:05:14	4:19:57	4:40:56	4:33:47	5:47:35	4:54:04	3:44:26	3:56:27	36:02:26	6	6	6	6	6
1º	2014_UGA1i	22	20	2	4:28:23	4:14:44	4:33:50	3:57:51	3:40:15	x	x	3:18:15	24:13:18	4	3	3	3	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	2:45:31	2:15:50	0:51:12	x	x	3:11:00	11:55:24	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
<b>TOTAIS</b>		<b>400</b>	<b>282</b>	<b>91</b>	<b>86:36:43</b>	<b>82:40:24</b>	<b>90:49:26</b>	<b>81:13:34</b>	<b>89:46:59</b>	<b>63:14:53</b>	<b>29:29:50</b>	<b>57:27:18</b>	<b>581:19:07</b>	<b>108</b>	<b>105</b>	<b>101</b>	<b>98</b>	<b>93</b>
Obs:	Nesta planilha "0" indica arquivo não foi salvo e "x" indica que para a turma não havia sido previsto a realização da tarefa em questão.																	
Geral		Semi-integrado	Integrado															

Fonte: Elaborado pela autora para o MulTeC

Seguindo a padronização que estabeleci para as planilhas no MulTeC, nessa planilha, entre os tipos de dados diferentes há colunas com a somatória dos dados. A inserção da fórmula que realiza o cálculo torna menor a possibilidade de erro no resultado da soma. A partir das linhas e colunas que apresentam a soma, o pesquisador sabe quantas horas de SOTis estão salvas no *corpus*, bem como a quantidade de diários de aprendizagem escritos sobre a SOTi.

Todos os procedimentos descritos até aqui seguiram o mapa do ciclo dos dados, sobre o qual tratará a próxima subseção.

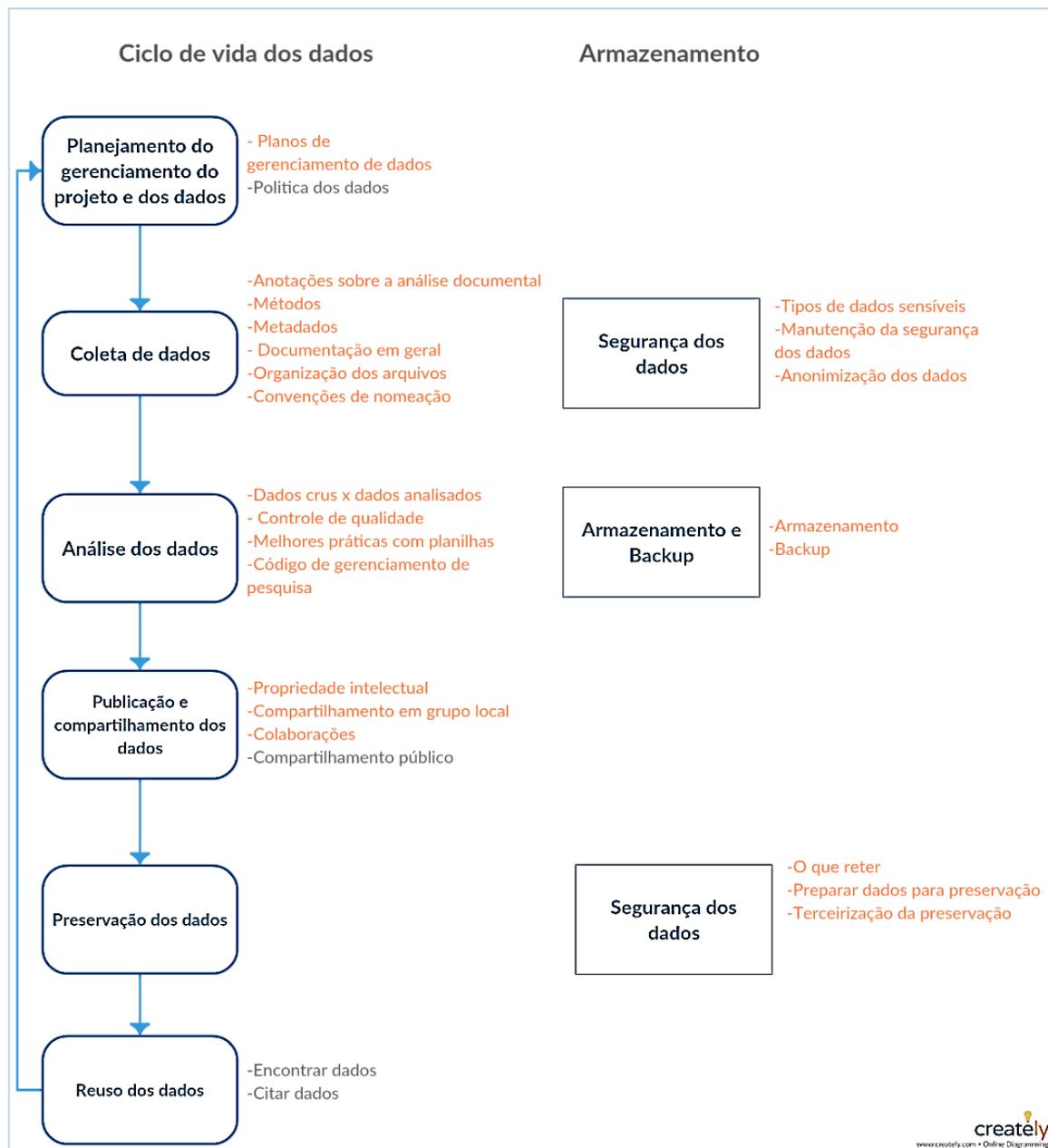
#### 2.2.4 O CICLO DOS DADOS NO MULTEC

Considerando os procedimentos apresentados adotados para a construção do MulTeC e apresentados neste capítulo, e ainda, levando em consideração o ciclo de vida dos dados, gostaria de destacar sumariamente como as fases desse ciclo ocorreram no trabalho, lembrando que uma das primeiras fases do ciclo, a coleta de dados, não foi realizada nesta pesquisa, haja vista que a construção se deu a partir do Banco TTDii como discuti anteriormente. Contudo, procedimentos realizados na etapa de seleção dos dados precisaram ser repensados em atendimento às exigências para a construção de um *corpus* como vimos no capítulo teórico, a partir dos trabalhos de Berber Sardinha (2004, 2009), Chanier e Wigham (2016), (BRINEY, 2015), Eynden (2013) e Eynden et al. (2011).

Assim, a partir da análise documental dos arquivos constantes no Banco TTDii se seguiu uma fase de planejamento do gerenciamento dos dados. Nessa fase, estabeleci quais procedimentos utilizaria para a coleta de assinatura retroativa dos TCLEs e que tipos de códigos precisaríamos criar para auxiliar na fase seguinte que inclui a anonimização dos dados. Muito embora a anonimização dos dados possa ser confundida com a nomeação, vale lembrar que a nomeação é um dos procedimentos para a anonimização os dados, no entanto, não se constituindo o único, como apresento neste capítulo. Assim, houve necessidade de criação de códigos para a anonimização e de outros para a nomeação dos arquivos.

O mapa proposto por Briney (2015) na figura 21 tem como destaque os procedimentos utilizados para a construção do MulTeC em cada etapa.

**Figura 21 – Mapa dos dados no MulTeC**



Fonte: Briney (2015, p. 28) destaques pela autora desta tese.

Entre os procedimentos de cada etapa do ciclo, incluindo aqueles voltados ao armazenamento dos dados no *corpus*, apenas quatro deles não foram executados para a construção do MulTeC, mais precisamente em decorrência do contexto: a política dos dados, o compartilhamento público, encontrar os dados, citar os dados. Esses são procedimentos que estão em fase de planejamento pela coordenação do projeto de pesquisa financiado pela FAPESP ao qual este trabalho está vinculado. Vale lembrar que as transcrições das SOTs estão em andamento.

Os dados do MulTeC foram utilizados por Rampazzo (2017) e Freschi (2017), como Toledo (2018), mesmo antes de passarem pela organização proposta nesta tese. Além dessas pesquisadoras do grupo de pesquisa TTD de São José do Rio Preto, Del Monte (em andamento) tem trabalhado com os dados pós-organização em sua tese de doutoramento que objetiva investigar a autonomia dos aprendizes a partir da escrita dos diários de aprendizagem.

### 2.2.5 CABEÇALHOS

Os cabeçalhos, conforme Berber Sardinha (2004), são “parte do arquivo de cada texto do *corpus* que contém informações sobre o texto, tais como a origem, a data de coleta, o grupo responsável. Acrescentei o cabeçalho a cada um dos arquivos de textos que compõem o MulTeC seguindo o padrão estabelecido, seguindo o padrão como explano a seguir.

Para a criação dos cabeçalhos dos arquivos de textos no MulTeC, dentre os formatos, Cocoa e SGML, que discutimos no capítulo teórico a partir de Sardinha (2004), optamos pelo formato Cocoa. Os cabeçalhos no MulTeC apresentam informações gerais, bem como a relação entre os dados presentes no *corpus*. A apresentação da relação dos arquivos entre si ocorreu em função do contexto de produção de dados. Segundo apontam as pesquisas de Aranha (2014), Rampazzo (2017) bem como Luvizari-Murad (2011) e Araújo (2012), todas realizadas no contexto TTD, o teletandem estabelece-se como um sistema. Embora essas pesquisadoras tenham tido objetivos de estudo diferentes, haja vista que as duas primeiras dedicaram-se ao estudo da SOTi como um gênero, Luvizari-Murad (2011) dedicou sua pesquisa à compreensão da atividade colaborativa de aprendizagem de língua via TTD a partir de uma parceira de Português-Alemão e Araújo (2012) teve o intuito de investigar a sistematização da formação de parcerias institucionais para o TTD. Mesmo que cada uma dessas pesquisas tenha optado ainda por embasamentos teóricos ímpares<sup>92</sup>, elas são uníssonas no que se refere à observação de que o TTD constitui-se um sistema, como, por exemplo, conclui Luvizari-Murad (2011) no resumo de sua tese “a análise dos dados mostrou a formação de um sistema complexo” (LUVIZARI-MURAD, 2011, p. ix).

Considerando, portanto, o contexto de produção dos dados constitui-se relevante aparecer no cabeçalho dos arquivos a indicação de relação entre os dados, assim o pesquisador

---

<sup>92</sup> Aranha (2014) e Rampazzo (2017) optam pela abordagem retórica e sócio-retórica de gêneros para realizar suas investigações, enquanto Luvizari-Murad (2011) e Araújo (2012) utilizam o arcabouço teórico da Teoria da Atividade.

ao abrir cada um dos 2.108<sup>93</sup> arquivos de texto poderá saber a quais outros arquivos presentes no *corpus* o dado em questão está relacionado.

O cabeçalho do arquivo do primeiro texto reescrito em inglês de uma turma de TTDii de 2014 conterá as seguintes informações:

<Nome do arquivo 2014\_I9M4\_UGA1i\_TReI1>  
 <Relacionado a 2014\_I9M4\_UGA1i\_SOTin1 2014\_I9M4\_UGA1i\_TOI1>  
 <Local de gravação Laboratório TTD Rio Preto>  
 <Data de produção 27 de março de 2014>  
 <Data do arquivamento 30 de abril de 2014>  
 <Tipo de autoria individual>  
 <Língua Inglês>  
 <Autor I9M4>  
 <Nº de palavras 485>  
 <Fonte Teletandem Rio Preto>

O item “Relacionado a” indica a quais outros arquivos, constantes do MulTeC, o texto em questão se relaciona. No caso do texto revisado na turma TTDii de 2014, há relação entre esse texto com a SOTin1, pois nessa sessão de interação oral em teletandem o texto original (TOI1) foi discutido. Assim, de modo semelhante, o texto revisado está relacionado por sua vez ao texto original 1 (TOI1), que foi o ponto de partida para a discussão e consequente revisão textual. Essas informações são relevantes para o usuário do MulTeC, afinal os dados se interligam e a compreensão de um conjunto de dados se aprimora tanto pelo acesso aos outros textos, produzidos pelo mesmo par, quanto pelo resgate dos indícios de dinâmicas de aprendizagem que emergem em teletandem e que envolvem os dados em questão (no caso usado nesta exemplificação, dados de produção textual).

Procurei manter a ordem e quantidade de informações nos cabeçalhos independentemente do tipo de dado, se diário, texto, questionário ou chat. No entanto, há variações nos cabeçalhos. Duas delas consistindo em: i) a ausência do item “Relacionado a”<sup>94</sup>; ii) acréscimo de itens nos cabeçalhos específicos para as transcrições das SOTs.

No cabeçalho da transcrição, sugeri o acréscimo de informações como: nome do transcritor e indicação de troca de turnos, ficando a sugestão como segue:

---

<sup>93</sup> Esse número refere-se à somatória dos arquivos: questionários (inicial e final), diários, chats e textos produzidos para a troca entre os parceiros. Os arquivos das transcrições não foram contabilizados porque ainda não foram concluídos. As transcrições das 581 horas de vídeo/áudio estão sob a responsabilidade de bolsistas, coordenados pela Prof<sup>a</sup>. Dr<sup>a</sup> Solange Aranha.

<sup>94</sup> Haja vista trabalharmos com dados coletados posteriormente e por não haver qualquer registro sobre a motivação para a ausência de alguns dados, não podemos afirmar o que ocorreu em algumas parcerias para termos, por exemplo, as SOTs e os chats, em um cenário onde fora planejado haver troca de textos, como se pode verificar no cronograma apresentado no tutorial da turma e ainda como pode-se perceber ao verificar as demais parcerias da mesma turma.

<Nome do arquivo 2015\_I27F12\_UGA1i\_SOTin1\_T>  
 <Relacionado a 2015\_I27F12\_UGA1i\_SOTin1 2015\_I27F12\_UGA1i\_D1  
 2015\_I27F12\_UGA1i\_C1>  
 <Local de gravação Laboratório TTD Rio Preto>  
 <Data de gravação 12 de junho de 2015>  
 <Data da transcrição 14 de outubro de 2016>  
 <Tipo de autoria dupla>  
 <Línguas Português e Inglês>  
 <Autores I27F12 e U0M12>  
 <Duração do arquivo 40'32''>  
 <Nº de palavras 5.132>  
 <Fonte Teletandem Rio Preto>  
 <Indicação de turnos B=Brasileiro E=Estadunidense>  
 <Transcritora Giovanna Pomaro>

A partir do sistema de transcrição apresentado nas pesquisas de Rampazzo (2017) e Freschi (2017) e do que está proposto por Sardinha (2004) sobre os cabeçalhos, no MulTeC, os cabeçalhos apresentam indicação de troca de turno, facilitando a compreensão do texto transcrito. A decisão de não incluir um sistema de transcrição em cada cabeçalho objetivou não ampliar muito o arquivo que, pelo material transcrito, já pode apresentar tamanho razoável dado que as SOTs têm duração média de 40-50 minutos.

A inserção dos cabeçalhos ocorreu concomitante à conversão dos dados, de modo que tão logo o cabeçalho de cada arquivo fosse concluído, os arquivos eram convertidos para o formato livre como apresentado a seguir.

## 2.2.6 CONVERSÃO DOS DADOS

Nesta subseção, discuto dois procedimentos relevantes para o processamento dos arquivos independentemente do sistema operacional do computador utilizado: a conversão dos dados e a (re)nomeação dos arquivos. A conversão dos dados procurando atender ao ciclo de vida dos dados citado por Briney (2015), a saber: 1) o planejamento do gerenciamento do projeto e dos dados; 2) coleta; 3) análise; 4) publicação e compartilhamento; 5) preservação; e 6) reuso dos dados. Converter os dados em formato aberto, ou seja, de formato legível por qualquer sistema operacional e pelos softwares de processamento linguístico procura contemplar precisamente duas fases do ciclo: a preservação e o compartilhamento.

A preservação porque, como dito anteriormente, caso os arquivos sejam mantidos no *corpus* como de um determinado software pago, como por exemplo o Word®, atualizações podem resultar na não legibilidade de arquivos produzidos em versões anteriores. A legibilidade pode ser inviabilizada ainda em sistemas operacionais diversos do que foi utilizado na produção

dos arquivos, o que dificulta a fase de compartilhamento. Logo, a conversão dos arquivos em formato livre atende as fases de preservação e compartilhamento.

Procedimento voltado para a garantia da leitura dos arquivos em máquinas com softwares diversos, a conversão dos dados no MulTeC ocorreu em arquivos de texto. Os arquivos em áudio e vídeo foram considerados como não necessitando de conversão, pois estavam salvos em mp3 e mp4 respectivamente cuja leitura é possível independentemente do sistema operacional, não havendo, até o presente momento formatos mais atualizados. No Banco TTDii, todos os arquivos estavam salvos em Word®, cuja leitura só é factível por máquinas com o pacote Microsoft Office® instalado.

No processo de conversão, dada a natureza dos dados, concluí haver necessidade de dois formatos diferentes do mesmo arquivo a depender da informação contida no texto: .txt e/ou .pdf. Embora não se refira à necessidade de conversão do mesmo arquivo em dois formatos diversos, Briney (2015) e Eynden (2013) citam o formato pdf como recomendável. Há textos, por exemplo, com inserção de imagens, que não aparecem em txt, por se caracterizarem como arquivos de texto sem formatação. Assim, textos com imagens, indicações de revisão com utilização de cores e destaques com negrito foram salvos em duas vias: uma em .txt para processamento por softwares linguísticos e em pdf para leitura pelos pesquisadores com a informação como está no arquivo original. Na imagem a seguir, é perceptível a diferença entre um texto salvo no formato de doc e o texto sem formatação. Há a perda de destaques feitos durante a revisão pelo parceiro, por exemplo.

**Figura 22 – Texto produzido por participante salvo com formatos diversos**

Nas últimas décadas, a sociedade tem mudado. No 1960s e 1970s, a sociedade era famosa para o movimento dos "hippies." Embora eu não vivei durante o tempo , eu gostava ( because today do you don't like?) destas décadas porque a musica do tempo era muito bom. As bandas famosas eram Pink Floyd e Led Zepplin. Carly Rae Jepsen não existia. NO início de 1990 começava o tempo (or ERA) do (da- informática is feminino) informática, e as pessoas começavam usar os computadores e a Internet cada dia. O mundo ganhava muito (feminino – a velocidade) velocidade. Eu acho que os computadores mudavesse a sociedade. As pessoas podiam acessar mais (internet – informática é o conjunto das

Nas últimas décadas, a sociedade tem mudado. No 1960s e 1970s, a sociedade era famosa para o movimento dos "hippies." Embora eu não vivei durante o tempo , eu gostava ( because today do you don't like?) destas décadas porque a musica do tempo era muito bom. As bandas famosas eram Pink Floyd e Led Zepplin. Carly Rae Jepsen não existia. NO início de 1990 começava o tempo (or ERA) do (da- informática is feminino) informática, e as pessoas começavam usar os computadores e a Internet cada dia. O mundo ganhava muito (feminino – a velocidade) velocidade. Eu acho que os computadores mudavesse a sociedade. As pessoas podiam acessar mais (internet – informática é o conjunto das ciências que estudam a ciência

**Fonte: MulTeC**

Como é possível visualizar na figura, há destaques feitos em vermelho e azul no texto original. Essas marcações podem ser relevantes para análise por pesquisadores interessados em investigar a troca de textos entre os participantes do TTD. Assim optamos por salvar o mesmo arquivo em dois formatos, um em .pdf e outro em .txt. O excerto acima foi retirado de um texto escrito em português por um participante da UGA e revisado por seu parceiro brasileiro em uma turma de TTDisi em 2012.

### 2.3 TIPOLOGIA DO MULTEC

Tomando em consideração a classificação apresentada por Berber Sardinha (2000; 2004), concluí que o *corpus*, cuja constituição apresento aqui, trata-se de: a) um *corpus* multimodal – por incluir tanto registros (modos) escritos quanto falados; b) diacrônico e histórico – pois está constituído do que fora produzido em uma parceria de TTD em 8 semestre letivos, compreendendo vários períodos de tempo passado; c) dinâmico, podendo a ele ser acrescido novos registros de semestres vindouros, seguindo o protocolo de coleta e organização propostos; d) bilíngue – por contar com produções tanto em português quanto em inglês; e e) de aprendiz e de língua nativa, pois tanto há momentos em que o participante produz na língua na qual é proficiente, como na língua-alvo. As informações sobre os participantes contribuem na classificação tipológica do *corpus*.

## CAPÍTULO 3: RESULTADOS E DISCUSSÃO

O preparo dos dados do Banco TTDii para a constituição do MulTeC garantiu reflexões acerca do processo de coleta e de organização dos dados a partir dos pressupostos teóricos que embasam esta tese. Neste capítulo, apresento os resultados alcançados a partir dos procedimentos propostos. Discorro assim sobre uma proposta de coleta e organização para corpora futuros a partir da experiência de organização dos dados do Banco TTDii para a constituição do MulTeC. O capítulo está dividido em quatro sessões: na primeira apresento os resultados no que se refere à organização do MulTeC, na segunda seção trato a questão de documentos que registrem o contexto de produção dos dados; na terceira, em que apresentaremos uma proposta de coleta embasados na experiência de organização dos dados do Banco TTDii e na última seção discuto uma relação diversa entre macro e microtarefas no TTD, considerando essa discussão uma reflexão que emergiu do que observei nos dados utilizados para esta pesquisa e que foi relevante na organização das pastas e subpastas do MulTeC.

### 3.1 RESULTADOS: O MULTEC E A PROPOSTA DE COLETA PARA O TTD

Obtive dois resultados nesta pesquisa: i) o primeiro foi a organização do MulTeC à qual o leitor terá acesso um pouco mais detalhado a partir do QRCode®<sup>95</sup> encontrado no anexo 8 desta tese, que é fruto dos seguintes resultados: a padronização do cabeçalho, da anonimização e da nomeação dos dados que o constituem; ii) o segundo é a proposta de coleta de dados para o TTD que também pode ser estabelecida a partir das reflexões suscitadas a partir das dificuldades que encontrei junto ao Banco TTDii durante o processo de torná-lo compartilhável.

#### 3.1.1 A ORGANIZAÇÃO DAS PASTAS

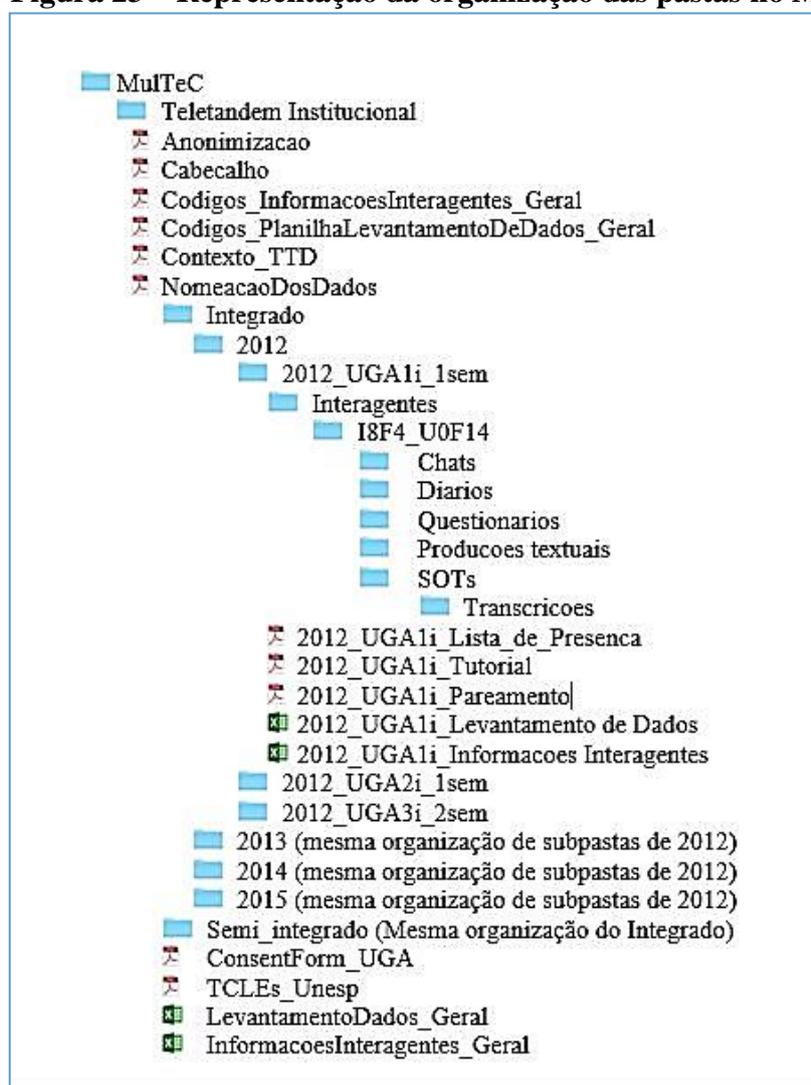
A organização das pastas foi feita a partir dos tipos de modalidade nas quais as turmas TTD foram desenvolvidas e a relação entre tarefa e macrotarefa no contexto TTD. As modalidades separando os dados em duas pastas: Integrado e semi-integrado e a relação entre

---

<sup>95</sup> Optei pelo QRCode em detrimento de um CD, considerando que em alguns notebooks/laptops não há mais leitor de CD. Assim, a partir de um leitor de QRCode qualquer pessoa interessada pode ter acesso ao material. No mesmo anexo há também o link para o arquivo de modo que a ausência de um leitor de QRCode impeça a visualização pelo leitor.

macrotarefa e tarefa determinando a ordem dessas pastas. Desse modo, separamos os dados de modalidades diferentes em pastas diferentes. O que já alterou um pouco a representação de Aranha, Luvizari-Murad e Moreno (2015), chegando a uma organização representada na figura 20.

**Figura 23 – Representação da organização das pastas no MulTeC**



Fonte: Elaborado pela autora

Além da identificação e organização de dados de semi-integrado e integrado em pastas diferentes, aquelas que contém os dados produzidos pelos participantes também foram dispostas de maneira diversa daquela proposta por Aranha, Luvizari-Murad e Moreno (2015). A organização do Banco TTDii apresenta divisão dos dados em pastas por tipo de dado enquanto que no MulTeC os dados estão organizados nas pastas de cada par que constituiu parceria na turma, considerando esses dados como tarefas que são realizadas pelos participantes

para realizar aprendizagem no contexto teletandem. Como na figura, acima o par constituído por I8F4 e U0F14 produziram chats, diários, questionários, produções textuais e SOTs, sendo essas subpastas da pasta I8F4\_U0F14 (pasta do par) naquela turma. As turmas, nessa organização são consideradas cenários de aprendizagem, outro conceito que embasa a organização das pastas.

Essa organização foi feita considerando principalmente o cenário de aprendizagem como a paisagem encontrada quando da execução do cenário pedagógico, que se constitui na arquitetura pedagógica projetada para aquela turma, pois, como apresentado por Foucher (2010), a relação entre cenário pedagógico e cenário de aprendizagem é a mesma entre o que fora previsto e o que fora realizado de fato. Desse modo, as tarefas são meios para o alcance da aprendizagem de línguas no TTD, como degraus de uma escada auxiliam o alcance do próximo andar em um edifício. Além das pastas com os dados, há documentos complementares que servem como lentes para que haja compreensão do contexto de produção desses dados, de modo que possam ser compartilháveis com quem não atuou na coleta de dados.

### 3.5 DOCUMENTOS PROPOSTOS: CENÁRIOS

Os cenários pedagógicos e de aprendizagem aparentemente são bastante claros no contexto TTD. Para Aranha e Leone (2016; 2017) a relação cenário pedagógico e de aprendizagem são, no contexto TTD, conceitos profícuos e cuja documentação precisa ser detalhada de modo que se compreenda melhor o contexto de produção dos dados. Rampazzo (2017) concluiu que, a partir dos dados coletados no Banco TTDii, se pode observar que cada turma pode constituir um cenário diferente, embora as turmas tenham sido realizadas na mesma modalidade de TTD. A autora investigou quatro turmas de TTDii e constatou em cada turma um cenário diverso “[...] as SOTis dos diferentes anos pertencem a quatro cenários de aprendizagem com características próprias” (RAMPAZZO, 2017, p. 84), embora a pesquisadora tenha investigado a mesma modalidade de teletandem o que poderia representar cenários iguais. Nesta pesquisa, pude constatar que entre as turmas de TTD presentes no MulTeC, não é a modalidade que delimita tarefas e configurações de cenários de aprendizagem. Essa definição ocorre durante o planejamento da turma quando as mediadoras elaboram o percurso da turma TTD para o semestre.

Na construção do MulTeC, considerando os conceitos de cenários, tanto o de aprendizagem quanto o pedagógico, bem como o contexto de aprendizagem TTD no qual foram produzidos os dados, logo em pares, optei por organizar os dados produzidos por cada par em

uma pasta com os dados produzidos e salvos<sup>96</sup> no período na turma. Desse modo, o pesquisador que acessar as pastas dos pares poderá encontrar quantidade diversa, em alguns pares cinco pastas – diários, questionários, SOTs, produções textuais e chats – enquanto em outra parceira quatro ou mesmo três pastas, a depender dos dados produzidos e salvos pela equipe responsável.

### 3.6 A PROPOSTA DOS CENÁRIOS

Dadas as dificuldades encontradas para garimpar informações referentes ao que ocorreu e como foi planejada cada turma TTD, apresentei ao grupo de pesquisa TTD da Unesp de São José do Rio Preto a elaboração de documentos que têm como função receber o registro de dois momentos cujo conhecimento interfere na compreensão dos dados produzidos no TTD: o momento do planejamento (cenário pedagógico) e o momento da realização (cenário de aprendizagem) da turma TTD. A partir dos conceitos de cenários pedagógico e de aprendizagem (FOUCHER, 2010; ARANHA, LEONE, 2016; 2017), sugeri a utilização de dois documentos que podem ser chamados de cenário pedagógico e cenário de aprendizagem dada a natureza de seus registros. O primeiro deve conter o registro do que foi planejado para a turma e o segundo contendo o que foi realizado, deixando registrada qualquer ocorrência no período de atividades TTD da turma.

#### 3.6.1 CENÁRIO PEDAGÓGICO

Compreendendo cenário pedagógico como o cenário planejado (FOUCHER, 2010) para a realização de um espaço de aprendizagem sugeri a utilização de um documento no qual fosse inserido, pelas mediadoras da turma TTD, tanto brasileira quanto estrangeira, o que foi planejado para a turma. O documento, criado no formato Excel® está composto por três abas, salvo na conta Google Drive do grupo de pesquisa em teletandem da Unesp de São José do Rio Preto: uma aba para as características gerais da turma, a segunda constando os dados de pareamento e a terceira com o cronograma das tarefas planejadas com suas respectivas datas de ocorrência.

---

<sup>96</sup> É possível inferir que houve dado produzido e não salvo no Banco TTDii. Levanto essa hipótese porque há textos revisados sem o seu respectivo texto original e reescrito, por exemplo. Como não participei da coleta, posso apenas inferir razões pelas quais os dados não foram salvos, como por exemplo, o participante brasileiro pode ter deixado de enviar algum dado por esquecimento. Um plano de gerenciamento dos dados bem elaborado diminuiria a possibilidade de ocorrências similares.

Importante ressaltar que, como o cenário pedagógico apresentado acima, o cenário de aprendizagem proposto foi elaborado de modo a contemplar dois critérios: auxiliar a coleta de metadados e a organização das atividades de uma turma pelas mediadoras. A opção por uso de documento em Excel®, possibilitando a inserção de abas, tem o intuito de facilitar o trabalho das mediadoras quando da realização das turmas, haja vista que inserir essas informações em um documento de texto demandaria mais tempo e na maioria das vezes essa inserção se realiza durante a realização das SOTs.

A primeira aba do documento nomeado como cenário pedagógico pode ser visualizada na figura a seguir.

**Figura 24 – Cenário pedagógico - Aba 1**

	A	B	C	D	E	F	G	H	I	J	K
	teletandem brasil línguas estrangeiras para todos										
MODALITY	INSTITUCIONAL SEMI INTEGRATED										
INSTITUTIONS											
CLASSES											
PROFESSORS											
MEDIATORS											
MEDIATION											
PERIOD											
DAY											
TOSs #											
PLACE											
DISCOURSE TYPE											
TYPOLOGY											
	TIME										
MONTH											
MONTH											
MONTH											

**Fonte: Elaborado pela autora em parceria com Laura Rampazzo –  
Salvo no Google Drive do grupo TTD da Unesp de São José do Rio Preto**

Nessa aba, as mediadoras deverão registrar as definições de modalidade de TTD na qual a turma será realizada, se teletandem institucional integrado ou semi-integrado<sup>97</sup>. O nome das instituições envolvidas na parceria, das mediadoras, professores regentes (no caso da modalidade de TTDii), dia da semana e o período de realização das SOTs, tipo de discurso (se conversação livre, realização de tarefas durante a SOT, ou se haverá direcionamento temático), e a tipologia. Nesse último item as mediadoras devem registrar se as SOTs ocorrerão em monolingualismo alternado ou bilingualismo, embora no contexto TTD a partir do princípio da separação de línguas, os participantes seguem o monolingualismo alternado, no qual se alternam no uso das línguas em estudo pelo par.

<sup>97</sup> Atualmente essas são as duas modalidades em ocorrência na Unesp de São José do Rio Preto.

Outro detalhe relevante a ser considerado é o fuso horário no período de realização das SOTs. As mediadoras deverão, nesse momento de planejamento, identificar quando e por quanto tempo haverá realização de SOTs em horário diverso do inicial em decorrência de horário de verão, por exemplo. Assim, as mediadoras avisarão os participantes sob sua responsabilidade sobre alterações quaisquer no horário de modo que todos possam estar presentes pontualmente no laboratório para realizar a SOT.

Vale enfatizar a relevância desse documento como registro do momento do planejamento das SOTs e demais tarefas do TTD para que o pesquisador que irá acessar os dados possa comparativamente observar o que ocorreu de fato e o que foi planejado, pois alterações podem afetar a produção dos dados em algum modo.

Na segunda aba do cenário pedagógico, há espaço para a inserção do pareamento dos participantes, caso esse ocorra previamente. Em turmas nas quais o pareamento será aleatório<sup>98</sup>, as mediadoras devem inserir uma observação nessa aba, registrando que, durante o planejamento foi decidido que o pareamento seria realizado a partir das cabines escolhidas pelos participantes no dia da SOTi, estando definidos apenas os usuários Skype a serem utilizados para aquela turma.

A terceira aba do documento está destinada a inserção do cronograma, como podemos ver na figura a seguir.

---

<sup>98</sup> Gostaria de ressaltar que essa nomenclatura “aleatório” e “prévio” é adotada pelo grupo de pesquisas em TTD de São José do Rio Preto, mas ambos são realizados aleatoriamente, ou seja, não há estabelecimento de critério para parear os candidatos a participantes do TTD. O uso dos termos é pragmático para diferenciar quando ocorrerá o pareamento, se antes da SOTi ou não.

**Figura 25 - Cenário pedagógico – Aba 3**

teletandem brasil		CRONOGRAMA		
línguas estrangeiras para todos				
Mês	Data/Descrição			

**Fonte: Elaborado pela autora em parceria com Laura Rampazzo e aprovado pela equipe TTD de São José do Rio Preto**

As mediadoras inserem, na tabela apresentada na figura acima, as datas de realização de cada tarefa prevista para a turma em questão, inclusive as datas em que haverá recesso, feriados e, também, informarão se haverá alguma SOT a ser realizada a partir das residências dos participantes.

A realização de SOTs é prioritariamente realizada no laboratório TTD (no Brasil) e nos laboratórios das instituições estrangeiras; no entanto, quando as mediadoras decidem que no período deverão ter 8 SOTs, mas nos calendários institucionais a incidência de feriados não o permite (logo os participantes não poderão usar os respectivos laboratórios), as mediadoras optam pela realização de uma SOT extra, normalmente, em horário e dia combinado pelos participantes, como descrito por Aranha, Luvizari-Murad e Moreno (2015). Urge lembrar que a gravação dessa SOT, caso esteja ocorrendo coleta de dados, é dificultada, considerando-se que o participante pode não possuir a ferramenta de gravação da sessão. Então, encerrado o período de planejamento, a mediadora responsável pela turma brasileira de participantes começa a inserir as informações no documento que deverá registrar as ocorrências da turma: o cenário de aprendizagem.

### 3.6.2 CENÁRIO DE APRENDIZAGEM

Como documento que registrará o que acontecerá no período de 4 a 12 semanas, período em que costumam ocorrer as sessões, o cenário de aprendizagem proposto ao grupo de pesquisa TTD de São José do Rio Preto contém quatro abas para o registro de: a) informações gerais sobre a turma TTD, igual à que está no cenário pedagógico; b) frequência dos participantes; c) informações sobre as SOTs; e d) dados sobre as sessões de mediação. Como a primeira aba desse documento é exatamente igual a primeira aba do cenário pedagógico, discorreremos aqui sobre as três abas seguintes.

No cenário de aprendizagem as duas primeiras abas devem conter registro quanto a informações gerais da turma. Nelas, o pesquisador tem conhecimento quanto a modalidade na qual a turma foi realizada, que dias da semana as SOTs ocorreram, qual o período de atividades da turma, bem como o registro da frequência de cada participante.

As duas abas seguintes, estão divididas seguindo a discussão que apresentei no capítulo teórico quanto a natureza das tarefas no TTD, no qual o teletandem é a macrotarefa que ocorre a partir da realização de tarefas de duas naturezas: i) mediação e ii) interação. Como tarefas que têm o objetivo realizar mediação elenco: a reunião tutorial, os questionários (inicial e final), as aulas (quando o TTD está integrado a uma disciplina de LE), as sessões de mediação e os diários de aprendizagem. Em função desse entendimento, a aba ‘Mediation’ apresentará informações sobre a realização dessas tarefas. Essas tarefas objetivam tanto promover a reflexão quanto fornecer orientação quanto a aprendizagem no contexto TTD, como discutimos quando nos referimos a mediação anteriormente, neste trabalho.

Vale destacar que a aba nomeada como ‘Frequência’ apresenta uma tabela com espaço para as mediadoras inserirem informações que têm como objetivo facilitar o gerenciamento da turma, mas também a coleta de dados sociolinguísticos sobre os participantes. Assim, a tabela contém colunas para a inserção do nome dos participantes, contato telefônico<sup>99</sup>, gênero, e-mail e proficiência. Vale lembrar que quando esse documento for inserido em um *corpus*, devem ser apagadas as informações de nome, e-mail e contato telefônico dos participantes.

---

<sup>99</sup> O contato telefônico solicitado costuma ser o número de celular, para entrar em contato mais célere com algum participante, e/ou para criação de grupos de WhatsApp, por exemplo. O grupo de WhatsApp de uma turma TTD serve tanto para as mediadoras, quanto aos participantes para divulgação de informações referente a turma ou avisos de alteração, ausência por parte dos participantes, etc.

**Figura 26 - Cenário de aprendizagem – Aba 2**

FAIR		STUDENT NAME	Phone#	Email	IT	COURSE	PROFICIENCY	Gender (M/F)	TOSs Dates				
1													
2													
3													
4													
5													

**Fonte: Elaborado pela autora em parceria com Laura Rampazzo – Salvo no Google Drive do grupo TTD de São José do Rio Preto**

As colunas intituladas *TOSs (Teletandem Oral Sessions) dates* estão destinadas ao registro da frequência dos participantes. As mediadoras alimentam esse documento a cada SOT, ficando aberto na máquina utilizada pela mediadora, principalmente para a inserção das ocorrências, que devem ser registradas na terceira aba.

A aba das SOTs está destinada a inserção de informações detalhadas sobre as sessões, a saber: o tipo de discurso (conversação livre, discussão de temas, realização de tarefa), a tecnologia utilizada para a comunicação, se foi prevista troca de texto (quantos serão trocados e quando se iniciará a troca), se durante as sessões deve ocorrer o monolingualismo alternado ou a intercompreensão e ainda o detalhamento do que foi planejado para cada SOT ao lado das ocorrências de cada dia, como apresento na figura abaixo.

**Figura 27 - Cenário de aprendizagem – Aba 3**

Session typology		Intercomprehension	Communication scenario	VoIP technology use:
Text exchange	Alternate monolingualism		Discourse	WhatsApp:
	Expected:			Free conversation:
	Beginning:			Specific theme discussion:
	Quantity:			Perform a task:
TOSs				
TOS	Date	WORK PLAN		OCCURENCES
1				
2				
3				
4				
5				

**Fonte: Elaborado pela autora em parceria com Laura Rampazzo – Salvo no Google Drive do TTD de São José do Rio Preto.**

A visualização da tabela dessa aba permite ao leitor a percepção de que a inserção das ocorrências está na mesma linha do planejamento, de modo que facilite tanto o trabalho da mediadora quanto a compreensão do contexto pelo pesquisador que vá usar os dados produzidos nesse cenário. A aba seguinte do documento está destinada a mediação.

**Figura 28 – Cenário de aprendizagem – Aba 4**

MEDIATION						
Tutorial:	Date:					
	Observation:					
Learning diary	Expected #	Observations:				
Initial Questionnaire	Date:	Answered:				
Final Questionnaire	Date:	Answered:				
Mediation Sessions	Expected #	Environment	Synchronous	Instruments:		
	Done #		Asynchronous			
MEDIATION SESSIONS						
Session	DATE	DESCRIPTION			OBSERVATIONS	
1						
2						
3						
4						
5						

Fonte: Elaborado pela autora em parceria com Laura Rampazzo – Salvo no Google Drive do TTD de São José do Rio Preto

As informações na quarta aba estão relacionadas à mediação realizada no TTD. Desse modo, haverá inserção de informação quanto ao tutorial, diário de aprendizagem, questionários inicial e final e as sessões de mediação.

Referente ao tutorial deve ser registrado a data de realização, havendo espaço para a inserção de alguma observação que a mediadora entender como pertinente para a compreensão do contexto de aprendizagem. No item diário de aprendizagem, a mediadora informa quantos estão previstos serem escritos, lembrando que para cada SOT deve ser produzido um diário pelo participante. Quanto ao que se refere aos questionários (inicial e final) solicita-se que a mediadora informe a data de solicitação de preenchimento desses documentos, bem como a quantidade de respostas recebidas<sup>100</sup>.

A respeito da sessão de mediação, a mediadora deverá informar quantas estão previstas para o período de realização da turma, quantas foram realizadas, bem como se foi utilizada

<sup>100</sup> Como percebemos nos dados do Banco TTDii, mesmo em turmas de teletandem institucional integrado no qual há aferição de nota pela realização das tarefas do TTD os questionários foram os dados salvos menor quantidade. Com a inserção da informação de quantas respostas foram recebidas pela mediadora, caso haja quantidade incompatível de questionários o pesquisador terá certeza de que não foram respondidos. Não podemos descartar a possibilidade de, no caso do Banco TTDii, não termos alguns dados no por não terem sido salvos.

alguma ferramenta assíncrona para a realização da sessão de mediação, como por exemplo, o uso do Google Groups<sup>101</sup>, que viabiliza a criação de fóruns nos quais a mediadora pode inserir uma questão reflexiva e os participantes são convidados a inserir suas reflexões a respeito. O lado positivo dessa ferramenta é o registro escrito, no entanto, diferentemente da sessão de mediação presencial em que a mediadora pode conversar livremente, logo após uma SOT, com os participantes e temas podem surgir e serem discutidos imediatamente, na discussão via fórum, algumas questões podem ser esquecidas pelo participante e sequer serem citadas.

As demais linhas da tabela estão destinadas às sessões de mediação *per si*, o detalhamento do que fora planejado e as observações quanto ao que ocorreu no decorrer de cada sessão. Nesse espaço, a mediadora pode, por exemplo, registrar que a mediação planejada não ocorreu porque os participantes não puderam ficar para participar.

O cenário de aprendizagem contém duas abas para inserção de informações referentes a presença dos participantes e do espaço de interação. No entanto, apenas essas duas abas não forneceriam os metadados suficientes para a compreensão do contexto de produção dos dados. Assim, foram adicionadas mais duas abas referentes à: mediação e SOT, como discutimos acima. A separação dessas duas novas abas se deu em decorrência da compreensão de que há uma relação entre tarefas previstas no TTD a partir de seu propósito: a) tarefas que objetivam estabelecer espaço de aprendizagem através da interação virtual, que categorizei como tarefas da mediação, e b) tarefas que objetivam auxiliar o participante no processo de aprendizagem telecolaborativo proposto no TTD, categorizadas aqui como tarefas da interação.

Discuto a seguir a proposta de coleta que apresentamos ao grupo TTD de São José do Rio Preto a partir do que experienciei na organização do MulTeC.

### 3.7 A PROPOSTA DE COLETA AO TTD

A proposta de coleta foi elaborada tendo como ponto de partida os desafios encontrados para a construção do MulTeC bem como o ciclo de vida dos dados discutido nesta tese, pois o planejamento e execução de algumas fases do ciclo precisam ser considerados desde o momento da elaboração do plano de gerenciamento dos dados de pesquisa, como discutido por Briney (2015), Eynden (2013) e Eynden *et al* (2011). É um resultado da experiência de organização dos dados coletados de 2012 a 2015 e tem relevância levando em consideração o contexto TTD, cuja produção de dados por semestre é ampla.

---

<sup>101</sup> Para mais informações sobre o Google Groups acessar: <https://groups.google.com/forum/#!overview>

Embasada na experiência de organização dos dados para a construção do MulTeC, apresentei uma proposta de coleta constituída por 14 procedimentos que, sugiro, sejam lembrados antes do início do processo de coleta, considerando que os primeiros demandarão planejamento e definições de responsabilidades na equipe que almeja realizar a coleta dos dados. A seguir a figura do texto do protocolo de coleta apresentado a equipe TTD de São José do Rio Preto.

**Figura 29 – Protocolo de coleta apresentado ao grupo TTD de São José do Rio Preto**

- 1) Definir os responsáveis pela coleta. Ideal que sejam duas pessoas minimamente por turma TTD. Uma delas encarregada da revisão dos procedimentos de coleta.
- 2) Estabelecer nome da turma de acordo com a ordem de ocorrência no ano; (UGA1; IO1; ...)
- 3) Definir locais de armazenamento dos dados salvos, preferencialmente em três locais (HDs externos e um computador, por exemplo);
- 4) Enviar os TCLEs por e-mail para os interagentes;
- 5) Coletar assinatura dos TCLEs até a SOTi, salvando-os no Drive nas pastas das respectivas turmas;
- 6) Criar o IT dos participantes; (**Universidade**, **curso**, **gênero**, **usuário Skype – I9F6**) que autorizaram o uso dos dados, criando uma lista de conferência;
- 7) Criar pastas para salvar os dados (textos, diários, chats) no Drive. A pasta para as SOTs deve ser criada na área de trabalho (pasta para as SOTs). No encerramento das atividades da turma, o organizador deve baixar as pastas nomeando-as com os ITs dos participantes (I9F6\_U0M12, por exemplo);
- 8) Realizar a coleta semanalmente dos arquivos de texto. Os arquivos de vídeo devem ser averiguados semanalmente quanto a sua qualidade, objetivando não organizar vídeos que não serão utilizados dada a baixa qualidade de áudio, por exemplo;
- 9) Nomear seguindo o padrão de nomeação dos dados; (ano\_IT\_TURMA\_Dado)  
Ex.: 2017\_I9F18\_UGAII\_C1
- 10) Acompanhar o envio dos questionários;
- 11) Organizar os dados nas respectivas pastas por turma;
- 12) Estar atento ao preenchimento das informações nos cenários de aprendizagem e pedagógico, procurando sanar dúvidas a respeito junto a mediadora;
- 13) Por ocasião do encerramento da turma, utilizar padrões de anonimização estabelecidos no MulTeC;
- 14) Baixar arquivos já em .txt e/ou .pdf a depender do tipo de texto produzido, em havendo imagens, salvar nos dois formatos.

**Fonte:** Elaborado pela autora para o grupo de pesquisa TTD de São José do Rio Preto

As orientações elencadas no protocolo de coleta proposto visam, portanto, minimizar ou mesmo dirimir quaisquer prejuízos nos dados, havendo o máximo de aproveitamento, bem como realizando um controle de qualidade nos dados produzidos e autorizados para uso em pesquisa pelos participantes. Os procedimentos 1 a 4 devem ser realizados no período de planejamento da turma TTD que terá seus dados coletados. O intuito dessa preparação é procurar facilitar o trabalho de coleta, diminuindo a possibilidade de perdas de dados. A ordem

dos procedimentos deve ser seguida, assim, tão logo haja a coleta das assinaturas dos TCLEs, os mesmos devem ser escaneados e salvos na respectiva pasta nos locais definidos para arquivamento do *corpus* como discute Briney (2015). No entanto, esses procedimentos terão o efeito esperado se houver uma equipe responsável por realizar o processo, havendo pelo menos duas pessoas responsáveis, revisando seus trabalhos mutuamente, em colaboração.

A criação dos ITs por participante segue como procedimento seguinte, pois, após a coleta dos dados, os responsáveis sabem que participantes autorizaram o uso dos dados em pesquisa, coletando apenas os dados produzidos por eles. Essa prática visa diminuir a quantidade de espaço ocupado sem necessidade com dados sem autorização, além de reduzir o trabalho dos responsáveis pela coleta, que, como esclareço na proposta, devem ser duas pessoas por turma TTD de modo que uma possa revisar o trabalho da outra, podendo haver divisão de atividades nesse processo, como por exemplo, um coleta os vídeos e as planilhas elaboradas pelas mediadoras e o outro coleta os arquivos de texto produzidos pelos participantes. Como sugiro na proposta, essa coleta deve ocorrer semanalmente, no intuito de, caso surja alguma dúvida quanto ao dado, o responsável possa entrar em contato com a mediadora e dirimir tal dúvida quanto ao dado produzido.

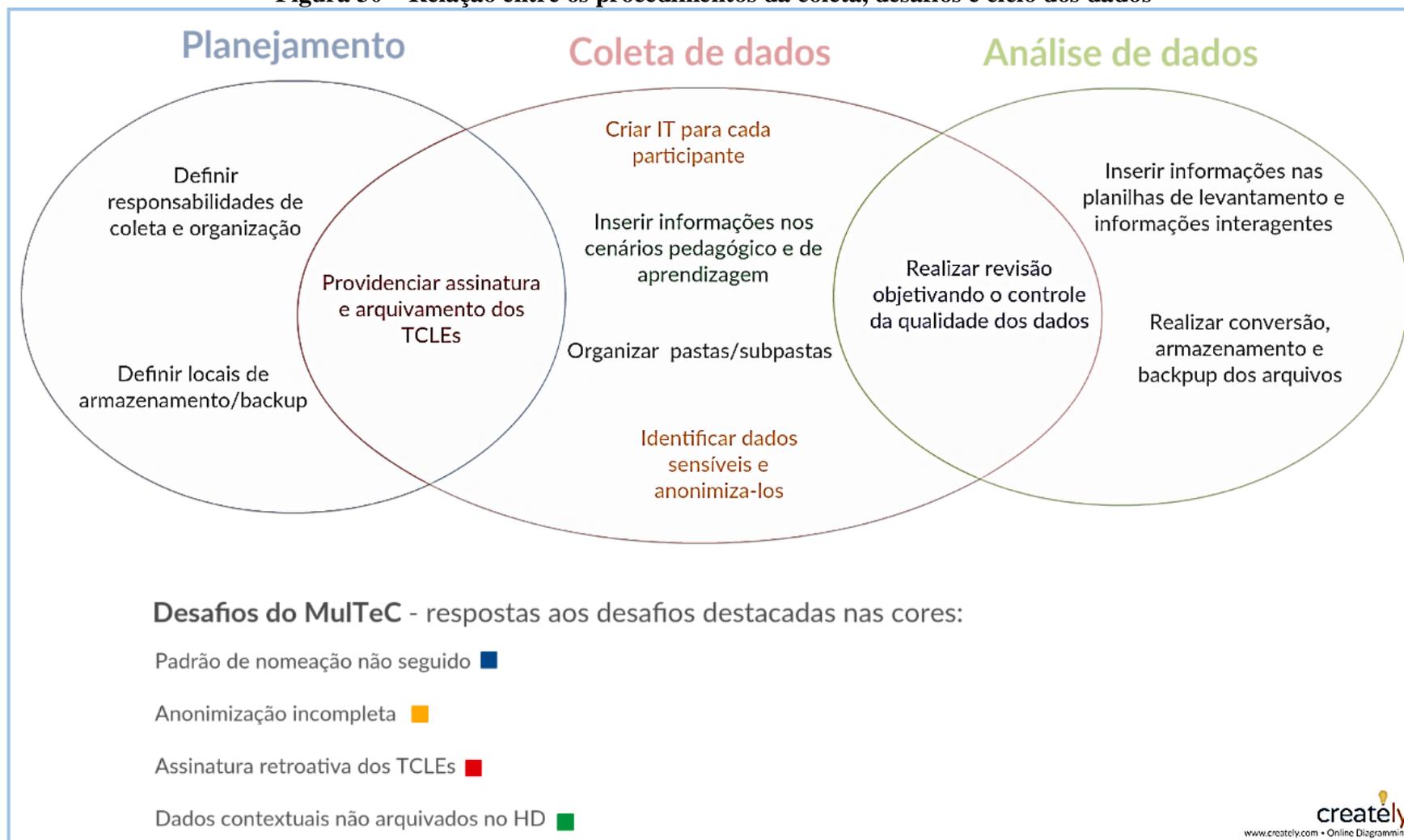
Referente a nomeação, apenas relembro na proposta o quanto é importante manter a consistência na nomeação dos dados. Como essa atividade é feita ainda manualmente, é salutar que uma outra pessoa a revise, como afirmei anteriormente nesta seção. Os procedimentos referentes a organização também devem ser realizados semanalmente, facilitando a realização do trabalho.

A orientação de anonimizar os textos apenas quando concluído o período de realização da turma TTD se deve ao fato de que, durante o processo os participantes e mesmo a mediadora, podem precisar acessar um desses documentos e a coleta não deve intervir no percurso de aprendizagem do participante ou sequer no trabalho da mediadora. Essa intervenção pode ocorrer pela substituição de um nome pelo IT quando a mediadora não tem domínio dos códigos utilizados na anonimização do MulTeC e talvez não tenha interesse em conhecer cada um deles, por ventura em decorrência da complexidade do contexto e das relações entre os dados produzidos.

Encerrada a anonimização, como os dados de textos estão sendo produzidos no Google Drive®, os responsáveis pela coleta podem realizar o download dos arquivos na formatação ‘texto sem formatação’ e arquivá-los no local apropriado.

As orientações apresentadas nesta proposta de protocolo de coleta estão relacionadas às três primeiras fases do ciclo de vida dos dados supracitado e essa relação procurei representar em uma visualização na figura a seguir.

**Figura 30 – Relação entre os procedimentos da coleta, desafios e ciclo dos dados**



Fonte: Elaborado pela autora

Os procedimentos propostos para a coleta procuram fazer com que os problemas encontrados para a construção do MulTeC sejam minimizados, havendo a necessidade de revisão de cada um dos procedimentos por uma segunda pessoa, pois, como a produção dos dados é fecunda no TTD, uma única pessoa responsável pela coleta, mesmo coletando semanalmente, poderá deixar de seguir algum procedimento ou padrão estabelecido. Os procedimentos que na figura estão na cor automática (preto) referem-se a questões que não representaram grande desafio para a construção do MulTeC e/ou que são resultado da inserção de novos documentos como os cenários de aprendizagem e pedagógico. Tais documentos não eram previstos antes deste trabalho, logo um procedimento que enfatize a relevância do preenchimento das informações neles não podia estar listado. Seguidos os procedimentos apresentados na proposta, o pesquisador que precisar usar os dados, poderá dedicar mais tempo a análise e elaboração do texto de apresentação dos seus resultados de pesquisa.

## CONSIDERAÇÕES FINAIS

Nesta parte do trabalho, apresento uma breve retomada dos objetivos que conduziram a realização desta pesquisa, passando então a discutir as limitações do trabalho e encerrando com os encaminhamentos futuros a partir do que foi detectado durante o processo de realização da pesquisa e escrita deste texto.

## OBJETIVOS E RESULTADOS ALCANÇADOS

Iniciei este trabalho com a hipótese de que seria possível construir um *corpus* multimodal em teletandem com as produções coletadas no TTD da Unesp de São José do Rio Preto, a partir do levantamento, anonimização e organização dos dados produzidos no período de 2012 a 2015. Para alcançar a realização desse objetivo, precisei dos seguintes objetivos específicos: a) investigar a organização dos dados; b) padronizar a nomeação e anonimização e c) organizar e analisar os dados. No percurso houve necessidade de seguir os padrões que estabeleci quando da elaboração do plano de gerenciamento dos dados. No plano de gerenciamento de dados (Anexo 7), defini como os dados deveriam ser organizados e anonimizados, estabelecendo os padrões de nomeação, inclusive.

Cito como resultados, além do próprio MulTeC, sua proposta de organização de dados para o TTD, a proposta de uso de documentos que representem o cenário pedagógico e o cenário de aprendizagem. Esses documentos se configuram como relevantes para a organização de *corpus* de coleta no TTD da Unesp de São José do Rio Preto, pois documentos fornecerão informações indispensáveis para a compreensão do contexto de produção dos dados. Como informações necessárias para a construção do contexto no *corpus* tais documentos devem apresentar: i) a modalidade, ii) a turma, iii) os professores e os mediadores, iv) registro das ocorrências em cada SOT; v) registro sobre fatos ocorridos durante a coleta que possam influenciar na apresentação dos dados no *corpus*; e vi) nome da turma, que seguindo os padrões estabelecidos para o MulTeC, vai informar a modalidade de TTD na qual a turma está ocorrendo.

Os dados indicaram que os tipos de dados produzidos no teletandem dependem mais das decisões tomadas durante o planejamento da turma (cenário pedagógico) do que da modalidade na qual o mesmo ocorre.

## LIMITAÇÕES DA PESQUISA

Encontrei entraves que impediram a conclusão da construção do MulTeC, conforme as intenções expressas em meus projetos iniciais. Elenco como percalços: i) o tempo, considerando que transcrever mais que 500 horas de vídeo não pode ser realizado no âmbito temporal de um doutorado e ii) a inexistência de algumas informações, sobretudo no tocante a ocorrências durante as SOTs, haja vista que não houve registro na maioria das turmas, o que levou, por exemplo a exclusão de um ano de coleta de dados realizada em 2011.

A ausência de documentos com registros sobre o contexto de produção mais detalhado tornou-se um obstáculo para o preenchimento de todas as informações das instituições constituintes da parceria em algumas planilhas, como no caso da planilha que apresenta as informações sobre os participantes (idade, gênero, curso), pois, no que se refere aos participantes da universidade estrangeira que fez parceria com a Unesp, cujos dados foram utilizados na construção do banco, não foi possível coletar informações sobre o curso e a idade, pois não havia documento com tal registro.

## ENCAMINHAMENTOS FUTUROS

Os dados sugerem a necessidade de ampliação dos dados incluindo mais tipos de dados, como a reunião tutorial, que pode ser gravada em áudio ou vídeo, bem como as sessões mediação. Essa necessidade advém da compreensão de que ambos constituem tanto o cenário pedagógico quanto o de aprendizagem, podendo influir nas produções dos participantes no TTD. Ainda apresento como sugestão, a partir deste trabalho, uma ampliação do questionário inicial. Entendo que esse documento pode servir para coletar informações sociolinguísticas que contribuirão na construção do perfil sociolinguístico do participante sem descuidar da preservação de sua identidade, podendo haver a inserção da informação de tempo de estudo da língua alvo, por exemplo, bem como quantas línguas fala/estuda. No contexto do TTD questionar apenas sobre L1 e L2 me parece um tanto quanto reduzido considerando que, nos dados utilizados para este trabalho, mais especificamente nas SOTs que precisei realizar oitiva, identifiquei falantes de língua inglesa que informavam aos parceiros que sabiam falar espanhol além do português, língua-alvo para eles na parceria em questão. Outro levantamento relevante é a idade do participante, além da data de nascimento, para que aqueles que estejam responsáveis pela coleta e organização dos dados não precisem realizar o cálculo, como precisamos fazer para este trabalho, considerando que esse é um fator relevante na análise dos

dados pelos pesquisadores que podem, inclusive, optar por selecionar dados produzidos por um faixa etária definida.

Considerando que a participação dos aprendizes pode ocorrer em mais de um semestre, como detectei nos dados, o ideal na anonimização seria a utilização de um IT único por participante. Ao participar pela primeira vez do contexto TTD seria criado o IT seria utilizado em todos os dados produzidos por ele quando participasse do TTD. A pergunta no questionário inicial sobre a participação em oportunidade anterior no teletandem seria o ponto de partida, para a equipe responsável pela coleta, procurar o IT criado para aquele participante e então nomear os dados produzidos por ele com o mesmo IT. Uma proposta de IT único poderia ser, por exemplo, constituída pelo ano que participou pela primeira vez, instituição, semestre, curso, usuário Skype e gênero declarado: 2018I127S13F – participante declarada feminina, cursando o doutorado em estudos linguísticos, participou pela primeira vez em 2018 como aluna do Ibilce, no primeiro semestre, utilizando o usuário Skype 13. Para registro e para que sequenciação dos ITs criados tenha consistência, o ideal seria a elaboração e manutenção de uma planilha com os ITs criados com: nome, curso, data de nascimento e e-mail do participante de modo que possa ser facilmente localizado quando for realizar TTD novamente.

A análise dos dados ainda aponta para a necessidade de uma equipe responsável pela coleta, havendo pelo menos alguém da área de TI e uma coordenadora do processo de coleta e organização dos dados no intuito de manter a consistência do padrão de anonimização e nomeação dos dados. O papel dessa pessoa, que estará responsabilizada pela compreensão do plano de gerenciamento de dados e por quem estiver responsável pela coleta e organização, deveria consistir na orientação e revisão de cada dado coletado, observando se está se seguindo o padrão convencional, bem como procurar dirimir dúvidas durante o processo de gerenciamento de dados por aqueles que estão responsáveis pela coleta.

Outra questão relevante a ser considerada é que os documentos produzidos para esta pesquisa, como as planilhas, cenário pedagógico e de aprendizagem<sup>102</sup>, não terão qualquer utilidade se não forem criteriosamente preenchidos pelas mediadoras/coordenadoras do TTD que integram a equipe de modo que se consiga levantar os dados contextuais indispensáveis para a construção de *corpus* de pesquisa que atenda aos pesquisadores interessados em investigar o TTD. Gostaria de destacar que os dados que constituem o MulTeC podem ser de

---

<sup>102</sup> Foi disponibilizado para o grupo de pesquisa em TTD da Unesp de Rio Preto um modelo de cada um desses documentos. No caso das planilhas, já com as fórmulas, de maneira que quem for utilizá-las não precise dedicar tempo a esse trabalho, apenas precise inserir as informações coletadas.

interesse não apenas para cientistas que pesquisam a aprendizagem de línguas em projeto telecolaborativo, mas investigações que discutam:

- i) Interculturalidade
- ii) Identidade
- iii) Escrita de textos em língua estrangeira
- iv) *Feedback*
- v) Processo autorreflexivo da aprendizagem (a partir dos registros nos diários de aprendizagem)
- vi) Uso de ferramentas digitais para a intercompreensão;
- vii) Construção da aprendizagem autônoma.

Desse modo, acredito que envidar tais esforços deva contribuir para o fortalecimento das pesquisas em TTD em âmbito internacional, isso considerando o direcionamento que a ciência no mundo tem dado para o compartilhamento de dados de pesquisa objetivando tornar cada vez mais comum a *Open Science*.

## REFERÊNCIAS

- ANDREU-FUNO, L. *Teletandem: um estudo sobre identidades culturais e sessões de mediação da aprendizagem*. Tese de Doutorado. Programa de Pós-Graduação em Estudos Linguísticos da Universidade Estadual Paulista “Júlio de Mesquita Filho” – São José do Rio Preto, 2015.
- ASSANTE, M.; CANDELA, L.; CASTELLI, D.; TANI, A.; *Are Scientific Data Repositories Coping with Research Data Publishing?* *Data Science Journal* 15, no. 6, p. 1-24., 2016. Disponível em: <http://doi.org/10.5334/dsj-2016-006> Acesso em: 12 de jul, 2017.
- ALBAGLI, S. Open science in question. In: ALBAGLI, S.; MACIEL, M. L.; ABDO, A. H. (Org.) *Open Science, open issues*. Rio de Janeiro: Unirio, 2015. p. 9–22.
- ARANHA, S.; CAVALARI, S. M. S. A trajetória do Projeto Teletandem Brasil: Da modalidade institucional não-integrada à institucional integrada. *The ESpecialist*, v. 35, p. 183–201, 2014.
- ARANHA, S. Os gêneros na modalidade de teletandem institucional integrado: a primeira sessão de interação. *Gêneros de texto/discurso e os desafios da contemporaneidade*. [S.l.: s.n.], 2014. .
- ARANHA, S.; LEONE, P. DOTI: Databank of Oral Teletandem Interactions. *New directions in telecollaborative research and practice: selected papers from the second conference on telecollaboration in higher education*. [S.l.]: Research-publishing.net, 2016a. p. 327–332.
- ARANHA, S.; LEONE, P. DOTI: Databank of Oral Teletandem Interactions. *New directions in telecollaborative research and practice: selected papers from the second conference on telecollaboration in higher education*. [S.l.]: Research-publishing.net, 2016b. p. 327–332.
- ARANHA, S.; LEONE, P. The development of DOTI (Data of oral teletandem interaction). *Investigating Computer-Mediated Communication: Corpus-Based Approaches To Language in the Digital World*. Ljubljana: University of Ljubljana Press, Faculty of Arts, 2017. p. 172–190. Disponível em: <<https://e-knjige.ff.uni-lj.si>>.
- 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.
- ARAUJO, N. R. P. *A formação de parcerias de teletandem: da organização ao sistema de atividade*. 2012. Universidade Estadual Paulista - UNESP/ Ibilce, 2012.
- BERBER SARDINHA, T. Pesquisa em linguística de *Corpus* com WordSmith Tools. *Mercado de Letras*, p. 299, 2009. Disponível em: <[http://sis.posugf.com.br/AreaProfessor/Materiais/Arquivos\\_1/13879.pdf](http://sis.posugf.com.br/AreaProfessor/Materiais/Arquivos_1/13879.pdf)>.
- BERBER SARDINHA, T. Tamanho de *Corpus*. *the ESpecialist*, v. 23, n. 2, p. 103–122, 2002.
- BERBER SARDINHA, T. *Linguística de Corpus*. Barueri: Editora Manole, 2004.
- BOGDAN, R. C.; BIKLEN, S. K. *Investigação qualitativa em educação - Uma introdução à teoria e aos métodos*. [S.l.: s.n.], 1994.

BRAMMERTS, H. Tandem language learning via internet and international e-mail tandem network. *A guide to language learning in tandem via internet*. [S.l.]: CLCS Occasional Paper, 1996. p. 9–22.

BRINEY, K. *Data management for researchers: Organize, maintain and share your data for research success*. Exeter: Pelagic Publishing, 2015.

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

CHANIER, T. *et al.* 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. A scientific methodology for researching CALL interaction data. *Language-learner computer interactions: Theory, methodology and CALL applications*, p. 215–240, 2016. Disponível em: <<https://benjamins.com/catalog/lse.2.10cha>>.

COSCARELLI, C. V. *Tecnologias para aprender*. 1 ed ed. São Paulo: Parábola Editorial, 2016.

DELFANTI, A.; PITRELLI, N. Open science: revolution or continuity? In: ALBAGLI, S.; MACIEL, M. L.; ABDO, A. H. (Org.). *Open Science, open issues*. Rio de Janeiro: IBICT, 2015. p. 59–68.

DETOMINI, A. C. S. M. *Dúvidas Linguísticas de Aprendizizes de PLE: avaliação de proficiência oral de professores no EPPLE-PLE*. Dissertação de Mestrado. Programa de Pós-Graduação em Estudos Linguísticos. Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP, 2018.

DOOLY, M.; O’DOWD, R. Telecollaboration in the foreign language classroom : A review of its origins and its application to language teaching practice. In *This Together: Teachers’ experiences with transnational, telecollaborative language learning projects*. [S.l.: s.n.], 2018. p. 11–34. Disponível em: <<https://www.peterlang.com/view/9783034335348/chapter01.xhtml>>.

DÖRNEY, Z. *Research Methods in Applied Linguistics*. Oxford: [s.n.], 2009. v. 19. Disponível em: <<http://doi.wiley.com/10.1111/j.1473-4192.2009.00223.x>>.

ELLIS, R. *Task-based language learning and teaching*. New York: Oxford University Press, 2003.

EYNDEN, V. V. D. *Data Life Cycle & Data Management Planning*. UK Data Archive University of Essex, n. April, p. 24–25, 2013. Disponível em: <<https://www.ukdataservice.ac.uk/media/187718/dmplanningdm24apr2013.pdf>>.

EYNDEN, V. V. D. *et al.* *Managing and Sharing Data - Best Practice For Researchers*. [S.l.: s.n.], 2011. Disponível em: <<http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>>.

FINATTO, M. J. *et al.* *Linguística de Corpus: perspectivas*. Porto Alegre: Instituto de Letras - UFRGS, 2018.

FOUCHER, A. *Didactique des Langues-Cultures et Tice : scénarios , tâches , interactions*. . [S.l: s.n.], 2010.

FRESCHI, A. C. *A avaliação por pares no teletandem institucional integrado: um estudo de caso sobre o feedback linguístico nas sessões orais em português*. Dissertação de Mestrado. Programa de Pós-Graduação em Estudos Linguísticos. Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP, 2017.

FRESCHI, A. C.; LOPES, Q. B. Potenciais Sequências De Aprendizagem Intercultural No Teletandem: a Importância Da Mediação. *Revista do GEL*, v. 13, n. 3, p. 49, 2016. Disponível em: <<https://revistadogel.gel.org.br/rg/article/view/1470>>.

GUICHON, N. *Langues et TICE: méthodologie de conception multimédia*. [S.l: s.n.], 2006.

GUICHON, N. Sharing a Multimodal *Corpus* To Study Webcam-Mediated Language Teaching. v. 21, n. 1, p. 56–75, 2017.

GÜNTHER, H. Pesquisa qualitativa versus pesquisa quantitativa: esta é a questão? *Psicologia: Teoria e Pesquisa*, v. 22, n. 2, p. 201–209, 2006. Disponível em: <[http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S0102-37722006000200010&lng=pt&tlng=pt](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-37722006000200010&lng=pt&tlng=pt)>.

KNIGHT, D. The future of multimodal corpora. *RBLA, Belo Horizonte*, v. 11(2), p. 391–415, 2011. Disponível em: <[http://www.google.com/url?sa=t&rct=j&q=the future of multimodal corpora+pdf&source=web&cd=1&ved=0CFAQFjAA&url=http://www.periodicos.letras.ufmg.br/rbla/arquivos/332.pdf&ei=VRMIUJiJE4yU0QXLrqnzBA&usq=AFQjCNHeGzq-mj102Z2f9p\\_3yqEJDpaqNA&cad=rja](http://www.google.com/url?sa=t&rct=j&q=the future of multimodal corpora+pdf&source=web&cd=1&ved=0CFAQFjAA&url=http://www.periodicos.letras.ufmg.br/rbla/arquivos/332.pdf&ei=VRMIUJiJE4yU0QXLrqnzBA&usq=AFQjCNHeGzq-mj102Z2f9p_3yqEJDpaqNA&cad=rja)>.

LUVIZARI-MURAD, L. H. *Aprendizagem de alemão e português via teletandem: um estudo com base na Teoria da Atividade*. 2011. Universidade Estadual Paulista (Unesp), Instituto de Biociências, Letras e Ciências Exatas (Ibilce), Câmpus de São José do Rio Preto, 2011.

MANGENOT, F. *La question du scénario de communication dans les interactions pédagogiques en ligne*. JOCAIR 2008 (Journées Communication et Apprentissage Instrumentés en Réseau), v. 2008, n. Ea 609, p. 13–26, 2008.

O'DOWD, R. *Emerging Trends and New Directions in Telecollaborative Learning*. CALICO Journal, v. 33, n. 3, p. 291–310, 2016.

PRIMO, A. *O aspecto relacional das interações na Web 2.0*. E-Compós: Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação, v. 9, p. 1–21, 2007. Disponível em: <<http://www.compos.org.br/seer/index.php/e-compos/article/view/153>>.

RAMPAZZO, L. *Gêneros textuais e telecolaboração : uma investigação da sessão oral de teletandem inicial*. Dissertação de Mestrado. Programa de Pós-Graduação em Estudos Linguísticos. Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP, 2017.

REFFAY, C.; BETBEDER, M.; CHANIER, T. *Multimodal learning and teaching corpora exchange : lessons learned in five years by the Mulce project* Marie-Laure Betbeder Thierry Chanier. 2013.

SALOMÃO, A. C. B. *Gerenciamento e estratégias pedagógicas na mediação dos pares no*

*teletandem e seus reflexos para as práticas pedagógicas dos interagentes*. 2008. Universidade Estadual Paulista “Júlio de Mesquita Filho” - São José do Rio Preto, 2008.

SHEPHERD, T. M. G. *O Estatuto Da Linguística De Corpus: Metodologia Ou Área Da Linguística ? Matraca*, v. 16, n. 24, p. 150–172, 2009.

SINCLAIR, J. *Corpus and Text - Basic principles*. Disponível em: <<https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm>>. Acesso em: 20 out. 2017.

SOUZA, M.G. *Os primeiros contatos de professores de línguas estrangeiras com a prática de teletandem*. 2012. Universidade Estadual Paulista “Júlio de Mesquita Filho” - São José do Rio Preto, 2012. Disponível em: <<https://repositorio.unesp.br/handle/11449/93887>>.

TOLEDO, L. D. L. *Um estudo sobre o uso de vocabulário rico por aprendizes de inglês na sessão oral do teletandem institucional integrado*. Dissertação (Mestrado em Estudos Linguísticos) - Universidade Estadual Paulista Júlio de Mesquita Filho, 2017.

WIGHAM, C. R.; CHANIER, T. *LEarning and TEaching Corpora ( LETEC ): data-sharing and repository for research on multimodal interactions*. Global perspectives on Computer-Assisted Language Learning, p. 10–13, 2013. Disponível em: <<https://hal.archives-ouvertes.fr/edutice-00778274/document>>.

ZAKIR, M. de A. *Cultura e(m) telecolaboração: uma análise de parcerias de teletandem institucional*. 2015. Universidade Estadual Paulista “Júlio de Mesquita Filho” - Campus de São José do Rio Preto, 2015.



## Anexo 2 – Pareamento

					
<b>Teletandem Unesp – UGA</b> <b>2012</b> <b>Cronograma e Pareamento</b>					
<b>Modalidade</b>	Teletandem Institucional integrado				
<b>Turmas</b>	PORT2002 (USA) – Letras III (BRA)	TURMA TTD	2012_UGA2i (Quinta-feira)		
<b>Professoras</b>	Suzi e Cristiane	<b>Mediadoras:</b> Suzi			
<b>Horários</b>	Março 13h (USA) - 13h (BRA)	Abril 13h (USA) - 13h (BRA)			
<b>Mês</b>	<b>Data/Descrição</b>				
Fevereiro			14 - 18 Tutorial com alunos da UGA	20 - 24 Carnaval	27/02 – 03/03 Tutorial com alunos da Unesp
Março	05 - 09 Interação I  (Teste: alunos se conhecem) Alunos da UGA enviam redação para alunos da UNESP	12 - 16  Spring Break - UGA	19 - 23 Interação II  (Discussão do tema sugerido pela UGA) Alunos da UNESP enviam redação para alunos da UGA	26 - 30 Interação III  (Discussão do tema sugerido pela UNESP) Alunos da UGA enviam redação para alunos da UNESP	
Abril	01 - 06 Interação IV  (Discussão do tema sugerido pela UGA) Alunos da UNESP enviam redação para alunos da UGA	09 - 13 Interação V  (Discussão do tema sugerido pela UNESP) Alunos da UGA enviam redação para alunos da UNESP.	16 – 20 Interação VI  (Discussão do tema sugerido pela UGA) Alunos da UNESP enviam redação para alunos da UGA	23 - 27 Interação VII  (Discussão do tema sugerido pela UNESP) Alunos da UGA enviam redação para alunos da UNESP.	Interação VIII  Obs.: Deverá ocorrer fora do horário de aula – a qualquer momento, conforme horário mais conveniente para os parceiros.
<b>Observações:</b> Calendário igual ao grupo de terça. <b>02/21 and 02/22</b> – Test interaction: introduce yourself, get to know your partner, tell him/her about the material you are going to be working on next week; <b>02/24 and 02/25</b> – Send text in Portuguese to be revised by your partner; <b>02/28 and 01/march</b> – Interaction I: conversation about linguistic details (revised text in Portuguese) and about the material you both read/watched; your partner will tell you about the material you are going to be working on next week; <b>Two – week break</b> <b>17/03 and 18/04</b> – Receive the text in English to be revised by you; <b>03/21 and 03/22</b> – Interaction II: Conversation about linguistic details (revised text in English) and about the material you both read/watched; tell your partner about the material you are going to be working on next week; <b>03/24 and 03/25</b> – Send text in Portuguese to be revised by your partner (...)					

**Pairing**

Pair	INTERACTANTS ITs	SKYPE USERS	Course
1	I9M1	unespriopreto01	Letras
	U0F29	Teletandem29	X
2	I9F2	unespriopreto02	Letras
	U0M12	teletandem12	X
3	I9F3	unespriopreto03	Letras
	U0F13	teletandem13	X
4	I9M4	unespriopreto04	Letras
	U0F14	teletandem14	X
5	I9M5	unespriopreto05	Letras
	U0F15	teletandem15	X
6	I9M6	unespriopreto06	Letras
	U0F16	teletandem16	X
7	I9M7	unespriopreto07	Letras

## Anexo 3 – Tutorial

### Institutional Teletandem

# Teletandem partnership between UNESP and UGA students: a first talk



teletandem brasil  
línguas estrangeiras para todos

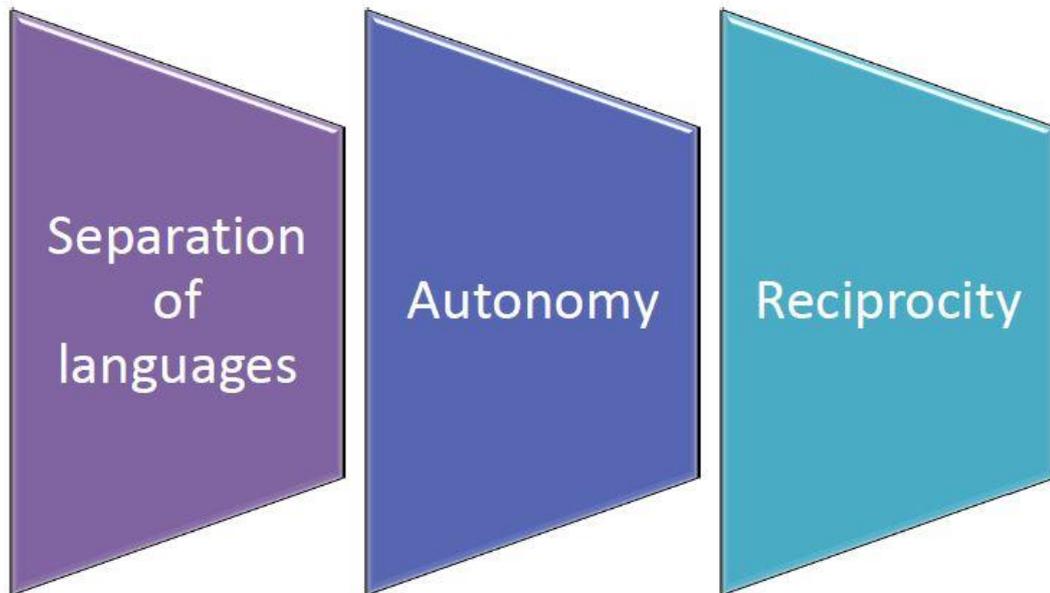


What is Teletandem (TTD)??



**TTD involves pairs of native speakers of different languages who meet periodically via chat, audio and videoconferencing tools (*Skype, Oovoo, Windows Live Messenger*) in order to teach-learn each other's language.**

TTD practice relies on three principles:



*What does that mean?*



## In practice, what is going to happen?

Semana/ Mês	1ª	2ª	3ª	4ª	5ª
Fev.			14 - 18 Tutorial com alunos da UGA	20-24 <b>Carnaval</b>	27/02-03/03 Tutorial com alunos da UNESP
Março	05-09 <b>interaction I</b> (teste: alunos se conhecem)  alunos da UGA enviam redação para alunos da UNESP	12-16 <b>Spring break - UGA</b>	19 - 23 <b>interação II</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	26 - 30 <b>interação III</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	
Abril	01-06 <b>interação IV</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	09-13 <b>interação V</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	16-20 <b>interação VI</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	23-27 <b>interação VII</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	Obs: a <b>interação VIII</b> deverá ocorrer fora do horário da aula - a qualquer momento, conforme o horário mais conveniente para os parceiros

Semana/ Mês	1ª	2ª	3ª	4ª	5ª
Fev.			14 - 18 Tutorial com alunos da UGA	20-24 <b>Carnaval</b>	27/02-03/03 Tutorial com alunos da UNESP
Março	05-09 <b>interaction I</b> (teste: alunos se conhecem)  alunos da UGA enviam redação para alunos da UNESP	12-16 <b>Spring break - UGA</b>	19 - 23 <b>interação II</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	26 - 30 <b>interação III</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	
Abril	01-06 <b>interação IV</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	09-13 <b>interação V</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	16-20 <b>interação VI</b> (discussão do tema sugerido pela UGA)  alunos da UNESP enviam redação para alunos da UGA	23-27 <b>interação VII</b> (discussão do tema sugerido pela UNESP)  alunos da UGA enviam redação para alunos da UNESP	Obs: a <b>interação VIII</b> deverá ocorrer fora do horário da aula - a qualquer momento, conforme o horário mais conveniente para os parceiros

•03/06 (Tue) and 03/08 (Thu) – test interaction: introduce yourself, get to know your partner; tell him/her about the material you are going to be working on next week

• up to 03/15 – send text in Portuguese to be revised by your partner

•03/20 and 03/22 – interaction I : conversation about linguistic details (revised text in Portuguese) and about the topic addressed in the text ;

• up to 03/22 – receive text in English to be revised by you

•03/27 and 03/29 – interaction II : conversation about linguistic details (revised text in English) and about the material you both read/watched; tell your partner about the material you are going to be working on next week

Besides a language learning opportunity...

- TTD is also a research project that investigates how such an arrangement can help foreign language students develop their linguistic abilities. For this reason, it's important also for your teachers that both you and your partner:
  - record the interactions;
  - save the comments you make on the chat;
  - organize and keep the texts or any other material you exchange.
- The data collected will be used ethically and responsibly by the researchers involved in the project as long as you agree to do it.

**We hope you...**

- learn a lot;
- enjoy the experience;
- keep practicing after April 27.

### Anexo 4 - Tabela de Cursos Unesp

Curso	Nível	Número
Ciências Biológicas	Graduação	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 diurno		9
Licenciatura em Letras noturno		10
Pedagogia		11
Biofísica Molecular	Mestrado	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	Mestrado Profissional	23
Biofísica molecular	Doutorado	24
Biologia Animal		25
Engenharia e Ciência de Alimentos		26
Estudos Linguísticos		27
Genética		28
Letras		29
<b>Matemática</b>		30
Microbiologia		31
Química		32

## Anexo 5 - Contexto TTD

### O CONTEXTO TELETANDEM DE APRENDIZAGEM

Contexto telecolaborativo de aprendizagem, o teletandem, tem o intuito de oportunizar aprendizagem de línguas assim como se constitui um ambiente de pesquisa. Os princípios básicos são: a reciprocidade, a autonomia e a separação de línguas. Esses princípios objetivam nortear os interagentes para o melhor aproveitamento do tempo de interação.

1 - Seguindo o princípio da *reciprocidade* o interagente deve auxiliar seu parceiro a aprender a língua alvo conforme suas necessidades e decisões, respeitando a autonomia do parceiro.

2 - A *autonomia* do interagente no teletandem visa garantir que os objetivos e o modo como deseja aprender a língua sejam respeitados. O interagente é quem deve decidir como e o que deseja aprender sobre a língua/cultura alvo.

3 - A *separação de línguas* objetiva a divisão igualitária do tempo dedicado a cada língua alvo. Assim, em uma sessão oral de teletandem que deverá durar 50 minutos, os parceiros devem dedicar 25 minutos para cada uma das línguas alvo.

No intuito de aprender línguas nesse ambiente o aprendiz participa impreterivelmente em tarefas de duas naturezas: de mediação e de interação.

As tarefas de natureza da mediação têm como característica primordial: suscitar a reflexão e auxiliar na tomada de decisão pelos interagentes em seu processo de aprendizagem. Nessas tarefas os envolvidos são mediadoras e interagentes. As mediadoras são professoras de língua, ou estudantes de graduação ou pós-graduação, que são responsáveis por auxiliar<sup>103</sup> as turmas<sup>104</sup> de interagentes nos períodos de realização das sessões orais de teletandem. Vale destacar que, No MulTeC, os interagentes são estudantes da Unesp – Universidade Estadual Paulista “Júlio de Mesquita Filho” – Campus de São José do Rio Preto (falantes de Português) e da UGA - University of Georgia (falantes de Inglês) que pretendem aprender um a língua do outro e possuem proficiência tal que os possibilita manter uma conversa de, pelo menos, 30 minutos na língua alvo.

---

<sup>103</sup> O auxílio pode ser de ordem técnica- para sanar questões referentes a problemas com áudio ou vídeo, substituição de máquinas – como pode estar voltado à questões próprias do teletandem como questões interculturais, dificuldades linguísticas e acompanhamento da realização das tarefas propostas para a turma.

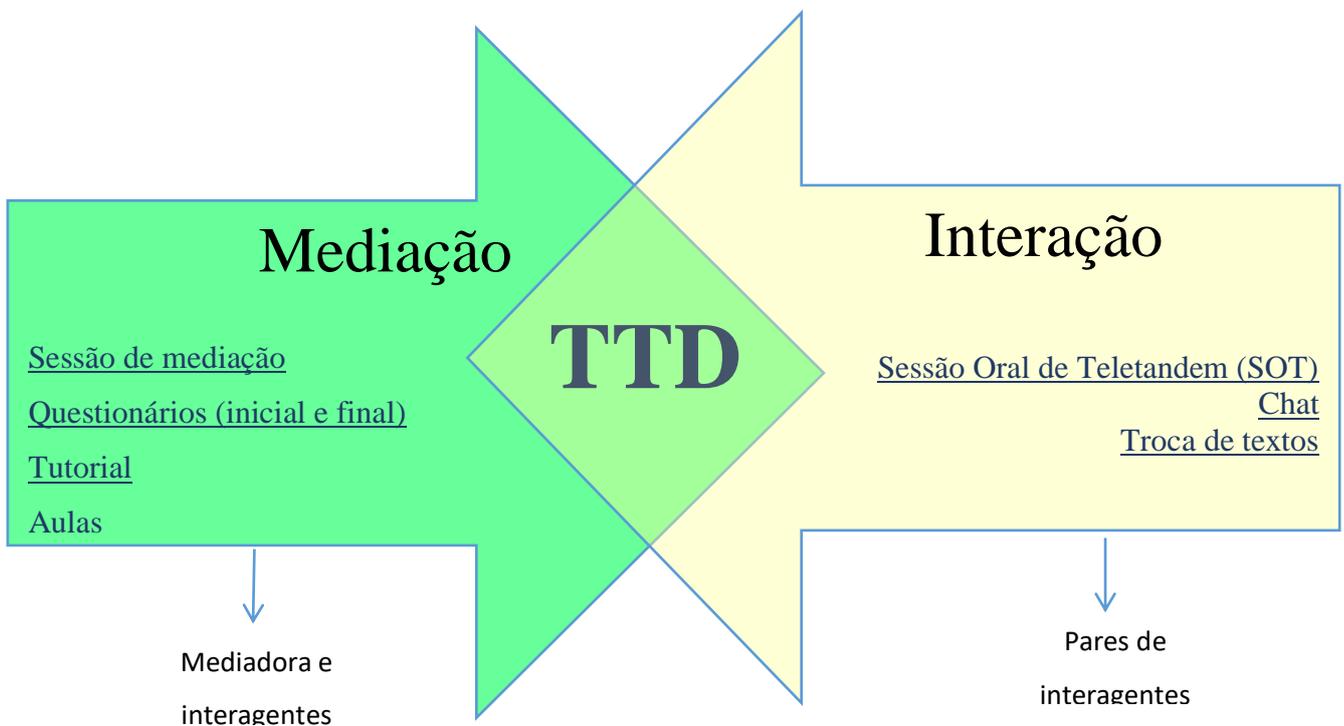
<sup>104</sup> Nomeamos ‘turma’, no MulTeC, o grupo de interagentes (tanto da universidade brasileira quanto da universidade estrangeira) que se reúne em determinado local e horário para, semanalmente, realizar uma sessão oral de teletandem via ferramentas síncronas de comunicação *on-line*, contando com o auxílio de, pelo menos, uma mediadora.

As tarefas de mediação podem ser, a depender da modalidade e do planejamento das mediadoras: o tutorial, os diários de aprendizagem, os questionários (inicial e final) e a sessão de mediação.

O chat, a troca de textos e a Sessão Oral de Teletandem (SOT) estão circunstritas nas tarefas de natureza da interação. Essas tarefas consistem, primordialmente, na telecolaboração via ferramentas síncronas de vídeo para aprender a língua do parceiro. A troca de textos é realizada através de envio de texto produzido na língua alvo para o par com quem foi designado para interagir no período, discutindo essa produção textual durante a SOT que sucede o envio do texto. Procuramos representar na figura abaixo a relação entre as tarefas do teletandem a partir de sua natureza na constituição do que atualmente é o teletandem.

Obs.: Na figura, pressionando a tecla *control* e clicando na tarefa, o leitor será direcionado a uma breve descrição da tarefa.

**Figura 1 - O Teletandem e suas tarefas**



A cronologia das microtarefas<sup>105</sup> pode variar conforme a modalidade de teletandem, se integrado ou semi-integrado<sup>106</sup>, e ainda de acordo com o planejado para cada turma de TTD.

Em trabalho de 2002, Brammerts previu modalidades para o tandem. Essas modalidades se aplicam ao teletandem a saber:

1 – Teletandem institucional - ambos os parceiros possuem um vínculo com instituição de ensino que promove e reconhece o teletandem como espaço de interação que oportuniza aprendizagem de línguas estrangeiras. O TTD institucional pode ocorrer como:

- a) Integrado – Reconhecido pela instituição, integrado a um curso, tendo caráter obrigatório.
- b) Semi-integrado – Apenas em uma das instituições o TTD é parte integrante do currículo, como obrigatória. Na outra instituição da parceria o TTD ocorre como opcional e os interagentes são voluntários que participam das atividades do TTD em contra turno às suas atividades acadêmicas relacionadas ao seu curso de graduação/pós-graduação.
- c) Não-integrado – Há apoio institucional para a realização do TTD pelos alunos da instituição, mas não há reconhecimento oficial, não havendo vinculação com um curso.
- d) Complementar – Pode ser escolhido dentre as iniciativas opcionais e, neste caso, é reconhecido pela instituição como parte integrante do curso.
- e) Opcional – Não é considerado parte do curso, se apresenta como uma opção reconhecida pela instituição.

2 – Teletandem não-institucional – os parceiros não possuem vínculo com instituição de ensino;

3 – Teletandem semi-institucional – um dos parceiros possui o vínculo com instituição de ensino.

Os dados que constituem o MulTeC foram produzidos na modalidade de teletandem institucional integrado e teletandem institucional semi-integrado. Nessas modalidades as tarefas são estabelecidas no momento do planejamento da turma TTD, quando as mediadoras chegam a um consenso quanto a quantidade de interações, de participantes e prazos para a realização das tarefas.

No período em que as coordenações preparam a realização de uma turma TTD podem ocorrer: videoconferências, emails de planejamento e negociação de tarefas e prazos, elaboração de

---

<sup>105</sup> Todas as palavras destacadas em azul são links que levam o leitor para figura ou modelo da tarefa.

<sup>106</sup> Há outras modalidades de teletandem, no entanto, no MulTeC, há apenas as modalidades de teletandem institucional integrado (TTDii) e teletandem institucional semi-integrado (TTDisi).

listas de presença, listas de pareamento e email de apresentação, cuja descrição breve se segue abaixo.

### **Emails de negociação/planejamento**

Durante o período de planejamento de realização de uma turma TTD, as coordenações do teletandem nas univesidades parceiras (no caso do MulTeC Unesp – Rio Preto e UGA) trocam emails com a finalidade de discutir prazos, tarefas e quantidade de sessões orais de teletandem a ocorrerem para a turma. Caso percebam que houve algum ruído na comunicação, as coordenações costumam marcar uma videoconferência.

### **Videoconferência**<sup>107</sup>

Normalmente só é utilizada quando a troca de emails não está fluindo para que os acordos sejam concluídos. Durante a videoconferência são esclarecidas as dúvidas.

### **Listas de pareamento**

A elaboração de uma lista de pareamento ocorre caso as coordenações tenham chegado a um acordo de que, para aquela turma, haverá pareamento prévio. Essa lista então, acrescida do calendário para a turma, constitui o documento nomeado de “lista de pareamento” no MulTeC. Há turmas em que o pareamento ocorre aleatoriamente, não aparecendo no referido documento a lista apresentando a formação dos pares. Quando as coordenações optam pelo pareamento aleatório os pares são formados a partir da escolha da cabine no primeiro dia de sessão oral de teletandem. Isso significa que, o número da cabine no primeiro dia se manterá até o final do período de interações da turma. O número da cabine corresponde ao número de usuário Skype que será utilizado para as sessões orais daquele período. As listas de pareamento podem aparecer tanto em uma turma de TTD integrado quanto de semi-integrado. Vale ressaltar que, quando o cenário é de uma turma de TTD semi-integrado há um período de inscrições<sup>108</sup>, sendo selecionados para participar da turma aqueles que primeiro se inscreveram e participaram da reunião tutorial. A diferença no caso do semi-integrado é que, como aqueles que vão participar de fato são selecionados apenas após a reunião tutorial, a lista de pareamento é elaborada após a reunião tutorial, descrita mais adiante.

---

<sup>107</sup> Como nas sessões orais de teletandem, na Unesp Rio Preto tem sido utilizado para as videoconferências o Skype®.

<sup>108</sup> As inscrições, desde 2015, têm sido feitas através de formulários da plataforma Google (GoogleForms).

### **Listas de presença**

Durante a reunião tutorial, os candidatos a participarem de uma turma de TTD semi-integrado devem assinar uma lista de presença, elaborada por uma das mediadoras do TTD – Rio Preto a partir das inscrições recebidas. A partir das assinaturas da lista e da ordem de inscrição é que são selecionados os participantes que são notificados por email (email de pareamento ou de confirmação da participação), quando selecionados para aquela turma. Caso não tenha sido planejado o pareamento prévio, os interagentes de turmas de TTD semi-integrado recebem email informando que foram selecionados. Nesse email as mediadoras enviam calendário e lembretes quanto a horários e local de realização das atividades. Para os interagentes cuja turma foi planejada com pareamento prévio há o envio de um email de pareamento.

### **Email de pareamento**

O email de pareamento é enviado para os interagentes que participarão de turmas para as quais foi definido que haveria pareamento prévio. Os interagentes recebem nesse email informações como o nome, email e, em algumas casos, também o usuário Skype do parceiro com o qual interagirá no período. Nesse email os interagentes são informados ainda da necessidade de escrever um email de apresentação, caso as coordenações da turma tenham decidido que assim seria.

### **Email de apresentação**

Nesse email os interagentes devem se apresentar para o parceiro. Deve ser enviado, quando foi planejado pelas coordenações, antes da primeira sessão oral de teletandem.

Independente da modalidade, a reunião tutorial, realizada com o auxílio de uma apresentação em Powerpoint®, no entanto, as demais tarefas terão sua ocorrência a depender do planejado pelas coordenações do teletandem nas instituições que estabelecem parceria.

### **Tutorial**

O tutorial no teletandem precede o início das sessões orais de teletandem (SOT). Trata-se de uma reunião na qual uma mediadora, com o auxílio de uma apresentação em powerpoint, apresenta o contexto teletandem como um ambiente de aprendizagem, assim como de pesquisa. Quanto a aprendizagem a mediadora apresenta os princípios básicos do contexto teletandem de aprendizagem, além do calendário das tarefas para aquela turma especificamente. No tocante a pesquisa, a mediadora discute a relevância da coleta dos dados produzidos pelos pares para

pesquisa, esclarecendo que os dados são anonimizados e que os interagentes poderão solicitar a exclusão de seus dados coletados a medida que assim o desejar.

### **Questionários (inicial e final)**

No contexto teletandem os questionários são instrumentos reflexivos para os interagentes. Eles podem auxiliar os interagentes na reflexão sobre seu processo de aprendizagem, mas também representam aos mediadores uma forma de avaliar a realização das tarefas naquela turma e sua organização.

No questionário inicial, as questões buscam auxiliar o interagente no estabelecimento de suas metas de aprendizagem para o período de interações. Assim, os interagentes devem responder quanto a sua proficiência, considerando o Quadro Comum Europeu (CEFR), autoavaliando-se e estabelecendo seus objetivos para o semestre a partir dessa autoavaliação.

Ao responder o questionário final, o interagente avalia sua participação no TTD durante o semestre, apresenta “regras de ouro” para oportunizar uma experiência bem sucedida no contexto e pode sugerir alterações que possam aprimorar a experiência no teletandem para futuras turmas.

No período de 2012 a 2014 os questionários eram respondidos e disponibilizados a mediador na plataforma Teleduc. A partir de 2015, com o uso do GoogleForms do GoogleDrive, os interagentes recebem um link que os direciona aos formulários dos questionários. As respostas são automaticamente salvas em pasta apropriada, facilitando a organização dos dados coletados.

### **Aulas**

Tarefa que acontece apenas nas turmas de teletandem institucional integrado, as aulas são momentos em que o mediador (professor-regente da turma de língua estrangeira) discute questões que surgem durante as SOTs, mas também apresenta questões a serem discutidas nas SOTs. Não foram coletados dados referentes a essa microtarefa no *corpus* atual, havendo necessidade também de pesquisas sobre o modo como essa microtarefa se integra no contexto teletandem.

### **Sessão de mediação**

A sessão de mediação costuma ter periodicidade definida pelo mediador, normalmente ocorrendo a cada quinzena. Na sessão de mediação os interagentes são instigados pela mediadora a refletirem sobre suas interações via TTD, suas impressões e dificuldades na

parceria que está participando naquele período. Não houve coleta desse dado no período de 2012 a 2015 que fosse armazenada no banco TTD.

### **Diários de aprendizagem**

Após cada uma das SOTs, os interagentes são convidados a escrever um diário de aprendizagem. Nesse documento, compartilhado com a mediadora, o interagente deve descrever suas descobertas, dificuldades e reflexões quanto aquela SOT específica. O diário alimenta o preparo da sessão de mediação pela mediadora, fornecendo subsídios para o levantamento de possíveis “soluções” para as questões levantadas nos diários.

De 2012 a 2014, os diários eram escritos em Word. Os interagentes faziam upload dos documentos na plataforma Teleduc. A partir de 2015 os diários foram compartilhados no GoogleDrive como GoogleDocs, facilitando o compartilhamento e conseqüentemente o acesso pela mediadora.

### **Sessão Oral de Teletandem (SOT)**

A Sessão Oral de Teletandem (SOT) consiste na interação via vídeo entre aprendizes de língua estrangeira em que um objetiva aprender a língua na qual o parceiro é considerado competente linguisticamente. Utilizando as redes telemáticas, os interagentes conversam por cerca de 60 minutos respeitando os princípios de autonomia, reciprocidade e separação de línguas.

Segundo o princípio da autonomia, os interagente tem autonomia para decidir o que quer aprender e como. No tocante a reciprocidade, ele deve estar atento a necessidade do parceiro e procurar ajuda-lo a alcançar seus objetivos independente do que considere ser mais adequado, respeitando assim a autonomia do parceiro. O princípio da separação de línguas objetiva que o tempo de interação seja igualmente dividido entre as duas línguas. Assim, em uma SOT de 50 minutos, devem dedicar 25 minutos para conversar em casa uma das línguas. O período de realização, assim como a quantidade de SOTs de cada turma, depende do calendário letivo nas instituições parceiras, podendo variar entre 4 e 12 SOTs.

Durante as SOTs, pode haver o uso da ferramenta de mensagem escrita - o chat.

### **Chat**

Ferramenta utilizada durante as SOTs normalmente para sanar alguma dificuldade de compreensão oral. Para tal, os interagentes costumam enviar via chat links de vídeo, imagens e mesmo páginas na web/dicionários *on-line* que podem auxiliar seu par a compreender sobre o

que estão falando. A ferramenta também é utilizada quando há dúvida quanto a ortografia de alguma palavra.

A coleta deste dado foi feita pelo interagente brasileiro, copiando e colando em documento Word todo a conversa da SOT do respectivo dia.

### **Troca de textos**

Tarefa cuja ocorrência não é obrigatória em todas as turmas, a troca de textos surge no contexto teletandem como um instrumento auxiliar para os parceiros aproveitarem ao máximo a SOT, pois fornece subsídios para a conversa. Assim, os ‘silêncios constrangedores’ durante a SOT podem ser minimizados.

Na troca de textos os parceiros enviam, um a cada semana, um texto para seu parceiro revisar. Essa revisão deve gerar um texto final (texto reescrito) a partir das sugestões dadas pelo parceiro linguisticamente competente na língua alvo do autor do texto.

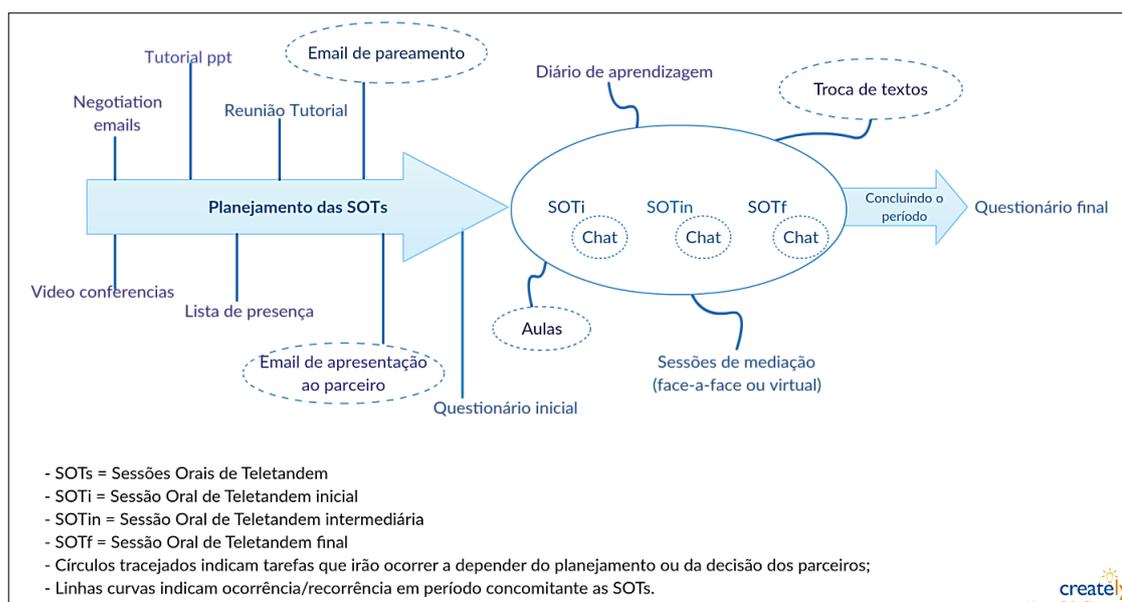
Antes do uso do GoogleDrive pelo grupo de pesquisa Teletandem Rio Preto, entre 2012 e 2014, os interagentes enviavam seus textos aos parceiros via email.

A coleta desses textos, assim como dos demais arquivos na extensão .doc apresentados acima, foram inicialmente disponibilizados pelos interagentes na plataforma Teleduc entre 2012 e 2014.

A partir de 2015 os textos começaram a ser criados como GoogleDoc no GoogleDrive e compartilhados tanto com o parceiro quanto com o grupo de pesquisa Teletandem Rio Preto, responsável por organizar os dados.

### **Cronologia das microtarefas**

A cronologia de ocorrência dessas microtarefas descritas neste documento podem ser vista sinopticamente na figura a seguir.



## Questionário inicial



### **TELETANDEM - Initial Questionnaire**

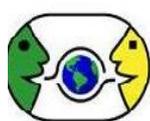
1. Have you had any experience with teletandem?
2. Write in your own words what you think teletandem is.
3. What is your proficiency level in English? How well can you read, write, listen, speak English?

Check this self-assessment grid and write down what your level is in different abilities.

[http://www.langcen.cam.ac.uk/culp/CEF\\_sa\\_grid.pdf](http://www.langcen.cam.ac.uk/culp/CEF_sa_grid.pdf)

4. Based on your self-assessment, establish a learning goal for your participation in teletandem this semester. Think about what you need to learn so that you may go to the next level.

## Questionário final



**teletandem brasil**

línguas estrangeiras para todos

### **QUESTIONÁRIO FINAL**

Caros participantes, esse questionário é uma atividade voluntária e o único propósito é nos auxiliar a aprimorar a experiência dos alunos que querem aprender via teletandem. Se você puder, por favor, responda às questões a seguir e envie para [teletandem.unespriopreto@gmail.com](mailto:teletandem.unespriopreto@gmail.com). Desde já, agradecemos por sua colaboração.

1. Por quanto tempo você interagiu (em meses e em horas de interação)?
2. Como você define o teletandem (TTD) após sua participação nesse projeto?
2. Você alcançou suas metas de aprendizagem ou, pelo menos, considera que aprendeu alguma coisa com essa experiência? Justifique sua resposta.
3. Descreva de que maneiras seu parceiro de TTD contribuiu (ou poderia ter contribuído mais) para sua aprendizagem.
4. Com base em sua experiência, quais são as três “regras de ouro” para uma experiência bem-sucedida de TTD?
5. De que maneira a equipe do TTD do IBILCE (professores e monitores) poderia ter auxiliado sua experiência para que você aprendesse mais?

## Anexo 6 – Nomeação dos dados

### Criação dos IT

✓ Interagente aluno da Unesp – Campus de São José do Rio Preto (Ibilce – Instituto Biociências, Letras e Ciências Exatas) tem a formação do IT com as seguintes informações: instituição, curso, gênero identificado, número de usuário Skype para aquela turma.

- Exemplo:

I9F3

<b>I</b> (IBILCE)	<b>9</b> (Curso de Letras)	<b>F</b> (Identificado como feminino)	<b>3</b> (Usuário skype para a turma)
----------------------	-------------------------------	--	--

✓ Interagente aluno da UGA– University of Georgia tem a formação do IT com as seguintes informações:

U0M4

<b>U</b> (University of Georgia)	<b>0</b> (indica a não identificação do curso do interagente)	<b>M</b> (identificado como masculino)	<b>4</b> (usuário skype)
-------------------------------------	--	---	-----------------------------

**Tabela 1**

Numeração dos cursos do Campus Ibilce – Unesp seguindo a ordem de apresentação na página da instituição<sup>109</sup>

<b>Curso</b>	<b>Nível</b>	<b>Número</b>
Ciências Biológicas	Graduação	<b>1</b>
Bacharelado em Ciência da Computação		<b>2</b>
Engenharia de Alimentos		<b>3</b>
Física		<b>4</b>
Matemática (Bacharelado)		<b>5</b>
Matemática (Licenciatura)		<b>6</b>
Química		<b>7</b>
Bacharelado em Letras – Tradutor		<b>8</b>
Licenciatura em Letras diurno		<b>9</b>
Licenciatura em Letras noturno		<b>10</b>
Pedagogia		<b>11</b>
Biofísica Molecular	Mestrado	<b>12</b>
Biologia Animal		<b>13</b>
Ciência da Computação		<b>14</b>
Engenharia e Ciência de Alimentos		<b>15</b>
Ensino e Processos Formativos		<b>16</b>
Estudos Linguísticos		<b>17</b>
Genética		<b>18</b>
Letras		<b>19</b>
Matemática		<b>20</b>
Microbiologia		<b>21</b>
Química		<b>22</b>
Matemática em Rede Nacional – PROFMAT	Mestrado Profissional	<b>23</b>
Biofísica molecular	Doutorado	<b>24</b>
Biologia Animal		<b>25</b>
Engenharia e Ciência de Alimentos		<b>26</b>
Estudos Linguísticos		<b>27</b>
Genética		<b>28</b>
Letras		<b>29</b>
<b>Matemática</b>		<b>30</b>
Microbiologia		<b>31</b>
Química		<b>32</b>

<sup>109</sup> <http://www.ibilce.unesp.br/#!/graduacao/cursos/> e <http://www.ibilce.unesp.br/#!/posgraduacao>

### Constituição do nome da turma TTD

**Universidade parceira** – UGA, SHEFFIELD, HARVARD, ...

**Ordem da turma** – de acordo com o dia e horário da turma. Por exemplo, em 2020 há 5 turmas com a UGA: 1 turma na terça (10h), 1 turma na terça (12h), 1 turma na terça (15h), 1 turma na terça (19h), e 1 turma na quinta (12h). As turmas serão designadas UGA1, UGA2, UGA3, UGA4 e UGA5 respectivamente.

**Modalidade** – Indicada com **i** – integrado e **si** – semi-integrado.

Exemplo: UGA1i

### Nomeação dos arquivos

A nomeação dos arquivos produzidos durante o período de interação de uma turma serão nomeados da seguinte forma:

Diário produzido pelo interagente I9M13 a partir da primeira interação realizada em uma parceria com a UGA, turma 1 de semi-integrado de 2019 será nomeado apresentado nesta ordem:

## 2019\_I9M13\_UGA1si\_D1

2019	I9M13	UGA1si	D1
<b>Ano</b> (2014, 2012,2013, ...)	<b>IT</b> (elaborado conforme padronização descrita acima)	<b>Turma</b> (UGA1i, UGA2i, UGA3si, ...)	<b>Dado</b> C - Chat; D - Diários; <b>TOI</b> - Texto original em inglês; <b>TRevI</b> - Texto revisado em inglês, <b>TReI</b> - Texto re-escrito em inglês; <b>TOP</b> - Texto original em português; <b>TRevP</b> - Texto revisado em português; <b>TReP</b> - Texto re-escrito em português

- SOTi – Sessão Oral de Teletandem Inicial
- SOTin – Sessão Oral de Teletandem intermediária (1, 2, 3, ...)
- SOTf – Sessão Oral de Teletandem final
- QF – Questionário final

➤ QI – Questionário inicial

✓ Lembrar de numerar o dado conforme sua ocorrência, excetuando-se as SOTs que seguem numeração própria.

✓ Esta nomeação viabilizará o uso de ferramentas de busca para encontrar dado conforme o ano, modalidade e tipo de dado.

### **Especificidades a serem descritas na nomeação**

Em dados no qual houve mais de uma gravação por interação em decorrência de problemas técnicos a SOT está gravada em duas partes, ou em uma das SOTs houve a substituição de um dos interagentes do par, deverá ser acrescido um *underline* ( \_ )seguido de um código conforme definido abaixo:

**Substituição de interagente –  
SOT em duas partes – 1, 2, ...**

Exemplo:

2012\_I9F15\_UGA1i\_SOTin1\_sf (Sessão Oral intermediária 1 da turma UGA1 de integrado com parceiro ausente que fora substituído)

2015\_I9F15\_UGA1si\_SOTi  
2015\_I9F15\_UGA1si\_SOTi\_1



Sessão Oral de Teletandem inicial gravada em
---

## **Anexo 7 – Plano de gerenciamento de dados - MulTeC**

### **Coleta dos dados**

As Sessões Orais de Teletandem, os chats, os diários de aprendizagem escritos pelos interagentes, os textos produzidos para troca entre eles e os questionários inicial/final foram coletados conforme descrito por Aranha, Luvizari-Murad e Moreno.

Além deles, precisaremos das listas de presença, documento de pareamento e Powerpoint® dos tutoriais, todos serão dados coletados nos arquivos pessoais das coordenadoras do TTD em São José do Rio Preto.

### **Documentação e metadados**

A documentação que acompanhará os dados será:

- Planilha com levantamento de dados (geral e por turma) (Necessário criar “nomes” de turmas para diferencia-las por ano);
- Planilha com informações sobre os interagentes (geral e por turma) (Necessário criar código para cada interagente – IT- Identificação Teletandem);
- Documento com breve descrição do contexto TTD;
- Documento com os códigos das planilhas (Necessidade de criação de códigos significativos);
- Documento descritivo do padrão de nomeação;
- Documento descritivo do padrão de anonimização.

### **Ética**

Assinatura retroativa do Termo de Consentimento Livre e Esclarecido pelos participantes cujos dados estão no Banco TTDii.

Criação de código por interagente por turma para anonimização e nomeação. Anonimização deverá ser feita em cada um dos arquivos, substituindo nomes próprios, telefone e demais dados pessoais que possam levar a identificação por um código que possibilite a compreensão pelo pesquisador.

Os dados serão salvos em dois HD externos com acesso restrito a equipe do TTD.

Os dados constantes do MulTeC estarão sob a responsabilidade das coordenadoras do TTD da Unesp de São José do Rio Preto.

O acesso aos dados será decisão das coordenadoras do grupo supracitado.

**Arquivo e Backup**

Dados serão salvos em HD com capacidade de 2T para salvar 1 cópia do Banco TTDii e 1 cópia do MulTeC.

O backup será feito na máquina pessoal da pesquisadora.

O MulTeC estará salvo no HD externo do grupo e no HD externo da pesquisadora, bem como na máquina da coordenação do projeto TTD da Unesp de São José do Rio Preto.

O controle de acesso aos dados será realizado pela coordenação do TTD.

**Preservação dos dados**

Todos os dados organizados no MulTeC poderão ser disponibilizados para pesquisas por tempo indeterminado, havendo, no entanto, exclusão de dados apenas por solicitação do participante, como informado a ele no Termo de Consentimento Livre e Esclarecido.

Os dados ficarão salvos em 2 HDs externos em diferentes locais e 2 computadores pessoais.

**Compartilhamento dos dados**

Os primeiros usuários potenciais do MulTeC são as pesquisadoras do grupo Teletandem da Unesp de São José do Rio Preto, mas os dados poderão ser disponibilizados a qualquer pesquisador interessado em investigar o TTD desde que atenda aos requisitos estabelecidos pela coordenação do TTD supracitada.

A coordenação do TTD definirá as possíveis restrições de acesso aos dados.

**Reponsabilidades e recursos**

Responsável pelo compartilhamento e política de reuso dos dados: Solange Aranha

Responsável pelo gerenciamento do dados: Queila Barbosa Lopes

Na primeira fase do projeto, apenas necessidade de aquisição de dois HDs externos de 2T para arquivamento e organização dos dados no MulTeC.

## Anexo 8 – Organização do MulTeC



## TERMO DE REPRODUÇÃO XEROGRÁFICA

Autorizo a reprodução xerográfica do presente Trabalho de Conclusão, na íntegra ou em partes, para fins de pesquisa.

São José do Rio Preto, 03/04/2019

---

Assinatura do autor