O BANCO DE DADOS DO CRITT PARA A PESQUISA DO PROCESSO TRADUTÓRIO[1]

THE CRITT TRANSLATION PROCESS RESEARCH DATABASE

Michael CARL[2]
Moritz SCHAEFFER[3]
Srinivas BANGALORE[4]
Tradução de: Leonardo Penha MESQUITA e Leonardo Lima Beschizza dos SANTOS[5]

Resumo: Desde sua criação, há dez anos, o Center for Research and Innovation in Translation and Translation Technology (CRITT), da Copenhagen Business School (CBS), Dinamarca, está envolvido em pesquisas do processo tradutório. Os dados dessas pesquisas foram coletados inicialmente pela ferramenta Translog e publicados em 2012 como um banco de dados das pesquisas do processo tradutório (em inglês, Translation Process Research Database, ou TPR-DB). Desde 2012, outros experimentos foram realizados, e mais dados foram adicionados ao TPR-DB. Em particular, dentro do projeto CASMACAT – do inglês, Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation (SANCHIS-TRILLES et al., 2014) –, foi gravada uma grande quantidade de dados do processo de pós-edição de tradução automática, e o TPR-DB foi disponibilizado sob uma licença Creative Commons. No momento da redação deste artigo, o TPR-DB contém quase 30 estudos com tarefas de tradução, pós-edição, escrita (autoral) e/ou cópia, todas gravadas com o Translog e/ou com a ferramenta CASMACAT. Cada estudo é constituído por cerca de oito a mais de 100 sessões de gravação, envolvendo mais de 300 tradutores. Atualmente, os dados somam mais de 500 horas de produção textual reunidas em mais de 1.400 sessões, com mais de 600.000 palavras traduzidas para mais de dez línguas-alvo diferentes. Este artigo descreve os recursos e as opções de visualização do TPR-DB. Esse banco de dados contém registros do processo, bem como informações derivadas e anotadas organizadas em sete tipos de unidades simples e compostas do processo e do produto, adequadas para investigar os processos de tradução humana e assistida por computador, bem como para realizar modelagem avançada dos usuários.

Palavras-chave: pesquisa empírica do processo tradutório, banco de dados de pesquisas do processo tradutório.

---

[1] Agradecemos a Michael Carl e à editora Springer a gentileza de cederem os direitos para a publicação desta versão traduzida. Referência do artigo original: CARL, Michael; SCHAEFFER, Moritz; BANGALORE, Srinivas. The CRITT translation process research database. In: _____. *New directions in empirical translation process research:* exploring the CRITT TPR-DB. Nova York: Springer, 2015. p. 13-56.

[2] Center for Research and Innovation in Translation and Translation Technology (CRITT), Departamento de Comunicação em Negócios Internacionais, Copenhagen Business School (CBS), Frederiksberg, Dinamarca.

[3] Center for Research and Innovation in Translation and Translation Technology, Departamento de Comunicação em Negócios Internacionais, Copenhagen Business School (CBS), Frederiksberg, Dinamarca.
Instituto de Linguagem, Cognição e Computação, Universidade de Edimburgo, Edimburgo, Reino Unido.

[4] Interactions Corporation, New Providence, NJ, Estados Unidos.

[5] Leonardo Penha Mesquita: Tradutor de inglês e analista de TI. Possui Bacharelado em Tradução pela Universidade Federal de Uberlândia (2015) e experiência no mercado de tecnologia da informação desde 2003. Atua como tradutor de inglês para o portal web <www.wikihow.com> desde 2014 e como analista de TI na LGTi Tecnologia desde 2016. E- mail: lpmesquita@gmail.com. Leonardo Lima Beschizza dos Santos: Estudante do curso de Bacharelado em Tradução na Universidade Federal de Uberlândia (UFU) desde 2016. E-mail: leonardobeschizza@live.com. Tradução supervisionada pelo Prof. Dr. Igor Antonio Lourenço da Silva.

1 Introdução

A pesquisa empírica do processo tradutório demanda a disponibilidade de dados adequados do processo. A fim de permitir uma pesquisa empiricamente fundamentada do processo de tradução, Jakobsen e Schou (1999) idealizaram, em 1995, uma ferramenta para registrar os acionamentos de teclas e *mouse* (*key logging*), o Translog, que permite a gravação de sessões de tradução, a visualização dos dados e a realização de análises estatísticas. Desde então, o Translog foi submetido a inúmeras mudanças, tanto no que diz respeito à aquisição de dados quanto no que tange ao formato e à representação dos processos coletados (cf. JAKOBSEN 2011, traduzido neste volume), para permitir uma análise mais robusta dos dados. O Translog-II (CARL, 2012a) atual foi complementado com a interface do CASMACAT – do inglês, Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation (SANCHIS-TRILLES et al., 2014) –, uma ferramenta de pós-edição de tradução automática utilizando um navegador; e os dados brutos obtidos ao final das sessões gravadas de tradução podem ser complementados com anotações e convertidos em um banco de dados de pesquisas do processo tradutório (em inglês, Translation Process Research Database, ou TPR-DB). Até o momento, o TPR-DB acumulou uma grande quantidade de dados do processo, com o objetivo de:

1. representar, de forma consistente, dados de atividade obtidos na pesquisa do processo tradutório de modo a viabilizar a investigação a partir de centenas de sessões de tradução, considerando línguas diferentes e modos distintos de tradução; e
2. implementar e disponibilizar um grande número de recursos para todo o conjunto de dados coletados, algo que seria difícil ou praticamente impossível de se calcular individualmente para cada sessão.

Dessa forma, o objetivo do TPR-DB é estimular a investigação e reduzir as barreiras de entrada para a pesquisa, em grande escala, do processo tradutório, facilitada por um formato consistente de banco de dados e um conjunto de recursos bem definidos.[6]

O TPR-DB é organizado em estudos e sessões. Um estudo é uma coleção de sessões conduzidas no mesmo contexto experimental. O Translog e o CASMACAT geram um único arquivo de registro para cada sessão. Em seguida, esses dados brutos do registro são anotados e processados em um conjunto em tabelas que contêm ampla gama de recursos e atributos.

[6] O banco de dados está disponível gratuitamente sob uma licença Creative Commons e pode ser baixado em: <https://sites.google.com/site/centretranslationinnovation/tpr-db>.

Este artigo descreve as tabelas e os recursos extraídos dos dados registrados e anotados. A Seção 2 fornece uma visão geral do TPR-DB; descreve o processo de anotação dos dados registrados em uma sessão de tradução, explica seu mapeamento para o TPR-DB e fornece uma visão geral das tabelas do TPR-DB. As Seções 3 a 5 descrevem as tabelas mais detalhadamente: a Seção 3 apresenta as tabelas que codificam os acionamentos de teclas e de *mouse* e as fixações; a Seção 4 ilustra as tabelas das unidades de produção e unidades de fixação, unidades essas que configuram como propriedade especial a sinalização de comportamentos de leitura e digitação paralelas e/ou alternadas, os quais apontam a carga de trabalho cognitivo de um tradutor; a Seção 5 expõe as tabelas das unidades do produto tradutório, ou seja, as unidades derivadas do produto final da tradução: itens-fonte (*source tokens*), itens-alvo (*target tokens*) e unidades de alinhamento (*alignment units*). A Seção 6 mostra as possibilidades de visualização dos dados do processo. A Seção 7 sugere as possibilidades de adição de recursos gerados externamente ao TPR-DB. Por fim, três apêndices complementam este artigo: os Apêndices 1 e 2 fornecem uma visão geral dos estudos no TPR-DB; o Apêndice 3 apresenta uma lista completa dos recursos disponíveis.

2 Visão geral do TPR-DB

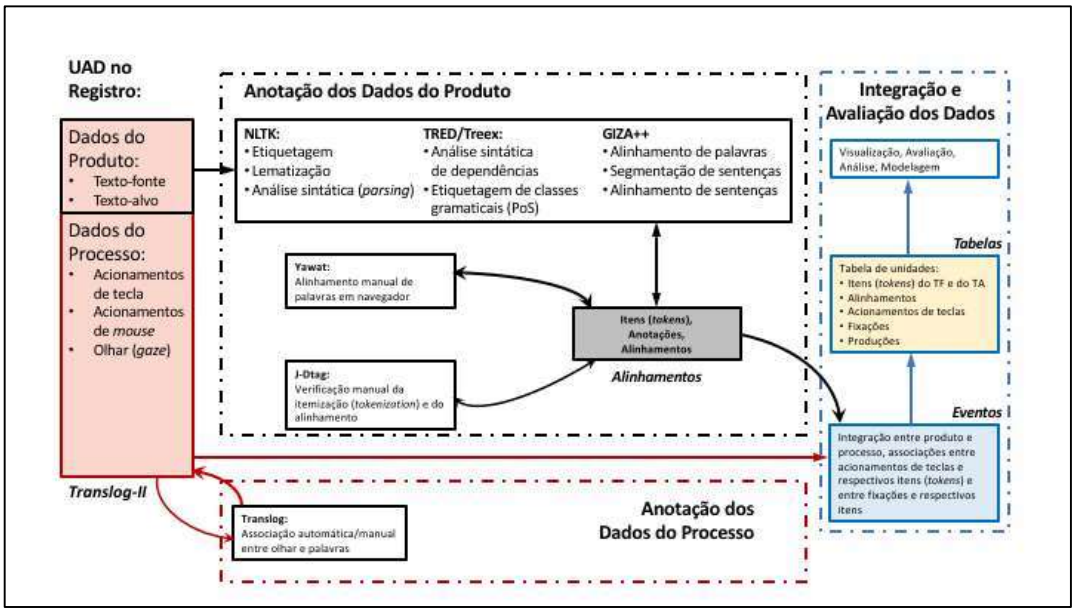
O banco de dados disponibilizado ao público geral pelo CRITT (Center for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, Copenhagen, Dinamarca) reúne seções gravadas de tradução (e de outras formas de produção textual) para fins de pesquisa do processo tradutório. Contém dados de atividade do usuário (do inglês, *user activity data*, ou UAD) referentes aos comportamentos de tradutores coletados em quase 30 estudos envolvendo tarefas de tradução, pós-edição, revisão, escrita (autoral) e/ou cópia, todas gravadas com a interface CASMACAT (SANCHIS-TRILLES et al., 2014) e/ou o *software* Translog-II (CARL, 2012a). Cada estudo é constituído por cerca de oito a mais de 100 sessões de gravação. Atualmente, o banco de dados soma mais de 500 horas de produção textual reunidas em mais de 1.400 sessões, contendo, ao todo, mais de 600.000 palavras traduzidas para mais de dez línguas-alvo. O *website* do TPR-DB disponibiliza, sob uma licença Creative Commons, todos os dados registrados durante os processos tradutórios (>20 GB), bem como uma série de anotações adicionais aos conjuntos de sessões (arquivo de 170 MB compactado em formato .zip). Nesta seção, descreve-se como os dados registrados durante uma sessão de tradução são inseridos no TPR-DB.

2.1 Compilação do TPR-DB

Os dados brutos de atividade do usuário (UAD) – que incluem os dados do processo tradutório (*e.g.*, fixações e acionamentos de teclas e *mouse*) e os dados do produto tradutório (*i.e.*, o texto-fonte e o produto final da tradução) – são armazenados e mantidos em um repositório de especifico. Dentro do processo de compilação do TPR-DB[7] (CARL, 2012b), diversas tabelas são geradas a partir dos UAD brutos, podendo então ser utilizadas como base para análises e visualizações mais detalhadas.

A Figura 1 mostra um fluxograma do processo de compilação do TPR-DB. Os UAD registrados (rotulados como *Translog-II* na figura) são processados em dois fluxos independentes para anotar os dados do produto (parte superior) e os dados do processo (parte inferior).

Figura 1 - Arquitetura do processo de compilação do TPR-DB



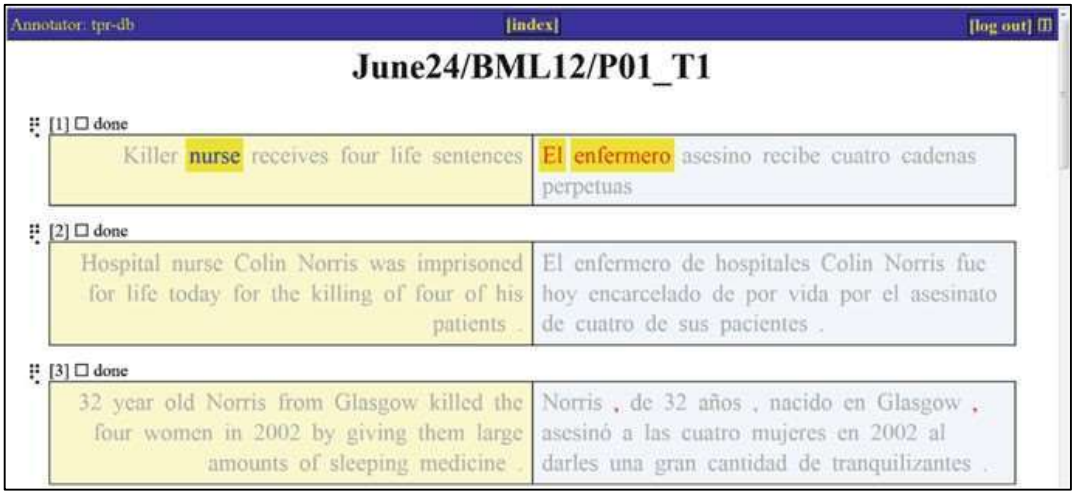
As anotações dos dados do produto, isto é, dos textos-fonte e do texto-alvo, incluem a itemização (*tokenization*), o alinhamento das sentenças e das palavras (*tokens*), bem como, de forma opcional, a lematização e a etiquetagem de classes gramaticais (PoS – do inglês, *parts of speech*), dentre outros processos. O Translog-II também oferece a possibilidade de ajustar e anotar os dados do processo, como o remapeamento das associações entre os dados do olhar e as respectivas palavras fixadas. Em seguida, uma etapa de integração dos dados registra as

[7] Embora os registros dos UAD sejam um pouco distintos no Translog-II e no CASMACAT, a estrutura das tabelas geradas é idêntica.

associações dos acionamentos de teclas e das fixações às respectivas palavras de que fazem parte (cf. CARL, 2012a). Por fim, são produzidas diversas tabelas diferentes, as quais disponibilizam um grande número de recursos e descrevem vários tipos de unidades de produto e de processo, conforme descrito na Seção 2.3.

O processo de compilação do TPR-DB é completamente automático, mas fornece uma interface gráfica do usuário (do inglês, *graphical user interface*, ou GUI), na qual o alinhamento das palavras pode ser ajustado para semiautomático. A Figura 2 mostra a interface gráfica do usuário do alinhador YAWAT (GERMANN, 2008), na qual os alinhamentos por palavra podem ser realçados, verificados e alterados.

Figura 2 - Capturas de tela do YAWAT, ferramenta de alinhamento de palavras que roda na interface de um navegador



2.2 Estudos do TPR-DB

Cada estudo é uma coleção coerente de sessões de tradução ou de produção textual, podendo envolver diferentes tarefas. Todos os estudos do TPR-DB contêm um registro dos acionamentos de teclas e *mouse*, e uma grande quantidade deles também contém dados de rastreamento ocular (*eye tracking*). Cada estudo do TPR-DB foi conduzido com (um conjunto de) pergunta(s) de pesquisa em mente, mas os estudos do banco de dados podem ser brevemente resumidos da seguinte forma:

- nove estudos foram conduzidos com as diferentes interfaces do CASMACAT para testar e avaliar suas diferentes funcionalidades;
- sete estudos são parte de um experimento multilíngue que visou comparar tradução do zero, pós-edição bilíngue e pós-edição monolíngue feitas do inglês

para uma de seis línguas-alvo por mais de 120 tradutores diferentes.

- alguns experimentos individuais foram realizados com o Translog-II com os propósitos descritos no Apêndice 1.

O Apêndice 1 fornece uma visão geral detalhada sobre diversos estudos reunidos no TPR-DB, bem como seus propósitos, participantes e durações. A Tabela 1 é um excerto do Apêndice 1 e mostra uma informação resumida do estudo CFT14.[8]

Tabela 1 - Excerto da tabela do Apêndice 1

Study	Task	Part	Sess	Texts	SL	TL	FDur	KDur	PDur	SLen	TLen
CFT14	R	4	14	14	en	es	4.54	1.21	0.28	2901	40,614
CFT14	P	7	7	2	en	es	16.51	7.90	3.41	2901	20,273
CFT14	PIO	7	7	2	en	es	15.68	7.98	3.49	2901	20,341

Cada estudo é constituído por uma ou mais sessões (Sess) de gravação, com diferentes números de participantes (Part), diferentes textos, diferentes tarefas e diferentes direções da tradução, considerando a língua-fonte (SL) e a língua-alvo (TL). Por exemplo, o estudo CFT14 na Tabela 1 possui três tarefas distintas (R, P e PIO)[9], nas quais dois textos-fonte em inglês (SL = en) foram pós-editados em espanhol (TL = es). Além disso, sete participantes produziram sete traduções (sessões) para cada P e para cada POI; e quatro participantes revisaram posteriormente os 14 textos pós-editados.

O tempo total de produção é indicado por FDur, KDur e PDur, siglas que representam a soma das durações de todas as sessões, menos as pausas encontradas antes do acionamento da primeira tecla e após o acionamento da última tecla, bem como as pausas entre os acionamentos de teclas consecutivos (dependendo da duração da pausa):

- Dur representa a duração total da sessão;
- FDur exclui as pausas maiores que 200 s entre os acionamentos de teclas e *mouse* consecutivos;
- KDur é o tempo da sessão menos as pausas maiores que 5 s entre os acionamentos de teclas e *mouse*; e
- PDur indica o tempo de digitação sem pausas maiores que 1 s entre os

[8] NT: Observe-se que os termos e os algarismos, nesta e em outras tabelas, são dispostos conforme o original em língua inglesa, para manter a forma como se visualizam as informações no banco de dados.

[9] O exemplo foi tirado de um estudo utilizando o CASMACT. As tarefas são: R (revisão), P (pós-edição) e POI (pós-edição interativa com aprendizado *online*). Há uma lista com as descrições completas de cada uma delas no Apêndice 1.

acionamentos de teclas e *mouse*.

Na Tabela 1, a duração total (FDur) das pós-edições (P) dos sete textos pelos sete pós-editores foi de 16,51 h. Dois atributos adicionais de duração indicam as durações da digitação. De acordo com o valor de KDur, os pós-editores digitaram cerca de 50% do tempo total (7,90 h), ao passo que, com base no valor de PDur, esse tempo foi de apenas 3,41 h, ou aproximadamente 20% do tempo total de pós-edição.[10] A Tabela 1 também mostra o comprimento médio do texto-fonte (SLen) e o número total de palavras produzidas na língua-alvo (TLen).

2.3 Tabelas de Resumo do TPR-DB

O processo de compilação do TPR-DB projeta os UAD brutos em diferentes unidades de produto e de processo que são reunidas nas tabelas do banco de dados. Cada linha de uma tabela descreve uma unidade particular com diversos atributos que serão descritos detalhadamente nas Seções 3 a 5[11]:

As tabelas com as unidades de produto básicas são:

1. itens-fonte (do inglês, *source tokens*, ou ST): essa tabela lista os itens (*tokens*) do texto-fonte, junto com seus correspondentes no texto-alvo, o número de inserções e exclusões necessárias para produzir a tradução, informações das microunidades etc. (Seção 4.6);
2. itens-alvo (do inglês, *target tokens*, ou TT): essa tabela lista os itens do texto-alvo, junto com seus correspondentes no texto-fonte, número de inserções e exclusões necessárias para produzir o texto-alvo, informações da microunidade, quantidade de atividade de leitura paralela à digitação etc. (Seção 4.6).

As tabelas com as unidades de produto compostas são:

3. sessão (do inglês, *session*, ou SS): essa tabela descreve as propriedades relacionadas à sessão, como língua-fonte, língua-alvo, duração total da sessão, início e final da redação (*drafting*) (Seção 3.1);

[10] Diferentes valores para delimitação de pausas já foram sugeridos e utilizados. Vandepitte et al. (2015) segmentam as sequências de acionamentos de teclas considerando 200 ms, enquanto Lacruz e Shreve (2014, p. 250) sugerem que "eventos completos de edição são separados por pausas longas (de 5 s ou mais). Eles geralmente contêm pausas curtas (superiores a 0,5 s, mas inferiores a 2 s), e os eventos completos de edição que demandam mais esforço geralmente incluem múltiplas pausas curtas. Os pós-editores podem fazer pausas de duração intermediária (superiores a 2 s, mas inferiores a 5 s) durante um evento completo de edição". Jakobsen (2005) sugere 2,4 s para a sua definição de "desempenho de pico" em tradução.

[11] As letras entre parênteses na lista representam a extensão do arquivo no TPR-DB. As seções indicam onde a tabela é descrita com maiores detalhes.

4. segmentos (do inglês, *segments*, ou SG): essa tabela lista as propriedades dos segmentos alinhados entre o texto-fonte e o texto-alvo, dentre as quais estão o número de inserções e exclusões, bem como o número de duração das fixações (Seção 3.2);
5. unidades de alinhamento (do inglês, *alignment units*, ou AU): essa tabela lista as unidades de alinhamento entre o texto-fonte e o texto-alvo, junto com o número acionamentos de teclas e *mouse* (inserções e exclusões) necessários para produzir a tradução, informações sobre as microunidades, quantidade de atividades de leitura paralela à produção das AU etc. (Seção 4.1).

As tabelas com as unidades de processo básicas são:

6. dados dos acionamentos de teclas (do inglês, *keystroke data*, ou KD): essa tabela enumera as operações de modificação do texto (inserções e exclusões), junto com o tempo dos acionamentos de teclas e *mouse*, bem como a palavra no texto final para a qual o acionamento de tecla contribui (Seção 5.1);
7. dados de fixação (do inglês, *fixation data*, ou FD): essa tabela enumera as fixações no texto-fonte ou no texto-alvo definidas pelo tempo inicial, tempo final e sua respectiva duração, bem como a posição do caractere e da palavra fixada na janela do texto-fonte ou do texto-alvo (Seção 5.2).

As tabelas com as unidades de processo compostas são:

8. unidades de produção (do inglês, *production units*, ou PU): essa tabela lista as unidades de sequência contínua de atividade de digitação da sessão, definidas pelo tempo inicial, tempo final e sua respectiva duração, porcentagem de atividade de leitura paralela à produção da unidade, duração da pausa na produção antes do início da digitação, bem como o número de inserções e exclusões (Seção 5.3);
9. unidades de fixação (do inglês, *fixation units*, ou FU): essa tabela lista as sequências contínuas de atividade de leitura, caracterizadas pelo tempo inicial, tempo final e sua respectiva duração, bem como os índices do percurso do olhar pelas palavras fixadas (Seção 5.4);
10. unidades de atividade (do inglês, *activity units*, ou CU): essa tabela fornece uma lista dos fragmentos da sessão, sendo cada fragmento definido por atividades de digitação ou leitura do texto-fonte ou do texto-alvo (Seção 5.5).

Além disso, o uso de fontes externas está resumido em:

11. fontes externas (do inglês, *external resources*, ou EX): essa tabela lista os dados

de acionamentos de teclas e *mouse* registrados com o Inputlog (LEITEN; WAES 2013) (Seção 7.1).

3 Informações do resumo das sessões e dos segmentos

Dependendo do desenho, um estudo compreende uma ou mais sessões. Durante uma sessão, um texto é traduzido, copiado, editado ou revisado, sendo que cada texto possui diversas sentenças (segmentos). Esta seção descreve as informações do resumo das sessões e dos segmentos.

3.1 Informações do resumo das sessões

As informações do resumo da sessão estão contidas na tabela "sessão" (SS), conforme consta nas Tabelas 2 a 4. Essas tabelas apresentam as informações conforme descrito a seguir.

a. As informações gerais da sessão incluem:
   - o nome do estudo (*Study*) e da sessão (*Session*), ou seja, o nome do diretório e do arquivo de registro;
   - a língua-fonte e a língua-alvo (SL e TL, respectivamente);
   - um identificador dos participantes exclusivo para o estudo (Part);
   - um identificador do texto exclusivo para o estudo (Text);
   - o tipo da tarefa (Task) (conforme apresentado na Tabela 1);
   - um identificador do segmento exclusivo para o estudo, tanto no texto-fonte (SegST) quanto no texto-alvo (SegTT), conforme discutido abaixo;
   - TokS e TokT indicam, respectivamente, o número de itens (*tokens*) no texto-fonte e no texto-alvo;
   - LenS e LenT indicam, respectivamente, o comprimento, em caracteres, do texto-fonte e o comprimento do texto-alvo.

b. As informações sobre as durações da sessão indicam quanto tempo levou o processamento dos seguintes itens:
   - a duração total da sessão (Dur) e a duração da pausa (Pause) caso hajam ocorrido interrupções na sessão;
   - o tempo de início da redação (*drafting*) (TimeD) e o tempo de início da revisão (TimeR). TimeD indica o tempo transcorrido desde o início da sessão até o primeiro toque do teclado, coincidindo com o final da fase de orientação.

TempoR indica o tempo em que a fase de redação terminou, dando início à fase da revisão; seu início coincide com o final da primeira microunidade (cf. a seguir) em que é traduzida a última palavra do texto-fonte (cf. JAKOBSEN, 2002, traduzido neste volume);

- as durações FDur, KDur e PDur já foram apresentadas anteriormente. O intervalo PDur fragmenta os UAD em unidades de produção (PU), que serão discutidas na Seção 5. Pnum representa o número de PU em uma sessão.

c. As informações sobre o processamento da sessão indicam os acionamentos de teclas e *mouse* e o comportamento do olhar:

- FixF e FixD são, respectivamente, os números de fixações no(s) iten(s)-fonte e no(s) iten(s)-alvo, enquanto TrtS e TrtT representam o tempo total de leitura, ou seja, a soma das durações de todas as fixações no texto-fonte e texto-alvo, respectivamente;
- Mins e Mdel representam, respectivamente, os números de caracteres inseridos e excluídos de forma manual, enquanto Ains e Adel representam, respectivamente, os números de caracteres inseridos e excluídos de forma automática. Ains e Adel dizem respeito à pós-edição no CASMACAT, onde o texto editado pode ser programaticamente alterado no modo de interatividade;
- Scatter indica a frequência com a qual a digitação não ocorreu em sequência, isto é, com que frequência o tradutor ou editor sucessivamente digitou teclas que foram parte de duas ou mais palavras diferentes.

Tabela 2 - Informações gerais da sessão, incluindo comprimento do texto-fonte e texto-alvo em termos de palavras e caracteres

Study	Session	SL	TL	Part	Text	Task	SegST	SegTT	TokS	LenS	TokT	LenT	…
BML12	P01_E5	en	es	P01	5	E	6	6	139	788	153	840	…
BML12	P01_P4	en	es	P01	4	P	5	5	110	668	131	763	…
BML12	P01_T1	en	es	P01	1	T	10	10	160	838	180	964	…

Tabela 3 - Informações sobre as durações da sessão

…	Dur	TimeD	TimeR	Pause	Fdur	Kdur	Pdur	Pnum	…
…	310.234	114.140	232.656	0	167.110	80.374	23.366	29	…
…	268.328	71.234	264.765	0	193.531	29.407	14.485	15	…
…	757.281	92.016	290.391	0	654.812	314.378	210.415	72	…

Tabela 4 - Informações sobre o processamento da sessão no estudo BML12

…	FixS	TrtS	FixT	TrtT	Scatter	Mins	Mdel	Ains	Adel
…	3	167	661	68.214	17	85	93	0	0
…	551	78.224	236	18.668	9	77	62	0	0
…	1122	115.692	392	26.605	30	1152	186	0	0

As Tabelas 2 a 4 apresentam três sessões do estudo BML12 conduzidas pelo participante P01. O texto (Text) 5 foi editado (Task = E); o texto 4, pós-editado (Task = P); e o texto 1, traduzido do zero (Task = T). A tradução levou mais tempo em termos de todas as medidas de duração disponíveis, Dur, FDur, KDur e PDur, enquanto a pós-edição foi mais rápida que a edição em relação aos valores de Dur, KDur e PDur, mas mais lenta em relação ao valor de FDur. Observe-se que a edição foi uma atividade mais fragmentada do que a pós-edição, pois foi produzida uma quantidade maior de PUs.

3.2 Informações do resumo dos segmentos

O nome da sessão P01_T1 significa que o participante P01 traduziu (T) o texto 1. Os 11 segmentos do texto-fonte (STseg) foram traduzidos em dez segmentos no texto-alvo (TTseg). As propriedades desses segmentos são exibidas com maior detalhamento nas tabelas de resumo dos segmentos (SG), conforme ilustrado na Tabela 5.

Tabela 5. Informações processuais das unidades de alinhamento

STseg	TTseg	Study	Session	Nedit	Dur	…	Scatter	Literal	HTra	HSeg	CrossS	CrossT
1	1	BML12	P01_T1	2	20,028	…	2	27.93	2.16	1.18	2	1.29
2	2	BML12	P01_T1	3	38,951	…	5	48.24	1.23	0.67	2	1.29
3	3	BML12	P01_T1	2	83,452	…	5	67.41	1.7	0.95	1.57	1.08
4	4	BML12	P01_T1	4	73,292	…	4	29.45	1.74	0.8	1	1.29
5	5	BML12	P01_T1	3	24,373	…	3	31.67	1.84	0.79	1.14	1.5
6	6	BML12	P01_T1	2	14,030	…	2	33.3	2.43	1.36	1.3	1.09
7	7	BML12	P01_T1	2	58,966	…	4	19.65	0.97	0.46	1.47	0.94
8	8	BML12	P01_T1	2	40,779	…	4	151.9	2.9	1.59	2.94	1.19
9	9	BML12	P01_T1	1	32,812	…	1	31.6	1.38	0.72	1.21	1.1
10 + 11	10	BML12	P01_T1	3	61,326	…	6	29.11	1.24	0.61	1.67	1.28

As tabelas de resumo dos segmentos contêm diversas informações parecidas com aquelas das tabelas das sessões, mas cada uma de suas linhas representa um segmento, e não uma sessão. Sendo assim, Dur, FDur, KDur e PDur representam o tempo de tradução do segmento, e não o tempo de tradução da seção, conforme visto na Tabela 5. As colunas STseg e TTseg indicam as informações de alinhamento do segmento (no caso da Tabela 5, todos os segmentos, estão alinhados um a um, à exceção dos dois últimos, dado que os segmentos-fonte – sentenças – 10 e 11 foram traduzidos como o segmento-alvo 10).

Nedit indica com que frequência o segmento foi revisado. Um número maior que 1 sinaliza que o tradutor primeiro fez um "rascunho" da tradução para depois retomá-lo e revisá-lo. Por exemplo, o segmento 4 foi esboçado e depois revisado três vezes, ao passo que o STseg 9 foi o único não revisado durante o processo tradutório.

Literal, HTra, HSeg, CrossS e CrossT serão abordados mais detalhadamente nas Seções 4.4 e 4.6. CrossS e CrossT medem a grau de similaridade sintática entre o texto-fonte e o texto-alvo. HTra e HSeg representam, respectivamente, a média de palavras traduzidas e a entropia média de segmentação, enquanto Literal corresponde à soma do produto de HTra e CrossS.

4 Informações do resumo dos dados no nível da palavra

Esta seção introduz as unidades de alinhamento (AU) de nível inferior (baseada em palavras), os itens (*tokens*) do texto-fonte (ST) e os itens do texto-alvo (TT). Como a maioria dos atributos das AU também ocorre nas unidades do texto-fonte e do texto-alvo, começaremos pela apresentação das AU na Seção 4.1. Na Seção 4.2, discutiremos as representações das microunidades; e, na Seção 4.3, uma métrica de (in)eficiência na digitação. A Seção 4.4 apresenta o atributo *Cross*, que quantifica as distorções sintáticas entre o texto-fonte e texto-alvo. Os itens do texto-fonte e do texto-alvo são introduzidos na Seção 4.5, seguidos de informações mais detalhadas sobre o olhar (Seção 4.6) e da entropia na tradução das palavras (Seção 4.7).

4.1 Unidades de alinhamento

Os itens-fonte e os itens-alvo correspondem a sequências de caracteres, geralmente separadas por um espaço em branco, enquanto as unidades de alinhamento (AU) são correspondências, de m para n, entre itens-fonte e itens-alvo. As tabelas de unidades fornecem

informações similares para esses três tipos de unidades. Essas tabelas contêm:

- informações gerais: o nome do estudo (*Study*) e da sessão (*Session*), o tipo de tarefa (Task), o identificador do participante (Part), o número do texto (Text), o número do segmento-fonte e dos segmentos-alvo (SFseg e TTseg, respectivamente), a língua-fonte e a língua-alvo;
- informações do produto: as cadeias de caracteres da língua-fonte e da língua-alvo (SAU e TAU, respectivamente), o número de itens (*tokens*) dessas cadeias (SAUnbr e TAUnbr) e a relação entre os itens do texto-fonte e os do texto-alvo em termos de valores do atributo Cross (cf. Seção 4.5);
- informações do processo: o número acionamentos de teclas (inserções e exclusões), a duração da produção, o comportamento do olhar em termos de números de fixações nas cadeias de caracteres-fonte e alvo da AU (FixS e FixT), o tempo total de leitura (TrtS e TrtT) e a duração da primeira passagem do olhar nas cadeias de caracteres-fonte e alvo (FPDrS e FPDurT). Esses atributos serão explicados na Seção 4.6

Tabela 6 - Informações gerais das unidades de alinhamento

AUid	STseg	TTseg	Study	Session	SL	TL	Task	Text	Part	SAU	TAU	SAUnbr	TAUnbr
44	3	3	BML12	P01_T1	en	es	T	1	P01	of	de	1	1
45	3	3	BML12	P01_T1	en	es	T	1	P01	sleeping_ medicine	tranquili- zantes	2	1

Tabela 7 - Informações do processo das unidades de alinhamento (parte 1)

Ins	Del	Dur	FixS	FPDurS	TrtS	FixT	FPDurT	TrtT
24	21	11,407	2	167	167	18	50	1232
15	0	1610	27	631	1896	8	465	615

Tabela 8. Informações do processo das unidades de alinhamento (parte 2)

Cross	InEff	Munit	Edit
1	15	2	de_medicinas_para_dormir[rimrod_arap_sanicidem]
2	0.94	1	tranquilizantes

As Tabelas 6 a 8 exibem a tradução inglês → espanhol das AU_{44} e AU_{45}, relativas respectivamente a "*of* ↔ *de*" e "*sleeping medicine* ↔ *tranquilizantes*". Conforme indicado nas colunas SAUnbr e TAUnbr, AU_{44} consiste em uma correspondência direta, ao passo que AU_{45} apresenta uma correspondência de dois para um. A coluna Edit (Tabela 8) traça a sequência de acionamentos de teclas realizados para produzir a tradução. Na AU_{44}, primeiramente foi digitada a expressão "*de medicinas para dormir*", que depois teve todos seus itens excluídos à exceção da preposição "*de*", enquanto, na AU_{45}, a palavra "*tranquilizantes*" foi digitada sem qualquer revisão. A Tabela 7 mostra os números gerais dos acionamentos de teclas: na AU_{44}, ocorreram 24 inserções, das quais 21 caracteres (a sequência entre parênteses) foram excluídos posteriormente. Observe-se que as exclusões devem ser lidas na direção inversa, ou seja, a leitura de "[*rimrod_arap_sanicidem*]" de trás para frente informa a cadeia de caracteres excluída. Mesmo que "*medicinas para dormir*" e "*tranquilizantes*" sejam paráfrases, a primeira cadeira de caracteres excluída é parte da AU_{44}; e a segunda, da AU_{45}. A associação de múltiplas palavras excluídas àquelas encontradas no texto final para as quais elas de alguma forma contribuíram só pode ser aproximada, de modo que se deve esperar uma margem de erro para as palavras vizinhas. Seguindo Alves e Vale (2011), referimo-nos a essas revisões como microunidades, as quais serão discutidas na Seção 4.2.

O tempo necessário para digitar a tradução é fornecido pelo atributo Dur. No exemplo anterior, mais de 11 s (11.407 ms) foram necessários para todas as atividades de digitação na AU_{44}, enquanto a digitação de "*tranquilizantes*", na AU_{44}, precisou de 1.610 ms.

A Tabela 8 exibe o tempo total de leitura (TrtS e TrtT) e o número de fixações (FixS e FixT) nos itens-fonte e nos itens-alvo. De acordo com essa informação, a SAU da palavra "*of*", na AU_{44}, foi fixada duas vezes, com um tempo total de leitura igual a 167 ms, enquanto a tradução *"de"* foi fixada 12 vezes, com um tempo total de 1.232 ms. A cadeia de caracteres-fonte na AU_{45} foi fixada 27 vezes, com um TrtS de 1.896 ms, ao passo que a cadeia de caracteres-alvo recebeu oito fixações, com um TrtT de 615 ms.

4.2 Microunidades

Alves e Vale (2011) denominam microunidades as atividades recorrentes de edição para a tradução de uma mesma palavra ou grupo de palavras. Para Alves e Vale (2011, p. 107), "define-se uma microunidade de tradução como o fluxo de produção contínua do

texto-alvo [...] separado por pausas durante o processo tradutório". Por sua vez, uma macrounidade consiste em um conjunto de microunidades "que compreende todas as produções de texto provisórias que correspondem ao foco do tradutor em um mesmo segmento do texto-fonte" (ALVES; VALE 2011, p. 107).

O TPR-DB computa as unidades de "produção contínua do texto-alvo" como unidades de produção (cf. Seção 4.5) e lista, nas tabelas, os detalhes das duas primeiras microunidades que contribuíram para a produção de dada tradução. A coluna Munit da Tabela 8 indica quantas microunidades contribuíram para a produção de uma AU. Embora possa haver, a princípio, qualquer número de microunidades – um tradutor pode revisar dada parte de um texto com frequência –, indicam-se, conforme a seguir, as informações detalhadas apenas das duas primeiras microunidades.

As Tabelas 9 e 10 exibem as informações das microunidades da AU_{44} e da AU_{45}.Uma microunidade conta com informações sobre o tempo inicial (Time) e sobre a duração (Dur) da atividade de sua digitação, um valor de pausa que precede essa atividade (Pause) e a quantidade de atividade de leitura concomitante no texto-fonte (ParalS) e no texto-alvo (ParalT). E o importante: uma microunidade é caracterizada pela atividade real de digitação, representada pela cadeia na coluna Edit.

Tabela 9 - Primeiras microunidades referentes a "tranquilizantes"

AUid	Edit1	Time1	Dur1	Pause1	FixS1	ParalS1	FixT1	ParalT1
44	de_medicinas_para_dormir	225.703	11.110	187	10	716	2	116
45	tranquilizantes	570.250	1.610	172	0	0	9	536

Tabela 10 - Microunidades 1 e 2

AUid	Edit2	Time2	Dur2	Pause2	FixT2	ParalS2	FixT2	ParalT2
44	[rimrod_arap_sanicidem]	569.781	297	22.937	0	0	4	214
45	–	0	0	0	0	0	0	0

As Tabelas 9 e 10 decompõem a atividade de produção da Tabela 8 em duas microunidades: no tempo 225.703 ms (Time), o tradutor produziu a primeira microunidade na AU_{44}, digitando "*de medicinas para dormir*"; no tempo 569.781 ms, em uma fase de revisão que ocorreu mais de quatro minutos depois, na microunidade 2 (Tabela 10) foi excluída a cadeia de caracteres "*medicinas para dormir*", havendo, no tempo 570.250, sua substituição por "*tranquilizantes*", que é parte da microunidade 1 da AU_{45} (Tabela 9). A duração dessas

atividades é fornecida juntamente com as pausas que as seguem e com as atividades concomitantes do olhar. Com base nas informações da Tabela 3, sabemos que a fase de revisão (TimeR) começou no tempo 290.391 ms da correspondente sessão de tradução. Sendo assim, observamos que a microunidade 1 da AU_{44} ocorre durante a fase de redação (*drafting*), enquanto a microunidade 2 da AU_{44} e a microunidade 1 da AU_{45} constituem eventos de revisão (*revision*).

4.3 Ineficiência de edição

A ineficiência de edição (*InEff)* mede a razão entre o número de caracteres produzidos e o comprimento da tradução final, o que corresponde, aproximadamente, à soma do número de inserções com o número de exclusões dividida pela diferença entre esses dois números, conforme explicitado na Equação Eq1.

(Eq1) *InEff = número de caracteres digitados / comprimento da tradução final*
≈ (*inserções* + *exclusões) / (inserções–exclusões)* + 1.

Na maioria dos casos, o comprimento de uma palavra equivale ao número de inserções de caracteres menos o número de exclusões de caracteres mais "1". Adiciona-se "1" porque o espaço em branco após a palavra é contato como parte dela. No entanto, há vezes em que não há espaço em branco após a palavra; nesses casos, o valor de InEff pode ser menor que "1". Dessa forma, para a AU_{44}, na Tabela 8, o somatório de inserções e exclusões é igual a 45, valor esse que, dividido por "3" – número de caracteres na palavra final "*of"* (incluindo o espaço em branco) –, resulta em uma ineficiência de edição igual a 15; por sua vez, o número de teclas digitadas em cadeia para produzir a palavra "*tranquilizantes*" na AU_{45} corresponde ao comprimento da tradução final, de modo que o esforço de edição é igual 0,94. Observe-se que, para a pós-edição, o valor de InEff pode ser zero se a proposta da tradução automática for aceita sem qualquer modificação, mas seria igual a "2" se a palavra fosse excluída e outra de comprimento idêntico fosse digitada novamente.

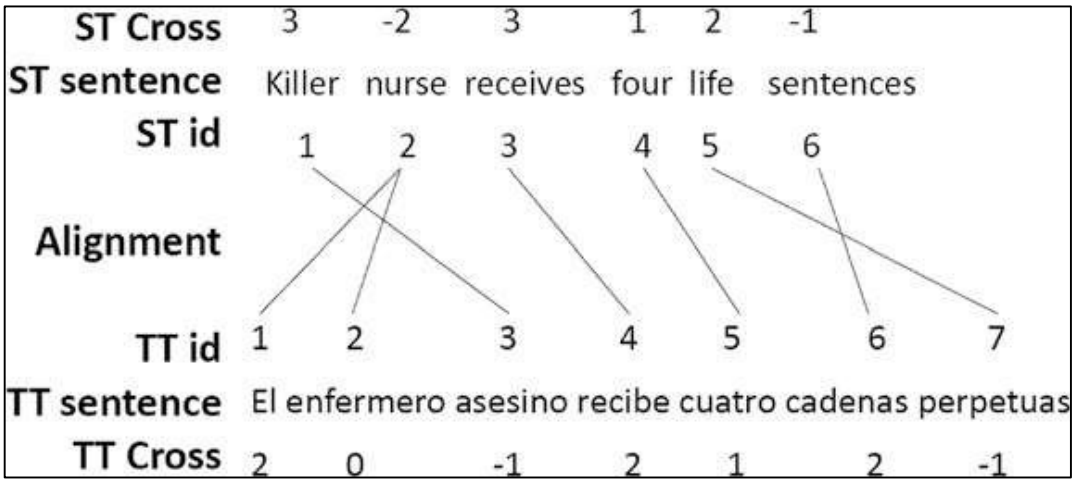
4.4 Atributo Cross

O atributo Cross representa as informações do alinhamento entre as palavras do texto-fonte e aquelas do texto-alvo como uma distorção local entre as línguas em questão. Essa distorção tem uma direção dependente, partindo do texto-fonte em direção ao texto-alvo

(CrossS), ou vice-versa (CrossT). Talvez os valores desse atributo possam ser mais bem compreendidos como uma referência ao método de geração de uma sentença – da esquerda para a direita – por meio de ligações de alinhamento em que se contam quantas palavras devem ser "saltadas" em uma língua para se produzir a próxima palavra na outra língua.

A Figura 3 exemplifica uma tradução inglês → espanhol. Ela mostra duas sentenças alinhadas, os valores de Cross, a enumeração dos itens (*tokens*) nas duas sentenças e as ligações entre o texto-fonte e o texto-alvo. Para produzir a primeira palavra na tradução em espanhol ("*El*"), duas palavras em inglês ("*Killer*" e "*nurse*") tiveram de ser "saltadas" no texto-fonte, resultando em um valor *Cross* igual a "2". Como a segunda palavra no texto-fonte ("*nurse*") produz duas palavras adjacentes no texto-alvo, nenhuma palavra no texto-fonte deve ser "saltada" para produzir "*enfermero*", resultando em um valor *Cross* igual a zero. Para produzir a terceira palavra em espanhol, "*asesino*", uma palavra do texto-fonte, localizada à esquerda de "*nurse*", deve ser processada, levando ao valor *Cross* de "–1". Já a próxima palavra em espanhol, "*recibe*", é a tradução de duas palavras à direita da atual posição do cursor no texto-fonte, levando a um valor *Cross* igual a "2". Dessa forma, os valores TT Cross indicam uma reordenação relativa das palavras do texto-fonte para se chegar à tradução do texto-alvo.

Figura 3 - Valores Cross das unidades do texto-fonte e texto-alvo



Também se computam valores de Cross para o texto-fonte. Os valores de ST Cross presumem que o texto-fonte é o texto de saída e o texto-alvo é o texto de entrada. Portanto, esses valores indicam a reordenação relativa das palavras do texto-alvo para se chegar ao texto-fonte.

Línguas com similaridade na ordem das palavras apresentam, em média, valores Cross baixos. Em uma tradução uniforme de "1" para "1", todos os valores Cross são iguais a "1". Quanto maior o número de reordenações sintáticas entre o texto-fonte e o texto-alvo, maior é a média dos valores do atributo Cross.

4.5 Itens (tokens) do texto-fonte e do texto-alvo

As tabelas de resumo dos itens do texto-fonte (ST) e dos itens do texto-alvo (TT) contêm basicamente os mesmos dados que aqueles presentes nas tabelas de AU. Em especial, as informações gerais, como o nome do estudo (*Study*) e o nome da sessão (*Session*), são idênticas. No entanto, no lugar dos atributos SAL e TAU (como vistos na Tabela 6), as tabelas ST e TT apresentam os atributos SToken e TToken, bem como uma etiquetagem de Lemma (lemas) e outra de PoS (classes gramaticais – do inglês, *part of speech*) dos itens da língua-fonte e da língua-alvo, respectivamente.

A Tabela 11 apresenta três itens do texto-fonte. Os atributos Prob1e Prob2 são probabilidades, em $\log_{10}$, de monogramas e bigramas. No caso do inglês, utilizou-se o BNC[12] como corpus de referência. No caso, existe uma probabilidade de $10^{-3,4339} = 1$ em 2.715 e de $10^{-6,1669} = 1$ em 1.468.588 para a ocorrência de, respectivamente, "*four*" e de "*life sentence*" no BNC.

Tabela 11 - Informação dos itens do texto-fonte (ST)

STid	STseg	Study	Session	…	SToken	Lemma	Prob1	Prob2	PoS	TToken	TTid
5	1	BML12	P01_T1	…	four	four	–3,4339	–50	CD	cuatro	5
6	1	BML12	P01_T1	…	life	life	–3,3508	–50	NN	perpetuas	7
4	1	BML12	P01_T1	…	sentences	sentence	–4,64	–6,1669	NNS	cadenas	6

Também existem mais informações detalhadas sobre o olhar nas tabelas ST e TT. Além disso, as tabelas ST contêm informações sobre a entropia na tradução das palavras.

[12] BRITISH NATIONAL CORPUS. Disponível em: <http://www.natcorp.ox.ac.uk/>. Acesso em: 16 nov. 2017.

4.6 Informação do olhar (gaze)

Diversas novas medidas de tempo de leitura foram adicionadas ao TPR-DB 2.0. Seguimos sugestões propostas pelo Kertz Lab[13], as quais foram implementadas da seguinte maneira:

- FFTime, ou tempo da primeira fixação (do inglês, *first fixation time*): é o tempo (em ms) despendido até a primeira fixação em dado item (*token*);
- FFDur, ou duração da primeira fixação (do inglês, *first fixation duration*): é a duração da primeira fixação no item;
- FPDurS, ou tempo de leitura do item-fonte quando da primeira passagem por ele (do inglês, *first pass source token reading duration*): é a soma das durações de todas as fixações no item-fonte desde a primeira fixação até o momento em que os participantes olham para um item diferente;
- FPDurT ou tempo de leitura do item-alvo quando da primeira passagem por ele (do inglês, *first pass target token reading duration*): a soma das durações de todas as fixações no item-alvo desde a primeira fixação até o momento em que os participantes olham para um item diferente;
- RPDur, ou duração do percurso de regressão (do inglês, *regression path duration*): é a quantidade de tempo necessária desde o FFTime até que os olhos se movam para a direita no texto (em outras palavras, inclui todas as regressões para a esquerda);
- Regr, ou regressão (do inglês, *regression*), é um valor booleano que indica se uma regressão sucedeu a primeira passagem de leitura;
- FixS, ou número total de fixações no item-fonte (do inglês, *fixations on the source token*);
- FixT, ou número total de fixações no item-alvo (do inglês, *fixations on the target token*);
- TrtS, ou tempo total de leitura no item-fonte (do inglês, *total reading time on source token*): é a soma de todas as fixações no item-fonte em toda a sessão;
- TrtT, ou tempo total de leitura no item-alvo (do inglês, *total reading time on*

[13] KERTZ LAB. Disponível em: <https://wiki.brown.edu/confluence/display/kertzlab/Eye-Tracking+While+Reading>. Acesso em: 17 nov. 2017.

target token): é a soma de todas as fixações no item-alvo em toda a sessão.

A Tabela 12 exibe exemplos dessas medições do olhar. De acordo com as definições, FFDur é sempre menor que RPDur ou Trt.

Tabela 12 - Informações do olhar nas unidades do texto-fonte e texto-alvo (parte 1)

STid	Study	Session	...	FFTime	FFDur	RPDur	Regr	FixS	FPDurS	TrtS	FixT	FPDurT	TrtT
4	BML12	P01_T1	...	280	50	1317	1	9	50	1567	3	133	183
5	BML12	P01_T1	...	1843	650	650	0	8	650	1600	24	149	1945
6	BML12	P01_T1	...	2515	283	866	1	12	666	1498	0	0	0

4.7 Entropia e perplexidade de tradução

A perplexidade de tradução indica quantas escolhas de tradução um tradutor tem para determinado ponto do texto-fonte, ou seja, quantas palavras igualmente prováveis podem ser produzidas para certa palavra-fonte em dado contexto. Nós assumimos que a escolha das traduções segue certa distribuição de probabilidades de tradução ***p*** e estimamos essas probabilidades a partir de um *corpus* de traduções alinhadas. As probabilidades de tradução $p(s \rightarrow t_i)$ para dada palavra ***s*** do texto-fonte e suas possíveis correspondências na tradução $t_i...n$ são computadas como a razão entre o número de alinhamentos $s \rightarrow t_i$ contados nos textos-alvo e o número total de itens observados no texto-alvo, conforme mostra a Equação Eq2.

$$\text{(Eq2)} \quad p\ (s\ \rightarrow\ t_{\mathrm{i}}) = \frac{número\ de\ (s \rightarrow t_{\mathrm{i}})}{número\ de\ traduções}$$

A informação de uma distribuição com probabilidades iguais, ***p***, é definida por $I(p) = -\log_2(p)$. Enquanto a probabilidade expressa a expectativa para um evento, essa informação indica a quantidade mínima de *bits* com a qual essa expectativa pode ser codificada. A entropia ***H*** indica a expectativa para essa informação, também entre distribuições de probabilidades desiguais, conforme visto na Equação Eq3.

$$\text{(Eq3)} \quad H(s) = \sum_{i=1}^{n} p(s\ \rightarrow\ t_{\mathrm{i}})^{*} - log_2(p\ (s \rightarrow t_i))$$

A entropia de tradução, ***H(s)***, é a soma de todas as probabilidades (*i.e.*, expectativas)

observadas para a tradução de dada palavra-fonte, ***s*** , nas palavras-alvo, $t_i...n$, multiplicada pelo seu conteúdo informacional. Isso representa a quantidade média de informações contidas em uma escolha de tradução. Sendo assim, se dada palavra-fonte ***s*** tiver somente uma possibilidade de tradução ***t*** em determinado contexto, sua tradução é $p(s \rightarrow t) = 1$, sua informação é $I(p(s \rightarrow t)) = 0bit$ e, portanto, sua entropia $H(s) = 0$ é mínima. Quanto mais traduções diferentes igualmente prováveis uma palavra-fonte tiver, maior será sua entropia ***H(s)*** de tradução.

A perplexidade (***PP)*** está relacionada com a entropia ***H*** seguindo uma função exponencial, conforme revelado na Equação Eq4.

(Eq4) $PP\ (s) = 2^{H(s)}$

Quanto maior for a perplexidade, maior será a probabilidade de existirem escolhas semelhantemente possíveis e, portanto, mais difícil será a tomada de decisão.

As tabelas ST fornecem algumas dessas informações. CountT representa o número de alinhamentos ($s \rightarrow t_i$) observados em SToken → $TToken_i$ alignments count. AltT aponta o número de diferentes $TTokens_i$. ProbT refere-se à probabilidade de ocorrência do item. HTra é a entropia de tradução do SToken.

Por exemplo, considere-se $STid_4$ na Tabela 13. A tradução "*four → cuatro*" ocorreu 25 vezes no *corpus*, com uma probabilidade de 0,8. Com isso, podemos reconstruir o número total de traduções no *corpus* como sendo 31 ≈ 25/0,8, sendo que as seis traduções restantes (31 – 25) foram distribuídas como três palavras diferentes.

Tabela 13 - Informações do olhar nas unidades do texto-fonte e texto-alvo (parte 2)

STid	Study	Session	:::	SToken	TToken	SAUnbr	TAUnbr	AltT	CountT	ProbT	HTra	HSeg
4	BML12	P01_T1	:::	four	cuatro	1	1	4	25	0,8065	0,9511	0,7088
5	BML12	P01_T1	:::	life	perpetuas	1	1	8	17	0,5484	1,9385	0,6595
6	BML12	P01_T1	:::	sentences	cadenas	1	1	8	18	0,5806	1,899	0,4587

O atributo HSeg indica a entropia da segmentação do alinhamento das palavras. Por exemplo, uma expressão como "*life sentences*" poderia ser alinhada como uma única unidade multipalavra ou como duas unidades distintas. O número de palavras da língua-fonte e da

língua-alvo da unidade de alinhamento (AU) da qual "*life*" faz parte está refletido nos valores SAUnbr e TAUnbr, respectivamente. O atributo HSeg leva em consideração esse contexto de segmentação do alinhamento, sendo calculado de forma parecida com o atributo HTra, mas com a diferença de que depende da contagem de TAUnbr idênticos, em vez da contagem de TToken.

O atributo *Literal*, na Tabela 5, é, pois, basicamente a literalidade média na tradução das palavras. É calculada por $Literal = \frac{1}{n} * \sum_j^n abs(cross^* \boldsymbol{Htra})$, em que $\boldsymbol{n}$ é o comprimento da sentença do texto-fonte.

5 Unidades de processamento

Esta seção começa com a descrição das unidades básicas de processamento, dados de acionamentos únicos de teclas (KD – em inglês, *keystroke data*) e de fixações (FD – em inglês, *fixation data*). As Seções 5.3 e 5.4 introduzem, respectivamente, as unidades de produção (PU – em inglês, *production units*) e as unidades de fixação (FU – em inglês, *fixation units*). A Seção 5.5 apresenta a noção de unidades de atividade (CU – em inglês, *activity units*), que fragmenta exaustivamente o processo tradutório em oito tipos de segmentos.

5.1 Dados dos acionamentos de teclas

Dentro do TPR-DB, cada acionamento de tecla feito por um tradutor humano é caracterizado por sete elementos, quais sejam:

1. Time: o tempo (ms) despendido até o acionamento de tecla;
2. Type: a distinção entre os tipos acionamentos (de inserção e de exclusão);
3. Cursor: a posição do texto-alvo em que se dá o acionamento de tecla;
4. Char: o caractere (UTF8) que é alvo do acionamento (para inserção ou exclusão);
5. TTseg: o segmento-alvo (sentença-alvo) que está sendo produzido(a);
6. STid: a identificação da palavra do texto-fonte da qual a palavra-alvo produzida é uma tradução;
7. TTid: a identificação da palavra do texto-alvo que está sendo processada pelo acionamento de tecla.

O exemplo da Tabela 14 mostra os dados de acionamentos de teclas para a produção

de duas palavras em espanhol, "*El enfer[e]mero*", como tradução da palavra-fonte $STid_2$. Essas são as duas primeiras palavras do primeiro segmento da tradução. A tabela registra somente os acionamentos de teclas para modificação do texto, ou seja, para inserções e exclusões – as informações de navegação, como cliques do *mouse*, são ignoradas. As inserções e exclusões podem ser produzidas manualmente (Mins e Mdel) ou automaticamente (Ains e Adel). A Tabela 14 contém, na linha 9, um exemplo de exclusão manual.

Tabela 14 - Informações de acionamentos de teclas na sessão P01_T1 do estudo BML12

KDid	Time	Type	Cursor	Char	TTseg	STid	TTid
0	92.016	Mins	0	"E"	1	2	1
1	92.172	Mins	1	"l"	1	2	1
2	92.313	Mins	2	"_"	1	2	1
3	92.375	Mins	3	"e"	1	2	2
4	92.563	Mins	4	"n"	1	2	2
5	92.828	Mins	5	"f"	1	2	2
6	92.938	Mins	6	"e"	1	2	2
7	93.047	Mins	7	"r"	1	2	2
8	93.266	Mins	8	"e"	1	2	2
9	93.610	Mdel	8	"e"	1	2	2
10	93.797	Mins	8	"m"	1	2	2
11	93.875	Mins	9	"e"	1	2	2
12	93.938	Mins	10	"r"	1	2	2
13	94.078	Mins	11	"o"	1	2	2
14	94.203	Mins	12	"_"	1	2	2

5.2 Dados de fixação

Durante a fixação Dentro do TPR-DB, o centro de uma fixação é mapeado sobre o caractere mais próximo na tela e conectado aos dez atributos seguintes:

1. *Time*: momento em que a fixação teve início;
2. *Dur*: duração da fixação em ms;
3. *Win*: janela-fonte (1) ou janela-alvo (2) na qual a fixação é observada;
4. *Cursor*: mapeamento do centro da fixação no caractere mais próximo na janela;
5. *STid*: identificação do item (*token*) do texto-fonte que está sendo visualizado;
6. *TTid*: identificação do item do texto-alvo que está sendo visualizado;
7. *Seg*: identificação do segmento da palavra do texto-fonte (*STid*) que está sendo

visualizada;

8. *ParalK*: quantidade de atividades de digitação concomitante, ou seja, unidades de produção (PU, cf. Seção 5.3);
9. *Edit*: caractere(s) digitado(s) durante a fixação;
10. *EDid*: identificação do segmento-alvo produzido pela digitação dos caracteres.

A Tabela 15 mostra uma sequência de 13 fixações, FDid 507–519, que fazem parte da sessão P01_T1, introduzida anteriormente. Todas as fixações ocorrem na janela 1, no primeiro segmento e nos itens de STid 4, 6, 3 e 5, traduzidos, respectivamente, nos itens de TTid 5, 6, 4 e 7. Algumas fixações apresentam atividades de digitação concomitante: como a quantidade de atividade paralela de teclado (ParalK) equivale ao tempo de duração das fixações (Dur), as sete primeiras fixações (FDid 507–513) se sobrepõem a 100% da produção do texto. Não ocorreu atividade no teclado durante as fixações FDid 515–517, mas uma sobreposição parcial de 16% (124 ms/750 ms) foi encontrada entre a atividade de digitação e a fixação FDid 518. Durante as fixações 507–510, por exemplo, foi digitada (Edit) a sequência "*eno*", que é parte da produção de "*as**e**si**no***". A coluna EDid indica o número de identificação STid da tradução produzida, ou seja, "*as**e**si**no***" é uma tradução de STid 3.

Tabela 15 - Informações de fixação (.fd file)

FDid	Time	Dur	Win	Cursor	Seg	STid	TTid	ParalK	Edit	EDid
507	94.530	150	1	25	150	4	5	150	e	3C
508	94.749	67	1	24	1	4	5	67	–	–
509	95.077	67	1	25	1	4	5	67	n	3C
510	95.218	67	1	26	1	4	5	67	o	3C
511	98.952	50	1	36	1	6	6	50	i	4C
512	99.015	167	1	37	1	6	6	167	b	4C
513	99.202	50	1	36	1	6	6	50	–	–
514	99.265	83	1	25	1	4	5	1	e	4C
515	99.499	100	1	16	1	3	4	0	–	–
516	99.624	83	1	16	1	3	4	0	–	–
517	99.718	50	1	17	1	3	4	0	–	–
518	99.780	750	1	24	1	4	5	124	_	4C
519	100.546	250	1	30	1	5	7	250	–	–

Na Seção 5.3, mostramos que a sequência de teclas é parte de uma unidade de produção, PU_0, enquanto as fixações são parte de FU_{14}. A Seção 6 dispõe os dados em um

contexto mais amplo.

5.3 Unidades de produção

As unidades de produção (PU) são sequências de atividades contínuas de digitação. De acordo com a definição em Carl e Kay (2011), a delimitação de uma PU é definida como uma pausa de 1000 ms ou mais sem qualquer atividade no teclado. Presume-se que uma pausa superior a essa indique uma interrupção no fluxo de digitação, com um possível desvio da atenção em direção a um segmento diferente do texto. Como um segmento temporal/textual contínuo, as PU possuem um tempo inicial (Time) e uma duração (Dur) e podem englobar um ou mais acionamentos de teclas para inserção ou exclusão (Edit) que contribua para a produção de um ou mais itens do texto-alvo (TTid). A Tabela 16 disponibiliza três unidades de produção, uma das quais, PU_0, tem sua sequência de edição apresentada no exemplo E1.

Tabela 16 - Três unidades de produção da sessão P01_T1 do estudo BML12

PUid	Estudo	Sessão	Time	Dur	Pausa	Ins	Del	Edit
0	BML12	P01_T1	92.016	7250	92.016	34	7	El_enfere[e]mero_ asesiono_re[er_ono]no_ recibe
1	BML12	P01_T1	100.406	1313	1140	8	0	_cuatro_
2	BML12	P01_T1	103.594	4187	1875	23	3	sentencias_de_vida. __[__.]__

(E1) El_enfere [e] mero_asesiono_re [er_ono] no_recibe

E1 iniciou no tempo 92.016 ms e foi digitado em 7250 ms, sem interrupções superiores a 1000 ms entre os acionamentos das teclas. Esse exemplo foi precedido por uma pausa de 92.016 ms.

A próxima unidade, PU_1, começou com uma pausa de 1140 ms. Após essa pausa, a sequência de digitação começa no tempo 100.406 ms e dura 1.313 ms.

A Tabela 16 indica o número de inserções e exclusões das PUs. A unidade PU_0 contém 34 inserções (Ins) e sete exclusões (Del). As exclusões estão dentro de colchetes na coluna Edit, e devem ser lidas de trás para frente. Dessa forma, a subsequência "*[er_ono]*" na verdade reflete a exclusão "*ono_re*", conforme visto no exemplo E2.

(E2) asesiono_re → asesino_recibe

A Tabela 17 – que na verdade é uma continuação da Tabela 16 – contém informações adicionais do produto e do processo das três PUs. STseg e TTseg indicam que as três PU fazem parte da tradução do primeiro segmento. STid e TTid exibem as palavras-alvo e as palavras-fonte abrangidas pela tradução. Observe-se que TTid se refere às numerações das palavras na tradução final; sendo assim, a numeração de uma palavra em uma versão intermediária do texto pode não coincidir com aquela no texto final se houver inserção ou exclusão de palavras.

Tabela 17 - Três unidades de produção da sessão P01_T1 do estudo BML12

STseg	TTseg	STid	TTid	FixS	ParalS	FixT	ParalT	Scatter	CrossS	CrossT	PosS	PosT
1	1	1 + 2 + 3	1 + 2 + 3 + 4	10	735	0	0	0	2,67	1,25	NNP + VBP + NNS	ART + NC + NC + VLfin
1	1	3 + 4	4 + 5	4	504	0	0	0	2	1,5	NNS + CD	VLfin + CARD
1	1	4 + 5	5 + 7	4	216	0	0	1	1,5	1	CD + NN	CARD + ADJ

Conforme podemos constatar na sucessão da STid, a tradução se desenvolve sucessivamente na ordem das palavras do texto-fonte. A PU_1 "*_cuatro_*" diz respeito a duas palavra-fonte e a duas palavra-alvo ($STid_{3+4}$ e $TTid_{4+5}$), dado que o espaço em branco – representado pelo subtraço "_" – já conta como parte da próxima palavra. A PU_2 também compreende duas palavras, $TTid_{5+7}$, mesmo que a PU seja constituída por três palavras "*sentencias de vida*", haja vista que o grupo nominal foi reescrito posteriormente como "*cadenas perpetuas*", formando $TTid_6$ e $TTid_7$. Observe-se que essa descontinuidade também é o motivo pelo qual o valor de Scatter é igual a "1": há uma sequência de dois acionamentos sucessivos de teclas nessa PU que produz traduções separadas por mais de uma palavra.

Os atributos FixS e FixT representam os números de fixações contadas, respectivamente, no lado-fonte e no lado-alvo das PU. Observe-se que, devido a uma má qualidade do rastreamento ocular, nenhuma fixação foi registrada nas cadeias de caracteres do texto-alvo.

Os atributos ParalS e ParalT indicam as quantidades de tempo que o tradutor estava olhando, respectivamente, para a janela-fonte e para a janela-alvo enquanto produzia a tradução. Em outras palavras, durante os 7.250 ms despendidos para produzir a PU_0, o

tradutor olhou por quase 1 s (900 ms) para a janela do texto-fonte.

Os atributos CrossS e CrossT representam a média de distorção local entre o lado-fonte e o lado-alvo das PU. O cálculo dos atributos Cross é discutido em detalhe na Seção 4.4. Os atributos PosS e PosT indicam as etiquetas de classes gramaticais das palavras-fonte e das palavras-alvo envolvidas nas PU.

5.4 Unidades de fixação

As unidades de fixação (FUs) descrevem sequências de comportamento de leitura contínua. Com base em evidências experimentais (CARL; KAY, 2011), definimos a delimitação entre duas FU sucessivas como uma pausa do olhar com duração superior a 400 ms. Por exemplo, quando o olhar é desviado da tela por mais do que 400 ms, interrompendo assim a atividade de leitura contínua, delimitamos uma unidade de fixação. Uma FU tem um tempo inicial e uma duração, sendo seguida por uma pausa (com duração superior 400 ms) antes do início da FU subsequente.

O atributo *Path* descreve a sequência de palavras visualizadas na janela-fonte (1) ou na janela-alvo (2). O percurso do olhar consiste em uma ou mais fixações indicadas por uma sequência de variáveis "Win:WordID" com as fixações sucessivas sendo separadas por "+". A FU_{14}, na Tabela 18, possui um caminho de quatro fixações (1:4 + 1:4 + 1:4 + 1:4+) na palavra-fonte "*four*" (1:4). Por sua vez, a FU_{15}, apresentada na Figura 4, representa um padrão de leitura no que diz respeito às palavras em negrito no título "*Killer nurse* ***receives four life sentences***". Essa figura mostra como o olhar vai e volta entre as quatro palavras, o que levou 1844 ms.

Tabela 18 - Quatro unidades de fixação

FUid	Time	Dur	Pause	ParalK	Path
14	94.530	755	5293	755	1:4 + 1:4 + 1:4 + 1:4+
15	98.952	1844	3667	704	1:6 + 1:6 + 1:6 + 1:4 + 1:3 + 1:3 + 1:3 + 1:4 + 1:5+
16	101.577	1272	781	142	1:5 + 1:5 + 1:6 + 1:6 + 1:6 + 1:5 + 1:5+

Figura 4 - Captura de tela da repetição da situação FU_{13}

O atributo *ParalK* em uma tabela FU indica a quantidade de atividades paralelas de olhar e acionamento de teclas. Durante a FU_{14}, o tradutor escreveu enquanto estava lendo. Já na FU_{16}, houve uma sobreposição de 11% entre a atividade de digitação e a atividade de leitura.

Observe-se que a soma das durações de todas as FU pode ser maior que a soma das durações de todas as fixações. A razão é que as FU incluem intervalos interfixações com duração inferior a 400 ms que podem não integrar uma fixação.

5.5 Unidades de atividade

As unidades de atividade (CU) segmentam exaustivamente a sessão gravada em sequências de atividades que são levemente diferentes daquelas constantes nas tabelas PU e FU. Diferentemente dessas duas últimas, as CU dividem uma sessão em segmentos digitados. Para as tarefas de tradução, distinguem-se três tipos (em inglês, *Types*) de atividades do tradutor, a saber:

- tipo 1: leitura do texto-fonte;
- tipo 2: leitura do texto-alvo;
- tipo 4: digitação da tradução.

Como a leitura do texto-fonte ou do texto-alvo podem ocorrer em conjunto com a digitação (cf. Seção 4.3), também temos as seguintes atividades concomitantes:

- tipo 5: digitação da tradução concomitante com leitura do texto-fonte;
- tipo 6: digitação da tradução concomitante com leitura do texto-alvo;
- tipo 7: digitação da tradução concomitante com leitura do texto-fonte e do texto-alvo.

Define-se uma atividade de digitação contínua como um uso ininterrupto do teclado

(similar ao registrado nas PU), havendo apenas pausas iguais ou inferiores a 1 s entre dois acionamentos consecutivos de teclas. Se não forem registradas nem atividades do olhar nem atividades do teclado por mais do que 1 s, tem-se um segmento ocioso:

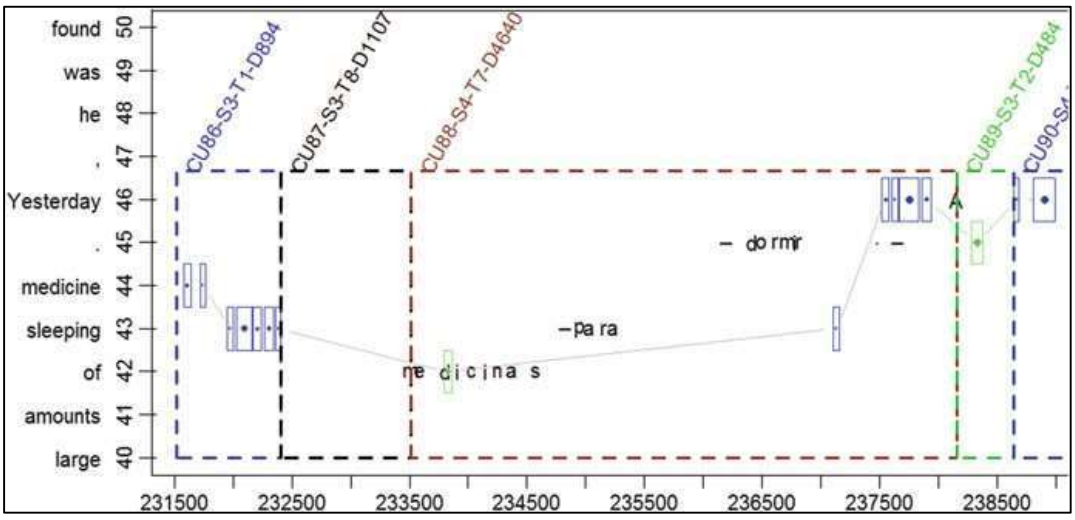
- tipo 8: nenhuma atividade registrada.

Uma CU é descrita por seu tempo inicial (*Time*), duração (*Dur*) e segmento (*Seg*) no qual se encontra. A Figura 5 mostra uma sequência de quatro unidades de atividade envolvidas na tradução apresentada no Exemplo E3.

E3 *sleeping medicine → medicinas para dormir*

A Figura 5 mostra as delimitações das CUs e seus respectivos rótulos. A primeira CU, com a marcação temporal 231.500–232.500 ms é uma atividade de leitura do texto-fonte de 894 ms, seguida por uma unidade "ociosa" (tipo 8) de 1107 ms, na qual nenhuma atividade foi registrada. Em seguida, tem-se uma CU de digitação (Tipo 7), com início na marcação temporal de 233.500 ms, com duração de 4640 ms, na qual podem ser observadas leituras concomitantes do texto-fonte e do texto-alvo. Durante esse intervalo de tempo, foi produzida a tradução "*medicinas para dormir A*". Essa unidade foi seguida por uma atividade de leitura do texto-alvo (tipo 2, com duração de 484 ms), na qual houve monitoramento da palavra recém-digitada ("*dormir*"). A figura representa um gráfico de progressão da tradução (do inglês, *translation progression graph*, ou TPG), que será discutido na Seção 6.

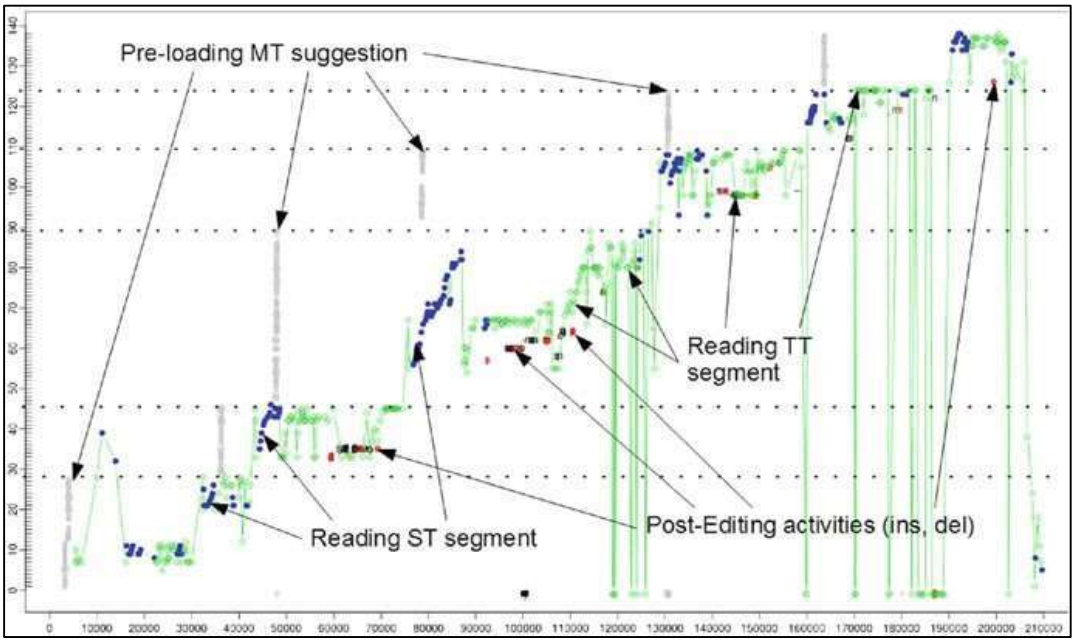
Figura 5 - Segmentação em unidades de atividades sucessivas



6 Visualização dos dados do produto e do processo em um gráfico de progressão

As informações de várias tabelas podem ser analisadas, avaliadas e visualizadas de diferentes formas. O gráfico de progressão da tradução (TPG) é um método de visualização que faz parte do TPR-DB. Os TPG permitem visualizar a forma como as traduções emergem ao longo do tempo, dispondo, de uma só vez, as informações parciais de diversas tabelas. A Figura 6 exibe um TPG para uma sessão de pós-edição no CASMACAT.

Figura 6 - Gráfico de progressão da tradução dispondo informações do olhar e dos acionamentos de teclas



O gráfico esboça as atividades de pós-edição de seis segmentos consecutivos. O eixo vertical enumera as palavras do texto-fonte (0 ... 140) com traços horizontais separando os segmentos, enquanto o eixo horizontal exibe o momento em que a tradução foi produzida para o respectivo texto-fonte. Os símbolos do gráfico são os seguintes:

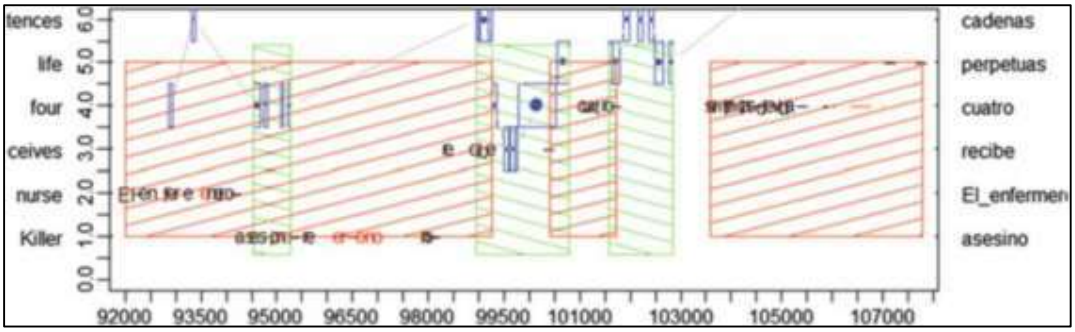
- losangos azuis: representam as fixações no texto-fonte;

- losangos verdes: representam as fixações no texto-alvo;
- caracteres pretos: representam as inserções;
- caracteres cinzas: representam as inserções automáticas;
- caracteres vermelhos: representam as exclusões.

O gráfico mostra quando os segmentos são traduzidos da língua-fonte para a língua-alvo, quando e onde o tradutor leu os segmentos-fonte e os segmentos-alvo e quando o texto foi modificado. Sendo assim, os gráficos de progressão são formas bastante úteis para avaliar os dados do TPR-DB quantitativamente.

Outro TPG é exibido na Figura 7. Esse gráfico relaciona o produto da tradução, considerando o texto-fonte (eixo vertical esquerdo) e o texto-alvo (eixo vertical direito), com os dados do processo tradutório em uma linha do tempo disposta no eixo horizontal, permitindo visualizar como a tradução emerge no tempo. As inserções são representadas por letras pretas, as exclusões estão em vermelho, e as fixações são pontos azuis dentro de retângulos que são mais largos ou mais estreitos conforme a duração.

Figura 7 - O gráfico de progressão apresenta as informações do produto e do processo das Tabelas 14 a 18



O TPG da Figura 7 esboça os dados dos acionamentos de teclas da Tabela 14, os dados de fixação da Tabela 15, bem como as três FU da Tabela 18 e as três PU das Tabelas 16 e 17. Os retângulos vermelhos preenchidos com traços indicam PU, enquanto os retângulos verdes preenchidos com traços indicam FU. A primeira parte (tempo aproximado de 92.000 ms até 94.000 ms) reproduz a produção das palavras 1 e 2 ("*El enfermero*") como delineado na Tabela 14.

Conforme discutido nas Seções 5.3 e 5.4, as atividades de leitura e escrita podem ocorrer em paralelo. Por exemplo, a FU_{14}, por volta do tempo 95.000 ms, aconteceu enquanto o tradutor produzia a palavra "*asesino*", tradução de "*Killer*", ao passo em que a FU_{15} e a FU_{16}, nas marcas temporais de, respectivamente, 99.000 ms e 101.500 ms, se sobrepuseram

apenas parcialmente à PU_0 e à PU_1 adjacentes. Os gráficos de progressão ilustram, de forma diagramática, a relação entre as atividades de leitura e escrita.

7 Fontes externas

7.1 Incorporação de dados externos do Inputlog

Os *softwares* Translog-II e CASMACAT somente registram os dados de acionamentos de teclas realizados dentro de uma interface gráfica do usuário. No entanto, em muitos casos, o tradutor utiliza fontes externas, como dicionários eletrônicos, ferramentas de colocação e pesquisas por expressões na internet. Essas atividades não são gravadas no Translog-II ou no CASMACAT, mas o comportamento relativo às pesquisas externas pode ser interessante para fins de investigação e correlação com os dados de atividade do usuário (UAD) do Translog-II.

O Inputlog (LEIJTEN; VAN WAES, 2013) é uma ferramenta para Windows que registra todos os tipos de entrada: acionamentos de teclas e *mouse*, bem como reconhecimento de fala. Diferentemente do Translog-II e do CASMACAT, o Inputlog não é um aplicativo independente, podendo registrar as atividades do teclado independentemente do aplicativo (do Windows) que está recebendo a entrada de dados. O Inputlog reconhece o aplicativo que está recebendo o foco e, em um arquivo IDFX, armazena essa informação juntamente com os dados da tecla pressionada (ou do clique de *mouse*) e sua respectiva a marcação de tempo.

Um *script*[14], InfuseIDFX.pl, pode ser usado para integrar os dados de registro do Inputlog, IDFX, com os arquivos do Translog-II. Primeiramente, o *script* "InfuseIDFX.pl" sincroniza os dados de registro do Inputlog e do Translog-II com base nos acionamentos de teclas comuns e, depois, insere, no arquivo de registro do Translog- II, os dados coletados fora da interface gráfica do usuário do Tranlog-II (ou do CASMACAT). Em seguida, os processos de compilação do TPR-DB geram uma tabela EX que indica o uso de fontes externas.

Por exemplo, o Inputlog reconhece a janela que está em foco em um navegador. Os sucessivos acionamentos de teclas e *mouse* podem ser associados à página *web* em foco. Dessa forma, pode-se acompanhar e reconstruir as pesquisas na internet.

Por um lado, o Inputlog é universalmente implementável em diferentes aplicativos do Windows; por outro lado, ele não tem como identificar em que parte de um texto o caractere foi digitado. O Inputlog informa quais teclas foram pressionadas, mas não necessariamente

[14] O *script* InfuseIDFX.pl faz parte do TPR-DB e pode ser baixado em: <https://sites.google.com/site/centretranslationinnovation/tpr-db>.

quais caracteres são produzidos ou excluídos, tampouco onde essas operações ocorrem em um texto – a menos que se utilize o Microsoft Word.

Como exemplo, a Tabela 19 exibe um excerto de uma tabela convertida do Inputlog mostrando que o Google Chrome foi usado como fonte externa principal em uma sessão do Translog-II. No instante 33.453 ms, uma aplicação com o nome TASKBAR foi ativada por aproximadamente 0,5 s, seguida por uma busca no Google Chrome, que durou pouco mais do que 32 s. Em seguida, o usuário usou o Menu Iniciar para retornar à janela do Translog-II User, mas ele a deixou novamente após 14.297 s.

Tabela 19 - Uso de fontes externas no TPR-DB

EXid	Study	Session	Focus	Time	Dur	STsegN	STsegL	STidN	STidL	KDidN	KDidL	Edit
0	N1	P02_P1	TASKBAR	33.453	547	2	1	18	8	27	26	
1	N1	P02_P1	Google Chrome	34.000	32.813	2	1	18	8	27	26	**EDIT**
2	N1	P02_P1	Menu Iniciar	66.813	812	2	1	18	8	27	26	
3	N1	P02_P1	Translog-II User	67.625	14.297	2	1	18	8	27	26	

A coluna Edit contém a concatenação das teclas digitadas que ocorreram durante o tempo em foco. Esse valor vai estar vazio caso nenhum acionamento de tecla tenha sido realizado. A seguir, constam três exemplos do que se pode encontrar na cadeia de caracteres representada por EDIT nos 32 s entre as marcas temporais de 34.000 ms e 66.818 ms, quando o Google Chrome estava em foco:

1. bring

2. emit[..]otional trra[..]adução

3. in [.] the arr[.]ticle
 tradução
presented
&#xA.

Uma pesquisa geralmente é ativada com o acionamento da tecla “Enter”, que nesse caso foi codificado como “
”, e as exclusões estão em colchetes [..]. Sendo assim, na linha (1), o tradutor digitou “*bring*” e pressionou a tecla “Enter”. Na linha (2), o tradutor excluiu dois caracteres duas vezes na cadeia inserida. Pelo arquivo IDFX do Inputlog, não é possível saber quais caracteres foram excluídos, mas é provável que tenham sido, nesta ordem, “*it*” e “*ra*”, mediante acionamento da tecla “Backspace”, para produzir a sequência de busca “*emotional tradução*”. Na linha (3), foram produzidas três sequências de busca, quais sejam, “*in the article*”, “*tradução*” e “*presented*”.

Embora possamos reconstruir a pesquisa realizada pelo tradutor na fonte externa, não sabemos quais foram seus resultados. No entanto, é possível rastrear a reação do tradutor

dentro do Translog-II User. O atributo KDidL indica a última tecla acionada (KDid) antes de o tradutor sair da tela do Translog-II, e o atributo KDidN representa a tecla acionada logo que o tradutor retorna ao programa. De forma parecida, STidL e STidN representam a identificação da palavra-fonte referente a esses acionamentos de teclas, enquanto STsegL e STsegN representam os segmentos-fonte. Sendo assim, a última tecla acionada antes de o tradutor sair do Translog-II User, em 33.453 ms, foi KDidL = 26; e o primeiro toque após seu retorno ao programa foi KDidN = 27. Esses dois acionamentos fazem parte da produção para a tradução de STidL = 8 e STidN = 18, que pertencem aos dois segmentos sucessivos 1 e 2. Apesar de não ser possível saber o que exatamente o tradutor descobriu ao visitar uma fonte externa, é possível inferir seu efeito ao investigar as ações executadas antes e após a consulta.

7.2 Adição de colunas às tabelas de resumo do TPR-DB

Em alguns casos, pode ser desejo do usuário adicionar a algumas tabelas do TPR-DB colunas com anotações próprias. Por exemplo, em um experimento sobre entropia sintática, cada segmento foi anotado manualmente com um conjunto de tríades (em inglês, *Triplets*), descrevendo a estrutura sintática da sentença. Tais anotações podem ser adicionadas automaticamente a uma tabela de resumo por meio de um *script* que faz parte do TPR-DB.[15] Um arquivo com a extensão da tabela especifica o estudo, a sessão e a identificação da unidade, assim como as colunas a serem adicionadas, como mostra a Tabela 20.

Tabela 20 - Anotações com informações suplementares sobre os segmentos em cinco colunas adicionais

Study	Session	STseg	SynH	STriplet	TTriplet	PrimeDiff	Prime Prob
default	default	default	0	–	–	DIFF	0
BML12	P03_P1.sg	2	0	TPI	TPI	PRIME	1
BML12	P06_P1.sg	2	0	TPI	TPI	PRIME	1
BML12	P03_P1.sg	3	0.721	TAI_DAD	TAI_DAD_ IAD	DIFF	0.2
BML12	P06_P1.sg	3	0,721	TAI_DAD	TAI_DAD_ IAD	DIFF	0.2
ML12	P28_P1.sg	3	0,721	TAI_DAD	TAI_DAD	PRIME	0.2
BML12	P32_P1.sg	3	0,721	TAI_DAD	TAI_DAD	PRIME	0.2
BML12	P03_P1.sg	4	0	TPI_TAD	MPI	DIFF	0

[15] O *script* AddExtColumns.pl pode ser baixado em: <https://svn.code.sf.net/p/tprdb/svn/> e invocado com os parâmetros "AddExtColumns.pl -C ExtraColumnsFile –S Study_name".

Agradecimentos

Este trabalho teve apoio do projeto CASMACAT, financiado pela Comissão Europeia (Sétimo Programa-Quadro de Investigação). Somos gratos a todos que contribuíram com o banco de dados e que nos autorizaram o uso de seus dados.

Referências

ALVES, F.; VALE, D. C. On drafting and revision in translation: a corpus linguistics oriented analysis of translation process data. *Translation: Corpora, Computation, Cognition*, v. 1, n. 1, p. 105-122, 2011. Disponível em: <http://www.t-c3.org/>. Acesso em: 8 nov. 2017.

CARL, M. Translog-II: A program for recording user activity data for empirical reading and writing research. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8., Istanbul, Tyrkiet, Department of International Language Studies and Computational Linguistics, 21-27 maio 2012a. *Proceedings…* Istanbul: [s.e.], 2012. p. 2-6.

CARL, M. The CRITT TPR-DB 1.0: A database for empirical human translation process research. In: O'BRIEN, S.; SIMARD, M.; SPECIA, L. (Ed.). *Proceedings of the AMTA 2012 workshop on post-editing technology and practice (WPTP 2012).* Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA), 2012b. p. 9-18.

CARL, M.; KAY, M. Gazing and typing activities during translation: a comparative study of translation units of professional and student translators. *Meta,* v. 56, n. 4, p. 952-975, 2011.

JAKOBSEN, A. L. Translation drafting by professional translators and by translation students. In: HANSEN, G. (Ed.). *Empirical translation studies:* process and product. Copenhagen: Samfundslitteratur, 2002. p. 191-204.

JAKOBSEN, A. L. Instances of peak performance in translation. *Lebende Sprachen,* v. 50, n. 3, p. 111-116, 2005.

JAKOBSEN, A. L. Tracking translators' keystrokes and eye movements with Translog. In: ALVSTAD, C.; HILD, A.; TISELIUS, E. (Ed.), *Methods and strategies of process research:* integrative approaches in translation studies. Amsterdam: John Benjamins, 2011. p. 37-55.

JAKOBSEN, A. L.; SCHOU, L. Translog documentation. In: HANSEN, G. (Ed.). *Probing the process in translation methods and results*. Copenhagen: Samfundslitteratur, 1999. p. 1-36.

GERMANN, U. Yawat: Yet Another Word Alignment Tool. In: ACL-08. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics:* HLT demo session. Columbus, OH: Association for Computational Linguistics, 2008. p. 20-23.

LACRUZ, I.; SHREVE, S. Pauses and cognitive effort in post-editing. In post-editing of machine translation: processes and applications. In: O'BRIEN, S.; SIMARD, M.; SPECIA, L.; CARL, M.; BALLING, L. W. (Ed.). *Expertise in post-editing:* Processes, technology and applications. Cambridge: Scholars Publishing, 2014. p. 246–274.

LEIJTEN, M.; VAN WAES, L. Keystroke logging in writing research: using Inputlog to

analyze and visualize writing processes. *Written Communication,* v. 30, n. 3, p. 358-392, 2013.

SANCHIS-TRILLES, G.; ALABAU, V.; BUCK, C.; CARL, M.; CASACUBERTA, F.; GARCÍA-MARTÍNEZ, M.; GERMANN, G.; GONZÁLEZ-RUBIO, J.; HILL, R. L.; KOEHN, P.; LEIVA, L. A.; MESA-LAO, B.; ORTIZ-MARTÍNEZ, D.; SAINT-AMAND, H.; TSOUKALA, C.; VIDAL, E. Interactive translation prediction versus conventional post-editing in practice: a study with the CASMACAT workbench. *Machine Translation*, v. 28, n. 3-4, p. 217-235, 2014.

VANDEPITTE, S.; HARTSUIKER, R. J.; VAN ASSCHE, E. Process and text studies of a translation problem. In: FERREIRA, A.; SCHWIETER, J. W. (Ed.). *Psycholinguistic and cognitive inquiries into translation and interpreting.* Amsterdã/Filadélfia: John Benjamins, 2015. p. 127-143.

Apêndice 1

Ao todo, o TPR-DB contém mais de 580 horas de produção textual considerando o atributo Fdur. Contribuíram para as 1689 sessões 132 tradutores diferentes, que produziram um total de mais de 660.000 palavras em nove línguas diferentes.

O par linguístico en → es (inglês-espanhol) é de longe aquele que conta com maior representatividade no TPR-DB, com 660 sessões, 500.000 palavras nas traduções e mais de 320 horas de produção considerando o atributo Fdur. O segundo par linguístico mais bem representado é o en → hi (inglês-hindu), com 161 sessões, mais de 20.000 palavras nas traduções e mais de 46 horas de produção pelo atributo Fdur. O terceiro par linguístico é o en → de (inglês-alemão), com 146 sessões, mais de 24.000 itens (*tokens*) nas traduções e mais de 24 horas de produção pelo atributo Fdur. Em seguida, vem o par en → da (inglês-dinamarquês), com 127 sessões, mais de 18.000 itens nas traduções e 12 horas de produção segundo o atributo Fdur. Os outros pares linguísticos do TPR-DB envolvem mais de 20 direções de tradução a partir da combinação de sete línguas-fonte diferentes com 16 línguas-alvo distintas (inclusive alguns pares linguísticos que não aparecem na Tabela 21). Remete-se o leitor à página do TPR-DB para uma versão atualizada do conteúdo do banco de dados.

Cada estudo no TPR-DB foi conduzido a partir de uma ou mais perguntas de pesquisa. Os estudos podem ser resumidamente agrupados da seguinte forma:

(A) O TPR-DB contém dez estudos conduzidos com as três diferentes interfaces do CASMACAT:

1. ALG14: compara tradutores profissionais e falantes bilíngues durante uma pós-edição com o terceiro protótipo do CASMACAT, apresentando visualizações dos alinhamentos entre as palavras;
2. CEMPT13: contém gravações de pós-edições com o segundo protótipo do CASMACAT, apresentando a tradução automática interativa;
3. CFT12: contém dados do primeiro teste de campo do CASMACAT em junho de 2012, comparando a pós-edição com a tradução do zero;
4. CFT13: contém dados do segundo teste de campo do CASMACAT em junho de 2013, comparando a pós-edição com a tradução automática interativa;
5. CFT14: contém dados do segundo teste de campo do CASMACAT em junho de 2014, comparando a tradução automática interativa com o aprendizado *online*;
6. EFT14: compara a aprendizagem ativa e a aprendizagem *online* para

previsões em traduções interativas;

7. JN13: foi gravado com o segundo protótipo do CASMACAT e apresenta o recurso de alinhamentos de palavras e tradução automática interativa;
8. LS14: investiga, com o terceiro protótipo do CASMACAT, os efeitos do aprendizado com a pós-edição interativa durante um período de seis semanas (estudo longitudinal);
9. PFT13: pré-teste que antecede o segundo teste de campo do CASMACAT;
10. PFT14: pré-teste que antecede o terceiro teste de campo do CASMACAT.

Tabela 21 - Tabela de resumo dos estudos do TPR-DB: continuação abaixo

Study	Sess	SL	TL	Task	Texts	Part	Fdur	Kdur	Pdur	Stok	Ttok
ACS08	30	en	da	T	4	17	4.6776	2.9704	1.9332	5085	5075
ACS08	30	en	en	C	4	17	2.0436	1.8316	1.6013	5099	5109
ALG14	8	en	es	P	2	8	2.6018	0.4854	0.1747	4460	4807
ALG14	8	en	es	PA	2	8	2.7954	0.4437	0.1692	4460	4801
BD08	10	en	da	T	1	10	1.4575	0.7493	0.448	1100	1056
BD13	8	en	da	T	2	8	0.8079	0.5368	0.3213	786	751
BD13	10	en	da	P	2	10	0.4412	0.1074	0.0569	970	1014
BML12	64	en	es	P	6	32	4.6394	0.9079	0.4418	9012	10,216
BML12	63	en	es	T	6	32	9.8032	5.9308	3.8062	8936	10,102
BML12	60	en	es	E	6	30	3.7009	0.9657	0.4729	8468	9594
CEMPT13	20	en	pt	PIA	2	20	6.634	1.823	0.5387	6706	6840
CEMPT13	20	en	pt	P	2	20	5.5943	1.2678	0.5732	6494	6585
CFT13	27	en	es	R	26	4	8.3388	0.9733	0.4413	26,919	28,738
CFT13	27	en	es	PI	9	9	30.0923	10.2351	3.3044	31,752	33,871
CFT13	27	en	es	P	9	9	28.167	8.0677	03.51	31,294	33,770
CFT13	27	en	es	PIA	9	9	35.5658	11.2626	3.9125	31,838	34,047
CFT14	7	en	es	RE	7	3	3.8435	0.2465	0.0586	20,341	22,015
CFT14	7	en	es	R	7	4	3.2497	0.3687	0.1485	20,273	22,251
CFT14	7	en	es	P	2	7	16.8321	7.9316	3.418	20,273	22,067
CFT14	7	en	es	PEI	2	7	15.8297	8.1574	3.4917	20,341	22,284
DG01	60	fr	pl	T	2	60	33.8564	17.5784	11.2075	25,380	20,329
EFT14	11	en	es	PIVO	3	11	10.221	5.2041	2.2521	12,437	13,549
EFT14	11	en	es	PI	3	11	11.9495	6.8647	3.2755	12,437	13,696
EFT14	10	en	es	PIVA	3	10	10.7885	5.1993	2.3594	11,327	12,472
GS12	8	es	en	P	4	4	4	4	2.1901	0.3586	0.1909
HLR13	15	en	et	T	3	5	2.5457	1.1214	0.673	1535	1186
JIN15	18	en	zh	S	1	18	2.0227	0.2641	0.0455	1947	1728

(continua)

Tabela 21 - (continuação)

Study	Sess	SL	TL	Task	Texts	Part	Fdur	Kdur	Pdur	Stok	Ttok
JIN15	18	en	zh	P	1	18	4.5318	0.8192	0.1442	1998	1845
JIN15	17	en	zh	R	1	17	2.594	0.3451	0.0567	1946	1833
JLG10	10	en	pt	T	3	5	5.6048	2.1218	1.2302	2577	2781
JLG10	10	pt	en	T	3	5	5.6391	2.0787	1.1718	2611	2621
JN13	4	en	de	PIA	2	4	2.7428	0.7284	0.2735	2590	2668
JN13	4	en	de	P	2	4	2.3311	0.6374	0.2189	2590	2571
KTHJ08	69	en	da	T	3	24	7.4469	5.6824	3.8183	10,571	10,667
LS14	60	en	es	PI	24	5	53.3764	22.0971	9.5166	72,109	80,278
LS14	60	en	es	P	24	5	51.7256	17.3211	7.4178	72,126	80,454
LWB09	40	da	en	T	3	18	3.7061	2.8926	2.0511	5652	6206
MS12	19	en	zh	P	6	11	2.6953	0.4817	0.0497	2708	2562
MS12	15	en	zh	T	5	10	3.7369	1.0512	0.1088	2061	1916
MS12	10	en	zh	E	5	8	0.7714	0.1564	0.0183	1295	1203
MS13	16	zh	pt	P	2	16	2.7139	0.9211	0.4443	1410	1648
MS13	16	pt	zh	T	2	16	2.3327	0.7687	0.1161	1386	1378
MS13	22	zh	pt	T	2	22	4.1631	2.1803	1.2265	1938	2216
MS13	18	pt	zh	P	2	18	2.555	0.6698	0.0934	1555	1507
NJ12	39	en	hi	T	6	20	14.4697	7.5368	3.3156	5505	5784
NJ12	61	en	hi	P	6	20	17.4402	6.8654	3.0615	8581	9365
PFT13	9	en	es	P	1	9	2.0861	0.3154	0.1406	3035	3144
PFT13	19	en	es	PI	1	19	5.2058	1.5351	0.4267	6689	7437
PFT13	16	en	es	PIC	3	16	2.7853	0.744	0.1518	5396	5147
PFT13	15	en	es	PEI	3	15	2.4784	0.4741	0.0669	4611	4666
PFT13	16	en	es	PIL	3	16	2.7226	0.6761	0.1511	5572	5344
PFT14	3	en	es	PIVO	2	3	2.1558	0.6775	0.1622	3245	3150
PFT14	2	en	es	PIVA	1	2	2.0228	0.7255	0.1843	2286	2184
PFT14	2	en	es	PIV	2	2	1.987	0.7667	0.1905	2161	2077
RH12	2	es	es	A	2	2	2.9849	0.9786	0.6398	1207	1207
ROBOT14	40	in	nl	P	8	10	10.8706	3.2467	1.5417	7375	7527
ROBOT14	40	en	nl	T	8	10	12.2457	5.1006	3.1753	7375	7329
SG12	46	en	de	E	6	23	7.0716	1.8571	0.9342	6522	6741
SG12	45	en	de	P	6	23	8.027	1.9976	1.055	6352	6470
SG12	47	en	de	T	6	24	11.7259	4.7344	2.9421	6632	6777
TDA14	48	en	en	C	6	8	3.8335	3.5653	2.6617	6792	6779
WARDHA13	34	en	hi	T	6	18	15.2298	3.6917	0.5553	4832	4790
WARDHA13	31	hi	hi	C	6	18	11.49	5.3097	0.7569	4365	4104
WARDHA13	27	en	hi	P	6	15	8.0582	1.9611	0.4418	3780	4016
ZHPT12	12	zh	pt	T	1	12	3.5244	1.4856	0.851	1104	1603
Total	**1689**	**7**	**9**	**15**	**132**	**418**	**586.769**	**217.2386**	**100.2227**	**702,701**	**660,595**

A tabela mostra informações do resumo do TPR-DB para cada sessão: tarefa, direcionalidade e par linguístico, número de textos distintos, número de participantes, duração da produção (Fdur, Kdur e Pdur), assim como o

comprimento total, em palavras (itens, ou *tokens*), do texto-fonte (STok) e do texto-alvo (TTok).

(B) O objetivo do experimento multilinguístico é comparar a tradução a partir do zero (T), a pós-edição bilíngue (P) e a edição, ou pós-edição monolíngue (E), entre diferentes tradutores e línguas. Os seis textos-fonte em inglês foram traduzidos por tradutores em formação e tradutores experientes. Três (1-3) são textos jornalísticos, e três (4-5) são textos de sociologia extraídos de uma enciclopédia. Os textos foram sistematicamente alternados de forma que todos eles fossem traduzidos por todos os tradutores e cada tradutor traduzisse dois textos diferentes em cada uma das modalidades. Os estudos desse grupo são:

11. BML12: contém dados de tradução, pós-edição e edição dos seis textos do inglês para o espanhol;
12. KTHJ08: contém apenas dados de tradução dos textos jornalísticos (1-3);
13. MS12: contém dados de tradução, pós-edição e edição dos seis textos do inglês para o chinês;
14. NJ12: contém dados de tradução, pós-edição e edição dos seis textos do inglês para o hindu que foram realizadas por tradutores profissionais;
15. SG12: contém dados de tradução, pós-edição e edição dos seis textos do inglês para o alemão;
16. TDA14: contém dados de participantes que fizeram cópias dos seis textos em inglês;
17. WARDHA13: contém dados de tradução, pós-edição e edição dos seis textos do inglês para o hindu que foram realizadas por estudantes.

(C) O BD-TTP contém, ainda, alguns experimentos individuais que foram conduzidos com o Translog-II. São eles:

18. ACS08: explora, por meio da análise dos tempos do olhar associados às expressões, a forma como os tradutores processam o significado de expressões não literais;
19. BD08: envolve tradutores profissionais dinamarqueses trabalhando com o par linguístico en → da;
20. BD13: envolve estudantes do Ensino Médio realizando tarefas de tradução e pós-edição do inglês para o dinamarquês;

21. DG01: compara estudantes, tradutores profissionais e tradutores não profissionais em termos de representação do texto;
22. GS12: contém dados de pós-edição de quatro notícias traduzidas do espanhol para o inglês;
23. HLR13: investiga as traduções de três textos distintos do inglês para o estoniano realizadas por cinco participantes;
24. JLG10: investiga traduções direta (L2-L1) e inversa (L1-L2) no par linguístico inglês-português (do Brasil);
25. LWB09: traz os resultados de um experimento com rastreamento ocular em que tradutores profissionais traduziram dois textos do dinamarquês (L1) para o inglês (L2);
26. MS13: investiga o comportamento do tradutor durante a tradução e a pós-edição no par linguístico português-chinês (em ambas as direções);
27. RH12: investiga a produção autoral de notícias por dois jornalistas espanhóis;
28. ROBOT14: investiga o uso de fontes externas durante a tradução e a pós-edição;
29. ZHPT12: investiga o comportamento do tradutor ao traduzir textos jornalísticos, tendo como foco o processo tradutório durante o processamento de expressões não literais (metafóricas).

Apêndice 2

Durante cada sessão, é conduzida uma tarefa (em inglês, *Task*) específica, que pode ser:

- A: escrita autoral de um texto jornalístico, com a língua-fonte e a língua-alvo sendo as mesmas;
- C: cópia (manual) de um texto da janela-fonte para a janela-alvo, com a língua-fonte e a língua-alvo sendo as mesmas;
- E: edição de uma tradução automática sem acesso ao texto-fonte (pós-edição monolíngue);
- P: pós-edição de uma tradução automática (com acesso ao texto-fonte e sem consultas externas);
- R: revisão de um texto pós-editado;
- T: tradução "do zero".

No contexto do CASMACAT, diversas situações de pós-edição foram investigadas, a saber:

- PA: pós-edição tradicional com acesso, ativado por *mouse* ou cursor, aos alinhamentos entre texto-fonte (ST) e texto-alvo (TT);
- PI: pós-edição avançada com o aporte de previsões interativas de tradução (do inglês, *interactive translation prediction*, ou ITP) / tradução automática interativa;
- PIA: pós-edição avançada com o aporte de ITP e com acesso aos alinhamentos ST-TT (opção de visualização);
- PIC: pós-edição avançada com o aporte de ITP e com acesso aos alinhamentos ST-TT (opção de visualização);
- PIO: pós-edição avançada com o aporte de ITP e de técnicas de aprendizagem *online*;
- PIL: pós-edição avançada com o aporte de ITP e com acesso ao texto pós-editado (sufixo) em cinza (opção de visualização);
- PIV: pós-edição avançada com o aporte de ITP e com acesso a uma barra de busca e substituição, a alinhamentos e a opções de ITP ativadas pelo *mouse*;
- PIVA: pós-edição avançada com o aporte de ITP e de técnicas de aprendizagem ativa;
- PIVO: pós-edição avançada com o aporte de ITP e de técnicas de aprendizagem

online.

Apêndice 3

Este apêndice lista todos os atributos utilizados no TPR-DB 2.0 em suas tabelas de unidades. Há, no total, 275 atributos, dos quais 111 são distintos, nos onze tipos de tabelas discutidos neste artigo. Esses atributos foram agrupados em doze tipos, de acordo com o que descrevem (sessão, segmento, item, acionamento de tecla, comportamento ocular etc.). Entre parênteses estão indicadas as tabelas de unidades em que os atributos aparecem.

1. Dados de sessão (informações relativas às sessões de um estudo):
    - Study: nome do estudo como consta no TPR-DB (AU, EX, PU, SG, SS, ST, TT);
    - Session: nome da sessão, um composto formado a partir da identificação dada ao participante, ao texto e à tarefa (AU, CU, EX, PU, SG, SS, ST, TT);
    - Text: identificação do texto utilizado no estudo (AU, SS, ST, TT);
    - Task: tipo de tarefa, cf. Apêndice 2 (AU, SS, ST, TT);
    - Part: identificação do participante do estudo (AU, ST, TT, SS);
    - SL: língua do texto-fonte (AU, SS, ST, TT);
    - TL: língua do texto-alvo (AU, SS, ST, TT);
    - Break: duração do intervalo da sessão (SS);
    - TimeR: tempo de início da fase de revisão (SS);
    - TimeD: tempo de início da fase de redação (SS).
2. Dados do segmento (informações relativas aos segmentos):
    - Seg: identificação do segmento-fonte ou do segmento-alvo, dependendo do atributo Win, ou seja, da janela (FD);
    - STseg: identificação do segmento-fonte (AU, PU, SG, SS, ST);
    - Nedit: número de vezes que o segmento foi editado (SG);
    - TTseg: identificação do segmento-alvo (AU, CU, KD, PU, TT, SG, SS);
    - LenS: comprimento, em caracteres, do segmento-fonte (SG, SS);
    - LenT: comprimento, em caracteres, do segmento-alvo (SG, SS);
    - LenMT: comprimento, em caracteres, do segmento traduzido automaticamente (SG);
    - TokS: número de itens-fonte (*tokens*) no segmento (SG, SS);
    - TokT: número de itens-alvo (*tokens*) no segmento (SG, SS);
    - Literal: grau de literalidade de segmento (SG).

3. Itens (informações relativas aos itens, ou palavras, do texto-fonte e do texto-alvo):
    - STId: identificação única do item (*token*) do texto-fonte (FD, KD, PU, ST, TT);
    - TTId: identificação única do item (*token*) do texto-alvo (FD, KD, PU, ST, TT);
    - SAU: cadeia de caracteres do texto-fonte (AU);
    - TAU: cadeia de caracteres do texto-alvo (AU);
    - SAUnbr: número de itens (*tokens*) do lado-fonte da unidade de alinhamento (AU, ST, TT);
    - TAUnbr: número de itens (*tokens*) do lado-alvo da unidade de alinhamento (AU, ST, TT);
    - SToken: item (*token*) do texto-fonte (ST, TT);
    - TToken: item (*token*) do texto-alvo (ST, TT);
    - Lemma: lema do item (ST, TT);
    - PoS: classe gramatical do item (ST, TT);
    - PosS: classe gramatical da sequência do item-fonte (PU);
    - PosT: classe gramatical da sequência do item-alvo (PU);
    - Prob1: probabilidade de ocorrência do monograma (ST, TT);
    - Prob2: probabilidade de ocorrência do bigrama (ST, TT).
4. Métrica de literalidade tradutória:
    - AltT: número de diferentes alternativas de tradução (ST);
    - CountT: número de escolhas de tradução atuais observadas (ST);
    - ProbT: probabilidade de escolhas de tradução atuais (ST);
    - HTra: entropia de tradução da palavra (SG, ST);
    - HSeg: entropia de segmentação da tradução (SG, ST);
    - Cross: valor *Cross* do item (AU, ST, TT);
    - CrossS: valor *Cross* para item-fonte (PU, SG);
    - CrossT: valor *Cross* para item-alvo (PU, SG);
    - Literal: grau de literalidade de segmento (SG).
5. Acionamentos de teclas (informações relativas às atividades no teclado):
    - KDid: identificação do acionamento de tecla (KD);
    - Del: número de exclusões manuais e automáticas (AU, PU, ST, TT);
    - Ins: número de inserções manuais e automáticas (AU, PU, ST, TT);
    - Adel: número de exclusões realizadas automaticamente (SG, SS);

- Ains: número de inserções realizadas automaticamente (SG, SS);
- Mdel: número de exclusões realizadas manualmente (SG, SS);
- Mins: número de inserções realizadas manualmente (SG, SS);
- Char: caractere (UTF8) inserido ou excluído (KD);
- Munit: número de microunidades (AU, ST, TT);
- Edit: sequência de acionamentos de teclas na produção da cadeia no TT (AU, EX, FD, PU, ST, TT);
- Edit1: sequência de acionamentos de teclas da primeira microunidade (AU, ST, TT);
- Edit2: sequência de acionamentos de teclas da segunda microunidade (AU, ST, TT);
- InEff: medida de ineficiência para a geração do segmento (AU, ST, TT);
- Scatter: quantidade de produção textual não linear (PU, SG, SS).

6. Olhar sobre a janela-fonte e sobre a janela-alvo:
   - Path: sequência de fixações na janela do texto-fonte ou do texto-alvo (FU);
   - FFTime: marcação temporal do início da primeira fixação (ST, TT);
   - FFDur: duração da primeira fixação (ST, TT);
   - FPDurS: duração da primeira passagem na unidade do texto-fonte (AU, ST, TT);
   - FPDurT: duração da primeira passagem na unidade do texto-alvo (AU, ST, TT);
   - FixS: número de fixações na unidade do texto-fonte (AU, PU, SG, SS, ST, TT);
   - FixT: número de fixações na unidade do texto-alvo (AU, PU, SG, SS, ST, TT);
   - TrtS: tempo total de olhar (*gaze*) na unidade do texto-fonte (AU, SG, SS, ST, TT);
   - TrtT: tempo total de olhar (*gaze*) na unidade do texto-alvo (AU, SG, SS, ST, TT);
   - FixS1: número de fixações na unidade do texto-fonte durante a produção da primeira microunidade (AU, ST, TT);
   - FixS2: número de fixações na unidade do texto-fonte durante a produção da segunda microunidade (AU, ST, TT);
   - FixT1: número de fixações na unidade do texto-alvo durante a produção da

primeira microunidade (AU, ST, TT);

- FixT2: número de fixações na unidade do texto-alvo durante a produção da segunda microunidade (AU, ST, TT);
- RPDur: duração do percurso de regressão (ST, TT);
- Regr: valor booleano que indica se uma regressão partiu de um item (ST, TT).

7. Atividades concomitantes de olhar e acionamento de teclas:
   - ParalK: atividade de teclado paralela à atividade ocular (FU, FD);
   - ParalS: atividade ocular no texto-fonte paralela à digitação (PU);
   - ParalT: atividade ocular no texto-alvo paralela à digitação (PU);
   - ParalS1: atividade ocular no texto-fonte paralela à digitação da primeira microunidade (AU, ST, TT);
   - ParalS2: atividade ocular no texto-fonte paralela à digitação da segunda microunidade (AU, ST, TT);
   - ParalT1: atividade ocular no texto-alvo paralela à digitação da primeira microunidade (AU, ST, TT);
   - ParalT2: atividade ocular no texto-alvo paralela à digitação da segunda microunidade (AU, ST, TT).
8. Marcações temporais de início e duração de unidades e fases:
   - Dur: duração da produção da unidade (AU, CU, EX, FD, FU, PU, SG, SS, ST, TT);
   - Dur1: duração da produção da primeira microunidade (AU, ST, TT);
   - Dur2: duração da produção da segunda microunidade (AU, ST, TT);
   - Fdur: duração da produção do segmento excluindo pausas de digitação iguais ou superiores a 200 s (SG, SS);
   - Kdur: duração de atividade contínua de teclado excluindo pausas de digitação iguais ou superiores a 5 s (SG, SS);
   - Pdur: duração de atividade contínua de teclado excluindo pausas de digitação iguais ou superiores a 1 s (SG, SS);
   - Pnum: número de unidades de produção (SG, SS);
   - Time: marcação temporal de início da unidade (CU, EX, FD, FU, KD, PU);
   - Time1: marcação temporal de início da primeira microunidade (AU, ST, TT);
   - Time2: marcação temporal de início da segunda microunidade (AU, ST, TT);
   - TimeR: tempo de início da fase de revisão (SS);

- TimeD: tempo de início da fase de redação (SS).

9. Pausa anterior ao início de uma unidade:
   - Pause: pausa entre o fim da unidade anterior e o início da unidade atual (FU, PU);
   - Pause1: pausa entre o fim da unidade anterior e o início da primeira microunidade (AU, ST, TT);
   - Pause2: pausa entre o fim da unidade anterior e o início da segunda microunidade (AU, ST, TT).
10. Informações relativas à interface gráfica do usuário (GUI):
   - Win: janela em que foi gravada a atividade ocular, sendo "1" para texto-fonte e "2" para texto-alvo (FD);
   - Cursor: posição do caractere em que se registraram a atividade, os acionamentos de teclas e as fixações (FD, KD).
11. Fontes externas:
   - Focus: nome da janela em primeiro plano (EX);
   - KDidL: identificação do último acionamento de tecla anterior à saída da tela do Translog-II (EX);
   - KDidN: identificação do primeiro acionamento de tecla posterior ao retorno à tela do Translog-II (EX);
   - STidN: identificação do primeiro item (*token*) do texto-fonte posterior ao retorno à tela do Translog-II (EX);
   - STidL: identificação do último item (*token*) do texto-fonte anterior à saída da tela do Translog-II (EX);
   - STsegL: identificação do segmento (do texto-fonte) do último evento (EX);
   - STsegN: identificação do segmento (do texto-fonte) do evento subsequente (EX).
12. Outros atributos:
   - Type: tipo de acionamento de tecla: [AM]ins, [AM]del (KD);
   - Type: tipo de unidade de atividade, como explicado na Seção 4.5 (CU);
   - Label: rótulo das unidades de atividade (CU).