

# CFT13: A Resource for Research into the Post-Editing Process

Michael Carl, Mercedes García Martínez, Bartolomé Mesa-Lao, Nancy Underwood

Centre for Research and Innovation in Translation and Translation Technology

Copenhagen Business School

Denmark

E-mail: mc.abc@cbs.dk, mgm.abc@cbs.dk, bm.abc@cbs.dk, nancy.underwood@gmail.com

## Abstract

This paper describes the most recent dataset that has been added to the CRITT Translation Process Research Database (TPR-DB). Under the name CFT13, this new study contains user activity data (UAD) in the form of key-logging and eye-tracking collected during the second CasMaCat field trial in June 2013. The CFT13 is a publicly available resource featuring a number of simple and compound process and product units suited to investigate human-computer interaction while post-editing machine translation outputs.

**Keywords:** eye-tracking, key-logging, post-editing, translation process research.

## 1. Introduction

In today's globalised world, machine translation (MT) is increasingly being adopted by companies worldwide. The growing demand for quick and accurate translations, triggered through the technological advances in the field of natural language processing, has put MT in the spotlight in modern translation workflows. However, despite the fact that post-editing of MT outputs increasingly plays a role in the translation industry, there are many open questions that need to be addressed in order to better understand post-editing as compared to translation (Mesa-Lao 2014).

Against this background, the CasMaCat project is building a computer-assisted translation workbench for post-editing of MT. Within the project lifetime, three major field trials are planned to assess and evaluate the development of the workbench. The collected user activity data (UAD), in the form of key-logging and eye-tracking data, is analysed to examine how different GUIs are used, to determine post-editing types and styles, as well as to build a cognitive model of the translation and post-editing process.

The UAD collected in the second CasMaCat field trial in June 2013 has been added to the CRITT Translation Process Research Database (TPR-DB) under the study folder CFT13. Apart from the raw key-logging and eye-tracking data, a number of product and process units have been extracted and stored in the form of tables. Units (i.e. rows) in these tables are described by features, which are instrumental for translation process research.

This paper describes the type of data collected during the second CasMaCat field trial in June 2013 (CFT13) using the second prototype of the workbench. Section 2 describes this version of the prototype and the setup of the second field trial to test its functionalities. Section 3 introduces the CRITT TPR-DB, explains the underlying data structure, the post-processing steps in which various kinds of units deriving from the CFT13 were extracted and describes some of the features of these units.

## 2. The CasMaCat workbench

CasMaCat (*Cognitive Analysis & Statistical Methods for Advanced Computer Aided Translation*)<sup>1</sup> is an EU-funded project that is currently developing a post-editing workbench in order to improve productivity, quality, and work practices in the translation industry (Alabau *et al.* 2013, Koehn *et al.* 2013, Mesa-Lao 2012). The CasMaCat workbench implements a number of innovative features aiming at supporting translators while post-editing MT. Some of such features are:

- *Interactive translation prediction:* the system interactively produces new MT outputs taking into account the input entered by the post-editor while typing.
- *Visualization of confidence measures:* the system visually informs the user at the word level about the likelihood of the MT provided (words painted in red have a low confidence measure in the MT system).
- *Word alignment information:* the system displays the correspondences between both the source and target words.

These features can be individually enabled or disabled according to the preferences of the post-editor.

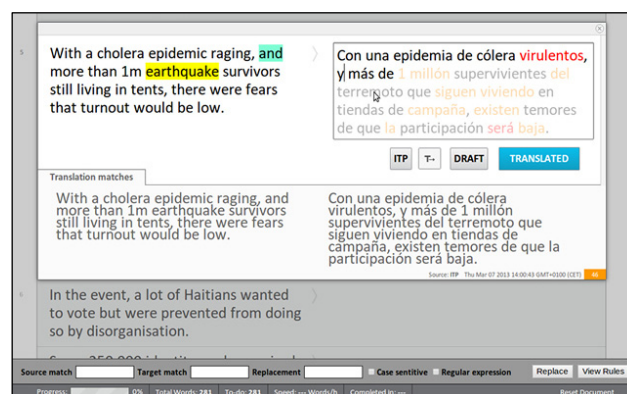


Figure 1: Graphical user interface of the second prototype of the CasMaCat workbench.

<sup>1</sup> Project 287576 (FP7 ICT-2011.4.2).

Figure 1 shows a screenshot of the second prototype of the CasMaCat workbench.

The window on the left side plots a source text segment, while the window on the right side contains the MT output provided by the system. Confidence scores are painted in the editing window (black: above confidence threshold; red: below confidence threshold). Alignment information is shown in the source text window highlighting in yellow the corresponding word over which the mouse in the target window hovers and in green the word corresponding to the cursor position in the target window.

An innovative feature of the CasMaCat workbench is the user activity logging module. Key-logging and eye-tracking data are logged allowing for detailed process studies. During the design and implementation of the workbench, the logged data is used to gain knowledge about the post-editing process. The primary goal is to investigate post-editing styles and to elaborate a cognitive model of the post-editing process. Based on these empirical data, novel types of assistance for human post-editors can be further implemented.

## 2.1 The CFT13 study

The three field trials in the CasMaCat project aim at (1) assessing the functionality and usability of the workbench being implemented, and (2) eliciting user behaviour and user feed-back for further development of successive workbench prototypes. The first field-trial, held in June 2012, compared from-scratch translation to post-editing with prototype-I of the CasMaCat workbench. Results of the first field trial are reported in Elming *et al.* (2014).

The main aim of this second field trial was to collect product and process data during a series of post-editing tasks using three different set-ups of the second prototype of the workbench. The three different GUI set-ups of the CasMaCat workbench used to perform nine different tasks were:

- Traditional post-editing with no assistance during the process (PE).
- Post-editing through basic interactive translation prediction (ITP), where the post-editors were presented with alternative ways to complete sentences as they typed their changes.
- Post-editing through advanced interactive translation prediction (AITP) featuring different visualisations (e.g. confidence measure and word alignments) while the translator interactively post-edited the raw MT output.

The field trial involved nine professional post-editors and four reviewers. Each of these GUI set-ups was used by the nine post-editors to complete nine different tasks. The order of presentation of these different GUIs was distributed among tasks and post-editors so as at least two of the nine tasks were performed in each of these three set-ups. The four reviewers only worked with the PE set-up and their role was to proofread the target texts produced by the post-editors.

Each task consisted of post-editing approximately 1,000 English source text words previously machine-translated into Spanish by a Moses system (Martínez-Gómez 2012). The nine texts involved in this CFT13 study were pieces of news and consisted of 460 different source segments which were post-edited by all nine different translators using the different GUI set-ups. Each text consisted of 30 to 63 segments. The quality expectations after completing each post-editing session were high, so that the post-editor should edit the text to produce an accurate and readable Spanish version of the source text. No time pressure was imposed, but all post-editors were instructed not to make preferential changes in the raw MT output in order to save as much time as possible during the task.

Three out of the nine post-editing tasks for each post-editor were recorded from CELER Soluciones SL<sup>2</sup>, where an eye-tracker device (Eyelink 1000) was used to record gaze behaviour in combination with the keystroke logging function in the CasMaCat workbench. The remaining six sessions were performed by the translators from home without access to an eye-tracker, but with keystroke logging enabled. For a detailed description of the second CasMaCat field trial see Underwood *et al.* (2014).

Due to logging problems, 74 segments were lost so that a total of 4,064 segments were finally logged and included in the CFT13 study. These segments amount to 94,865 English source tokens (average 23 source text words/segment) translated into 101,671 Spanish words (average 25 words/segment). 977 of the post-edited segments also have associated gaze data. A total of 1,115 post-edited segments were also reviewed by four different reviewers. 752 of these reviewed segments also have gaze information. Table 1 provides an overview of the collected data available in CFT13, indicating that almost all the 460 different segments have been translated by three translators in all three GUIs.

GUI	#Segments	Segments with gaze data	Segments reviewed
PE	1337	356	372
ITP	1359	296	371
AITP	1368	325	372

Table 1: Data overview of the CFT13 study

The CFT13 study has been added to the TPR-DB version 1.4, which will be described in the next section.

## 3. The CRITT TPR-DB

The Center for Research and Innovation in Translation and Translation Technology (CRITT) at the Copenhagen Business School has been involved for more than 10 years in Translation Process Research (TPR). Since 2012 different studies carried out in the framework of the CRITT research have been compiled and processed in the form of a publicly available TPR database. The aim of the

<sup>2</sup> CELER Soluciones SL is a language service provider based in Madrid and one of the partners of the CasMaCat project.

TPR-DB is to give rise to more grounded translation models and to extend understanding of the underlying human translation processes.

The CRITT TPR-DB contains a large number of recorded translation, post-editing, editing and authoring sessions with consistent and comparable feature representations. The user activity data (UAD) has been recorded either with the different versions of Translog (Carl 2012) or with the CasMaCat workbench (see section 2). Both sources of UAD have been processed and compiled into a consistent data format (Carl and Jakobsen 2009) and annotated with metadata about the sessions and the participants in each session (Hvelplund and Carl 2012). This consistent data format also makes possible the UAD to be compiled and processed by various visualization and analysis tools (e.g. R, SPSS, Excel, etc.).

The first version of the CRITT TPR-DB was released in May 2012 with a total of ten studies with 456 (translation, post-editing, editing or text copying) sessions (Carl 2012). At its present state, the CRITT TPR-DB contains a total of 24 studies with more than 1,000 sessions. While the raw logging data of different sessions is available in XML format and can be downloaded for analysis from the TPR-DB website<sup>3</sup>, an already processed version of the UAD is also provided with a number of features which can be directly used for a variety of investigations based on the studies in the database.

### 3.1 Linking process and product data

The CRITT TPR-DB links translation product and process data in such a way so that hypothesis about relations between text perception (reading), text production (writing) and the final translation product can be easily formulated and quantified.

The experimental setup in both Translog-II and CasMaCat recordings is similar in that a (chunk of) text is shown on a computer screen for which a translation must be produced from scratch or which is a first MT draft of the text which must be post-edited. While the recorded gaze data of the translator or post-editor allows us to investigate how a text is perceived, the keystroke data provides us with the trace of how a target text is produced (or post-edited). As we cannot look into the black box of the translator's mind to understand the hidden translation processes, the TPR-DB model approximates the translation process via linked product and process data.

The underlying data model of the TPR-DB is shown in Figure 2. The left side plots the translation product with

alignment links between the source text and target text on a word level. The right side of the figure shows the process data and a list of UAD, fixations and keystrokes. Each activity is assigned a time stamp and a position in the source or target text to which it refers. The TPR-DB, thus, maps the traces of text production (keystrokes) and the traces of text perception (gaze samples and fixations) on the textual items which they refer to, and relates the process activities via the alignment links between source text and target text tokens. In this way we can relate the process activities through the symbols in the product data.

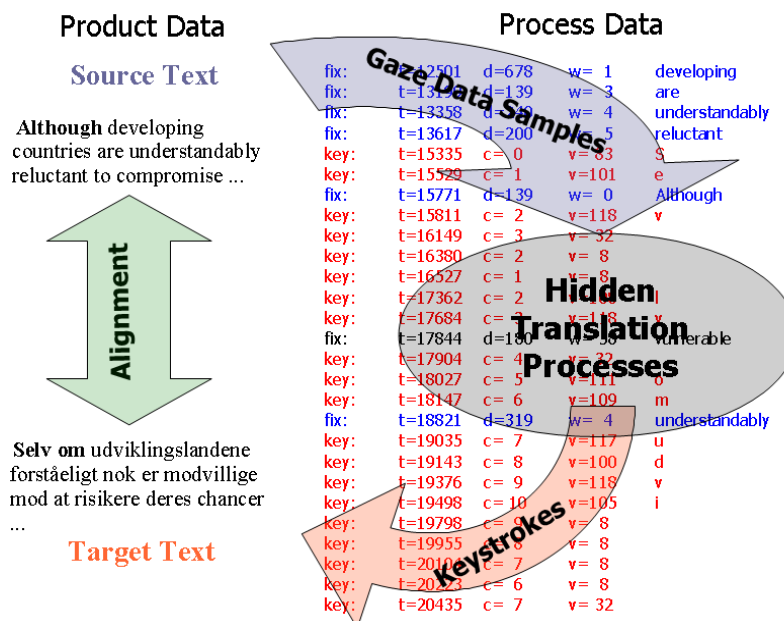


Figure 2: Links between product and process data in the CRITT TPR-DB

For instance, when investigating the eye-key span (Dragsted 2010), we need to correlate the temporal distance between gaze activities on a source text token and the production keystrokes of the translation of the word gazed at. This can only be achieved via an alignment of source and target words and by knowing which keystrokes contribute to the production of the translation in question. Similarly, if we want to investigate the relation between the production time for certain tokens and the gaze duration on these tokens, we need to align the tokens in source and the target texts (i.e. the product data) and link the events of the process data (keystrokes and gaze) to the product items. We can then relate the gaze events via the aligned product tokens with the keystroke events (Carl, Gutermuth and Hansen-Schirra 2014; Balling and Carl 2014). Linked product and process data is also required when mapping sequences of fluent text production (PUs) on the PoS tags of the words which are produced (Wilker, Koponen and Specia 2014).

As will be shown below, the CRITT TPR-DB compiles a large number of features that allow for correlating different modalities and dimensions easily with standard statistical tools. By linking product and process data across large amounts of translation sessions, it opens unprecedented avenues for empirical research into translation process research.

<sup>3</sup> This resource is available from the following website: CRITT Translation Process Database (TPR-DB) [on-line] <[http://bridge.cbs.dk/platform/?q=CRITT\\_TPR-db](http://bridge.cbs.dk/platform/?q=CRITT_TPR-db)>. [Last accessed: March 20, 2014]

### 3.2 The CRITT TPR-DB version 1.4

Rec	SL	TL	
40	da	en	The CRITT TPR-DB currently contains more than 1,000 text production sessions (translation, post-editing, editing, copying, authoring) in more than 10 different languages. This amounts to more than 300 hours of text production data. Table 2 shows the number of recordings (Rec) together with their source (SL) and target languages (TL).
111	en	da	
156	en	de	
30	en	en	
347	en	es	
15	en	et	
36	en	fa	
103	en	hi	
20	en	pt	
44	en	zh	
2	es	es	Table 2: Language pairs in the CRITT TPR-DB v.1.4
10	pt	en	
24	zh	pt	
			More than 400 of these recordings are translation sessions, and more than 300 are post-editing sessions. The TPR-DB v. 1.4

consists of post-processed tables (TPR data) which contain extracted product and process units and a large number of features. In contrast to TPR-DB version 1.0 (Carl 2012), the current version separates the raw UAD and the derived TPR data in two separate locations both of which can be downloaded independently and anonymously. The raw UAD is stored and maintained in a subversion repository<sup>4</sup>, and the post-processed TPR data, which can be downloaded as a zip file<sup>5</sup>.

While the structure and content of the UAD collected from Translog-II and from CasMaCat is different in terms of the attributes and containers, the data undergoes a similar compilation process (Carl 2012) to generate a number of tables, which can then be used as a basis for further analysis.

### 3.3 Compiling CRITT TPR-DB version 1.4

The raw logged UAD data retrieved from the translation sessions and stored in *Translog-II* folder. It is subsequently processed in two independent streams (1) to annotate the product data and (2) to annotate the process data. Figure 3 shows a processing workflow of the TPR-DB compilation process. Annotations of the product data, such as tokenized and lemmatized forms, PoS tags, word and sentence alignments are stored in an *Alignment* folder. The Yawat alignment tool (Germann 2008) can be used to correct semi-automatically word alignments.

<sup>4</sup> The raw logged and aligned data can be checked out via: [svn co https://130.226.34.13/svn/tpr-db/](https://130.226.34.13/svn/tpr-db/)

<sup>5</sup> The CRITT TPR-DB v. 1.4 can be downloaded from: <http://bridge.cbs.dk/resources/tpr-db/TPR-DBv1.4.3.zip>

A data integration step computes keystroke-to-token and fixation-to-token mappings, which are then stored in an *Events* folder. Finally, unit *Tables* are produced, which contain a large number of features allowing for advanced visualization, analysis and modelling of translation processes. As mentioned before, while the raw UAD and the alignment data are maintained in the SVN repository, information generated during the data integration and evaluation step (i.e. events and the table files) are available in the form of CRITT TPR-DB zip files.

### 3.4 Feature tables in the TPR-DB

In this section we outline the structure of the *Tables* that are produced in the post-processing step of the TPR-DB. As illustrated in Figures 2 and 3, the available process data consist of two basic types of information: (1) keystroke data (KD), i.e. information about the time and kind of text insertions and deletions, and (2) information

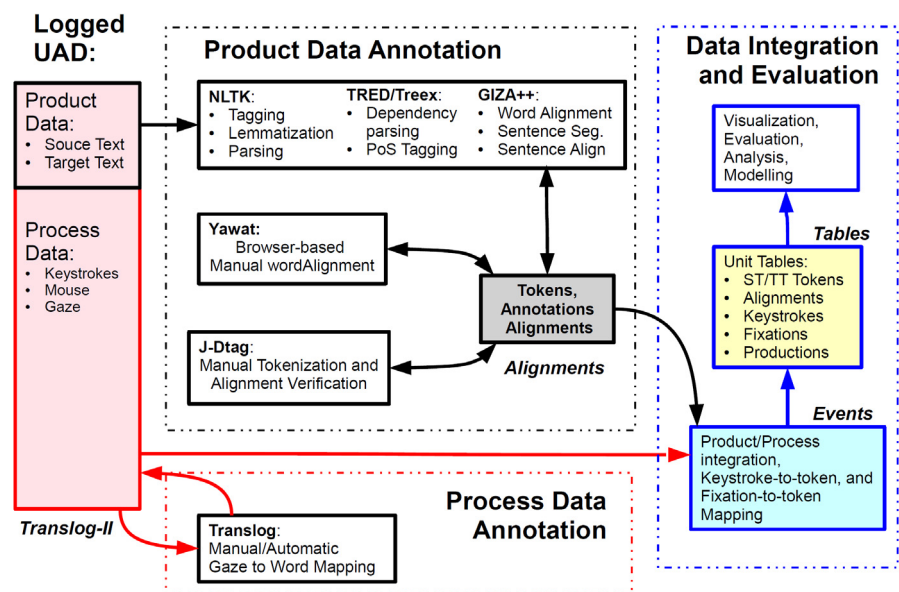


Figure 3: Architecture of the TPR-DB compilation process

From these data two complex process units are derived: text production units (PU) and fixation units (FU). PUs represent sequences of coherent sequences of writing and FUs represent sequences of coherent reading. From the final translation product are extracted three extra units: source text tokens (ST), target text tokens (TT) and alignment units (AU), which describe transitive closure between sets of aligned source and target tokens.

The following list presents the different features present in the data (the letters in brackets represent the file extensions in the database):

- **Keystrokes (KD):** basic text modification operations (insertions or deletions), together with time of stroke, and the word in the final target text to which the keystroke contributes.
- **Fixations (FD):** basic gaze data of text fixations on the source or target text, defined by starting time, end time and duration of fixation, as well as character offset and word index of fixated symbol in the source or target window.

- **Production units (PU)**: coherent sequence of typing (Carl and Kay 2011), defined by starting time, end time and duration, percentage of parallel reading activity during unit production, duration of production pause before typing onset, as well as number of insertion and deletions.
- **Fixation units (FU)**: coherent sequences of reading activity, including two or more subsequent fixations, characterized by starting time, end time and duration, as well as scan path indexes to the fixated words.
- **Activity units (CU)**: exhaustive segmentation of the session recordings into activities of typing, reading of the source or reading of the target text.
- **Source tokens (ST)**: as produced by a tokenizer, together with TT correspondence, number, and time of keystrokes (insertions and deletions) to produce the translation; micro unit information.
- **Target tokens (TT)**: as produced by a tokenizer, together with ST correspondence, number, and time of keystrokes (insertions and deletions) to produce the token; micro unit information.
- **Alignment units (AU)**: transitive closure of ST-TT token correspondences, together with the number of keystrokes (insertions and deletions) needed to produce the translation, micro unit information and amount of parallel reading activity during AU production.
- **Segments (SG)**: describe the source and target segments, annotated with the number of keystrokes, insertions, deletions, fixations on the source and the target side.
- **Session (SS)**: this table describes some properties of the sessions, such as source and target languages, total duration, beginning and end of drafting phase, etc.

In summary, these UAD can be analysed from different angles in a more comprehensive manner than the original logging data and may serve as a basis for further investigation for anyone wanting to use the resource.

### 3.5 Feature representation in CFT13

In this section we include further details about the features contained in the *Tables* folder of CF13 study. Each table describes a particular type of unit, where columns encode various features which characterize the event or unit.

The **keystrokes** table encodes keystroke events in time with no duration. Each keystroke is characterized by the *Time* at which it was produced, a *Type* indicating whether it was an insertion or deletion, and the *Cursor* offset at which the text was modified. The actual character (*Char*) which was inserted is also coded, as well as the target text token (*TTid*) to which the keystroke has contributed and the source text token (*STid*) for which the TTid is the translation.

Regarding **fixation data**, the following information is provided in the table: beginning of a fixation (*Time*) and its duration (*Dur*), the window (*Win*) in which the fixation was recorded, 1 for source and 2 for target window. *Cursor* encodes the offset of the closest character at which the center of the fixation was detected as it emerged, while *STid* and *TTid* refer to the final text.

Production units files are sequences of coherent typing activity (Carl and Kay 2011) and, as such, the **production units** tables provide a temporal beginning (*Time*), a duration (*Dur*) and they may cover one or more insertions or deletions (*Edit*) contributing to build up one or more target text tokens (*TTid*). Pauses (*Pause*) in milliseconds between production units are also registered. *ParalS* and *ParalT* record the amount of parallel keyboard activity.

Similarly to production units, **fixation units** indicate sequences of coherent reading behaviour. As with PUs, FUs tables include *Time*, duration (*Dur*), and boundaries between FUs (*Pause*). Under *Path* the table includes the sequence of words the post-editor looked at including the word ID. The column *ParalK* indicates the amount of parallel keyboard activity during reading.

For **activity units** (CUs) the following distinction between the following three basic types of activities is made:

- *Type 1*: source text reading.
- *Type 2*: target text reading.
- *Type 4*: translation typing.

Four composed types of activities are also considered:

- *Type 5*: translation typing while ST reading.
- *Type 6*: translation typing while TT reading.
- *Type 7*: translation typing while (alternating) ST and TT reading.
- *Type 8*: no activity recorded.

All CUs are presented with a starting time (*Time*), duration (*Dur*), and the segment number (*Seg*) in which it takes place.

The **alignment unit** table provide a similar kind of information for these different units concerning source/target correspondences (*TAU* and *SAU*), keystrokes to produce the translation (*Ins* and *Del*), Duration (*Dur*), as well as total gaze data on source and target tokens (*GazeS*, *GazeT*) and number of fixations (*FixS*, *FixT*). Under the *InEff* feature, an editing inefficiency score is provided calculating the number of insertions and deletions divided by their difference. The *Cross* feature represents alignment information from a procedural perspective. Source and target cross values represent the relative local distortion of the reference text with respect to the output text and indicate how many words need to be process in the source text to produce the next token(s) in the translation.

**Source tokens** and **Target tokens** provide similar information which relates to source and target tokens, instead of alignment units.

### 3.6 Segment features in CFT13

CFT13 features a special table due to the special setting of this study compared to other studies in the CRITT TPR-DB. As mentioned in section 2.1, the second CasMaCat field trial consisted of post-editing with different translation options and via the interactivity provided by the IPT and AITP GUIs.

SSeg	STseg	TTseg	Nedit	Dur	Kdur	Fdur	FixS	GazeS	FixT	GazeT	Mins	Mdel	Ains	Adel	Toks	Lens	TokT	LenT	SegLit	SegH	STCr2	TTCr2	RSTseg	RTTseg	Text	PE	Gui	RSegs	RE	TER	MTPE
0	1	1	2	21937	12924	21937	2	104	31	11596	39	40	61	0	12	53	14	59	19690	1272	123	220	4337	4337	1.1	1	PE	4337	10	0.55	
0	461	461	3	49480	10485	49480	6	1108	53	17880	20	21	187	126	12	53	14	59	9543	1272	115	60	5141	5141	1.1	5	PE	5141	13	0.55	
0	921	921	4	103319	16116	103319	79	18804	59	21196	31	18	183	122	12	53	17	73	49016	1272	196	198	5389	5389	1.1	9	PE	5389	12	0.43	
0	1381	1381	1	56493	14663	56493	1	36	21	5760	15	10	31	31	12	53	15	64	30995	1272	144	198	4690	4690	1.1	3	ITP	4690	11	0.33	
0	1841	1841	3	93748	9485	93748	66	16176	137	42164	14	3	83	24	12	53	15	69	10199	1272	116	170	4814	4814	1.1	4	ITP	4814	11	0.33	
0	2301	2301	2	43650	12631	43650	7	1496	52	20160	20	7	115	63	12	53	14	64	14574	1272	107	196	5062	5062	1.1	8	ITP	5062	12	0.27	
0	3221	3221	2	209546	41197	209546	9	1304	84	18200	43	35	216	160	12	53	14	63	17005	1272	131	200	5340	5340	1.1	6	AITP	5340	13	0.55	
0	3681	3681	3	117944	11521	117944	149	40072	169	49968	25	32	89	22	12	53	13	59	5757	1272	106	140	4968	4968	1.1	7	AITP	4968	12	0.6	
1	462	462	1	42452	37881	42452	26	2860	115	27024	38	38	250	128	24	112	25	122	20722	1699	159	108	5142	5142	1.1	5	PE	5142	13	0.43	
1	922	922	2	58338	0	58338	0	0	0	0	0	0	0	0	24	112	30	143	37054	1699	238	104	5390	5390	1.1	7	AITP	5390	12	0.5	
1	1382	1382	1	37803	11489	37803	29	6456	68	12984	4	5	127	2	24	112	27	124	15011	1699	117	82	4691	4691	1.1	3	ITP	4691	11	0.13	
1	1842	1842	2	90823	27158	90823	37	7136	199	65680	34	23	283	166	24	112	28	128	19165	1699	108	60	4815	4815	1.1	4	ITP	4815	11	0.16	
1	2302	2302	1	48206	17045	48206	39	8772	102	35348	32	11	150	45	24	112	26	126	13652	1699	78	80	5063	5063	1.1	8	ITP	5063	12	0.3	
1	2762	2762	1	111458	41536	111458	45	9460	184	74820	29	47	180	34	24	112	27	128	21155	1699	163	98	4597	4597	1.1	2	AITP	4597	10	0.32	
1	3222	3222	1	84337	31901	84337	0	0	0	0	33	40	205	75	24	112	26	123	16446	1699	100	74	5341	5341	1.1	6	AITP	5341	13	0.39	
1	3682	3682	1	140009	112139	140009	26	6132	265	102436	94	122	309	168	24	112	24	113	10731	1699	98	98	4969	4969	1.1	7	AITP	4969	12	0.57	
2	3	3	1	57083	17324	57083	6	764	62	28336	37	32	108	0	20	110	20	113	13816	2127	71	83	4339	4339	1.1	1	PE	4339	10	0.39	
2	463	463	1	76673	25065	76673	66	13196	183	50012	51	33	108	0	20	110	22	126	48968	2127	495	1010	5143	5143	1.1	5	PE	5143	13	0.65	
2	923	923	2	226012	24471	226012	337	56996	135	34432	191	72	0	0	20	110	24	119	68359	2127	473	988	5391	5391	1.1	9	PE	5391	12	0.64	
2	1383	1383	1	95399	34551	95399	65	14796	138	31436	23	20	403	294	20	110	19	112	16388	2127	277	856	4692	4692	1.1	3	ITP	4692	11	0.35	
2	1843	1843	1	157440	68767	157440	79	17876	252	82800	67	10	622	564	20	110	19	115	14559	2127	281	842	4816	4816	1.1	4	ITP	4816	11	0.53	
2	2303	2303	1	59500	22063	59500	59	13512	101	30808	27	18	361	259	20	110	20	111	17907	2127	311	824	4598	4598	1.1	8	ITP	4598	12	0.44	
2	2763	2763	1	110791	48667	110791	54	13374	186	57520	68	13	724	664	20	110	20	115	15477	2127	309	830	4598	4598	1.1	2	AITP	4598	10	0.5	
2	3223	3223	1	125283	29557	125283	0	0	1	20	57	17	557	487	20	110	22	110	31841	2127	357	854	5342	5342	1.1	6	AITP	5342	13	0.65	
2	3683	3683	2	187048	70768	187048	105	26448	351	120252	106	120	485	355	20	110	21	116	20019	2127	353	834	4970	4970	1.1	7	AITP	4970	12	0.53	

Table 3: CFT13 Segment information

This study included also a revision phase performed by a person different from the post-editor. These different features are taking into account in a separate segment table<sup>6</sup>.

Table 3 presents an excerpt of the CFT13 segment table. Each line represents a different segment which is identified by the source text segment number (*STseg*), target text segment number (*TTseg*), the source segment (*SSeg*) and the revision segment (*RSeg*). The *SSeg* identifier enumerates the different source segments. That is, since each source segment has been translated by nine different translators, identical *SSeg* will refer to the same source segment. As pointed out in section 2.1, there were 460 different source segments, numbered from 0 to 459. The *STseg* and *TTseg* numbers are internal identifiers in the CasMaCat system and refer to the indexed versions of the source and target segments. *RSeg* is also a CasMaCat internal identifier with points to the revised version of the data.

A number of features characterize the translation processes involved in the production of this segment, such as: the number of times the segment was open during the post-editing session (*Nedit*), time spent working on the segment (*Dur*), duration of coherent keyboard activity excluding pauses with no keystroke activities of 5 seconds or more (*Kdur*), duration of segment production time excluding keystroke pauses above 200 seconds (*Fdur*), fixations on source text (*FixS*), total gaze time in source text (*GazeS*), fixation count on target text (*FixT*), total gaze time on target text (*GazeT*), manual insertions (*Mins*), manual deletions (*Mdel*), automatic insertions triggered by ITP or AITP (*Ains*), automatic deletions triggered by ITP or AITP (*Adel*), source tokens (*Toks*), target tokens (*TokT*), entropy of lexical translation realisations (*SegH*), source text cross value (*STCr2*), target text cross value (*TTCr2*), segment literality in relation to the cross feature between source and target tokens (*SegLit*). These latter features are described in detail in (Schaeffer and Carl, 2014)

As many of the segments have also been revised, the same features are generated for the reviewing process. In this case they are prefixes with an *R* to become (*RSTseg*, *RTTseg*, *RNedit*, *RDur*, *RKdur*, *RFdur*, *RFixS*, *RGazeS*, *RFixT*, *RGazeT*, *RMins*, *RMdel*, *RTokT*, *RSegLit*, *RSTCr2*, *RTTCr2*).

Each of the segments in the CFT13 study has also information about the text to which the segment belongs to (*Text*), the GUI used while post-editing the segment (*Gui*), the ID of the post-editor who worked on the segment (*PE*) and the ID of the reviewer who proofread the segment (*RE*).

In addition, the segment table also contains TER<sup>7</sup> values between five different strings of text derived from the

<sup>6</sup> The CFT13 segment table can be downloaded from: [http://bridge.cbs.dk/resources/tpr-db/CFT13-PE-RE\\_seg](http://bridge.cbs.dk/resources/tpr-db/CFT13-PE-RE_seg)

<sup>7</sup> TER value: *Translation Error Rate* is an error metric for MT that measures the number of edits required to change a system output into one of the references.

CFT13 described in section 2.1:

- *MT*: output
- *PE*: post-edited version
- *RE*: revised version
- *R1*: a reference translation with the minimal essential changes introduced following the post-editing guidelines provided to the post-editors
- *R2*: a second reference translation produced without following post-editing guidelines, i.e. as it was found in the reference corpus

There are thus nine different edit distances comparing all possible combinations of these strings (*TER\_MTPE*, *TER\_MTRE*, *TER\_MTR1*, *TER\_MTR2*, *TER\_PERE*, *TER\_PER1*, *TER\_PER2*, *TER\_RER1*, *TER\_RER2*).

Finally, together with all these quantitative data, the segment table also includes a qualitative study of the changes made by the post-editors (Mesa-Lao *et al.* 2014). This human evaluation of post-editing changes includes information at the segment level for: the number of essential changes introduced (*ChEss*), number of preferential changes made (*ChPref*), number of essential changes not implemented (*ChNot*) and number of new errors introduced by the post-editor (*ChErr*).

### 3.7 Visualization of CRITT TPR-DB features

Figure 3 shows a visualization of the keystroke and fixation data, and segment boundaries as collected in a post-editing session of CFT13. This figure shows a fragment of 140 words distributed over six different segments (horizontal dotted lines). Translation progression graphs visualize how translations emerge in time, enumerating the source text words on the vertical axis and the post-editing time on the horizontal axis.

We have adapted the visualization of translation progression graphs, which was first developed for Translog-II logging data, to both the CasMaCat Prototype-1 (Elming and Bonk 2012) and the CasMaCat Prototype-2 (Alabau *et al.* 2013, Koehn *et al.* 2013). When plotting the CasMaCat logging data, automatic deletions and insertions which are generated in the interactivity mode are plotted in grey. Accordingly, the representation and visualization schema was extended to take these additional features into account.

The various symbols in Figure 3 represent:

- Blue diamonds represent fixations on the source text.
- Green diamonds represent fixations on the target text.
- Black characters represent insertions.
- Grey characters represent automatic insertions.
- Red characters represent deletions.

The graph shows the temporal sequence when segments are loaded into the CasMaCat target buffer, when and where translators read the source segments or the MT outputs, as well as when they make corrections in the form of insertions and deletions. Translation progression graphs show only a small fraction of the information in the CRITT TPR-DB tables, but are instrumental to qualitatively inspect and explore the UAD and to analyze and visualize translation processes.

## 4. Conclusion

The paper described the CFT13 translation data, consisting of 81 post-editing sessions from nine different post-editors. The data was collected using the second prototype of the CasMaCat workbench and now is part of the CRITT TPR-DB. By adding data from various languages, different working environments and user profiles, hypotheses about translation and post-editing styles, as well as translator types, can certainly become more reliable. Rather than describing, explaining or predicting isolated phenomena, the ultimate aim of translation process research should be to produce and verify hypotheses which will make it possible to describe, explain and predict translation and post-editing processes.

We believe that the CFT13 study can contribute to gain new insights into the post-editing process. Exploratory approaches investigating the latent relationships between the features included in this study can be used to detect, characterize and classify behavioural patterns, to uncover different styles of reading and writing in post-editing tasks, and to describe post-editing as a form of human-machine interaction.

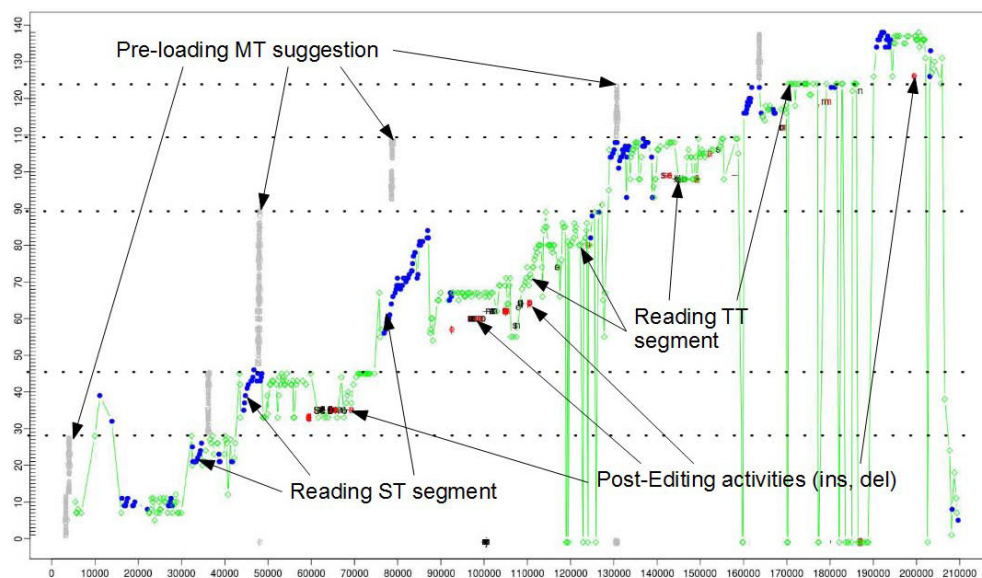


Figure 3: Progression graph plotting keystroke and gaze information during a post-editing session

## 5. Acknowledgements

This work has been supported by the CasMaCat project, co-funded by the European Union under the Seventh Framework Programme, project 287576 (ICT-2011.4.2). We are grateful to the nine post-editors and four reviewers who took part in the second field trial, whose UAD contributed to build the CFT13 study.

## 6. References

- Alabau, V., Bonk, R., Buck, Ch., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, Ph., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, C., Tsoukala, Ch. (2013). CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation. *The Prague Bulletin of Mathematical Linguistics*, 100, pp. 101--112.
- Aziz, W., Koponen, M., Specia, L. (2014). Sub-sentence Level Analysis of Machine Translation Post-editing Effort. In S. O'Brien *et al.* (Eds.), *Expertise in Post-Editing: Processes, Technology and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 170--200.
- Balling, L.W.; Carl, M. (2014). Production time across languages and tasks: a large-scale analysis using the CRITT translation process database, In J.W. Schwieter, A. Ferreira (Eds.), *Psycholinguistic and cognitive inquiries in translation and interpretation studies*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Carl, M. (2012). The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research. *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice*. Stroudsburg, PA: Association for Machine Translation in the Americas (AMTA), pp. 9--18.
- Carl, M., Gutermuth, S., Hansen-Schirra, S. (2014). Post-editing machine translation - a usability test for professional translation settings. In J.W. Schwieter, A. Ferreira (Eds.), *Psycholinguistic and cognitive inquiries in translation and interpretation studies*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Carl, M., Jakobsen, A.L. (2009). Towards statistical modelling of translators' activity data, *International Journal of Speech Technology*, 12(4), pp. 125--138.
- Carl, M., Kay, M. (2011). Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators. *Meta*, 56 (4), pp. 952--975.
- Dragsted, B. (2010). Coordination of reading and writing processes in translation: An eye on uncharted territory. In G.M. Shreve, E. Angelone (Eds.), *Translation and Cognition*. Amsterdam/Philadelphia: John Benjamins, 381, pp. 41--62.
- Elming, J., Balling, L.W., Carl, M. (2014). Investigating User Behaviour in Post-Editing and Translation Using the CasMaCat Workbench. In S. O'Brien *et al.* (Eds.), *Expertise in Post-Editing: Processes, Technology and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 148--169.
- Germann, U. (2008). Yawat: Yet Another Word Alignment Tool. *Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, pp. 20--23.
- Hvelplund, K.T., Carl, M. (2012). User Activity Metadata for Reading, Writing and Translation Research. In V. Arranz *et al.* (Eds.), *Proceedings of The Eighth International Conference on Language Resources and Evaluation. LREC 2012*. Paris: ELRA, pp. 55--59.
- Koehn, Ph., Carl, M., Casacuberta, F., Marcos, M. (2013). CASMACAT: Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation. In M. Simard *et al.* (Eds.), *Proceedings of the XIV Machine Translation Summit* (Nice, September 2--6, 2013), pp. 411.
- Martínez-Gómez, P., Sanchis-Trilles, G.; Casacuberta, F. (2012). Online adaptation strategies for statistical machine translation in post-editing scenarios. *Pattern Recognition*, 45(9), pp. 3193--3203.
- Mesa-Lao, B. (2012). The next generation translator's workbench: post-editing in CASMACAT v.1.0. In *Translating and the Computer Conference Proceedings*, 34.
- Mesa-Lao, B. (2014). Gaze Behaviour on Source texts: An Exploratory Study Comparing Translation and Post-editing. In S. O'Brien *et al.* (Eds.), *Expertise in Post-Editing: Processes, Technology and Applications*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 219--245.
- Mesa-Lao, B., Carl, M., Schaeffer, M. (2014). Human Evaluation of Machine Translation. An investigation of post-editing changes, To appear in *Proceedings of the FIT XXth World Congress* (Berlin, August 4--6, 2014).
- Schaeffer, M., Carl, M. (2014). Measuring the Cognitive Effort of Literal Translation Processes. *Proceedings of Human in the Loop: Workshop on Humans and Computer-Assisted Translation* (Gothenburg, April 26, 2014).
- Underwood, N., Mesa-Lao, B., García Martínez, M., Carl, M. (2014). Evaluating the effects of interactivity in a post-editing workbench. *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC 2014*. Paris: ELRA.