# CATaLog Online: Porting a Post-editing Tool to the Web

**Santanu Pal[1], Marcos Zampieri[1,2], Sudip Kumar Naskar[3], Tapas Nayak[3], Mihaela Vela[1], Josef van Genabith[1,2]**
Saarland University, Germany[1], German Research Center for Artificial Intelligence (DFKI), Germany[2]
Jadavpur University, India[3]
{santanu.pal, marcos.zampieri, josef.vangenabith}@uni-saarland.de, m.vela@mx.uni-saarland.de,
tnk02.05@gmail.com, sudip.naskar@cse.jdvu.ac.in

**Abstract**

This paper presents *CATaLog online*, a new web-based MT and TM post-editing tool. *CATaLog online* is a freeware software that can be used through a web browser and it requires only a simple registration. The tool features a number of editing and log functions similar to the desktop version of *CATaLog* enhanced with several new features that we describe in detail in this paper. *CATaLog online* is designed to allow users to post-edit both translation memory segments as well as machine translation output. The tool provides a complete set of log information currently not available in most commercial CAT tools. Log information can be used both for project management purposes as well as for the study of the translation process and translator's productivity.
**Keywords:** post-editing, machine translation, translation memories

## 1. Introduction

With the improvement of machine translation (MT) software, post-editing tools have become an important part of the translation workflow. In recent years, commercial computer-aided translation (CAT) tools have started to provide not only the popular translation memory (TM) matches but also MT segments to be post-edited by translators. The use of MT output for post-editing is regarded to increase translator's productivity and also to improve consistency in translation (Federico et al., 2012; Zampieri and Vela, 2014). In light of this, a recent trend in the field is to develop tools that integrate both MT and TM output providing translators a larger number of more useful and more accurate suggestions (Cettolo et al., 2013).

Contributing in this direction, this paper presents a new web-based CAT tool called *CATaLog online*[1] developed based on *CATaLog*, a recently-released desktop CAT tool (Nayek et al., 2015). The tool can be used to post-edit MT output as well as TM segments. *CATaLog online* records a wide range of logs that are not available in any commercial CAT tool making it a useful tool for project management and translation process research. We have observed a substantial increase in the number of online CAT tools available, both for commercial and non-commercial purposes. This includes tools such as WordFast Anywhere[2], MateCat[3] (Federico et al., 2014), Wordbee[4], and many others. In our opinion, this is a trend in the translation industry and it motivated us to release *CATaLog online*. Online CAT tools have a number of advantages over desktop tools, most notably: they do not require local installation; they can be used from any computer; projects can be easily shared with multiple translators; project managers can track the progress of projects on the fly.

This paper presents *CATaLog online* and summarizes the new features implemented in this tool as well as their usefulness for translators, project managers, MT developers, and researchers in translation studies who can use the log functions implemented in *CATaLog online* for translation process research.

The paper is organized as follows: Section 2 presents related studies focusing on the integration between TM matches and MT output to improve CAT tools; Section 3 describes in detail the main functions of *CATaLog online*; Section 4 presents the language pairs and data that are currently included in *CATaLog online*; Section 5 discusses the main functions of *CATaLog online* and their importance for translators, researchers and project managements; finally, Section 6 concludes this paper and presents avenues for future research.

## 2. Related Work

CAT tools are regarded to increase translator's productivity and improve translation quality (Lagoudaki, 2008). The core component of most commercial CAT tools are translation memories. TMs work under the assumption that previously translated segments are likely to be good examples for new translations. This is particularly true when translating documents from the same domain which share a similar structure and/or vocabulary. Two important aspects should be considered when working with TMs: 1) the quality and number of translated segments contained in the TM; 2) the quality of the TM matching and retrieval engine. To improve the latter, developers have been working on incorporating semantic knowledge to TMs by providing paraphrasing (Utiyama et al., 2011; Gupta and Orăsan, 2014; Gupta et al., 2015), as well as incorporating syntactic information (Clark, 2002; Gotti et al., 2005; Vanallemeersch and Vandeghinste, 2014).

To increase the number of suggestions presented to translators, a recent trend in state-of-the-art CAT tools is the aforementioned integration of TM segments and MT output (He et al., 2010; Kanavos and Kartsaklis, 2010). With the improvement of state-of-the-art MT systems, MT output is no

[1]The tool is available online. For more information, consult the following URL: http://ttg.uni-saarland.de/software/catalog
[2]https://www.freetm.com/
[3]http://www.matecat.com
[4]http://www.wordbee.com/

longer considered to be suitable just for *gisting* purposes and it has been used in real-world translation projects as well. CAT tools such as MateCat present MT output along segments retrieved from TMs in the list of suitable suggestions (Cettolo et al., 2013; Federico et al., 2014).
Substantial work has been carried out on improving translation recommendation systems which recommends post-editors either to use TM output or MT output (He et al., 2010). To optimize performance these systems use classifier trained to predict which output (TM or MT) requires less effort to be used for post-editing. Work on integrating MT with TM has also been done to make TM output more suitable for post-editing aiming to diminishing translators' effort (Kanavos and Kartsaklis, 2010).
Simard and Isabelle (2009) present the integration of Phrase-based Statistical MT (PB-SMT) with translation memories in a computer-aided translation environment in which the PB-SMT system exploits the most similar matches by making use of TM-based feature functions. Koehn and Senellart (2010) present another MT-TM integration strategy. In this study an Statistical MT (SMT) system is used to fill in the gaps in retrieved TM segments.

## 3. The Tool

*CATaLog online* is a language independent tool that enables users to upload their own translation memories on the platform of the tool. It provides three major functionalities:

- It provides a novel and user-friendly online CAT environment to post-editors and translators to reduce post-editing time and effort, as displayed in Figure 1.
- It collects post-editing logs which are a fundamental source of information for the translation process research. *CATaLog online* remotely monitors and records user activities generating a wide range of logs. It also provides on-demand MT output that automatically learns from post-editor feedback.
- It provides a straightforward way to compare various translation engines taking human evaluation into account.

A more detailed description of these functionalities is given in the following sections.

### 3.1. A Novel CAT Environment

In *CATaLog online*, users can choose between MT output and TM segments. The tool allows the user to choose either the background MT system (Pal et al., 2015a) integrated in the CAT tool or to upload the translations produced by third-party MT systems. A new feature in both *CATaLog* and *CATaLog online* is the ranking of matched TM segments based on their similarity given by Translation Error Rate (TER) (Snover et al., 2006). The system finds the matched and unmatched parts between the input segment and the five most similar TM segments from the TER alignment. It also finds out the correspondences between the source and target tokens in the matched TM segments and their corresponding translations using GIZA++ (Och and Ney, 2003) word alignments with grow-diag-final-and heuristics (Koehn, 2010). Matched parts and unmatched parts, both in the source and the target text, are colour-coded for better visualisation and displayed in green and red respectively.
*CATaLog online* provides facilities to translate either single sentences or in batch mode i.e., by uploading a file. As shown in Figure 1, for a given input sentence (English in this case), the current version of *CATaLog online* provides two alternative translation suggestions in the target language (German in this case): MT and TM. The TM suggestion is colour-coded. When the translator selects the colour-coded TM alternative (c.f., Figure 3), the given input sentence is also colour-coded to reflect the matching and unmatching parts in the input sentence. Additionally, the system also shows the corresponding matched fragments of the TM source sentence. Input sentence colouring deals with green for the matched parts and yellow for the unmatched parts with respect to the TM match. *CATaLog online* shows only the top-ranked TM suggestion with respect to the input text.
Comparing every input sentence against all the TM source segments in very large TMs makes tools very slow. To improve search efficiency, *CATaLog online* uses the Nutch[5] information retrieval (IR) system. Nutch follows the standard IR model of Lucene[6] with document parsing, document Indexing, TF-IDF calculation, query parsing and finally searching/document retrieval and document ranking. In this case each document contains (i) a TM source segment, (ii) its corresponding translation and (iii) the word alignments.
To generate the search query corresponding to an input segment, all the stop words are removed first from the input segment. After presenting an input segment as query, Nutch retrieves the most likely set of candidates complying with *a*, *b* and *c*. The set of relevant candidates are ranked by Nutch according to their similarity scores for each query and the retrieved documents are collected and stored in a file. The ranking process is also deals with dissimilarity measurement that provides a final fine-grained score to re-rank the retrieved matching segments.

#### 3.1.1. Dissimilarity Measurement

**Algorithm 1** Dissimilarity Measure

**procedure** DISSIMILARITY($(s_1, s_2)$
  $score \leftarrow 0$
  **for** all $n$-grams $n$ contained in $s_1$ or $s_2$ **do**
    $f_1 \leftarrow \begin{cases} frequency(n), & \text{if } n \in S_1 \\ 0, & \text{if } n \notin S_1 \end{cases}$
    $f_2 \leftarrow \begin{cases} frequency(n), & \text{if } n \in S_2 \\ 0, & \text{if } n \notin S_2 \end{cases}$
    $score \leftarrow score + \{\frac{2(f_1 - f_2)}{(f_1 + f_2)}\}^2$
  **end for**
  **return** $score$
**end procedure**

[5] http://nutch.apache.org/
[6] http://lucene.apache.org/

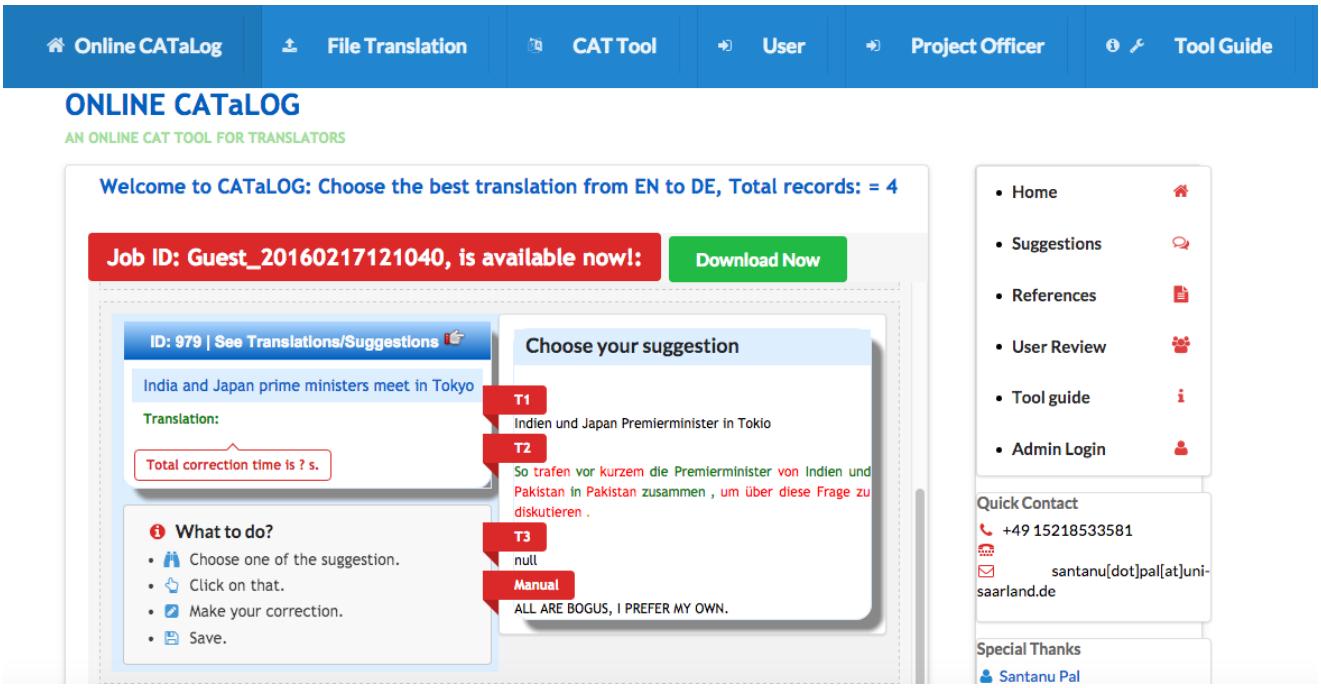


Figure 1: Working interface for *CATaLog online*

Algorithm 1 is based on Kešelj et al. (2003) and it provides dissimilarity measurement between the input segments and their corresponding retrieved candidate segments. For identical segments that have most identical n-grams, the dissimilarity score is 0. Consider two segments: $s_1$ and $s_2$. Here, $s_1$ contains the unigrams of an input segment whereas $s_2$ contains the unigrams of a retrieved candidate segments. The algorithm returns a positive dissimilarity score after being presented with $s_1$ and $s_2$.

#### 3.1.2. Re-ranking

The dissimilarity score returned by Algorithm 1 is subtracted from the similarity score assigned by Nutch for every candidate segment and a final fine-grained score is calculated. All retrieved candidates are re-ranked in accordance with their fine-grained scores. Only the top 10 ranked candidates are taken into consideration. These 10 ranked retrieved candidates are then re-ranked using TER score and only the 5 most similar segments are chosen. TER assigns equal costs to every edit operations (i.e., insertion, deletion, substitution and shifting). However, deletion is a much easier task than the other editing operations. Therefore, for the post-editing task, we set deletion cost much lower than insertion, substitution or shifting costs.

### 3.2. Recording Editing Logs

*CATaLog online* provides a web-based translation editing interface which is activated when users choose one of the possible translation suggestions which may come either from a TM or from the MT system. For a given input sentence, the user edits the best translation suggestion which may contain errors such as missing words, incorrect word order, wrong lexical choice, presence of irrelevant words, untranslated words or punctuation errors. The system records most of the user activities such as key strokes, cursor positions, text selection and mouse clicks. These logs are valuable source of information for translation process research and can later be used to derive training material for statistical automatic post-editing (Pal et al., 2015b). *CATaLog online* records the following logs.

- **Deletion log:** *CATaLog online* uses TER as well as Keystroke log information to record deletion logs. The log includes total number of words deleted, the deleted words and the corresponding token positions where they were originally belonging to in the translation suggestion.
- **Insertion log:** Like deletion logs, this log records how many words are inserted in the post-edited translation, the inserted words and the corresponding token positions in the post-edited (PE) translation.
- **Substitution log:** Each substitution log is associated with one deletion and one insertion operation. *CATaLog online* records both the token position and the corresponding deleted as well as inserted words This log information is very useful for identifying the lexical errors made by the MT system. This log also provides information regarding the morphological errors in the MT output. Thus, the substitution log can serve as a valuable resource for training a statistical automatic post-editing system.
- **Shifting or word re-ordering log:** Shifting of a word (or a sequence of words) essentially means that the system has made the right lexical choice (i.e., no lexical error), however, the placing of the word in the produced translation is not correct. *CATaLog online* takes logs of shifting of words or phrases from one position to another. The keystroke log records this kind of error by logging mouse selections, cursor positions, cut-

paste log, shift-arrow selection, mouse clicks, etc. In addition, *CATaLog online* records their original token position(s) in the selected translation suggestion and the new token position(s) in the PE translation.

The system calculates the translator (or post-editor) effort as measured by (I) keystrokes, (II) exact millisecond-level timing and (III) editing costs. In addition, an automatic post-editing (APE) system can use all these log information to improve word alignment between suggested MT segments and the post-edited segments.

### 3.3. Polling System

The final functionality discussed in this paper is the polling system. The advantage of a polling system lies in the evaluation of the different translations (here TM suggestion and MT) by the users. The polling scheme has three different options for each source segment as presented in Figure 2.

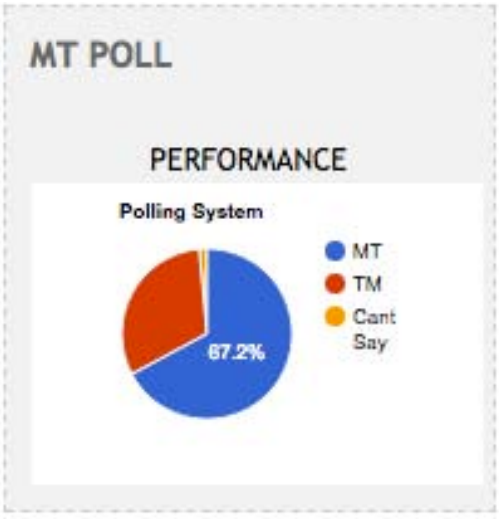


Figure 2: Polling system in *CATaLog online*

Translators act as voters and choose between the three options which refer to the quality of each output. Out of the three options, the first one is final edited translation by post editor, second and third options are for translations provided by TM and MT respectively and the final option is uncertain (U) which is applicable whenever the translators are uncertain about which translation is better i.e. both the MT and TM translations are equally good or worse to them. To avoid bias, we randomly swap the position of selective polling option between TM and MT translations so that the translators do not know to which system their votes are contributing to. They are simply asked to choose the best of the two translations.

## 4. Data

We collected parallel training data from the WMT 2015 translation shared task[7] for English–German translation. The training data includes Europarl, News Commentary and Common Crawl. The collected corpus is noisy and contains some non-German as well as non-English words and sentences. Therefore, we applied automatic language identification (Shuyo, 2010) on the English–German parallel data as well as monolingual German corpus. We discarded those parallel sentences from the bilingual training data which were detected as belonging to some different language by the language identifier. The same method was also applied on the monolingual data.

Successively, the corpus cleaning process was carried out first by calculating the global mean ratio of the number of characters in a source sentence to that in the corresponding target sentence and then filtering out sentence pairs that exceed or fall below 20% of the global ratio (Tan and Pal, 2014). We sorted the entire parallel training corpus based on their sentence length. Tokenization and punctuation normalisation were performed using Moses[8] (Koehn et al., 2007) scripts. In the final step of cleaning, we filtered the parallel training data on maximum allowable sentence length of 100 and sentence length ratio of 1:2 (either direction). Approximately 36% sentences were removed from the total training data during the cleaning process. The cleaned corpus has been used to build our SMT model. The TM database also consists of English sentences taken from the aforementioned cleaned corpus and their corresponding German translations.

## 5. Discussion

Colour coding of the input sentence, the TM source sentences and the corresponding translation suggestion(s) provides indications about which portions of matching TM source sentences as well as the translation suggestions match with the input sentence and which ones do not (c.f., Figure 3). The colour coding of the TM translation suggestions serve two purposes. Firstly, it makes the decision process easier for the translators as to which TM match to choose. Secondly, it guides the translators as to which fragments to edit. The reason behind colour coding both the TM source and the target translation suggestions is that a longer (matching or non-matching) source fragment might correspond to a shorter target fragment, or vice versa, due to language divergence. A translation suggestion which has more green fragments than red fragments is a good candidate for post-editing. Sometimes smaller sentences may get near 100% match (i.e., fully green), but they are not good candidates for post-editing, since post-editors might have to insert a lot of target words to turn the translation suggestion to an acceptable translation. In this context, while the system ranks the TM suggestions, it takes into consideration the fact that insertion and substitution are the most costly operations in post-editing, and thus, such sentences are given less priority by the TM. We assigned a higher cost for insertion than deletion so that such sentences get a higher editing cost and hence are automatically removed from the top candidates list of the TM.

*CATaLog online* incorporates a series of features that are similar to MateCat, but *CATaLog online* provides some additional features. As in MateCat, *CATaLog online* provides the translator with both MT and TM suggestions. The results in *CATaLog online* are re-ranked by TER as well as

[7] http://www.statmt.org/wmt15/translation-task.html

[8] http://www.statmt.org/moses/


Figure 3: Colour coded TM suggestion *CATaLog online*

Lucene retrieval score. Like MateCat, *CATaLog online* records editing logs such as keystrokes, exact millisecond-level timing and editing costs. An additional feature of *CATaLog online* is the colour-coding of the TM suggestions, both source and target, rendering the matching segments between the two sentences. An another additional feature of *CATaLog online* is the polling system which allows the translator to decide whether the MT or TM translation was the most appropriate alternative.

## 6. Conclusions and Future Work

This paper presented *CATaLog online*, a free online CAT tool. We discussed three main components of this tool and how they can be used in the translation workflow. We have successfully reduced the TM retrieval time in *CATaLog online* compared its desktop version. To the best of our knowledge, *CATaLog online* provides a wider range of logs than any commercial CAT tool in the market. This information is very important for translation process research and translation project management. The tool also provides a polling system developed as a resource for MT and TM evaluation. We plan to use the polling system and the information obtained in the log functions to investigate translation quality not only at the segment level but also at the document level (Scarton et al., 2015).

In future work we would like to enhance *CATaLog online* in terms of both user perspective and translation process. The user perspective includes: on-the-fly guidance during translation, analytical summaries of post-editing activities, and well structured XML formatted logs. The XML formatted logs can be customized according to the user's choice, for example the user can download entire logs or some specific logs for a particular translation job.

In terms of translation process research and development perspectives, we will implement functionalities in *CATaLog online* recording word alignments between Source–MT, MT–PE and source–PE, which will be beneficial for incremental MT and incremental APE. Using the post-editing information we would like to build and integrate a fully functional APE system into *CATaLog online* which can improve the background MT system output.

We will also deal with system level enhancement such as fuzzy based search facility within the Lucene search module. In addition, we would also like to explore contextual, syntactic and semantic features which can be included in similarity calculation to retrieve more appropriate TM match. Furthermore, Finally, we are implementing autocomplete suggestion in *CATaLog online* to accelerate the translation/post-editing time.

Finally, we would like to carry out a study to quantify the extent to which translators are faster or more productive using *CATaLog online* as compared to other CAT and post-editing tools.

## Acknowledgments

We would like to thank the anonymous reviewers for their feedback. Santanu Pal is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no 317471.

## Bibliographical References

Cettolo, M., Servan, C., Bertoldi, N., Federico, M., Barrault, L., and Schwenk, H. (2013). Issues in incremental adaptation of statistical MT from human post-edits. In *Proceedings of the Workshop on Post-editing Technology and Practice (WPTP-2)*, Nice, France.

Clark, J. (2002). System, method, and product for dynamically aligning translations in a translation-memory system, February 5. US Patent 6,345,244.

Federico, M., Cattelan, A., and Trombetti, M. (2012). Measuring User Productivity in Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of AMTA*.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., et al. (2014). The matecat tool. In *Proceedings of COLING*, pages 129–132.

Gotti, F., Langlais, P., Macklovitch, E., Didier Bourigault, B. R., and Coulombe, C. (2005). 3GTM: A Third-Generation Translation Memory. In *Proceedings of the CLiNE Workshop*.

Gupta, R. and Orăsan, C. (2014). Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In *Proceedings of EAMT*.

Gupta, R., Orăsan, C., Zampieri, M., Vela, M., and van Genabith, J. (2015). Can Translation Memories afford not to use paraphrasing? In *Proceedings of EAMT*.

He, Y., Ma, Y., van Genabith, J., and Way, A. (2010). Bridging SMT and TM with translation recommendation. In *Proceedings of ACL*, pages 622–630.

Kanavos, P. and Kartsaklis, D. (2010). Integrating Machine Translation with Translation Memory: A Practical Approach. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry*.

Kešelj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Proceedings of the conference pacific association for computational linguistics, PACLING*, 3:255–264.

Koehn, P. and Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Lagoudaki, E. (2008). The value of machine translation for the professional translator. In *Proceedings of AMTA*, pages 262–269, Waikiki, Hawaii.

Nayek, T., Naskar, S. K., Pal, S., Zampieri, M., Vela, M., and van Genabith, J. (2015). Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the NLP4TM Workshop*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Pal, S., Naskar, S., and van Genabith, J. (2015a). UdS-Sant: English–German Hybrid Machine Translation System. In *Proceedings of the WMT Workshop*, pages 152–157, September.

Pal, S., Vela, M., Naskar, S. K., and van Genabith, J. (2015b). USAAR-SAPE: An English–Spanish Statistical Automatic Post-Editing System. In *Proceedings of the WMT Workshop*, pages 216–221.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). Searching for context: a study on document-level labels for translation quality estimation. *Proceedings of EAMT*.

Shuyo, N. (2010). Language Detection Library for Java.

Simard, M. and Isabelle, P. (2009). Phrase-based machine translation in a computer-assisted translation environment. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, pages 120–127.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Tan, L. and Pal, S. (2014). Manawi: Using Multi-word Expressions and Named Entities to Improve Machine Translation. In *Proceedings of Ninth Workshop on Statistical Machine Translation*.

Utiyama, M., Neubig, G., Onishi, T., and Sumita, E. (2011). Searching Translation Memories for Paraphrases. pages 325–331.

Vanallemeersch, T. and Vandeghinste, V. (2014). Improving fuzzy matching through syntactic knowledge. In *Proceedings of Translating and the Computer*.

Zampieri, M. and Vela, M. (2014). Quantifying the Influence of MT Output in the Translators Performance: A Case Study in Technical Translation. In *Proceedings of the HaCaT Workshop*.