Project funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

Project reference: 317471

Project full title: EXPloiting Empirical appRoaches to Translation

D6.4: Improved Hybrid Translation Tool

Authors: Carla Parra Escartin (Hermes), Lianet Sepúlveda Torres (Pangeanic)

Contributors: Constantin Orasan (UoW), Anna Zaretskaya (UMA), Santanu Pal (USAAR), Hernani Costa (UMA), Rohit Gupta (UoW), Liling Tan (USAAR), Varvara Logacheva (USFD), Carolina Scarton (USFD), Liangyou Li (DCU), Chris Hokamp (DCU), Joachim Daiber (UvA), Hoang Cuong (UvA), Hanna Bechara (UoW)

Document Number: EXPERT_D6.4_20160920

Distribution Level: Public

Contractual Date of Delivery: 30.06.16

Actual Date of Delivery: 20.09.16

Contributing to the Deliverable: WP6

WP Task Responsible: Pangeanic

EC Project Officer: Ioanna Peppa

D6.4 Improved Hybrid Translation Tool

Carla Parra Escartín[1], Lianet Sepúlveda Torres[2]

[1]Hermes Traducciones, Spain; [2]Pangeanic, Spain

carla.parra@hermestrans.com, lisepul@gmail.com

September 20, 2016

Contents

1 Introduction 2

2 Translation Memories 2

2.1 Translation Memory Leveraging 3

2.2 Translation Memory Cleaning 7

3 Computer Assisted Translation tools 8

4 Terminology Management Tools 9

5 Conclusion 10

1 Introduction

The translation industry is facing nowadays more challenges than ever. On the one hand, translators are required to deliver high-quality professional translations, but on the other hand, they are also being imposed lower rates and high time pressure as clients expect to get the translations they demand as fast as possible and for the lowest possible rate. This new trend has provoked the need to look for new ways of speeding the translation process up and reducing the costs. The most obvious way of doing so is the introduction of Machine Translation Post-Editing (MTPE) tasks in the translation work flow. However, translators are still reluctant to accept MTPE tasks and the discounts applied to them are still a big issue of debate.

The EXPERT project aimed at building a bridge between Academia and Industry and thus new ways of improving the translation work flow have been investigated in the project. One possible way of doing so is by researching hybrid translation tools that integrate recent advances in Computational Linguistics and Machine Translation research. More concretely, the advances in this respect have been in three main areas: Translation Memories, Computer Assisted Translation Tools and Terminology Management Tools. Although some overlapping with previous deliverables is unavoidable[1], here we aim at offering an overview of all the advances made across the different Work Packages in the project which are related to one of these three main areas[2].

The remainder of this Deliverable is structured as follows: Section 2 summarizes the work related to incorporating Natural Language Processing Techniques into TM leveraging (Subsection 2.1) and TM cleaning (Subsection 2.2). Section 3 reports on the new Computer Assisted Translation (CAT) tools developed within the project, and Section 4 summarizes the work with regard to Terminology Managament tools. Finally, Section 5 summarizes and discusses the main innovations achieved within the EXPERT project.

2 Translation Memories

Translation Memory (TM) systems have become very important tools for professional translators and constitute the key component of most Computer Assisted Translation (CAT) tools. Translation Memories are the backbone

[1] See http://expert-itn.eu/?q=content/deliverables.

[2] For an overview of the work done with regard to corpus-based approaches (including hybrid MT approaches), please see Deliverable *D6.3. Improved corpus-based Approaches* [Parra Escartín and Sepúlveda Torres, 2016a]. All work involving user evaluations has been summarized in Deliverable *D6.5. Final User Evaluation* [Parra Escartín and Sepúlveda Torres, 2016b].

of TM systems. They are the resource where past translations are stored and thus contain high-quality, domain-specific bilingual data.

TM systems exploit this data in numerous ways during the translation process. One such use is the retrieval of past translations of source sentences that are either identical or very similar to the new sentences to be translated. This process is called TM leveraging. In order to leverage past translations, TM systems have an implemented algorithm that measures the similarity between the new sentence to be translated and that stored in the TM. For each new sentence, the TM system computes this similarity and assigns all identical or similar translations a score called the Fuzzy Match Score (FMS).

Translators are used to post-edit the so-called fuzzy-matches and the TM leverage is also used to compute rates and allocate resources at the planning stage of a translation project. As the FMS gets lower, sentences are more difficult to post-edit and at some point they are not worth post-editing. That is why in the translation industry a threshold of 75% FMS is used. Segments getting a 75% FMS or higher undergo TM Post-Editing (TMPE), and segments below that threshold are translated from scratch.

Within the EXPERT project, several ways of improving TM systems have been researched. More concretely there have been efforts in two main directions. While some researchers have focused on improving the way in which TM systems carry out the TM leveraging process, others have focused on the task of curating TMs to ensure their high quality and automatically cleaning them. These two research directions are reported in the following Subsections 2.1 and 2.2, respectively.

2.1 Translation Memory Leveraging

Translation Memory leveraging is key for professional translators, as it determines the amount of segments that can be re-used in a new translation task. Given a segment to be translated, CAT tools look for such segment in the available TMs. As previously explained, TMs will not only retrieve the exact matches found, but also fuzzy-matches (i.e. similar segments to the one that needs to be newly translated). Fuzzy matches are retrieved using some flavor of an Edit Distance Metric such as Levenshtein Distance.

As part of the EXPERT activities, ESR4 explored ways of improving the retrieval of fuzzy matches by relying on a system that calculates the semantic similarity between sentences. In Gupta et al. [2014b], he and his co-authors explored ways of incorporating "features based on surface form, parts of speech information, lemma, typed dependency parsing, named entities, paraphrasing, machine translation evaluation, and corpus pattern analysis [Hanks, 2013]". They used the Stanford CoreNLP3 toolkit [Manning et al., 2014] to retrieve the lemmas, parts of speech, named entities and dependencies relations of words. Additionally, they identified paraphrases in the segments using the PPDB paraphrase database [Ganitkevitch et al., 2013]. The system they used is described in more detail in Gupta et al. [2014a]. They ran experiments on two different test sets and their results revealed that work in this direction could yield positive results.

Gupta and Orăsan [2014] explore the integration of paraphrases in matching and retrieval from TMs using Edit Distance in an approach based on greedy approximation and dynamic programming. Their strategy consists of paraphrasing the existing TM to allow for offline processing of data and alleviate translators from the need to install additional software. Their system works following the following 5-step pipeline:

1. Read the TMs.
2. Collect all paraphrases from the paraphrase database and classify them in classes:
   (a) Paraphrases involving one word on both the source and target side.
   (b) Paraphrases involving multiple words on both sides but differing in one word only.
   (c) Paraphrases involving multiple words but the same number of words on both sides.
   (d) Paraphrases with differing number of words on the source and target sides.
3. Store all the paraphrases for each segment in the TM.
4. Read the file to be translated.
5. Get all paraphrases for all segments in the file to be translated, classify them and retrieve the most similar segment above a predefined threshold.

They report a significant improvement in both retrieval and translation of the retrieved segments. This research was further expanded with a human centered evaluation in which they assessed the quality of semantically informed TM fuzzy matches based on post-editing time or keystrokes Gupta et al. [2015][3]. The tool has been publicly released under an Apache License 2.0 and is available via a github repository[4].

ER3a and ER3b have also worked on improving the TM leveraging. More concretely, they have worked on an iterative way that involves two main steps:

1. Improvement of the Fuzzy Match computation to create an enhanced TM; and
2. Integration of the enhanced TM in an SMT system.

[3] This user evaluation is further described in the EXPERT Deliverable *D6.5: Final User Evaluation* [Parra Escartín and Sepúlveda Torres, 2016b]

[4] https://github.com/rohitguptacs/TMAdvanced

In the first step of their system, ER3a and ER3b aim at maximizing the reuse of already existing human translations by improving the fuzzy matching. This process includes several string transformation and segmentation rules with the aim to obtain new pairs of segments which then will be stored in the TM for future reuse. In order to accomplish the creation of new TM segments, they analyze and store multiple TM files in TMX (Translation Memory eXchange) format, that contain segments pairs of different domains and languages.

To overcome the limitations of current TM systems in terms of storage and concordance searches, ER3a and ER3b have designed a fast and scalable TM system called *ElasticTM*, where all the new segment pairs and their metadata can be successfully stored[5]. When designing the system, the following list of requirements to be fulfilled was used:

- *Great storage capacity*: The system shall have the capacity of storing large numbers of segments (over 10M), along with their corresponding metadata (source and target language, segment creation and modification date, Part-of-Speech tags for all tokens in a given segment and for both languages, domains, etc.).
- *Fast retrieval*: The system shall be very fast at retrieving the fuzzy matches, notwithstanding their FMS score.
- *Reasonable import time of new segments*: From time to time, it is foreseen that the existing TMs will be updated importing massive numbers of new segments in TMX or similar formats. The TM system shall be able to do such tasks within a reasonable time.
- *Effective segment filtering, retrieval and export*: The system shall be able to retrieve sets of segments fulfilling certain criteria used as filters (e.g. domain, date, time span, file name, terms appearing in the source or target language, etc.). Such subsets will also be exported as one or several TMs in TMX format that will subsequently be used as input to train models for the PangeaMT platform[6].

With the aim of maximizing the number of 100% matches retrieved from the existing TMs, the Fuzzy Match retrieval that has been implemented exploits already existing Natural Language Processing technologies. As a result of this, a tokenizer and a Part-of-Speech (PoS) tagger have been integrated in the system. New segmentation rules making use of regular expressions have also been put in place. The performance of this innovative approach for Fuzzy Match retrieval is currently being tested for several language pairs.

[5] This system is currently being deployed and will be released upon the end of the EXPERT project.

[6] http://pangeamt.com/en

The implemented system is divided into the following three main components:

- **TM preprocessing module**
  This module is in charge of preventing inconsistencies across the source and target sentences in a TM segment. It first tokenizes the sentences and subsequently runs the PoS tagger. Different PoS taggers have been used depending on the language to be analyzed. To allow for comparisons across languages, the Universal PoS tagset [Petrov et al., 2012] has been used.
- **TM storage and query interface**
  The TM database consists of two modules: a *Search Engine* and a *MapDB*[7]. The purpose of the *Search Engine* is to hold monolingual indices of segments and provide the user with a flexible search interface. The system is based on *Elasticsearch*[8]. The TF-IDF coefficient is used to rank the search results by their relevance. The *MapDB* is used to complement the *Search Engine* by storing bidirectional mappings between two languages using a unique id. Once it has been fully deployed, the system will be able to create on-the-fly new TMs for different language pairs by pivoting on the already existing ones.
- **Fuzzy match improvement**
  As it mentioned earlier, the main innovation of *ElasticTM* is that it incorporates linguistic knowledge to improve the TM leveraging. The linguistic knowledge was added with two aims: to enrich the TM database with artificial segments, and to improve the TM retrieval algorithm.
  To improve the fuzzy matching algorithm several language dependent and independent features are used: *Regular Expressions*, *PoS tag matches* and *segmentation rules*. *Regular Expressions* are used to improve the recognition of placeable and localizable elements (e.g. numbers, urls, etc.). *PoS tag matches* are used to detect similarities between source and target segments. Currently, only segments having a simple grammatical structure benefit from PoS tag matches. *Segmentation rules* are used to improve the fuzzy match retrieval and recall. The new rules include features such as delimiters and the length of the segments. In the case of some language pairs, more complex sentence patterns have been included (e.g. coordinating conjunctions, relative pronouns, verb phrases, etc.).
  These three features are also used to create new segments. The newly created artificial segments are stored in the database using a separate index. The storage of artificial segments in a different database is also

[7] http://www.mapdb.org/

[8] https://www.elastic.co/products/elasticsearch

used to assess the performance of the different methods used to create the new segments.

2.2 Translation Memory Cleaning

During the EXPERT project, ER1 focused on two processes: data compilation to create bilingual corpora for training MT systems[9], and Translation Memory cleaning to help curate existing TMs in a more efficient and automatic way. He has worked on a tool which has been released to the community in September 2016[10]. The tool, called for now on *TM cleaner*, is based on the code presented in Barbu [2015]. ER1 has re-written most of the code to make it easily usable by the translation industry and has additionally added new features:

1. Integration of the HunAlign aligner [Varga et al., 2005]: This component is meant to replace the automatic translation component as not every company can translate huge amounts of data. The score given by the aligner is smoothly integrated in the training model.

2. Addition of two operating modes: the *train modality* and the *classify modality*: In the *train modality*, the features are computed and the corresponding model is stored. In the *classify modality* a new TM is classified based on the stored model.

3. Passing arguments through the command line: It is now possible, for example, to indicate the machine-learning algorithm that will be used for classification.

4. Implementation of hand written rules for keeping/deleting certain bilingual segments: These hand written rules are able to decide in certain cases with almost 100% precision if a bilingual segment should be kept or not. This component can be activated/deactivated through an argument passed through command line.

5. Integration of an evaluation module: When a new test set is classified and a portion of it is manually annotated, the evaluation module computes the precision/recall and F-measure for each class.

The tool has been evaluated using three new data sets coming from aligned websites and TMs. Moreover, the final version of the tool has been implemented on an iterative process based on annotating data and evaluating it using the evaluation module. This iterative process has been followed to boost the performance of the cleaner.

[9] See Deliverable *D6.3. Improved corpus-based Approaches* [Parra Escartín and Sepúlveda Torres, 2016a] for a report on this work.

[10] The tool is freely available on GitHub and includes a series of tutorials in its "Documentation" folder (`https://github.com/SoimulPatriei/TMCleaner`).

3 Computer Assisted Translation tools

Computer Assisted Translation (CAT) tools are nowadays a key element in the translation workflow. CAT tools typically integrate several features used simultaneously by professional translators. Typically, all CAT tools include the following features:

- Leverage of previous translations including:
  - Segment-level fuzzy matching
  - Highlighting of differences
  - Translation Memory match information (who, when, ...?)
- Analysis for quoting, planning and keeping track of progress
- Concordance for sub-segment searches
- Maintenance to perform global changes, import/export content, etc.
- Quality Assessment: terminology control, typos, etc.

As part of his work on the project, ESR2 has developed a new Computer Assisted Translation (CAT) tool called *CATaLog*. This tool is described in Nayek et al. [2015b], and Nayek et al. [2016]. Pal et al. [2016] describes its online version, *CATaLog online*. The main innovation of *CATaLog* consists of integrating a color coded scheme both in source and target to highlight the chunks in a particular segment that should be changed. While CAT tools highlight fuzzy match differences, this is done only on the source side and it is left to the translator to locate the part of the target segment that needs to be changed. *CATaLog* alleviates this task by facilitating the translator the pointer at target level. The target chunks are identified using TER alignments.

In the online version of *CATaLog* [Pal et al., 2016], new interesting features have been included. The Translation Memory (TM) leverage is computed using the Nutch[11] information retrieval system which includes document parsing, document indexing, TF-IDF calculation, query parsing and searching/document retrieval and document ranking. Besides the TM segment, the word alignments are also used. The other big innovation is that now the tool also includes the automatic logging of user activity. It automatically records key strokes, cursor positions, text selection and mouse clicks together with the time spent post-editing each segment. These logs can be very useful for research on post-editing and can also be used as training materials for Automatic Post-Editing tasks.

Another approach to CAT tools is the one taken by ESR9, who focused on the design of tools centered on the Human-Computer Interaction. Thus, in his approach the role of the interface is precisely to provide the translator

[11] http://nutch.apache.org/

with intelligent interactions which facilitate the mapping of source language sequences into target language sequences. In his research, ESR9 assumed that CAT tools are used with the aim of helping translators to produce high quality translations and that the translators interact with one or more "intelligent agents" to achieve such goal more quickly and/or effectively than the unassisted user.

The result of his work is an implementation of a flexible, web based CAT tool specifically designed with interoperability and extensibility in mind: HandyCAT [Hokamp and Liu, 2015]. HandyCAT is an open source CAT tool[12] that allows the user to easily add or remove graphical elements and data services to/from the interface. Morever, new components can be directly plugged into the relevant part of the translation data model. These features make HandyCAT an ideal platform for developing prototypes and conducting user studies with new components[13].

4 Terminology Management Tools

Terminology Management tools can be extremely useful for translators. They are used both to extract terminology from already existing TMs or bilingual corpora, and to identify the terms appearing in a new translation project. As such, they can either be independent programs or programs that are integrated into CAT tools as one of their component. In some cases, the same Terminology Management tool can be used both as an independent program and as a component in a CAT tool[14], and in some others the tool may be running in the background and allow translators to do quick checks[15].

As part of his work on developing the CAT tool HandyCAT [Hokamp and Liu, 2015], ESR9 developed an end-to-end prototype of a dynamic linked terminology component implemented as part of the HandyCAT platform. This work is described in Hokamp [2015]. The component was created with a twofold aim: to demonstrate a potential use case for linked data within the localisation workflow; and to evaluate the effort needed to build such a system. The terminology component combines the entity linking approach proposed by Mihalcea and Csomai [2007] and the statistical models for extracting and disambiguating entities in the source and target languages proposed in Daiber et al. [2013].

The workflow proposed by Hokamp [2015] is illustrated in Figure 1. It

[12] http://handycat.github.io/

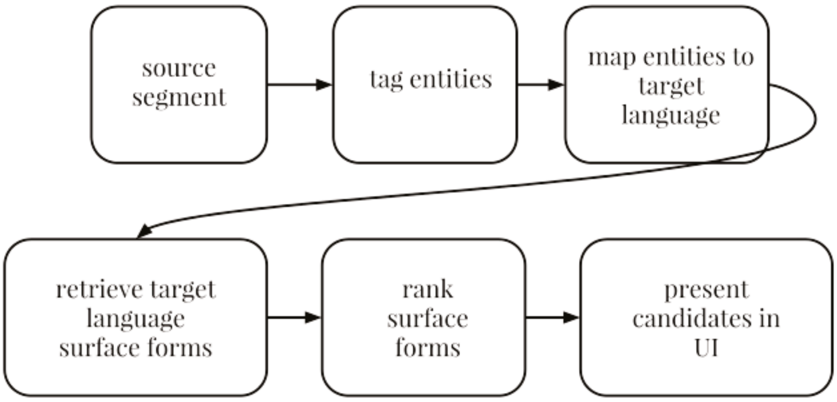
[13] See Section 4 for a description of a prototype integrating a terminology component developed as part of his work.

[14] SDL Multiterm (www.sdl.com) is an example of this.

[15] Xbench (www.xbench.net/) and VerifiKa (https://e-verifika.com/) are two examples of Quality Assurance tools with a Terminology Management component that works in this way.

starts by identifying the entities in the source segment and tagging them. After that, they are mapped to their translation candidates in the target language, for which the target language surface forms are retrieved and subsequently ranked before being presented to the translator in the interface.

Figure 1: Dynamic linked terminology workflow implemented by Hokamp [2015].


source segment
tag entities
map entities to target language
retrieve target language surface forms
rank surface forms
present candidates in UI

While describing the implementation of the tool, Hokamp [2015] also demonstrates how existing Natural Language Processing technologies can be successfully integrated into CAT tools as new components. Hokamp [2015] also concludes highlighting the fact that this new terminology component does not interfere with the translators work, as it does not force them to use it. In fact, translators are free to either ignore the additional markup and terminology options or use them, whatever suits their needs and wishes best.

5 Conclusion

In this Deliverable, we have summarized the main innovations of the EXPERT project dealing with the hybridization of translation tools. These innovations have been on three interrelated axes: Translation Memory Systems, Computer Assisted Translation Tools and Terminology Management tools.

While some researchers have focused on improving already existing algorithms with linguistic information, others have researched how to create new tools that can be used in the translation industry. Thus, the *TMAdvanced* tool developed by Gupta et al. [2014b], Gupta and Orăsan [2014] can already be used by any translator or translation company, as do the soon to be released *ElasticTM* developed by ER3a and ER3b and the *TM Cleaner* that ER1 has developed improving the algorithm he proposed in Barbu [2015].

The CAT tools CATaLog [Nayek et al., 2015b,a, Pal et al., 2016] and HandyCAT [Hokamp and Liu, 2015] and the terminology management system proposed by Hokamp [2015] are also samples of how research in Academia can result in Open Source tools that aim at fulfilling all the features in existing CAT tools and adding new functionalities with the sole purpose of helping translators translate better and focus on the task at hand: delivering high quality translations in a timely manner.

Several EXPERT researchers have explored ways of integrating new advances in Computational Linguistics and Machine Translation in the translation workflow. As shown by the work reported here, there is room for a successful hybridization of the translation workflow and such hybridization may be implemented in different components with a unique goal: allowing the end users (i.e. the translators) to work faster and more effectively benefiting from research.

References

Eduard Barbu. Spotting false translation segments in translation memories. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 9–16, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W15-5202`.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1972-0. doi: 10.1145/2506182.2506198. URL `http://doi.acm.org/10.1145/2506182.2506198`.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `http://cs.jhu.edu/~ccb/publications/ppdb.pdf`.

Rohit Gupta and Constantin Orăsan. Incorporating paraphrasing in translation memory matching and retrieval. In *Proceedings of the European Association of Machine Translation (EAMT-2014)*, pages 3–10, 2014.

Rohit Gupta, Hanna Béchara, Ismail El Maarouf, and Constantin Orasan. Uow: Nlp techniques developed at the university of wolverhampton for semantic similarity and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 785–789, Dublin, Ireland, August 2014a. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/S14-2139.

Rohit Gupta, Hanna Béchara, and Constantin Orăsan. Intelligent translation memory matching and retrieval metric exploiting linguistic technology. In *Proceedings of the Translating and Computer 36*, pages 86–89, 2014b.

Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Can translation memories afford not to use paraphrasing? In *Proceedings of the 2015 Conference on European Association of Machine Translation (EAMT-2015)*, Antalya, Turkey, 2015.

Patrick Hanks. *Lexical Analysis: Norms and Exploitations*. MIT Press, 2013.

Chris Hokamp. Leveraging NLP technologies and linked open data to create better CAT tools. *Localisation Focus - The International Journal of Localisation*, (14), 2015.

Chris Hokamp and Qun Liu. Handycat: The Flexible CAT Tool for Translation Research. In *Demo presented at EAMT 2015*, pages 15–19, Istanbul, Turkey, May 2015.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321475. URL http://doi.acm.org/10.1145/1321440.1321475.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the 1st Workshop on Natural Language Processing for Translation Memories. Workshop on Natural Language Processing for Translation Memories (NLP4TM), located at RANLP 2015, September 11, Hissar, Bulgaria*, pages 36–43, 2015a.

Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 36–42, Hissar, Bulgaria, September 2015b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-5206.

Tapas Nayek, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay, and Josef van Genabith. Beyond translation memories: Generating translation suggestions based on parsing and pos tagging. In *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, Portorož, Slovenia, May 2016.

Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. Catalog online: Porting a post-editing tool to the web. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.3. Improved corpus-based Approaches. EXPERT Project, 2016a.

Carla Parra Escartín and Lianet Sepúlveda Torres. Deliverable D6.5. Final User Evaluation. EXPERT Project, 2016b.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A Universal Part-of-Speech Tagset. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596, 2005.