

Uzbek-English and Turkish-English Morpheme Alignment Corpora

Xuansong Li, Jennifer Tracey, Stephen Grimes, Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania

Philadelphia, PA 19104 USA

Email: {xuansong,garjen,sgrimes,strassel}@ldc.upenn.edu

Abstract

Morphologically-rich languages pose problems for machine translation (MT) systems, including word-alignment errors, data sparsity and multiple affixes. Current alignment models at word-level do not distinguish words and morphemes, thus yielding low-quality alignment and subsequently affecting end translation quality. Models using morpheme-level alignment can reduce the vocabulary size of morphologically-rich languages and overcomes data sparsity. The alignment data based on smallest units reveals subtle language features and enhances translation quality. Recent research proves such morpheme-level alignment (MA) data to be valuable linguistic resources for SMT, particularly for languages with rich morphology. In support of this research trend, the Linguistic Data Consortium (LDC) created Uzbek-English and Turkish-English alignment data which are manually aligned at the morpheme level. This paper describes the creation of MA corpora, including alignment and tagging process and approaches, highlighting annotation challenges and specific features of languages with rich morphology. The light tagging annotation on the alignment layer adds extra value to the MA data, facilitating users in flexibly tailoring the data for various MT model training.

Keywords: machine translation, morpheme alignment, word alignment

1. Introduction

MT alignment has been an active research area for enhancing MT quality. Modern statistical machine translation (SMT) systems typically use “word” as the atomic unit for translation. Word-alignment is the initial step in SMT pipeline, aiming to identify word correspondence of source and target languages. As current word alignment models do not address morphology below the word level, it can be difficult to divide the MT problem into sub-problems and tackle each sub-problem in isolation to improve the overall quality of MT. The problem is particularly outstanding for morphologically rich languages as statistical correspondences between source and target words are diffused over many morphological forms (multiple surface forms for a morpheme). The fact that rare words and multiple affixes often occur in highly inflected languages exacerbates this problem. Morpheme-based alignment is useful in the translation process of highly inflected languages. Morphological inflections indicate tense, gender or number which are normally expressed as separate words in uninflected languages. Capturing such sub-word alignments can yield better word alignments. Recent SMT research has found that utilizing information from morphology improves the quality of word alignments. Eyig'oz et al. (2013) developed a MT model using a two-level Turkish-English alignment, achieving significant improvement of BLEU scores. Luong and Kan (2010) proposed a morphologically sensitive approach to word alignment for language pairs involving a highly inflected language, addressing morpheme alignment issues which are peculiar to highly inflected languages. Toutanova et al. (2008) improved the quality of SMT by applying inflection generation models that predict word forms from their stems using extensive morphological and syntactic information from both the source and target languages, Russian and Arabic. Their model improves the quality of SMT over both phrasal and syntax-based approaches.

Minkov et al. (2007) adopted a novel method for predicting inflected word forms for generating morphologically rich languages in machine translation. The use of morphological and syntactic features leads to large gains in prediction and alignment accuracy. Costa-jussà's work (2015) is a recent research effort at the level of morphology, focusing on differences between Spanish and Chinese.

As a part of the BOLT (Broad Operational Language Translation) program initiated by DARPA (the Defense Advanced Research Projects Agency), LDC created Uzbek-English and Turkish-English morpheme-level alignment corpora. The morpheme alignment task aims to identify correspondences between linguistic units at the morphological level in a set of parallel texts. The resulting morph alignment data can be used as gold standard training and testing data for developing machine translation systems. The task targets low-resource languages. This paper focuses on the creation of Uzbek-English and Turkish-English morph alignment corpora (Table 1) in the genres of newswire (NW), web and discussion forums (DF).

Language	Genre	Words	Morpheme Tokens	Sentence Segments
Uzbek	NW Web DF	9080	14290	777
Turkish	NW DF	7301	13348	515

Table 1: Uzbek and Turkish MA Corpora

The paper is organized into 6 sections. The first section describes the importance of morpheme-level alignment for MT technologies and LDC's contribution to advance this effort. Section 2 introduces the source data used for MA. Section 3 delineates various annotation approaches for the Uzbek-English and Turkish-English MA alignment data.

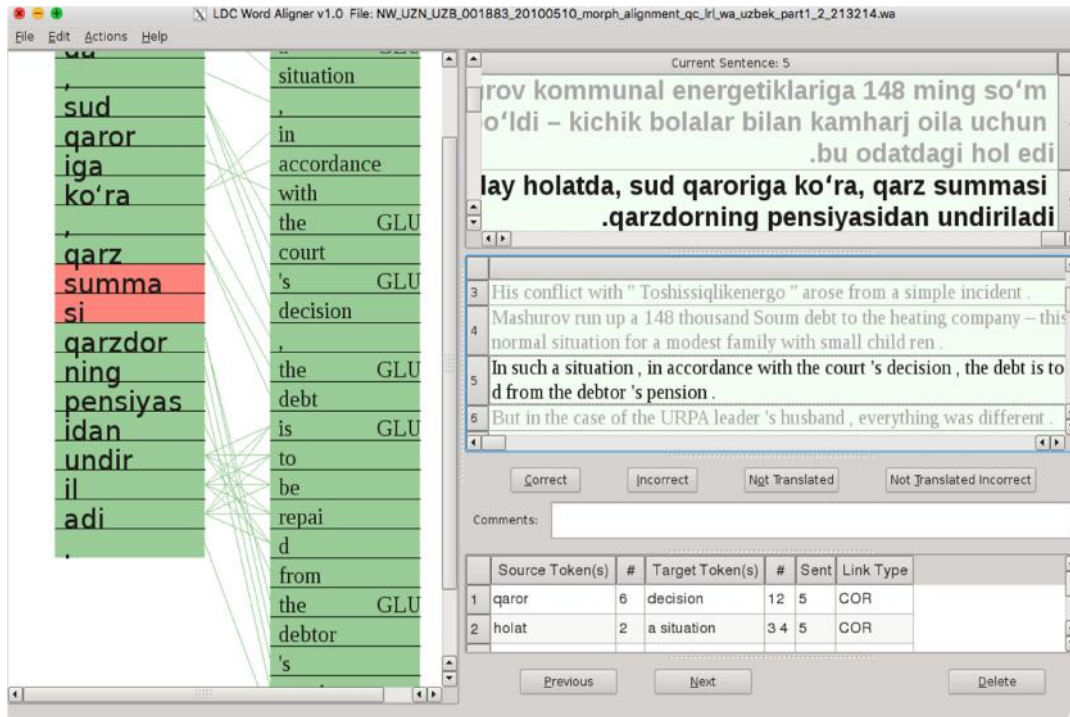


Figure 1: Morpheme Alignment Annotation Tool

Section 4 specifies data format and analyses annotation data features. Section 5 focuses on data use, with an emphasis on customizing data for different MT systems. Section 6 concludes the paper.

2. Source Data and Alignment Tool

The Uzbek and Turkish source data was harvested by LDC from web using a variety of methods. Native Uzbek and Turkish speakers first identify potential sources/websites of monolingual text in multiple genres by searching for general content sources on the web. Then the identified documents from each sites are automatically, semi-automatically or manually harvested utilizing LDC's WebCol infrastructure. Harvested text is further post-processed to desired source data format.

Source data are automatically sentence-segmented using a combination of open source tools and approaches developed by LDC. All data was converted to UTF-8 encoding. The sentence-segmented data was further selected for translation by professional translation agencies. Informal genre files (such as discussion forums) selected for translation or alignment annotation were manually reviewed by native speakers to verify that the text is in the expected language and that the content is acceptable (for instance, the content does not contain extremely offensive or vacuous content). Translation and annotation files in the newswire text genre are not subject to this manual check because news text is generally expected to be acceptable. Translators and annotators were instructed to reject any file that is not in the expected language or is otherwise problematic.

Morphological segmentation and tagging are performed on the sentence-segmented source and translation texts. For morpheme-level alignment, morpheme tokens are directly extracted from the morphologically annotated data, without any other tokenization added. In alignment, punctuations are treated as separate tokens.

A visualized tool was developed by LDC to facilitate the task (Figure 1). The tool was adapted from the word alignment tool which was originally used for word-level alignment annotation for the GALE (Global Autonomous Language Exploitation) and BOLT word alignment projects.

3. Annotation Methodologies

3.1 Annotation Process and Tasks

Morpheme alignment annotation is manually performed by LDC annotators. Annotation guidelines were developed based on guidelines used for the word alignment task of BOLT. The guidelines includes two major components. The first part addresses general annotation strategies for universal language features in alignment. The second part details specific alignment rules for Uzbek and Turkish alignment.

The annotation process is staged into two passes of annotation. The first pass is performed by junior annotators, followed by a pass of quality control by senior annotators.

Specifically, the alignment task includes:

- 1) Link morphemes in the source (Uzbek and Turkish)

language to those in the target language.

- 2) Make judgments on the linked alignment by tagging alignment types
- 3) Attach superficially-unmatched morphemes to their constituent heads according to attachment rules
- 4) Tag unaligned morphs with proper tags
- 5) Exclude noisy sentences which are improper for annotation via “Reject Segment” mechanism in the tool, such as blank sentences, unmatched sentences, half translated sentences or English sentences on both sides.
- 6) Add comments in case of any annotation problems

3.2 Annotation Approaches

3.2.1 Uzbek and Turkish as Synthetic Language

A language can be either analytic or synthetic. In analytic languages, such as English, one word usually equals one morpheme. In contrast, words of synthetic languages, such as Uzbek or Turkish, comprise several morphemes. “Word” is defined as a single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space. “Morpheme” is defined as a meaningful morphological unit of a language that cannot be further divided (e.g., in, come, -ing). It is the smallest morphological element representing functional relations in a linguistic system.

Uzbek, a Turkic language, is the official language of Uzbekistan. Uzbek is an agglutinative language lacking grammatical gender. Suffixes are added to a word in a fixed order, indicating morphosyntactic features. Uzbek is complex due to its high number of inflectional categories. For instance, nominals and verbs, as the two main morphosyntactic categories, can be suffixed with inflectional suffixes as well as suffixed with productive derivational suffixes to assume various grammatical functions. Syntactic functions (such as subject and object) and thematic relations (such as recipient, location, beneficiary) are identified by morphological case marking and verbal agreement. Uzbek is a Subject-Object-Verb order language. Subject is frequently omissible when the referent is obvious based on common cultural knowledge or the context of communication. Uzbek has no definite and indefinite articles, instead the word “bir” and the accusative case marker are used to express indefiniteness and definiteness.

Turkish belongs to the Altaic branch of the Ural-Altaic family of languages. Modern Turkish has several striking characteristic features. It is an agglutinating language. Due to this feature, it is possible to form long words by adding suffixes. On average, a speaker adds about two or three suffixes to a verbal or nominal stem. Turkish is also a harmonic language with complicated vowel harmony and consonant assimilation. When a suffix is added to a word, the word form will change according to the sound combinations of the word attached. Pro-drops and ellipsis are common in Turkish, ranging from suffixes and clitics to phrases. Subjects and/or objects can be dropped as well.

Turkish word order pattern, in contrast to English, is very flexible. Words can be arranged in various ways, such as the direct object before the verb, which is a typical sentence pattern in Turkish. Turkish is lack of some basic words/constituents which are essential to other languages, such as the copular words “am, is, are” or the determiner “the”. Instead, the meaning and grammatical relationship of such words are expressed by suffixes.

Dissimilarities in the granularity of part of speech and syntactic structures between analytic languages and synthetic languages prove to be problematic for machine translation. The disparity also poses challenge for the morpheme alignment task, which targets correspondences between morphemes. There are morphemes in the Uzbek and Turkish source that could be aligned with identical lexical forms in the target English language, while in some cases, a lot of source language morphemes cannot find their matched translation morphemes and remain unaligned. In our morpheme alignment task, we handle both aligned morphemes and unaligned ones.

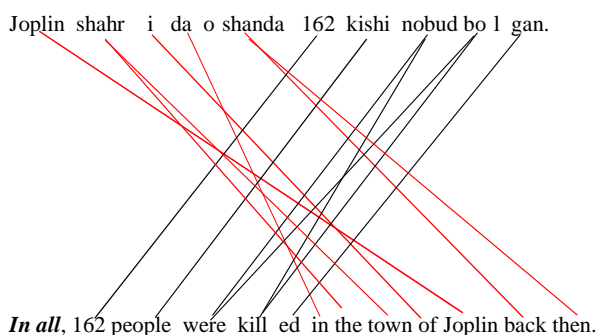
3.2.2 Alignment Link Types

We distinguish and tag two alignment types: translated-correct and translated-incorrect links. Translated-correct links are used when a morpheme is properly and semantically translated. Most of alignment links are translated and “correct” links, where the meaning is conveyed properly and they are grammatically correct, such as most morphemes in Example 1. If morphemes are translated incorrectly, either semantically or grammatically, they are aligned as translated-incorrect type. Typos or grammatical errors in target language are annotated as “incorrect” alignment type.

3.2.3 Tags for Unaligned Morphemes

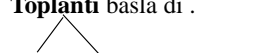
Two types of tags are designed for unaligned morphemes: not-translated-incorrect tag and not-translated-correct tag. A morpheme is tagged as not-translated-incorrect when it is both semantically and lexically missing from English translation, as “In all” in Example 1.

Example 1: Uzbek example for aligned and unaligned morphemes (lines indicating alignments; bold-italicized indicating unaligned)



A morpheme is tagged as not-translated-correct when it is not semantically missing but only missing its superficial lexical equivalent in English translation, as “The” Example 2. Another case under this not-translated-correct category includes morphemes that are neither semantically nor grammatically important, such as “ah” or punctuations, without which, the sentence is still grammatically-sound and meaningful.

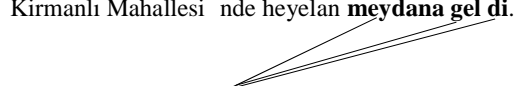
Example 2: Turkish example (“the” in English having no lexical match in Turkish)

Toplantı başladı .

The meeting has start ed.

3.2.4 Minimum-Match and Maximum-Match Approaches for Alignment

For matched morphemes between source and translation, two alignment approaches are adopted: minimum-match and maximum-match. The minimum-match approach is for literal translation where source morphemes are translated morpheme-for-morpheme, as “gan” in Uzbek is aligned to “ed” in English in Example 1. The maximum-match approach targets non-literal translation where meaning cannot be derived by decomposing it into its morphemes, such as idioms, set expressions, proper nouns or proverbs. In such cases, as many morphemes as needed should be selected for aligning in order to reach semantic equivalence, like “meydana gel di” in Turkish is aligned to “hit” in English in Example 3.

Example 3: Turkish example showing maximum-match for idioms, set expressions and non-literation translations.

Kirmanlı Mahallesi nde heyelan **meydana gel di**.

A landslide **hit** Kirmanli Quarter

3.2.5 Attachment Approach for Unaligned Morphemes Tagging

For unaligned and unmatched morphemes, we distinguish two types of unaligned morphemes and apply two different approaches accordingly: tag-and-attach and tag-not-attach. The tag-and-attach approach is employed for the unaligned morphemes which are either grammatically or contextually needed for language fluency or semantic completeness. Unaligned function morphemes are grammatically needed for fluency, while unaligned content morphemes are semantically needed for meaning completeness. Both content and function unaligned morphemes are tagged with the “glue” tag, and they are attached to head morphemes to show constituent dependency and morpheme relations. Table 2 illustrates some tag-and-attach rules with Uzbek/Turkish examples. The tag-not-attach approach finds it use with unaligned morphs which have no head morphemes to attach to and thus tagged as “not-translated correct”. Table 3 illustrates this approach.

Tag-and-Attach Approach	
Categories	Examples
Determiners (e.g.the/a/an)	(Turkish) Toplantı başladı . The meeting has start ed. (“The” is attached to “meeting” and aligned to “ Toplantı ”)
Auxiliary verbs (e.g. was, will)	(Turkish) Son olay lar ı bildir me di ler . They have not report ed the late st event s. (“have” is attached to “report”, and aligned to “bildir”)
Prepositions (e.g. in, at, of)	(Turkish) Ere li de ya mur su lar ı heyelan a neden oldu Rain water s cause d landslide in Eregli (“a” is attached to “heyelan”, and aligned to “landslide”)
Relative Clause marker (e.g. whom, which)	(Uzbek) Rasmii larga ko ra , bu shu paytgacha kuzat ilgan eng katta tahdid dir . According to official s , this is the biggest threat that has been observe d so far . (The unmatched “that” is attached to “threat”, and aligned to

Table 2: Tag-and-Attach Rules with Examples

Tag-not-attach Approach	
Categories	Examples
Copula BE (e.g. am, is, are)	(Uzbek) Bu hodisa chuda achinarli This event is very sad. (“is” in English is unaligned)
Expletives (e.g here, there, it)	(Turkish) Kim e oy verecek leri belli de il It is not certain whom they will vote. (“It” in English is unaligned)
Conjunctions (e.g and, but)	(Turkish) Bu deney daha sonra un yerine ba ka toz ya da granül madde ler le de tekrarla n di, sonuç ta hep orta ya bir elektrik enjri si çık tı . The experiment was later repeat ed with other kind s of granule s and powder s and produce d the same electrical energy. (“and” in English is unaligned)
“That” clause as object/subject	(Uzbek) Vaksina lar inson salomatlig i uchun zarar, de b hisobla ydi ba zi lar . Some people think that vaccinations are necessary for the health of human beings. (“that” in English is unaligned)
Punctuations	(Uzbek) Vaksina lar inson salomatlig i uchun zarar , de b hisobla ydi ba zi lar . Some people think that vaccination s are necessary for the health of human beings. (Comma in Uzbek is unaligned)

Table 3: Tag-not-attach Rules with Examples

3.3 Annotation Quality

To assure annotation is properly performed and to increase

annotation accuracy and efficiency, best-practices measures are developed for alignment annotation:

- Annotators should read through both the source sentence and target sentences before aligning a sentence
- Annotators first focus on aligning all the content morphemes.
- With all the content-morphemes aligned, annotators then shift to align function morphemes.
- With all content and function morphemes aligned, annotators can shift to tagging unaligned morphemes.
- All tokens in both source and target languages should be either aligned or tagged. No tokens should be left unattended.
- Annotators should reject a sentence if it is not suitable for annotation.

The annotation quality is further assured by a round of quality control by senior annotators and a round of corpus-wide alignment check via a special tool to assure alignment consistency across documents.

4. Annotation Data Format and Features

4.1 Annotation Data Format

Alignment annotation result is stored in .ma files. The format of alignment file is similar to GIZA++ word alignment format, but with some enhancements. Each line contains a list of space delimited alignments for the corresponding sentence. Each alignment is in the format of S-T(linktype) where S and T are a list of comma delimited source and translation token IDs respectively (as shown in the following sample output). S or T can be empty indicating a not-translated token.

```
1-5(COR) 2-(TIN) 6-3(COR) 3-(TIN) 4-(TIN) 7-2(COR)
5-4(COR) -1(TIN)
10-7(COR) 14[TOK]-12,13,14,15(COR) 2-(TIN) 6-
5(COR) 7-3(COR) 1-6(COR) 15-16,17(COR) 8-4(COR) 4-
(TIN) 3-(TIN) 16[TOK]-8,9,11(COR) -18(COR) 9-2(COR)
-1(TIN) 11-(TIN) 5-(TIN) 13-10(COR) 12-(TIN)
```

Valid values for linktype are COR (translated-correct alignment) and TIN (translated-incorrect alignment). Morph tags includes GLU (for morphs to be attached to other morphs), TYP (typos) and TOK (tokenization errors). For instance, in the alignment “2[TYP]-4,6(COR)”, source token #2 (a typo) is aligned to target tokens #4 and #6 to form a correct link. In the alignment “13[GLU],14-10(INC)”, source tokens #13 (tagged as so-called “glue”) and #14 are aligned to English token #10 to form an incorrect link. In the alignment “10-(COR)”, source token #10 having no target correspondent is a not-translated correct link. In the alignment “-19[TYP](COR)”, target token #19 (a typo) having no source correspondent is a not translated correct link.

4.2 Annotation Data Features

Table 4 summarizes the annotation results of Uzbek and

Turkish alignment corpora. An analysis of alignment data reveals that alignment and tagging annotation results are comparable between Uzbek and Turkish. The total occurrences of “alignment with attached morphs” (Uzbek/Turkish morphs having no matched morphs in English) are close to each other in Uzbek and Turkish corpora, indicating a high similarity between these two languages as well as their dissimilarity from English. The translation quality of the source data is perceivable from a high number of “translated-incorrect alignment” occurrence, where translators do not render a semantically-identical translations. The difference between “typo” occurrences of these two languages reflects messiness of Uzbek data of informal genres. As a special effort to capture tokenization/segmentation errors from upstream annotations, the tag “tokenization error” (TOK) is introduced for cases where a morpheme is not properly segmented and morphologically analysed.

Language	Alignment/tagging categories	Occurrence
Uzbek	Translated-correct alignment	10908
	Translated-incorrect alignment	1498
	Alignment with attached morphs	2468
	Not-translated correct morphs	449
	Tokenization errors	454
	Typo	500
Turkish	Translated-correct alignment	8594
	Translated-incorrect alignment	1495
	Alignment with attached morphs	1940
	Not-translated correct morphs	544
	Tokenization errors	343
	Typo	12

Table 4: Uzbek and Turkish Alignment Data Features

5. Data Use

The morpheme alignment annotation corpora are linguistic-orientated and supported by linguistic theories, aiming to reach a variety of users from NLP fields as well as other research fields, such as education or cultural studies. For MT research, the data is intended for all MT performers with varying MT models. The annotation data format is designed to allow MT users to flexibly tailor or customize the annotation for different use. Exploration into the annotation data is therefore crucial for properly customizing the data for training or tuning various MT models. The attachment and tagging approach is introduced to serve these multiple purposes. Exploration into unaligned words and attaching them to their dominating constituents can show valuable hidden grammatical as well as contextual information, which is valuable to

interpretation of semantic completeness in actual context. Unaligned morphemes without any tagging or attachment would lead to coherence loss in understanding communication completeness. The attachment-tagging approach helps to reveal important cultural-linguistic differences in achieving communication equivalency.

5.1 Re-attachment of Unaligned Morphemes for Various MT Model Schemes

In MT modelling, our linguistic-rule-based alignment annotation may not always be the ideal approach for all MT models. To satisfy different MT model preferences, users can modify current annotation data by automatically detaching and re-attaching tagged morphemes to derive their own MT-model-preferred annotations. For instance, your model may not favour the attachment of preposition “of” being attached and co-aligned to NP2 in the structure (NP1 (PNP2)), as annotated with the example “island (NP1) of Japan (NP2)” in current annotation. If your model favours “of” being attached to NP1 instead of NP2, then the annotation can be customized in two steps: 1) detach all preposition “of” from NP2 via the word tag “glue” and 2) re-attach “of” to NP1 based on morphological information. Such automatic customization of annotation is possible because we introduced the attachment and tagging approach so that all unaligned morphs are tagged. This provides various possibilities for MT researchers to tune parameters and look into subtle local features which might affect end translation quality.

5.2 Detachment of Glued Morphemes

If a MT model prefers less vocabulary size and would want to have all unmatched lexical units to be left unaligned, the annotation data can be automatically pre-processed to detach all tagged unaligned morphemes by making use of the “GLUE” tag. This can be conveniently realized as all unaligned morphemes in our annotation are tagged. Tagging unaligned morphemes provide users with more affordable alternative research approaches because a quick and automatic pre-processing of data is far less expensive than re-annotating data with a different linguistic scheme.

Morpheme alignment data can also be useful in fields other than MT. For instance, for lexicon development or language studies, it is important to extract pure semantic/content-morpheme alignments rather than those composite alignments with attached morphemes. Such semantic alignments can also be automatically derived by detaching all attached morphemes via the “GLUE” tag.

6. Conclusion

Recent research on translating morphologically rich languages has decomposed morphologically complex words into tokens of finer granularity and representation for MT. The resultant annotations including morphological alignment based on morpheme tokens has improved word alignment precision and MT quality, easing the problem of data sparsity for morphologically rich languages. Up to date, LDC has created Uzbek-English and Turkish-English

morpheme-level alignment. As described in this paper, both universal language features and idiosyncratic Uzbek and Turkish language peculiarities are addressed in alignment annotations. These corpora are valuable not only because they introduced morpheme as the minimum unit for alignment annotation, but also they are lower resource languages. Currently for BOLT and LORELEI (Low Resource Languages for Emergent Incidents) performers, the corpora will be prepared for broader distribution to LDC members and non-member licensees, through our usual mechanisms, including publication in the LDC catalogue.

7. Acknowledgements

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. References

- Costa-jussà, M. R. (2015). Ongoing Study for Enhancing Chinese-Spanish Translation with Morphology Strategies. In Proceedings of the ACL Workshop on Hybrid Approaches to Translation, HyTra. 2015, Beijing
- Eyig'oz, E. Gildea, D. and Oflazer, K. (2013). Simultaneous Word-Morpheme Alignment for Statistical Machine Translation. In *Proceedings of NAACL-HLT*. Atlanta, Georgia.
- Li, X., Ge, N., Grimes, S., Strassel, S. M. and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta.
- Luong, M. and Kan, M. (2010). Enhancing Morphological Alignment for Translating Highly Inflected Languages. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China.
- Minkov, E., Toutanova, K. and Suzuki, H. (2007). Generating Complex Morphology for Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech
- Toutanova, K., Suzuki, H. and Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH