Discours
Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

16 | 2015
Varia

A Quantitative Analysis of Discourse Phenomena in Machine Translation

Carolina Scarton and Lucia Specia

Electronic version
URL: http://journals.openedition.org/discours/9047
DOI: 10.4000/discours.9047
ISSN: 1963-1723

Publisher:
Laboratoire LATTICE, Presses universitaires de Caen

Electronic reference
Carolina Scarton and Lucia Specia, « A Quantitative Analysis of Discourse Phenomena in Machine Translation », *Discours* [Online], 16 | 2015, Online since 09 September 2015, connection on 01 May 2019. URL : http://journals.openedition.org/discours/9047 ; DOI : 10.4000/discours.9047

discours
Revue de linguistique, psycholinguistique et informatique
http://discours.revues.org/

A Quantitative Analysis of Discourse Phenomena in Machine Translation

Carolina Scarton
Department of Computer Science
University of Sheffield
c.scarton@sheffield.ac.uk

Lucia Specia
Department of Computer Science
University of Sheffield
l.specia@sheffield.ac.uk

Carolina Scarton et Lucia Specia, « A Quantitative Analysis of Discourse Phenomena in Machine Translation », *Discours* [En ligne], 16 | 2015, mis en ligne le 9 septembre 2015.

URL : http://discours.revues.org/9047

Titre du numéro : *Varia*
Coordination : Lydia-Mai Ho-Dac et Frédéric Landragin

revues.org
CENTRE POUR L'ÉDITION ÉLECTRONIQUE OUVERTE
CENTRE FOR OPEN ELECTRONIC PUBLISHING

discours

Presses universitaires de Caen

A Quantitative Analysis of Discourse Phenomena in Machine Translation

Carolina Scarton

Department of Computer Science
University of Sheffield

Lucia Specia

Department of Computer Science
University of Sheffield

State-of-the-art Machine Translation (MT) systems translate documents by considering isolated sentences, disregarding information beyond sentence level. As a result, machine-translated documents often contain problems related to discourse coherence and cohesion. Recently, some initiatives in the evaluation and quality estimation of MT outputs have attempted to detect discourse problems in order to assess the quality of these machine translations. However, a quantitative analysis of discourse phenomena in MT outputs is still needed in order to better understand the phenomena and identify possible solutions or ways to improve evaluation. This paper aims to answer the following questions: What is the impact of discourse phenomena on MT quality? Can we capture and measure quantitatively any issues related to discourse in MT outputs? In order to answer these questions, we present a quantitative analysis of several discourse phenomena and correlate the resulting figures with scores from automatic translation quality evaluation metrics. We show that figures related to discourse phenomena present a higher correlation with quality scores than the baseline counts widely used for quality estimation of MT.

Keywords: discourse in machine translation, document-level quality estimation, discourse features for quality estimation

1. Introduction

1 One challenge in Natural Language Processing (NLP) is how to automatically evaluate language output applications such as Machine Translation (MT) and Text Summarization. Although these tasks are very different, they are related in that a "target" text is produced given an input "source" text. The desiderata for evaluation metrics for these tasks is that they should measure quality with respect to different aspects (e.g., fluency and adequacy) and they should be fast and scalable. While human evaluation seems to be more reliable than completely automatic solutions, it often introduces biases from specific annotators. Human evaluation can be very subjective and annotators may have different perspectives on the same phenomena, especially if guidelines are not well defined or are vague. Human evaluation is also expensive and cumbersome for large datasets, and is not possible for certain scenarios, such

as *gisting*[1] in MT. Therefore, a significant amount of work has targeted measuring the quality of MT without direct human intervention.

2 BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and METEOR (Banerjee & Lavie, 2005) are examples of widely used automatic evaluation metrics for MT. These metrics compare the outputs of MT systems with human reference translations. The assumption is that the closer the MT output is to the reference translation, the better this output is. BLEU (Bilingual Evaluation Understudy) is a precision-oriented metric that compares n-grams (typically n = 1-4) in MT outputs against n-grams in human references, measuring how close the output of a system is to one or more references. TER (Translation Error Rate) measures the minimum number of edits required to transform the MT output into the closest reference text. METEOR (Metric for Evaluation of Translation with Explicit Ordering) scores MT outputs by aligning them with given references. This alignment can be done by exact, stem, synonym or paraphrase matching.

3 One limitation of reference-based metrics is that if the MT system outputs a translation that is considerably different from the references, this does not really mean that it is a bad output. Another problem is that human effort is still needed to produce the references. Finally, and importantly, these metrics cannot be used in scenarios where the output of the system is used directly by end-users, for example a user reading the output of Google Translate[2] for a given news article cannot count on a reference for that translated text.

4 Quality Estimation (QE) metrics aim to predict the quality of the output of MT systems without using human references. They rely on features related to quality that are extracted from source and target documents, and optionally from the MT system that produced the translations (Blatz et al., 2004; Specia et al., 2009a and b; Bojar et al., 2013; Bojar et al., 2014). They also require examples of translations with corresponding quality scores at system-building time (e.g., *likert*, Human-targeted Translation Error Rate – HTER [Snover et al., 2006] – or even BLEU-style metrics). These scores are used, along with the features, to train supervised Machine Learning (ML) models (e.g., regressors) to predict the scores for unseen translations. The advantage of these approaches is that we do not need to have all the words, sentences or documents of a task evaluated manually, we just need enough data points to build the ML model. QE systems predict scores that reflect how good a translation is for a given scenario. For example, a widely predicted score in QE is HTER, which attempts to measure the effort needed to post-edit a sentence. A user of a QE system predicting HTER could decide whether to post-edit or translate sentences from scratch based on the score predicted for each sentence.

---

1. *Gisting* refers to MT output used directly by end-users with the intention of "getting the general idea of the original document". Therefore, a perfect translation is not needed.

2. See: https://translate.google.com/.

5 QE for MT has a number of challenges, including the following:

- **Granularity level.** The vast majority of work done on QE is at sentence level. Not only the predictions but also the features refer to sentence-level information only. Sentence-level approaches are very useful in the post-editing scenario and in many others (e.g. gisting, combining MT systems). Document-level predictions are interesting in scenarios where one wants to evaluate the overall score of an MT system or where the end-user is interested in the quality of the document as a whole. However, there are several challenges in targeting QE at a different granularity level. Firstly, the requirements for quality labels are different. Whilst for sentence-level QE HTER and *likert* assessments are feasible, this is not the case for document-level. QE human *likert* scores are difficult to apply since, for a human, evaluating long units such as a document is much more subjective than evaluating a single sentence. Moreover, the use of HTER is not appropriate for scenarios where improving post-edition is not the final aim. Scarton et al. (2015) discuss the use of automatic metrics (such as BLEU) as quality labels for QE, reporting them as inappropriate and proposing a two-stage post-edition method to post-edit sentences in order to obtain a score for entire paragraphs (a first stage at sentence level, without context, and a second stage for entire paragraphs). They show that some editions were only done after the sentence context had been given to the annotators. Besides the quality label, the choice of features is also a challenge. It is not clear whether features should simply be a combination of sentence-level ones and whether document-wide information can help improve predictions. Another challenge is to design new features in order to capture phenomena at document level. Scarton and Specia (2014) and Soricut and Echihabi (2010) are examples of previous work that introduce document-level features. In this paper we further study and discuss the role of discourse-aware features for document-level QE.
- **Use of linguistic information.** The use of linguistic information to improve predictions in QE is a problem that goes beyond feature engineering for this particular task, as extracting linguistic information is generally a challenge in NLP. Modelling discourse is known to be more complex than shallower levels of linguistic processing such as syntax. As a result, the availability and reliability of resources and tools at this level is limited. Therefore, the study and evaluation of relevant (new) tools and methods needs to be carefully done in order to create linguistic-aware features. In the context of document-level QE, discourse is a linguistic phenomenon that often manifests across sentences. It is related to how sentences are connected, the genre and domain of a document, anaphoric references, etc. Since the state-of-the-art statistical MT (SMT) systems translate documents sentence by sentence, disregarding discourse information, it is likely that the outputs of these systems will contain problems related to discourse.

Using discourse for QE is a challenge mainly because tools and resources are few (or not available for several languages) and discourse is a linguistic phenomenon that depends on other phenomena (such as sentence-level semantics and syntax).

6 In this paper we focus on the study of several discourse-based features and their correlation with HTER for document-level QE. Two corpora (English-French and French-English) were used. Since some of the tools to extract discourse phenomena from corpora are only available for English, the evaluation was conducted only for this language (first as source language and second as target language). One of our hypotheses is that discourse information can be used to improve state-of-the-art QE models by detecting issues related to discourse due to the way machine translations are produced. Previous work has inspected the effects of MT in discourse phenomena. For example, Carpuat and Simard (2012) studied lexical consistency in MT, Marcu et al. (2000) compared Rhetorical Structure Theory (RST) (Mann & Thompson, 1987) trees of source and target documents and Li et al. (2014) studied the impact of discourse connectives in MT. They found that MT systems can harm various discourse aspects in the target language. There is also work that attempts to include discourse information in SMT (Marcu et al., 2000; Carpuat, 2009; Zhengxian et al., 2010; Le Nagard & Kohen, 2010; Meyer & Popescu-Belis, 2012; Ture et al., 2012; Ben et al., 2013; Hardmeier, 2014), uses discourse information for MT evaluation (Giménez & Màrquez, 2009; Giménez et al., 2010; Wong & Kit, 2012; Meyer et al., 2012; Guzmán et al., 2014) and that applies discourse features for QE (Rubino et al., 2013; Scarton & Specia, 2014). Our work however focuses on measuring the role of discourse in QE, relying on target and source documents only. We show that the majority of discourse-based features correlate better with HTER than simpler, sentence-level features. This corroborates our hypothesis that accurately detecting discourse phenomena can help predict the quality of MT outputs at document level. It is worth mentioning that although we are aware that HTER scores for document-level QE are probably not ideal, these were the only available corpora with some human-targeted score at document level that we could use for our experiments.

7 The remainder of this paper is structured as follows. In Section 2, related work is described. Section 3 introduces the features used and other experimental settings. Section 4 presents our results and an in-depth analysis of three discourse phenomena.

2. Related work

8 Various types of related work are presented here in order to motivate and contextualise our research. Section 2.1 introduces the standard framework for QE. Section 2.1.1 presents QE research that considers linguistic information in order to contextualise our use of discourse for QE and show that this phenomenon has been understudied. In Section 2.1.2, work on document-level QE is presented, showing that only basic lexical cohesion features have been applied so far. Finally, Section 2.2 describes work that aims to analyse, improve or automatically evaluate MT systems.

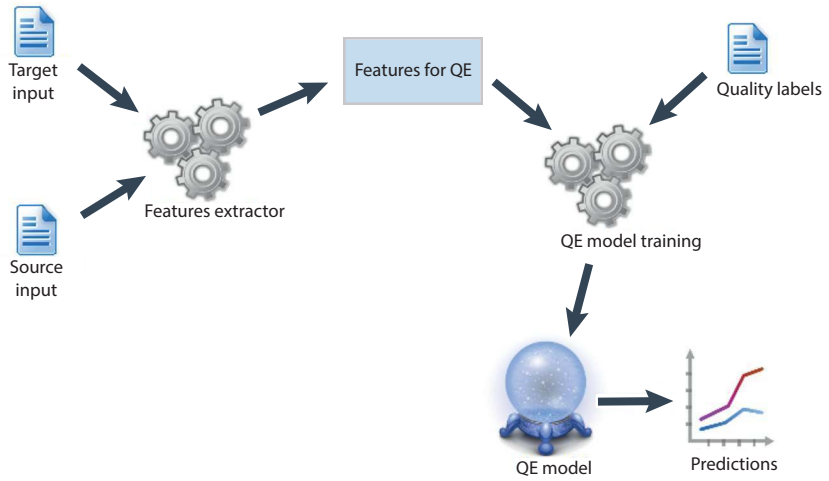

Figure 1. General framework of QE

2.1. QE for MT

9 The task of QE consists in predicting the quality of MT system outputs without the use of reference translations. To do this, previous work has used supervised ML approaches (mainly regression and classification algorithms). Besides the specific ML method adopted, the choice of features is also a design decision that plays a crucial role. The general framework of QE is illustrated in Figure 1. The sentences (or documents) in the source and target languages and optionally information from the MT system are used as input to design features. These features can be simple counts (e.g., length of sentences in source or target), or can explore linguistic information (e.g., percentage of nouns in source or target) and take into account internal information from the MT system that produced the translations (e.g., overall SMT model score, or n-best lists to build language models). The features extracted are used as input to train a QE model. The training is generally done using supervised ML, such as regression algorithms. In this case, a training set with quality labels is provided to the ML algorithm. These quality labels are the scores that the QE model will learn to predict. Therefore, the QE model will be able to predict a quality score for a new, unseen translation (sentence or document). The quality labels can be *likert* scores, HTER, or BLEU, to mention some widely used examples.

10 Most studies in the area (Blatz et al., 2004; Specia et al., 2009a and b; Specia & Farzindar, 2010; He et al., 2010) use shallow features (e.g., length of words and sentences) and features related to the MT system (e.g., SMT model score) to estimate quality at word or sentence levels. They experiment with different language pairs (e.g., Chinese-English, English-Spanish, French-English). Concerning what to predict, previous work predicts BLEU-style, HTER scores, or human scores. Different algorithms for supervised ML have been used: some treat the task as a

classification problem while others handle it as a regression problem. This depends mostly on the quality label used for training. Most of this work does not explore deep linguistic information.

11 QE frameworks such as QuEST [3] (Specia et al., 2013) are available for sentence-level prediction. QuEst has modules to extract several features from source and target documents and to experiment with various ML algorithms for predicting QE. The features are divided in two types: glass-box (dependent on the MT system) and black-box (independent from the MT system). In addition to the state-of-the-art QE features, the framework provides a simple but competitive set of 17 black-box features that has been widely used as baseline for Workshop on Machine Translation (WMT) QE shared tasks (Callison-Burch et al., 2012; Bojar et al., 2013; Bojar et al., 2014) [4]. These are basic features such as number and ratio of source/target tokens and punctuation marks, target and source language models, number of possible translations per word in source. An extended version of QuEst for word and document-level prediction and feature extraction has also been developed recently (Specia et al., 2015) [5].

2.1.1. Linguistic features for QE

12 Specia et al. (2011) predicted the adequacy of Arabic-English translations at sentence level. They grouped the features used in four classes: (i) source complexity features (such as average source word length); (ii) target fluency features (such as target language model); (iii) adequacy features (such as ratio of percentage of nouns in the source and target sentences); and (iv) confidence features (such as SMT model score). Linguistic features were used for the first three categories and covered different linguistic levels: lexical (such as ratio of percentage of nouns in the source and target sentences), syntactic (such as absolute difference between the depth of the syntactic trees of the source and target sentences) and semantic (such as difference in the number of "person"/"location"/"organization" named entities in source and target sentences).

13 Avramidis et al. (2011) considered syntactic features for ranking German-English SMT systems. The syntactic information was generated using a Probabilistic Context-Free Grammar (PCFG) parser on the target and source sentences. The best results were obtained when syntactic features were used. Similarly, Almaghout and Specia (2013) used Combinatory Categorical Grammar (CCG) in order to extract features for QE. They applied these features to the output of French-English and Arabic-English systems. The use of CCG features outperformed the PCFG features of Avramidis et al. (2011) when only these features were used. Hardmeier (2011) applied syntactic tree kernels to QE at sentence level for English-Spanish and

3. See: http://www.quest.dcs.shef.ac.uk.

4. See: http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17.

5. See: https://github.com/ghpaetzold/questplusplus.

English-Swedish machine translations. The syntactic information was encoded into the model as tree kernels, which measure the similarity of syntactic trees. The use of tree kernels led to improvements over a strong baseline.

14 Pighin and Màrquez (2011) used Semantic Role Labelling (SRL) to rank English-Spanish SMT outputs. They automatically annotated the SRLs in the source side and projected them into the target side, by using the word alignments of the SMT system. The evaluation was done by considering the human assessments available for the WMT 3 2007-2010 corpora and the dataset described in Specia et al. (2010). They trained the model in Europarl data, therefore, they separated the data into in-domain data (Europarl) and out-of-domain data (news). The results of using SRL features were better than the baseline for the in-domain data. Some results of out-of-domain data were comparable to in-domain data when SRL features were applied.

15 Felice and Specia (2012) introduced several linguistic features for English-Spanish QE. They covered three linguistic levels: lexical (e.g., percentage of nouns in the sentence), syntactic (e.g., width and depth of constituency trees) and semantic (e.g., number of named entities). Although the combination of linguistic features did not improve the results, an analysis of the contribution of features showed that linguistic features appeared among the top five. The application of a feature selection method resulted in a set 37 features, out of which 15 were linguistic. This selected set led to improved results.

16 Rubino et al. (2013) explored a Latent Dirichlet Allocation approach (LDA) (Blei et al., 2003) to create topic models in two different ways. First, the authors concatenated both sides of a large bilingual corpus at sentence level. Sentences of source and target language were treated as a single bag of words to generate a single topic model. Therefore, the topic model built contained the vocabulary of both languages. Second, they explored a polylingual model in which two monolingual topic models were built for each language. Features were extracted based on the distance between source and target languages at sentence level by using metrics such as Euclidean distance and the topic models generated. They experimented with Arabic-English, English-French and English-Spanish machine translations and reported improved results by using topic models.

2.1.2. Document-level QE

17 QE is traditionally done at sentence level. This happens mainly because the majority of MT systems translate texts at this level. Another reason is that sentence-level approaches have many applications. Document-level QE can be considered a harder problem because of possible variations in quality across the document. On average a given document may have a "good" quality score, but it could contain some particularly bad sentences, while another document with the same average quality score could contain average quality sentences across the whole document. Sentence-level approaches are more feasible as they can explore the peculiarities of each sentence

and are particularly useful for post-editing applications. However, these approaches do not consider the document as a whole and information regarding discourse is often disregarded. Moreover, for scenarios in which post-editing is not possible, for example, *gisting*, quality predictions for entire documents are necessary.

18 Many of the features developed for sentence-level QE can be directly used at document level (such as number of words in source and target sentences). However, other features that better explore the document as a whole or discourse-related phenomena can provide additional information. State-of-the-art features for document-level QE are based on pseudo-references. Pseudo-references are MT outputs generated by other MT systems, different from the MT system of which we want to predict the quality. These additional MT outputs (pseudo-references) are used in the same way as human references: automatic evaluation metrics are applied to measure the similarity between pseudo-references and MT outputs. These metrics can be used as features for QE. One important requisite is knowledge about the quality of the MT systems used to generate the pseudo-references. Taking BLEU as an example of feature, a high BLEU score against a pseudo-reference from a good MT system would mean that the MT output under investigation is probably good. Conversely, if the comparison is done against a bad MT system, a high BLEU score would mean that the MT output is probably a bad translation.

19 Soricut and Echihabi (2010) explored document-level QE prediction to rank documents translated by a given MT system, predicting BLEU scores. Features included text-based, language model-based, pseudo-reference-based, example-based and training-data-based. Pseudo-reference features were BLEU scores computed using pseudo-references from an off-the-shelf MT system, for both the target and the source languages. Soricut and Narsale (2012) also considered document-level prediction for ranking, proposing the aggregation of sentence-level features for document-level BLEU prediction. The authors claim that a pseudo-reference feature (based in BLEU) was the most discriminative in their experiments.

20 Scarton and Specia (2014) explored lexical cohesion and Latent Semantic Analysis (LSA) cohesion (Landauer et al., 1998) for document-level QE. The language pairs studied were English-Brazilian Portuguese, English-Spanish and Spanish-English. The lexical cohesion features were repetitions of words, lemmas or nouns, based on the work of Wong and Kit (2012). Repetition values were normalised by the number of content words in the document. LSA cohesion was achieved following the work of Graesser et al. (2004). A matrix of words versus sentences was built and Singular Vector Decomposition (SVD) was applied. Spearman's *rho* correlation was then applied between the word vectors of each sentence. High correlation between sentences is a sign that the sentences are connected. Therefore, LSA cohesion can capture word correlations that go beyond word repetitions. Pseudo-reference features (BLEU and TER) were also applied in this work. BLEU and TER were used as quality labels. The best results were achieved with pseudo-reference features. However, LSA cohesion features alone also showed improvements over the baseline.

2.2. Discourse in MT

21 To the best of our knowledge, Rubino et al. (2013) and Scarton and Specia (2014) are the only efforts reported to address discourse in QE. More generally in the MT area, several approaches have been proposed to use discourse information. The majority of them are very recent, but the need for document-level information to improve MT has been acknowledged for a long time. However, it is hard to integrate discourse information into traditional state-of-the-art sentence-level MT systems. It is also challenging to build a document-level or discourse-based MT system from scratch. One exception is the document-level decoder by Hardmeier et al. (2012), which performs decoding as a two-stage process, where the first stage uses a standard phrase-based SMT system to produce a draft translation, and the second stage applies a small set of operations to change parts of the translation (e.g., replacing a phrase by another in the phrase table), in which the resulting translations can be scored based on global features for the entire document as context. Other initiatives focus mostly on the design and integration of discourse informed features based on limited context into current decoding algorithms. These include lexical cohesion, co-reference, discourse relations and topic models.

22 Lexical cohesion refers to relations between lexical elements, such as words and phrases. It is achieved by vocabulary choices: word repetitions, use of synonyms, and collocations. Identifying cohesion devices at word level does not require the full interpretation of the document, as a simple search for repeated words can already identify lexical cohesion. Carpuat (2009) explored the "one translation per discourse" hypothesis (based on the "one sense per discourse" hypothesis) to post-processing SMT outputs. Ture et al. (2012) modified the decoder of an SMT system in order to model the "one translation per discourse" hypothesis into the system. Xiao et al. (2011) proposed a post-processing and a re-decoding method to consider lexical cohesion in document-level MT. Ben et al. (2013) also included lexical cohesion features in an SMT decoder. Wong and Kit (2012) proposed lexical cohesion metrics for the evaluation of MT systems at document level. They used these metrics alone and also combined them with traditional metrics, such as BLEU.

23 A document is considered coherent if its components, such as sentences, are well connected. Pronominal anaphora and connectives are explicit signals of coherence. One needs to take context into account, sometimes across sentences, in order to interpret the function of an element like a pronoun. Giménez and Màrquez (2009) and Giménez et al. (2010) explored the Discourse Representation Theory (Kamp, 1981) in automatic evaluation of MT. Le Nagard and Kohen (2010), Hardmeier and Federico (2010), Guillou (2012) and Hardmeier (2014) explored anaphora resolution techniques to improve SMT. Popescu-Belis et al. (2012), Meyer and Popescu-Belis (2012), Meyer et al. (2012) and Li et al. (2014) focused on modelling discourse connectives to improve SMT systems. Meyer et al. (2012) also proposed an evaluation metric (ACT: Accuracy of Connective Translation) to evaluate the translation of discourse connectives.

24 RST is a linguistic theory that correlates macro and micro units of discourse in a coherent way. A key step before applying RST is the text segmentation into Elementary Discourse Units (EDUs). EDUs are defined at sentence, phrase or paragraph-level. RST proposes discourse relations at EDU level; for example, two EDUs can be related by a *contrastive* relation. These relations are represented in tree form. If phrases are considered as EDUs, relations between phrases, sentences and paragraphs can be represented in the same tree. Marcu et al. (2000) explored RST to identify the feasibility of building a discourse-based SMT system. Guzmán et al. (2014) and Joty et al. (2014) used RST trees comparison for MT evaluation.

25 Similarly to lexical cohesion, topic models are related to vocabulary choices and cohesion of the document as they identify relationships between words in the document, although the methods used are considerably different. Topic models can be considered more general as they measure relationships between words beyond repetitions. For example, in documents related to *war*, words like *death* and *attack* will present a higher correlation between each other than each of them with the word *sport*. Words on the same topic are grouped together. Therefore, these methods can measure if a document that follows a given topic, is related to a genre or is part of a specific domain. Identifying such levels of correlation in a document can be used to measure coherence. However, the deep understanding of a text is not necessary for topic models. Zhao and Xing (2007) used a bilingual latent variable approach (called HM-BiTAM – Hidden Markov Bilingual Topic AdMixture) to model cohesion in SMT. Zhengxian et al. (2010) and Eidelman et al. (2012) used LDA to improve SMT.

3. Experimental settings

3.1. Corpora

26 The data used in the experiments conducted here are corpora with post-editions of MT outputs. With the post-editions, it is possible to calculate the HTER metric, which has been widely used as a quality label for QE at sentence level. Two corpora were used:

- LIG corpus (Potet et al., 2012): this corpus contains 10,881 French-English (FR-EN) machine-translated sentences (and their post-editions) from several editions of WMT translation shared tasks (news documents). The document boundaries were recovered and the HTER was calculated at document level[6]. 119 documents were analysed.
- Trace corpus (Wisniewski et al., 2013): this corpus contains 6,693 FR-EN and 6,924 English-French (EN-FR) machine-translated sentences with their post-editions. We used 38 documents recovered from the WMT and

6. Many thanks to Karin Smith for generating the document mark-ups.

Technology, Entertainment and Design (TED) Talks EN-FR sets[7]. Only the phrase-based SMT outputs were considered.

27 Due to lack of resources for French, we only considered evaluations for the English language. With these two corpora, however, we were able to evaluate discourse-related features for the source side (EN-FR) of the Trace corpus and target side of the LIG corpus.

28 It is worth mentioning that the MT systems used to translate the documents in both corpora were not discourse-aware. They were typical SMT systems that disregard discourse while producing translations. Unfortunately, we are not aware of any corpora of segments machine-translated by a discourse-aware MT system and assessed by humans, which would be ideal for our experiments. As previously mentioned, these systems are more likely to translate discourse phenomena incorrectly. However, it is worth emphasising that our main goal is to design document-level QE precisely because this problem is frequent and inherent to most MT systems. Although we are aware that HTER scores for document-level QE are not ideal, the LIG and Trace corpora were the only available ones from which we could extract some form of human-targeted scores at document-level. Moreover, HTER is expected to be better for QE than BLEU-style metrics, since it involves human corrections.

3.2. QE features

29 The discourse-related information we considered as features to evaluate translations were:

- **Lexical Cohesion:** lexical cohesion features at document level, following from Scarton and Specia (2014) (LC – Argument target/source, LC – Lemma target/source and LC – Noun target/source).
- **LSA Cohesion:** LSA cohesion features at document level, following from Scarton and Specia (2014) (LSA – All target/source and LSA – Adj target/source).
- **Connectives:** counts of connectives (Connectives).
- **Pronouns:** counts of pronouns (Pronouns).
- **Discourse unit segmentation (EDU break):** number of breaks (EDU). An example of a sentence broken into EDUs is the following:

  *"However, **EDU_BREAK** despite the economic success, **EDU_BREAK** it has become increasingly clear **EDU_BREAK** that Blair was over".*

- **RST relations:** number of *Nucleus* (RST – Nucleus) and number of *Satellite* (RST – Satellite) relations. An example of *Nucleus* and *Satellite* relations of

7. The other sets did not have document-level mark-ups.

an *Elaboration* relation is presented below (the Satellite – text in *leaf 7* – is in an *Elaboration* relation with the text in *leaf 6* – the Nucleus):

Elaboration relation:

- ***Nucleus*** *(leaf 6): "Brown has coveted the office of Prime Minister since May 12, 1994, the fateful day".*
- ***Satellite*** *(leaf 7): "when John Smith, the Labour leader in opposition, died of a heart attack".*

30 The analysis was done on the target side of the LIG corpus, and on the source side of the Trace corpus, since the tools needed for the identification of Connectives, EDU breaks and RST relations are only available for English.

31 The LC features were based on word repetitions. Following Scarton and Specia (2014), we counted the tokens, lemmas or noun repetitions across the document.

32 LSA features were extracted by building a matrix of word frequency per sentence. SVD was then applied to this matrix and correlations were computed between the word vectors of each sentence (Scarton and Specia, 2014). Both LC and LSA features can be applied to most languages as they only require a *PoS tagger* with lemma information.

33 In order to extract connectives, we used the connectives *tagger* developed by Pitler and Nenkova (2009). This tool automatically annotates connectives based on the output of the Charniak Syntactic Parser (Charniak, 2000). The tool was trained with Penn Discourse Treebank annotations (Prasad et al., 2008) and the connectives can fall into four classes: Expansion, Contingency, Comparison and Temporal. The *tagger* also outputs a label for lexical items of the same form as connectives but with a different function (e.g., prepositions), so these are not considered connectives in our experiments. In our experiments, we disregarded the distinction among different types of connectives, counting the total number of connectives.

34 Pronouns were identified by looking at the tag "PRP" of the Charniak parser. This tag marks only personal pronouns which are often found in the role of pronominal anaphora. If pronouns like *they*, *he* or *it*, for example, appear in a document, one expects that there will be an antecedent to resolve them.

35 EDU breaks were marked using the Discourse Segmenter module of the Discourse Parser developed by Joty et al. (2013). This module uses the outputs of the Charniak parser and the EDUBREAK module of the SPADE tool (Soricut and Marcu, 2003) in order to break sentences into smaller discourse units.

36 RST relations were extracted with the same discourse parser (Joty et al., 2013). This parser is able to annotate RST trees at sentence and document levels. At document level, the trees go from the smallest units (EDUs) to sentences and paragraphs, until they reach the full document. At sentence level, the trees model intra-sentence discourse relations.

37 For comparison, we also extracted the baseline features from QuEst, which deal with MT output at sentence level, averaged their values for the entire document

and used these averages as document-level features. It is worth mentioning that these features are often among the best for sentence-level QE (Shah et al., 2013). They include both target and source language features.

38 Target features:

- **QuEst 1:** number of tokens in the target sentence;
- **QuEst 2:** language model probability of target sentence;
- **QuEst 3:** number of occurrences of the target word within the target hypothesis (averaged for all words in the hypothesis – type/token ratio);
- **QuEst 4:** number of punctuation marks in the target sentence.

39 Source features:

- **QuEst 5:** average source token length;
- **QuEst 6:** number of tokens in the source sentence;
- **QuEst 7:** language model probability of source sentence;
- **QuEst 8:** average number of translations per source word in the sentence (as given by IBM 1 table thresholded such that prob(t|s) > 0.2);
- **QuEst 9:** average number of translations per source word in the sentence (as given by IBM 1 table thresholded such that prob(t|s) > 0.01) weighted by the inverse frequency of each word in the source corpus;
- **QuEst 10:** percentage of unigrams in quartile 1 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 11:** percentage of unigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 12:** percentage of bigrams in quartile 1 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 13:** percentage of bigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 14:** percentage of trigrams in quartile 1 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 15:** percentage of trigrams in quartile 4 of frequency (higher frequency words) in a corpus of the source language;
- **QuEst 16:** percentage of unigrams in the source sentence seen in a corpus (SMT training corpus);
- **QuEst 17:** number of punctuation marks in the source sentence.

3.3. Correlation analysis

40 We studied the impact of discourse phenomena on the quality of MT output by measuring the correlation of our discourse features, as well as QuEst features, against a quality label, and further manually analysing interesting cases. We used the HTER as quality label for each document and computed correlations with Spearman's *rho*

and Pearson's *r* correlation coefficients (a *p-value* smaller than 0.05 is considered significant in this analysis). HTER was used as quality label because it is widely used in QE. Besides that, Li et al. (2014) claim that HTER is more suitable as quality scores for discourse analysis than reference-based MT evaluation metrics, given that the comparison is made between target documents and post-edited documents.

41 The variation in sentence-level averaged HTER values for each document in the LIG corpus is low (only 0.07 points of standard deviation). Therefore, to better evaluate whether discourse features correlate with HTER scores, besides applying the analysis to the entire corpus, we also divided the corpus into four bins. The bins show how the features behave in the extreme quality cases: the best and worst parts of the corpora according to HTER. We sorted the documents according to HTER and split them into bins as follows:

- 10 documents: 5 documents with the best 5 HTER scores and the 5 documents with the worst 5 HTER scores;
- 20 documents: 10 documents with the best 10 HTER scores and the 10 documents with the worst 10 HTER scores;
- 40 documents: 20 documents with the best 20 HTER scores and the 20 documents with the worst 20 HTER scores;
- 80 documents: 40 documents with the best 40 HTER scores and the 40 documents with the worst 40 HTER scores.

42 The portion of the Trace corpus used here was too small to be split into bins (only 38 documents) and therefore in what follows we only present results per bin for the LIG corpus.

4. Impact of discourse phenomena on MT quality

4.1. Correlation of discourse features and HTER

43 The results of our analysis on the FR-EN LIG corpus are shown in Figure 2 (Pearson's *r* correlation) and Figure 3 (Spearman's *rho* correlation). Since the analysis was done in the target side only, the QuEst features used were QuEst 1-4. For the bin with 10 documents, the discourse features "RST – Nucleus", "RST – Satellite" and "EDU", together with the "QuEst 1" feature, show the best correlation scores according to both Pearson's *r* and Spearman's *rho*. For the bins of 20 documents, "QuEst 2" and "RST – Nucleus" show the highest Pearson's *r* correlation scores with HTER (above 0.37). The highest Spearman's *rho* correlation score is shown by "Pronouns". For bins with 40, 80 and all documents (119), the "LC – Argument target" feature shows the highest Pearson's *r* and Spearman's *rho* correlation scores (around –0.35, –0.23 and –0.20 respectively). Note that, in this case, the correlation scores are negative, but they still indicate correlation between quality and feature. In fact, a negative correlation is expected because higher values for the "LC – Argument target" feature mean higher document cohesion and thus lower HTER scores.

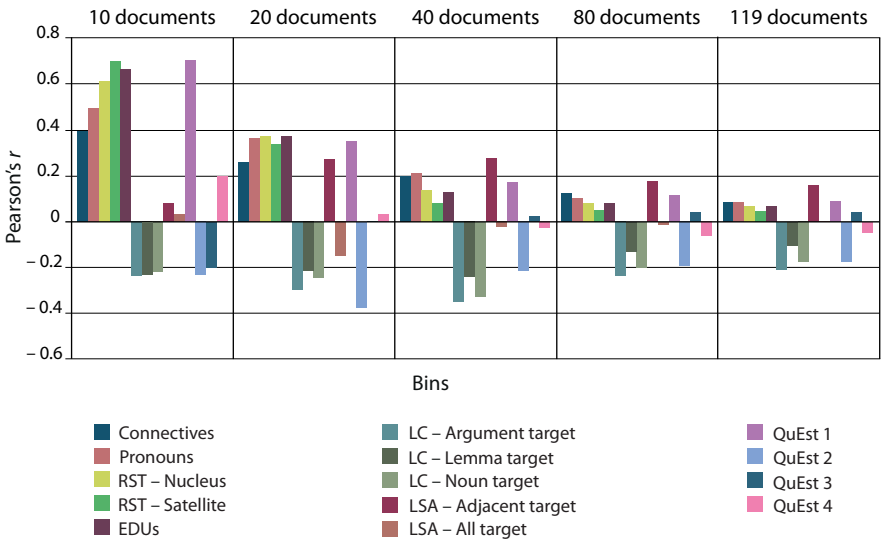

Figure 2. Pearson's *r* correlation between target features and HTER values on the LIG corpus

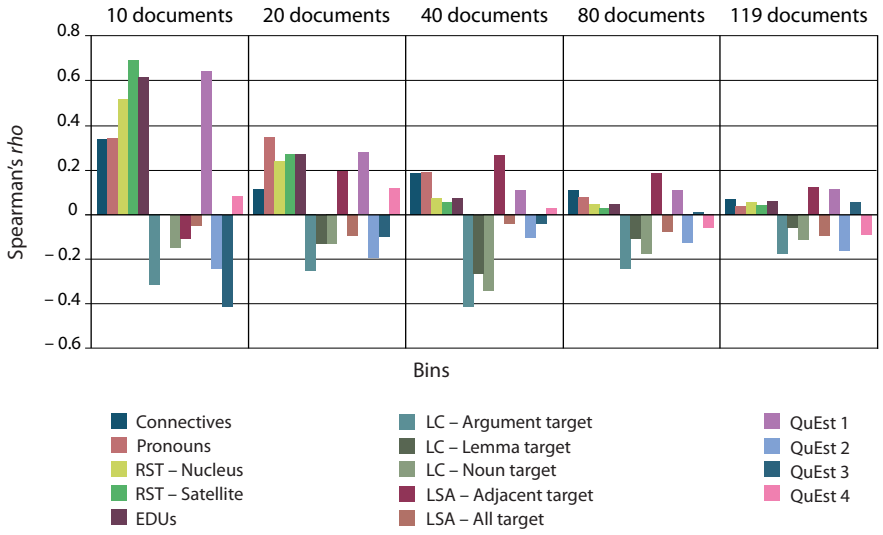

Figure 3. Spearman's *rho* correlation between target features and HTER values on the LIG corpus

44 As expected, both Pearson's *r* and Spearman's *rho* correlation scores are higher when moving from all documents to the 10-document bin. However, this was not the case for all features. In fact, it is possible to observe in Figures 2 and 3 that even taking the extreme quality values only leads to larger correlations in some discourse

phenomena. In the case of sentence-level features, only a feature that is known to perform very well for QE at this granularity level (QuEst 1 – target sentence length) achieves a high enough correlation score (above 0.6 for Pearson's *r* and Spearman's *rho*), comparable to "RST – Satellite" and "EDUs". All the other features achieved correlations of 0.4 or below. This provides evidence of how document-level QE differs from sentence-level QE.

45 Results for EN-FR documents from the Trace corpus (entire corpus, no bins) are shown in Figure 4 (Pearson's *r* correlation) and Figure 5 (Spearman's *rho* correlation). In this case, the analysis was done in the source side only, and the QuEst features used were QuEst 5-17.

46 For the analysis of English as source, the best feature is QuEst 5 with correlation scores below –0.4 for Pearson's *r* and below –0.5 for Spearman's *rho*, followed by "LC – Argument source" (with almost 0.4 points for both correlation metrics). However, all discourse features showed correlations above 0.2 or below –0.2 (with both metrics), higher than several QuEst features.

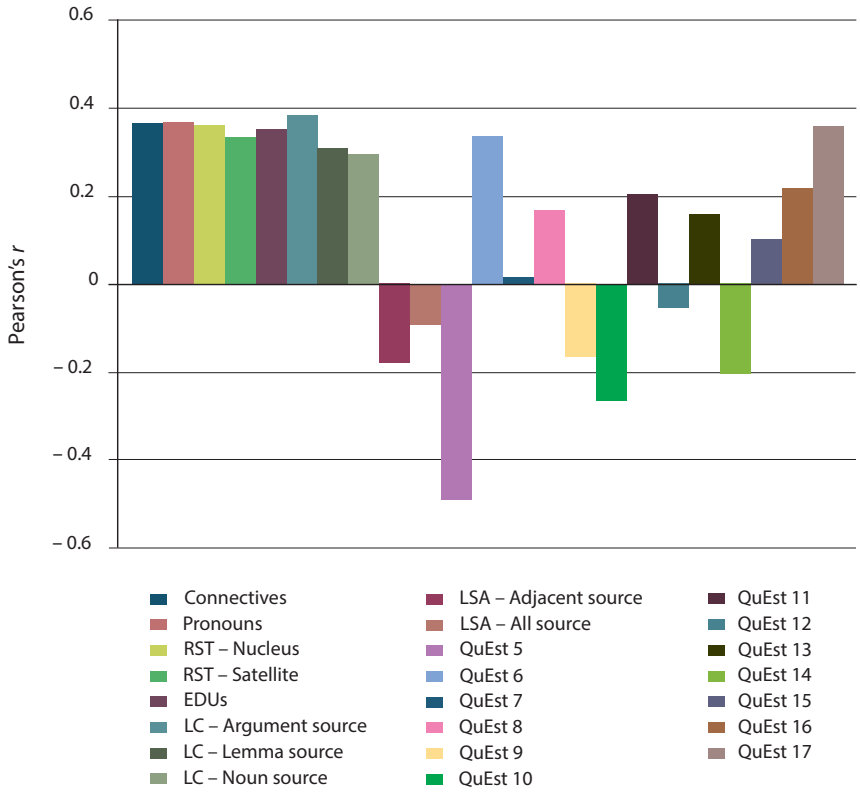

Figure 4. Pearson's *r* correlation between target features and HTER values on the Trace corpus

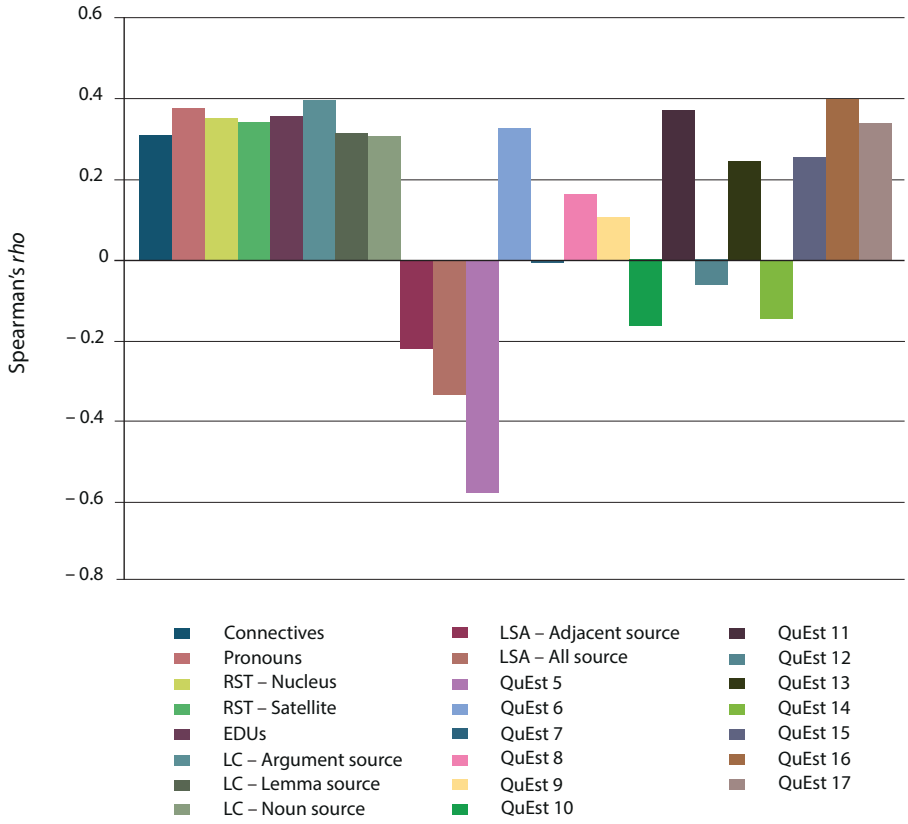

Figure 5. Spearman's *rho* correlation between target features and HTER values on the Trace corpus

47 Based on this analysis we can conclude that our discourse features do correlate with HTER scores and this correlation is often higher than the correlation presented by the basic, sentence-level QE features. Therefore, we believe that discourse information has the potential to improve state-of-the-art QE models.

4.2. Analysis of pronouns, connectives and EDUs

48 In order to better understand some of the discourse phenomena and the reasons behind their correlation with HTER, we conducted an analysis with the following features: number of pronouns, number of connectives and number of EDU breaks for the 10-document bin of the LIG corpus. Although these features do not correspond to all the best features identified in the previous section, they are the ones that it is feasible to analyse manually or semi-automatically. The pronoun count achieved 0.34 points of Spearman's *rho* correlation and 0.5 of Pearson's *r* correlation, but the *p-values* were higher than 0.05. This means that the correlation could be by chance. Pronouns were therefore analysed manually. An example of a problem

with pronouns found in the LIG corpus is the following, where **MT** is the original machine translation and **PE** its post-edited version:

> **MT:** "*Obviously, Gordon Brown wants to succeed Tony Blair as British prime minister. [...] Indeed, **it** has to renege on Blair's legacy, which, at least means promise to leave Britain for the Iraq war*".
> **PE:** "*Obviously, Gordon Brown wants to succeed Tony Blair as British Prime Minister. [...] Indeed, **he** absolutely has to disavow Blair's legacy, which at least means promising to get Great Britain out of the Iraq war*".

49 This example shows a change in the pronoun "it" in the MT output, corrected to "he" in the post-edition. Another example, where the pronoun "it" is removed in the post-edition, is the following:

> **MT:** "*It is the problem that **it** is the most urgent need to address: but for now, none of the main political parties has dared to touch it*".
> **PE:** "*This is a problem that must be addressed immediately, but for now, none of the major political parties has dared to touch it*".

50 Since the correlation between the number of pronouns against HTER was positive, the five documents with the highest HTER were manually evaluated looking for pronouns that were corrected from the MT version to the PE version. Figure 6 shows the total number of pronouns against the number of incorrect pronouns for the five documents. The number of incorrect pronouns is quite small compared to the total number of pronouns (proportionally, 23%, 10%, 16%, 33% and 34%, respectively in the five documents). This indicates that the number of pronouns showed a high correlation randomly. However, it could also be an indication that the presence of pronouns led to sentences that were more complicated and therefore more difficult to translate correctly (even if the pronouns themselves were correctly translated).

51 Connectives were analysed in terms of numbers of connectives in the MT and PE versions and also the number per class, considering the classification in (Pitler & Nenkova, 2009): *expansion*, *contingency*, *comparison*, *temporal* and *non-discourse*. As in the case of pronouns, connectives showed a positive correlation with HTER (0.4 Pearson's *r* and 0.33 Spearman's *rho*), but the *p-values* were also higher than 0.05. Figure 7 shows the results for connectives in the top five documents. As we can see, there is a change in the distribution of classes of connectives from the MT version to the PE version, i.e. the number of connectives in a given class changes from MT to PE. However, only document 4 showed significant changes. Therefore, it appears that the correlation between the number of connectives and HTER is by chance.

52 In the case of EDU's, the *p-values* for the Pearson's *r* and Spearman's *rho* correlation scores for the five documents with the highest HTER were below 0.05, meaning that the correlation is not by chance. Moreover, there is a change from the number of EDUs in the MT to the number of EDUs in the PE version. Therefore, we can infer that EDU breaks had an impact on the changes made to correct the documents, and thus on the MT quality of such documents.

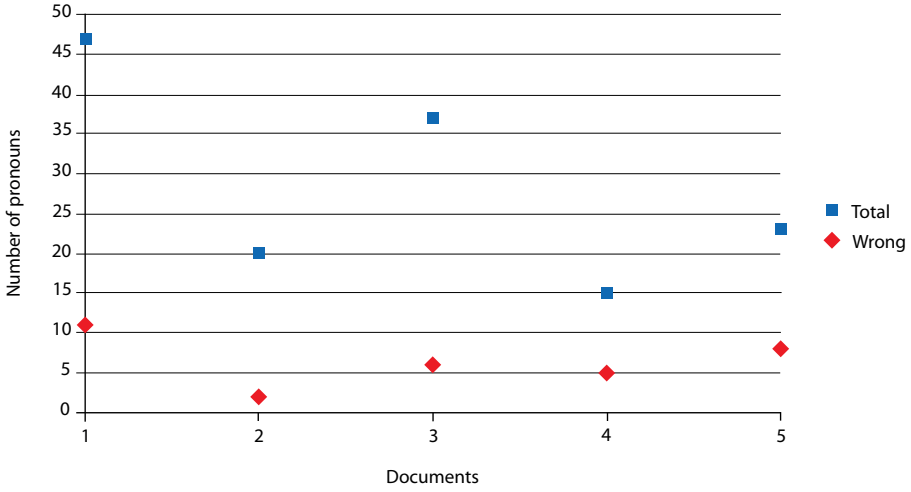

Figure 6. Total number of pronouns and number of incorrectly translated pronouns for the top five documents in the LIG corpus

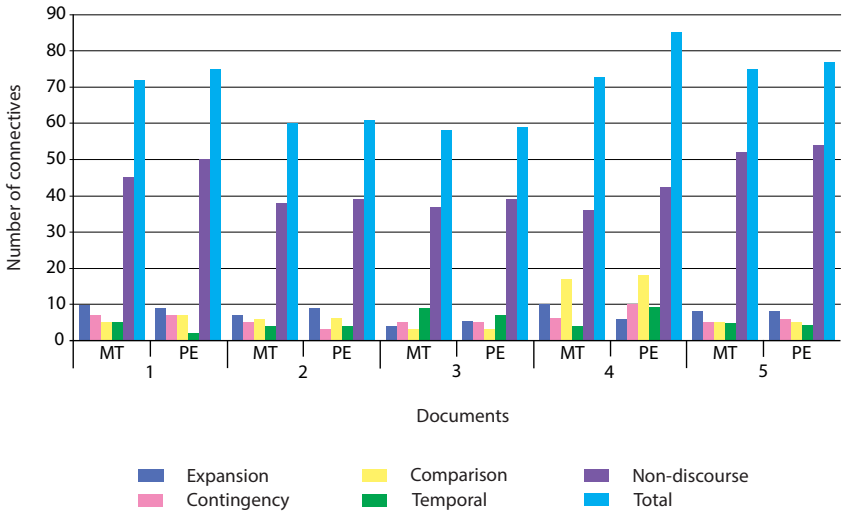

Figure 7. Number of connectives in the MT and PE versions of the top five documents in the LIG corpus

53 To avoid the bias of the top five documents, an additional analysis was done with 30 documents randomly selected from the 119 documents in the LIG corpus. We were interested in evaluating the impact of the same phenomena (number of pronouns, number of connectives and number of EDU breaks), but in a more general scenario. Figure 8 shows the percentage of incorrectly translated pronouns versus

HTER figures. Although the distribution of percentages of incorrectly translated pronouns is different from the HTER distribution, the correlation between number of pronouns and HTER was quite high: 0.45 for Pearson's *r* (*p-value* = 0.01) and 0.31 for Spearman's *rho* (*p-value* = 0.1). Therefore, we can conclude that there is a positive correlation between HTER scores and number of pronouns in this sample, and that it is not by chance.

54 For number of connectives, the correlation found was also high and significant: Pearson's *r* value of 0.52 (*p-value* = 0.0) and Spearman's *rho* value of 0.48 (*p-value* = 0.0), the same for the EDU breaks: the correlation found was 0.38 of Pearson's *r* (*p-value* = 0.04) and 0.44 of Spearman's *rho* (*p-value* = 0.01). This means that the correlation between HTER values and number of EDU breaks is also not by chance.

5. Conclusions

55 In this paper we have presented an analysis of several discourse phenomena in terms of their impact on MT quality. Discourse features were computed for English as source and target languages and their values were compared against HTER using correlation metrics. Results showed that discourse features achieve a high correlation with HTER (in terms of both Pearson's *r* and Spearman's *rho*) and that many discourse features reach a higher correlation than features from a strong baseline that uses sentence-level QE features averaged at the document level. Further analysis of the five worst documents showed that the correlation presented by two phenomena out of three analysed is likely to be by chance. However, an extended analysis of 30 documents randomly selected from the corpus showed that the correlations between HTER values and number of pronouns, connectives and EDU breaks are statistically significant.

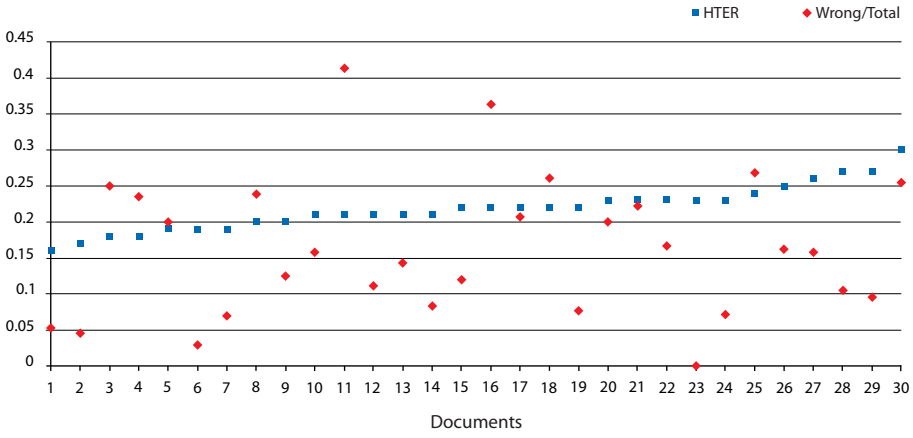

Figure 8. HTER values vs. percentage of incorrectly translated pronouns in a random sample of 30 documents from the LIG corpus

56 Future work includes more in-depth analyses to confirm our findings, in particular exploring more complex discourse phenomena such as other classes of discourse connectives, demonstrative and relative pronouns, anaphora resolution, and co-reference chains.

Acknowledgments

57 This work was supported by the EXPERT (EU Marie Curie ITN no. 317471) project.

References

Almaghout, H. & Specia, L. 2013. A CCG-Based Quality Estimation Metric for Statistical Machine Translation. In K. Sima'an et al. (eds.), *Proceedings of the 14th Machine Translation Summit (Nice, September 2-6, 2013)*. Allschwil: European Association for Machine Translation: 223-230. Available online: http://www.mt-archive.info/10/MTS-2013-Almaghout.pdf.

Avramidis, E. et al. 2011. Evaluate with Confidence Estimation: Machine Ranking of Translation Outputs using Grammatical Features. In *Proceedings of the 6th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 65-70. Available online: http://www.aclweb.org/anthology/W/W11/W11-2104.pdf.

Banerjee, S. & Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Stroudsburg: Association for Computational Linguistics: 65-72. Available online: http://aclweb.org/anthology-new/W/W05/W05-0909.pdf.

Ben, G. et al. 2013. Bilingual Lexical Cohesion Trigger Model for Document-level Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 382-386. Available online: http://www.aclweb.org/anthology/P/P13/P13-2068.pdf.

Blatz, J. et al. 2004. Confidence Estimation for Machine Translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 315-321. Available online: http://www.aclweb.org/anthology/C/C04/C04-1046.pdf.

Blei, D.M., Ng, A.Y. & Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning research* 3: 993-1022. Available online: https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf.

Bojar, O. et al. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 1-44. Available online: http://www.aclweb.org/anthology/W/W13/W13-2201.pdf.

Bojar, O. et al. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 12-58. Available online: http://www.aclweb.org/anthology/W/W14/W14-3302.pdf.

Callison-Burch, C. et al. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 10-51. Available online: http://www.aclweb.org/anthology/W/W12/W12-3102.pdf.

Carpuat, M. 2009. One Translation per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*. Stroudsburg: Association for Computational Linguistics: 19-27. Available online: http://www.aclweb.org/anthology/W/W09/W09-2404.pdf.

Carpuat, M. & Simard, M. 2012. The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 442-449. Available online: http://www.aclweb.org/anthology/W/W12/W12-3156.pdf.

Charniak, E. 2000. A Maximum-Entropy-Inspired Parser. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 132-139. Available online: http://www.aclweb.org/anthology/A/A00/A00-2018.pdf.

Eidelman, V., Boyd-Graber, J. & Resnik, P. 2012. Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 115-119. Available online: http://www.aclweb.org/anthology/P/P12/P12-2023.pdf.

Felice, M. & Specia, L. 2012. Linguistic Features for Quality Estimation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 96-103. Available online: http://www.aclweb.org/anthology/W/W12/W12-3110.pdf.

Giménez, J. et al. 2010. Document-Level Automatic MT Evaluation Based on Discourse Representations. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 333-338. Available online: http://www.aclweb.org/anthology/W/W10/W10-1750.pdf.

Giménez, J. & Màrquez, L. 2009. On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. In *Proceedings of the 4th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 250-258. Available online: http://www.aclweb.org/anthology/W/W09/W09-0440.pdf.

Graesser, A.C. et al. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, and Computers* 36 (2): 193-202.

Guillou, L. 2012. Improving Pronoun Translation for Statistical Machine Translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 1-10. Available online: http://aclweb.org/anthology-new/E/E12/E12-3001.pdf.

Guzmán, F. et al. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 687-698. Available online: http://www.aclweb.org/anthology/P/P14/P14-1065.pdf.

Hardmeier, C. 2011. Improving Machine Translation Quality Prediction with Syntactic Tree Kernels. In M.L. Forcada, H. Depraetere & V. Vandeghinste (eds.), *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011)*. Allschwil: European Association for Machine Translation: 233-240. Available online: http://www.ccl.kuleuven.be/EAMT2011/proceedings/pdf/eamt2011proceedings.pdf.

Hardmeier, C. 2014. *Discourse in Statistical Machine Translation*. PhD thesis. Uppsala University, Sweden.

Hardmeier, C. & Federico, M. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT)*. 283-289. Available online: http://uu.diva-portal.org/smash/get/diva2:420761/FULLTEXT01.

Hardmeier, C., Nivre, J. & Tiedemann, J. 2012. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 1179-1190. Available online: http://aclweb.org/anthology-new/D/D12/D12-1108.pdf.

He, Y. et al. 2010. Bridging SMT and TM with Translation Recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 622-630. Available online: http://www.aclweb.org/anthology/P/P10/P10-1064.pdf.

Joty, S. et al. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-Level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 486-496. Available online: http://www.aclweb.org/anthology/P/P13/P13-1048.pdf.

Joty, S. et al. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 402-408. Available online: http://www.aclweb.org/anthology/W/W14/W14-3352.pdf.

Kamp, H. 1981. A Theory of Truth and Semantic Representation. In J.A.G. Groenendijk, T.M.V. Janssen & M.B.J. Stokhof (eds.), *Formal Methods in the Study of Language*. Amsterdam: Mathematisch Centrum: 277-322.

Landauer, T.K., Foltz, P.W. & Laham, D. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25 (2-3): 259-284.

Le Nagard, R. & Koehn, P. 2010. Aiding Pronoun Translation with Co-reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Stroudsburg: Association for Computational Linguistics: 252-261. Available online: http://aclweb.org/anthology-new/W/W10/W10-1737.pdf.

Li, J.J., Carpuat, M. & Nenkova, A. 2014. Assessing the Discourse Factors that Influence the Quality of Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 283-288. Available online: http://www.aclweb.org/anthology/P/P14/P14-2047.pdf.

Mann, W.C. & Thompson, S.A. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*. Technical report ISI/RS-87-190. Los Angeles: Information Sciences Institute. Available online: http://www.sfu.ca/rst/05bibliographies/bibs/ISI_RS_87_190.pdf.

Marcu, D., Carlson, L. & Watanabe, M. 2000. The Automatic Translation of Discourse Structures. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 9-17. Available online: http://www.aclweb.org/anthology/A/A00/A00-2002.pdf.

Meyer, T. et al. 2012. Machine Translation of Labeled Discourse Connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*. Association for Machine Translation in the Americas. Available online: http://amta2012.amtaweb.org/AMTA2012Files/papers/119.pdf.

Meyer, T. & Popescu-Belis, A. 2012. Using Sense-Labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*. Stroudsburg: Association for Computational Linguistics: 129-138. Available online: http://aclweb.org/anthology-new/W/W12/W12-0117.pdf.

Papineni, K. et al. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 311-318. Available online: http://aclweb.org/anthology-new/P/P02/P02-1040.pdf.

Pighin, D. & Màrquez, L. 2011. Automatic Projection of Semantic Structures: An Application to Pairwise Translation Ranking. In *Proceedings of 5th Workshop on Syntax, Semantics and Structure in Statistical Translation*. Stroudsburg: Association for Computational Linguistics: 1-9. Available online: http://www.aclweb.org/anthology/W/W11/W11-1001.pdf.

Pitler, E. & Nenkova, A. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Stroudsburg: Association for Computational Linguistics: 13-16. Available online: http://www.aclweb.org/anthology/P/P09/P09-2004.pdf.

Popescu-Belis, A. et al. 2012. Discourse-Level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 2716-2720. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/255_Paper.pdf.

Potet, M. et al. 2012. Collection of a Large Database of French-English SMT Output Corrections. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*. Stroudsburg: Association for Computational Linguistics: 4043-4048. Available online: http://www.lrec-conf.org/proceedings/lrec2012/pdf/506_Paper.pdf.

Prasad, R. et al. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*. Stroudsburg: Association for Computational Linguistics: 2961-2968. Available online: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

Rubino, R. et al. 2013. Topic Models for Translation Quality Estimation for Gisting Purposes. In K. Sima'an et al. (eds.), *Proceedings of the 14th Machine Translation Summit (Nice, September 2-6, 2013)*. Allschwil: European Association for Machine Translation: 295-302. Available online: http://www.mt-archive.info/10/MTS-2013-Rubino.pdf.

Scarton, C. et al. 2015. Searching for Context: A Study on Document-Level Labels for Translation Quality Estimation. In *18th Annual Conference of the European Association for Machine Translation (EAMT 2015)* (11-13 May 2015, Antalya, Turkey). Available online: http://www.uni-koeln.de/~mzampier/papers/scartoneamt2015.pdf.

Scarton, C. & Specia, L. 2014. Document-Level Translation Quality Estimation: Exploring Discourse and Pseudo-references. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 2014)*. Allschwil: European Association for Machine Translation: 101-108. Available online: http://www.mt-archive.info/10/EAMT-2014-Scarton.pdf.

Shah, K., Cohn, T. & Specia, L. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In K. Sima'an et al. (eds.), *Proceedings of the 14th Machine Translation Summit (Nice, September 2-6, 2013)*. Allschwil: European Association for Machine Translation: 167-174. Available online: http://www.mt-archive.info/10/MTS-2013-Shah.pdf.

Snover, M. et al. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*. Association for Machine Translation in the Americas: 223-231. Available online: http://www.mt-archive.info/AMTA-2006-Snover.pdf.

Soricut, R. & Echihabi, A. 2010. TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 612-621. Available online: http://www.aclweb.org/anthology/P/P10/P10-1063.pdf.

Soricut, R. & Marcu, D. 2003. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 149-156. Available online: http://www.aclweb.org/anthology/N/N03/N03-1030.pdf.

Soricut, R. & Narsale, S. 2012. Combining Quality Prediction and System Selection for Improved Automatic Translation Output. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Stroudsburg: Association for Computational Linguistics: 163-170. Available online: http://www.aclweb.org/anthology/W/W12/W12-3121.pdf.

Specia, L., Cancedda, N. & Dymetman, M. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*. Stroudsburg: Association for Computational Linguistics: 3375-3378. Available online: http://www.lrec-conf.org/proceedings/lrec2010/pdf/504_Paper.pdf.

Specia, L. et al. 2009a. Estimating the Sentence-Level Quality of Machine Translation Systems. In L. Màrquez and H. Somers (eds.), *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*. Allschwil: European Association for Machine Translation: 28-35. Available online: http://www.mt-archive.info/05/EAMT-2009-Specia.pdf.

Specia, L. et al. 2009b. Improving the Confidence of Machine Translation Quality Estimates. In *Proceedings of the 12th Machine Translation Summit (26-30 August 2009, Ottawa, Canada)*. Association for Machine Translation in the Americas. Available online: http://www.mt-archive.info/05/MTS-2009-Specia.pdf.

Specia, L. et al. 2013. QuEst – A Translation Quality Estimation Framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 79-84. Available online: http://www.aclweb.org/anthology/P/P13/P13-4014.pdf.

Specia, L. & Farzindar, A. 2010. Estimating Machine Translation Post-editing Effort with HTER. In V. Zhechev (ed.), *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC 2010)*. 33-41. Available online: http://www.mt-archive.info/10/JEC-2010-Specia.pdf.

Specia, L. et al. 2011. Predicting Machine Translation Adequacy. In *Proceedings of the 13th Machine Translation Summit (19-23 September 2011, Xiamen, China)*. Tokyo: Asia-Pacific Association for Machine Translation: 513-520. Available online: http://www.mt-archive.info/MTS-2011-Specia.pdf.

Specia, L., Paetzold, G.H. & Scarton, C. 2015. Multi-level Translation Quality Prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations* (26-31 July 2015, Beijing, China). Stroudsburg: Association for Computational Linguistics: 115-120. Available online: http://www.aclweb.org/anthology/P/P15/P15-4020.pdf.

Ture, F., Oard, D.W. & Resnik, P. 2012. Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 417-426. Available online: http://www.aclweb.org/anthology/N/N12/N12-1046.pdf.

Wisniewski, G. et al. 2013. Design and Analysis of a Large Corpus of Post-edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-edition. In K. Sima'an et al. (eds.), *Proceedings of the 14th Machine Translation Summit (Nice, September 2-6, 2013)*. Allschwil: European Association for Machine Translation: 117-124. Available online: http://www.mt-archive.info/10/MTS-2013-Wisniewski.pdf.

Wong, B.T.M. & Kit, C. 2012. Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics: 1060-1068. Available online: http://aclweb.org/anthology-new/D/D12/D12-1097.pdf.

Xiao, T. et al. 2011. Document-Level Consistency Verification in Machine Translation. In *Proceedings of the 13th Machine Translation Summit (19-23 September 2011, Xiamen, China)*. Tokyo: Asia-Pacific Association for Machine Translation: 131-138. Available online: http://www.mt-archive.info/10/MTS-2011-Xiao.pdf.

Zhao, B. & Xing, E.P. 2007. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. *Advances in Neural Information Processing Systems* 20: 1-8. Available online: http://papers.nips.cc/paper/3365-hm-bitam-bilingual-topic-exploration-word-alignment-and-translation.pdf.

Zhengxian, G., Yu, Z. & Guodong, Z. 2010. Statistical Machine Translation Based on LDA. In *Proceedings of the 4th International Universal Communication Symposium (IUCS 2010)*. Institute of Electrical and Electronics Engineers: 286-290.