Comparabilty of Corpora in Human and Machine Translation

Ekaterina Lapshinova-Koltunski & Santanu Pal

Saarland University
Universität Campus A2.2,
66123 Saarbrücken, Germany
e.lapshinova@mx.uni-saarland.de, santanu.pal@uni-saarland.de

Abstract

In this study, we demonstrate a negative result from a work on comparable corpora which forces us to address a problem of comparability in both human and machine translation. We state that it is not always defined similarly, and comparable corpora used in contrastive linguistics or human translation analysis cannot always be applied for statistical machine translation (SMT). So, we revise the definition of comparability and show that some notions from translatology, i.e. registerial features, should also be considered in machine translation (MT).

Keywords: comparable corpora, paraphrases, machine translation, register analysis, registerial features

1. Introduction

Numerous studies and applications in both linguistic and language engineering communities use comparable corpora as essential resources, e.g. to compare phenomena across languages or to acquire parallel resources for training in statistical Natural Language Processing (NLP) applications, e.g. statistical machine translation.

Due to the fact that parallel corpora remain a scarce resource (despite the creation of automated methods to collect them from the Web) and often cover restricted domains only (political speeches, legal texts, news, etc.), comparable corpora have been used as a valuable source of parallel components in SMT, e.g. as a source for parallel fragment of texts, paraphrases or sentences (Smith et al., 2010).

In contrast to parallel corpora, which contain originals and their translations, comparable corpora can contain originals only, or translations only, and can thus be defined as a collection of texts with the same sampling frame and similar representativeness (McEnery, 2003). For example, they may contain the same proportions of the texts belonging to the same genres, or the same domains in a range of different languages.

However, the concept of 'comparable corpora' may differ depending on which measure is taken into account (register or domain), and what are the purposes of the analysis. In this paper, we present an experiment which demonstrates that comparability in human translation studies does not always coincide with what is understood under comparability in machine translation.

The remainder of the paper is structured as follows. In section 2., we outline the aims and the morivation of the present study. Section 3. presents related work on comparable corpora, the clarification of the notions of domain and register, as well as their definition applied in this work. Section 4. describes the resources at hand and the applied methodology. Here, we describe the resources at hand, and the methods used. In section 5., we show the results, and discuss the problems we face.

2. Aims and Motivation

The original aim of our experiment was to enhance the resources available for machine translation with the help of a paraphrase extraction from both parallel and comparable corpora at hand. The extracted paraphrases can then be used to improve statistical machine translation, as it was done in our previous studies. For example, in (Pal et al., 2013), multi-word expressions (MWE) were extracted from comparable corpora aligned on document level. These were aligned and used for the improvement in English-Bengali Phrase-Based SMT (PB-SMT) by incorporating them directly and indirectly into the phrase table. In another study, n-gram overlapping parallel fragment of texts were extracted from comparable corpora to serve as an additional resource to improve a baseline PB-SMT system, see (Gupta et al., 2013). Another possible application of such paraphrases is acquisition of parallel and comparable data from the web, which can also be used for MT enhancement.

For this experiment, we decide for English-German resources consiting of two parts: a baseline created for a PB-SMT system, and an existing comparable corpus, which was originally compiled to serve human translation tasks. Hence, comparability of its texts was stated according to criteria used in translatology, see sections 3.2. and 4.1. below.

The texts of the corpus belong to two genres – political speeches and popular science. The choice of these datasets for our experiment is motivated by the difference in the availability of resources. Whereas extensive parallel resources are available for political speeches, it is difficult to find parallel resources for popular-scientific texts. Therefore, we decide to apply procedures for both datasets, as on the one hand, we hope to enhance the resources available (improving machine translation with paraphrases), and on the other hand, we want to test how our procedures work on a dataset different to what is commonly used, e.g. news articles or political speeches.

Moreover, these two datasets are different not only in the amount of parallel resources available. They also differ in the correlation of the notions of domain vs. genre/register. In political speeches, the notion of domain correlates more with that of register, whereas in popular scientific texts, it doesn't. Therefore, we observet different results in the application of our procedures, which make us address the problem of corpus comparability in translation.

3. Related Work and Theoretical Issues

3.1. Comparable corpora

Comparable corpora in MT As already mentioned above, comparable corpora have become widely used in NLP, contrastive language analysis and translatology. In NLP, they found application in the development of bilingual lexicons or terminology databases, e.g. in (Chiao and Zweigenbaum, 2002; Fung and Cheung, 2004) or (Gaussier et al., 2004) and in cross-language information research, see e.g. (Grefenstette, 1998) or (Chen and Nie, 2000), as well as MT improvement, e.g. (Munteanu and Marcu, 2005) or (Eisele and Xu, 2010).

The methods used in these approaches are mostly based on context similarity: the same concept tends to appear with the same context words in both languages, the hypothesis that is also used for the identification of synonyms. Several earlier studies have shown that there is a correlation between the co-occurrences of words which are translations of each other in any language (Rapp, 1999) and that the associations between a word and its context seed words are preserved in comparable texts of different languages, cf. (Fung and Yee, 1998).

In most cases, the starting point is a list of bilingual "seed expressions" required to build context vectors of all words in both languages. This is either provided by an external bilingual dictionaries or databases, as in (Déjean et al., 2002), or is extracted from a parallel corpus, as in (Otero, 2007). We also start with a list of "seed expressions", which are paraphrases in our case. They are extracted from a bilingual parallel corpus, and enhanced with parapharses from a comparable corpus.

There are similar works with the application for automatic extraction of terms, e.g. in (Chiao and Zweigenbaum, 2002) and (Saralegi et al., 2008). The authors used specialised comparable corpora, e.g. English-French corpora in medical domain, or English-Basque corpora in popular science, for automatic extraction of bilingual terms. In both cases, comparability is accounted for by the distribution of topics (or also publication dates).

Comparable corpora and comparability In most works, comparability is correlated with the comparability of potential word equivalents and their contexts or collocates, which is reasonable for bilingual terminology extraction task. Although these criteria might be sufficient for creation of multilingual lexicons or terminology databases, translation of whole texts involve more influencing factors, as more levels of description, i.e. conventions of a register a text belongs to are at play. In translation studies, which are concerned with human translations, as well as human translator training, these aspects take on an important role. While translating a text from one language into another, a translator must consider the conventions of the text type to be translated.

In existing MT studies these conventions (specific register features) have not been taken into account so far. Describing comparable data collected for training, authors consider solely domains, i.e. topics described in the collected texts, ignoring the genre or the register of these texts. We claim that register features should also be considered in the definition of a comparable corpus in MT, as they are in human translation.

In the following, we define the notions of genre, register and domain, as well as their role in the definition of comparability in our analysis.

3.2. Genre, Register and Domain

We consider multilingual corpora comparable if they contain texts which belong to the same register.

In our analysis, we use the term *register*, and not *genre*, although they represent two different points of view covering the same ground, see e.g. (Lee, 2001). However, we refer to genre when speaking about a text as a member of a cultural category, about a register when we view a text as language, its lexico-grammatical characterisations, conventionalisation and functional configuration of a language which are determined by a context use situation, variety of language means according to this situation. Different situations require different configurations of a language.

This kind of register definition is used in human translation studies, e.g. corpus-based approaches as in (Teich, 2003; Steiner, 2004; Hansen Schirra et al., 2013; Neumann, 2013), and coincides with the one formulated in register theory, e.g. in (Quirk et al., 1985; Halliday and Hasan, 1989; Biber, 1995). In their terms, registers are manifested linguistically by particular distributions of lexico-grammatical patterns, which are situation-dependent. The canonical view is that situations can be characterised by the parameters of *field*, *tenor* and *mode* of discourse. Field of discourse relates to processes and participants (e.g., Actor, Goal, Medium), as well as circumstantials (Time, Place, Manner etc.) and is realised in lexico-grammar in lexis and colligation (e.g. argument structure). Tenor of discourse relates to roles and attitudes of participants, author-reader relationship, which are reflected in stance expressions or modality. Mode of discourse relates to the role of the language in the interaction and is linguistically reflected at the grammatical level in Theme-Rheme constellations, as well as cohesive relations at the textual level. So, the contextual parameters of registers correspond to sets of specific lexico-grammatical features, and different registers vary in the distribution of these features.

The definition of domain is also present in register analysis. Here, it is referred to as *experiential domain*, or what a text is about, its topic. Experiential domain is a part of the context parameter of field, which is realised in lexis, as already mentioned above. However, it also includes colligation, in which also grammatical categories are involved. So, domain is just one of the parameter features a register can have. Some NLP studies, e.g. those using web resources, do claim the importance of register or genre conventions, see e.g. (Santini et al., 2010). However, to our knowledge, register or genre features remain out of the focus in machine translation. Whereas there exist some works on domain adaptation, e.g. adding bilingual data to the training material of SMT systems, as in (Eck et al., 2004), or (Wu et al., 2008) and others, register features are mostly ignored.

In human translator training, on the contrary, the knowledge on lexico-grammatical preferences of registers plays an important role. A human translator learns to analyse

texts according to the register parameters both in a source and in a target language.

4. Resources and Methodology

4.1. Resources at hand

In our experiment, we use two types of dataset: (1) a big English-German parallel training corpus; (2) a small English-German comparable corpus. The first one is based on the English-German component of EUROPARL[1] (Koehn, 2005), used to build the baseline system and to create the initial paraphrase table, see section 4.3. below. The other dataset (2) is used for the enhancement of this paraphrase table. This dataset was extracted from the multilingual corpus CroCo (Hansen Schirra et al., 2013), which contains English and German texts, belonging to the same register. As already mentioned above, we decide for the registers of political speeches (SPEECH) and popular science (POPSCI), see section 1.

Data selection The texts in the corpus are selected according to the criteria of register analysis as defined in 3.2. above. According to the general register analysis, SPEECH belongs to the communication of an 'expert to expert' in a formal social distance, whereas the latter is rather 'expert to layperson' in a causal social distance. Both express an equal social and a constitutive language role. For popular-scientific texts in both languages, it is essential that texts are perceived as pleasurable, and not only informative reading. This means that author-reader relationship (the contextual parameter of tenor) is very important in this register, see (Kranich et al., 2012).

English originals (EO) in SPEECH are collected from the US public diplomacy and embassy web services, whereas German texts (GO) originate from German governmental, ministery and president websites. Both EO and GO texts have 'exposition', 'persuasion' and 'argumentation' as goal orientation, 'expert to expert' as agentive role, and include information on economic development, human security and other issues in both internal, foreign or global perspective.

Both EO and GO texts in POPSCI originate from popular-scientific articles, which have 'exposition' as goal orientation, 'expert to layperson' as agentive role. The information in the articles are on psychotherapy, biology, chemistry and others.

Although no attention was paid to the parallelity of topics discussed in both corpora (which could mean that their domains do not necessarily coincide), English and German registers are comparable along other features. Moreover, they have a number of commonalities in English and German. For example, popular-scientific texts show preference for particular process types, e.g. relation processes (expressed by transitivity), underspecified Agent (expressed by extensive use of passive constructions), and others in both languages (Teich, 2003).

Data processing We used Stanford Parser, see (Socher et al., 2013; Rafferty and Manning, 2008), and Stanford NER[2] for parsing and named entity tagging for the EO and GO texts. The experiments were carried out with the help of the standard log-linear PB-SMT model as baseline: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics as described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model trained with the SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007).

4.2. Paraphrase extraction

We start our experiment with the identification of paraphrases from the English-German parallel training corpus, (1) in section 4.1. above.

Paraphrase is a phrase or an idea that can be represented or expressed in different ways in the same language by preserving the meaning of that phrase or idea. Paraphrases can be collected from parallel corpora as well as from comparable corpora. Extraction of parallel fragments of texts, sentences and paraphrases from comparable corpora is particularly useful for any corpus-based approaches to MT, especially for SMT (Gupta et al., 2013). Paraphrases can be used to alleviate the sparseness of training data (Callison-Burch et al., 2006), to handle Out Of Vocabulary (OOV) words, as well as to expand the reference translations in automatic MT evaluation (Denoual and Lepage, 2005; Kauchak and Barzilay, 2006). Moreover, in SMT, the size of the parallel corpus plays a crucial role in the SMT performance. However, large volume of parallel data is not available for all language pairs or all text types (see section 1.).

A significant number of works have been carried out on paraphrasing. A full-sentence paraphrasing technique was introduced by (Madnani et al., 2007). They demonstrated that the resulting paraphrases can be used to drastically reduce the number of human reference translations needed for parameter tuning without a significant decrease in translation quality. (Fujita and Carpuat, 2013) describe a system that was built using baseline PB-SMT system. They augmented the phrase table with novel translation pairs generated by combining paraphrases where these translation pairs were learned directly from the bilingual training data. They investigated two methods for phrase table augmentation: source-side augmentation and target-side augmentation. (Aziz and Specia, 2013) report the mining of sense-disambiguated paraphrases by pivoting through multiple languages. (Barzilay and McKeown, 2001) proposed an unsupervised learning algorithm for identification of paraphrases from a corpus of multiple English translations of the same source text. A new and unique paraphrase resource was reported by (Xu et al., 2013), which contains meaning-preserving transformations between informal user-generated texts. Sentential paraphrases are extracted from a comparable corpus of (temporally and topically related) messages in Twitter which often express semantically identical information through distinct surface forms. A novel paraphrase fragment pair extraction method was proposed by (Wang and Callison-Burch, 2011) in which the authors used a monolingual comparable corpus containing different articles about the same topics or events.

[1] the 7th Release v7 of EUROPARL

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

The procedure consisted of document, sentence and fragment pair extraction.
Our approach is similar to the identification technique used by (Bannard and Callison-Burch, 2005). In our study, identification of paraphrases has been carried out by pivoting through phrases from the bilingual parallel corpus (1). We consider all phrases in the phrase table as potential candidates for paraphrasing.
After extraction of potential paraphrase pairs, we compute the likelihood of them being paraphrases. For a potential paraphrase pair (e1, e2) we have defined a paraphrase probability p(e2|e1) in terms of the translation model probabilities p(f|e1), that the original English phrase e1 is translated as a particular target language phrase f, and p(e2|f), that the candidate paraphrase e2 is translated as the same foreign language phrase f. Since e1 can be translated to multiple foreign language phrases, we sum over all such foreign language phrases. Thus the equation reduces to as follows:

$$\hat{e}2 = \underset{e_2 \neq e_1}{\arg\max} P(e_2|e_1) \quad (1)$$

$$= \underset{e_2 \neq e_1}{\arg\max} \sum_f P(f|e_1)P(e_2|f) \quad (2)$$

We compute translation model probabilities using standard formulation from PB-SMT. So, the probability p(e|f) is calculated by counting how often the phrases e and f were aligned in the parallel corpus as follows :

$$p(e|f) = \frac{count(e,f)}{\sum_f count(e,f)} \quad (3)$$

Using the equation (2) and (3) we calculate paraphrase probabilities from the phrase table.

4.3. Incorporation of paraphrases into PB-SMT System

The next step is to create additional training material using these extracted paraphrases. We initially found and marked the paraphrases in the source English sentences within the training data and then replaced each English paraphrase with all of its other variants, gradually creating more training instances. For example, consider the English phrase "throughout the year" and its two paraphrases "all year round" and "all around the year". Now we consider following sentences from our training data for each of these phrase and paraphrases.

(1) a. Events, parties and festivals occur throughout the year and across the country.
b. Weather on all of the Hawaiian islands is very consistent, with only moderate changes in temperature all year round.
c. There is an intense agenda all around the year and the city itself is a collection of art and history.

In example (1), the first sentence, the phrase "throughout the year" is replaced by its two paraphrases "all year round" and "all around the year" to create two additional sentences to be added to the existing training data. Similarly "all year round" and "all around the year" are replaced by the remaining two variants for the second and third sentence, respectively.
In this way, for these three training sentences, we can create six additional sentences from all combinations of replacement. Combining these additional resources with the existing training data, we enhance the existing baseline of the PB-SMT system.
We decode English original (EO) sentences from both SPEECH and POPSCI through our enhanced English-German PB-SMT system. The density of population of words for GO with respect to EO are measured through the decoded output provided by the enhanced system. The population measure is defined as how many translated German word words are corresponding to the GO words by measuring distance between them. For this, we use the following distance measure techniques: *Minimum Edit Distance Ratio* (MEDR) and *Longest Common Subsequence Ratio* (LCSR). Let, |W| be the length of the string W and ED is the minimum edit distance or levenshtein distance calculated as the minimum number of edit operations such as insert, replace, delete – needed to transform W1 into W2.
The definition of the Minimum Edit Distance Ratio is given in (4), and the definition of Longest Common Subsequence Ratio in (5).

$$MEDR(W1,W2) = 1 - \frac{|ED(W1,W2)}{max(|W1|,|W2|)} \quad (4)$$

$$LCSR(W1,W2) = \frac{|LCS(W1,W2)}{max(|W1|,|W2|)} \quad (5)$$

The training corpus was filtered with the maximum allowable sentence length of 100 words and sentence length ratio of 1:2 (either way). In the end, the training corpus contained 1.902,223 sentences. In addition to the target, side monolingual German corpus containing 2.176,537 sentences from EUROPARL was used for building the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length and found that a 5-gram language model and a maximum phrase length of 7 produced the optimum baseline result.
This baseline is now to be enhanced with additional paraphrases from comparable corpora at hand, which we describe in the following section.

4.4. Analysis of comparable corpora

To expand the paraphrase table, we first perform manual comparison of each corresponding comparable file in terms of token and part-of-speech (POS) alignment.
Then, we analyse density with the help of named entities (NE). Named entitites are identified on both EO and GO sentences separately with the help of English and German Stanford NER. So, using NEs we prove the comparability between the comparable parts of the corpus, i.e we check whether NEs are present on both its sides (English and German). We follow the same word similarity technique: MEDR and LCSR, as described in section 4.3. above. The comparability has been measured according to the population density (how many NEs correspond between the EO and GO) on both side of the comparable corpus.

5. Experiment Results

5.1. Comparison results

In tables 1 and 2, we present the results of the comparison for texts from the analysed corpus, including the total number of tokens (token) and NEs, as well as their population (pop) and population density (pop.dens) calculated as populated tokens/AVG (the sum of total EOand GO tokens), see section 4.3. for details.

	EO	GO	pop	pop.dens
token	13906	14598	5729	0.40
NE	369	263	8	0.02

Table 1: Similarities between EO and GO in POPSCI

	EO	GO	pop	pop.dens
token	9753	7094	3969	0.47
NE	387	297	149	0.43

Table 2: Similarities between EO and GO in SPEECH

Our results show that token alignment in SPEECH is much more reliable than that in POPSCI. The same results are obtained on the POS level: the total number of nouns are more probably matching between the comparable files in SPEECH. Moreover, we found more population density in the SPEECH data, if compared with the data in POPSCI.
This means that whereas we can prove the comparability of EO and GO in SPEECH using these measuring techniques, we are not able to do the same for POPSCI. Hence, we cannot extract paraphrases from the comparable corpus of POPSCI texts at hand. This shows that our method of paraphrase enhancement with the data from comparable corpora does not work with all types comparable corpora.
The reason for it is the nature of the comparable data. On the one hand, English and German texts are comparable in POPSCI if register settings in both languages are considered. On the other hand, they are not necessarily comparable in their domains. At the same time, SPEECH, which was also set up under same conditions of register analysis, seem to be comparable in both aspects. We assume that the notion of domain in SPEECH correlates with that of register, whereas in popular science it doesn't.

5.2. Discussion

Facing the negative results of our experiment, we decide to revise the notion of comparabilty, which does not always correspond in machine translation and in human translation. Defining comparability criteria for corpora, these scientific communities have often two different things in mind: (1) register in human translation (register-oriented perspective), (2) domain in machine translation (domain-oriented perspective). We assume that the relation between these two perspectives is inclusive: domain definition is implied in the register analysis as a part of 'experiential domain definition'. This is confirmed by the results of our experiment which demonstrates that in some cases, the definition of domain and register coincide. For instance, in political speeches, experiential domain is not that diverse as in popular-scientific texts, and thus, the texts identified as comparable according to the register-oriented perspective, are also comparable in terms of the domain-oriented perspective.
At the same time, if we define corpora as being comparable along the domain-oriented criterion only, they would not necessarily be comparable from the register-oriented perspective. For instance, for human translation, news reporting on certain political topics cannot be comparable with political speeches discussing the same topics as in the news texts. The latter would lack 'persuasion' and 'argumentation' in their as goal orientation, as well as 'expert to expert' as agentive role, which would be reflected in their lexico-grammatical features.
We believe that both perspectives are important for translation (both human and machine). The first one has an impact on the lexical level, e.g. terminology or general vocabulary used in a translated text. The other is important for lexico-grammar, i.e. morpho-syntactic preferences of registers and their textual properties, e.g. cohesive phenomena and information structure. Therefore, we claim that there is a need to define new measures of corpus comparability in translation, which can be measured e.g. by homegeneity[3], and would consider both domain and further registerial features.
In MT studies this problem has not been addressed so far. To our knowledgde, none of the existing MT studies integrate register features. As a result, machine-translated texts would (not) have features characteristic for the register they belong to. For example, German popular-scientific texts can be characterised by a high number of passive constructions, see section 3.2. above. We calculate the ratio of passive constructions[4] in German originals and compare it to the passive ratio in German translations from English, considering human (HU) and a statistical machine translation (SMT)[5]. Whereas human translations demonstrate a similar proportion of passives as in comparable originals, machine translations seem to underuse this verb construction type.

corpus	ratio
GO	6.62
HU	6.98
SMT	3.10

Table 3: Passive verb constructions in POPSCI

Undoubtably, we need to test more features to come to the final conclusion about the impact of registerial features on the translation output. However, it was not the original aim of the present paper. Moreover, we need to expand the parallel training corpus with additional genre to show possible differences in the resulting models. For future work, we also plan to experiment with another approach on MT enhancement, e.g. the one described in (Munteanu and Marcu, 2005).
However, the negative reults of our experiments made us raise the questions about (1) comparability, and (2) ad-

[3] see work on homogeneity measure by (Kilgarriff, 2001).

[4] We calculate the ratio of passives in all final verb constructions.

[5] the translations are available in VARTRA, see (Lapshinova-Koltunski, 2013).

ditional features which could have impact on translation, which we address to both communities and aim to raise a discussion in these issues.

Acknowledgments

The research leading to these results has received funding from the EU project EXPERT – the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013 under REA grant agreement no. [317471]. The resources available were provided within the project VARTRA supported by a grant from Forschungsausschuß of Saarland University.

6. References

W. Aziz and L. Specia. 2013. Multilingual WSD-like Constraints for Paraphrase Extraction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 202–211, Sofia, Bulgaria.

C. Bannard and C. Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL-2005*, pages 597–604.

R. Barzilay and K.R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of ACL-2001*, pages 50–57.

D. Biber. 1995. *Dimensions of Register Variation. A Cross-linguistic Comparison*. Cambridge University Press, Cambridge.

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Main Conference on HLT-NAACL-2006*, pages 17–24.

J. Chen and J-Y. Nie. 2000. Parallel web text mining for cross-language Ir. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, volume 1, pages 62–78, Paris.

Y. Chiao and P. Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 2*, COLING-02, pages 1–5.

H. Déjean, É. Gaussier, and F. Sadat. 2002. Bilingual Terminology Extraction: An Approach Based on a Multilingual Thesaurus Applicable Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, COLING-02.

E. Denoual and Y. Lepage. 2005. Bleu in characters: towards automatic Mt evaluation in languages without word delimiters. In *The Second International Joint Conference on Natural Language Processing*, pages 81–86.

M. Eck, S. Vogel, and A. Waibel. 2004. Improving statistical machine translation in the medical domain using the unified medical language system. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 792–798, Geneva, Switzerland.

A. Eisele and J. Xu. 2010. Improving Machine Translation Performance Using Comparable Corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 35–41, Malta. LREC-2010.

A. Fujita and M. Carpuat. 2013. Fun-nrc: Paraphrase-augmented Phrase-based Smt Systems for Ntcir-10 Patentmt. In *The 10th NTCIR Conference*, Tokyo, Japan.

P. Fung and P. Cheung. 2004. Mining Verynon-parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and Em. In *Proceedings of EMNLP*, pages 57–63.

P. Fung and L.Y. Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume volume 1 of *COLING-98*, pages 414–420.

E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Djean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL-04*, pages 527–534.

G. Grefenstette. 1998. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, London.

R. Gupta, S. Pal, and S. Bandyopadhyay. 2013. Improving Mt System Using Extracted Parallel Fragments of Text from Comparable Corpora. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora*, pages 69–76, Sofia, Bulgaria.

MAK Halliday and R. Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford University Press.

S. Hansen Schirra, S. Neumann, and E. Steiner. 2013. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

D. Kauchak and R. Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Main Conference on HLT-NAACL-2006*, pages 455–462.

A. Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL-2003*, volume 1, pages 48–54.

P. Koehn, H. Hoang, A. Birch, C. Callison Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL-2007*, pages 177–180.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

S. Kranich, J. House, and V. Becher. 2012. Changing conventions in English-german translations of popular scientific texts. In Kurt Braunmüller and Christoph Gabriel, editors, *Multilingual Individuals and Multilingual Societies*, volume 13 of *Hamburg Studies on Multilingualism*, pages 315–334. John Benjamins.

E. Lapshinova-Koltunski. 2013. VARTRA: A Compara-

ble Corpus for Analysis of Translation Variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria. Association for Computational Linguistics.

D. Y. Lee. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the bnc jungle. *Technology*, 5:37–72.

N. Madnani, N.F. Ayan, P. Resnik, and B.J. Dorr. 2007. Using Paraphrases for Parameter Tuning in Statistical Machine Translation. In *Proceedings of the Second Workshop on StatMT*, pages 120–127.

T. McEnery. 2003. Corpus Linguistics. In Ruslan Mitkov, editor, *Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, pages 448–463. Oxford University Press, Oxford.

D. S. Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

S. Neumann. 2013. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Trends in Linguistics. Studies and Monographs [Tilsm]. Walter de Gruyter.

P. G. Otero. 2007. Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In *Proceedings of MT Summit XI*, pages 191–198.

S. Pal, S. K. Naskar, and S. Bandyopadhyay. 2013. MWE Alignment in Phrase Based Statistical Machine Translation. In *Proceedings of the Machine Translation Summit XIV*, pages 61–68, Nice, France.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Anna Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *ACL Workshop on Parsing German*.

R. Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th ACL*.

M. Santini, A. Mehler, and S. Sharoff. 2010. Riding the rough waves of genre on the web. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–30. Springer.

X. Saralegi, I. S. Vicente, and A. Gurrutxaga. 2008. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of the 1rd Workshop on Building and Using Comparable Corpora*, Marrakesh. LREC-2008.

J. R. Smith, C. Quirk, and K. Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10*, pages 403–411.

R. Socher, J. Bauer, C.D. Manning, and A.Y. Ng. 2013. Parsing With Compositional Vector Grammars. In *Proceedings of ACL-2013*, Sofia, Bulgaria.

E. Steiner. 2004. *Translated texts: Properties, Variants, Evaluations*. Peter Lang, Frankfurt a. Main.

A. Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

E. Teich. 2003. *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin and New York.

R. Wang and C. Callison-Burch. 2011. Paraphrase Fragment Extraction from Monolingual Comparable Corpora. In *4th Workshop on Building and Using Comparable Corpora*, Portland, Oregon.

H. Wu, H. Wang, and C. Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In D. Scott and H. Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pages 993–1000, Manchester, UK.

W. Xu, A. Ritter, and R. Grishman. 2013. Gathering and Generating Paraphrases from Twitter with Application to Normalization. In *7th Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria.