The Effect of Translationese on Statistical Machine Translation

Gennadi Lembersky

A THESIS SUBMITTED FOR THE DEGREE
"DOCTOR OF PHILOSOPHY"

University of Haifa

Faculty of Social Sciences
Department of Computer Sciences

May, 2013

Winner of the 2013 Best Thesis Award
of the European Association for Machine Translation

The Effect of Translationese on Statistical Machine Translation

By: Gennadi Lembersky

Supervised By: Prof. Shuly Wintner

A THESIS SUBMITTED FOR THE DEGREE
"DOCTOR OF PHILOSOPHY"

University of Haifa

Faculty of Social Sciences
Department of Computer Sciences

May, 2013

Recommended by: _____ Date: _____
(Advisor)

Approved by: _____ Date: _____
(Chairman of Ph.D Committee)

Acknowledgements

First and foremost, I would like to express my deep gratitude to Professor Shuly Wintner, my research supervisor, for the patient guidance, encouragement and advice he has provided throughout my time as his student. I would also like to extend thanks to Dr. Noam Ordan for his ideas that made this research possible. In addition, I would like to thank Dr. Alon Lavie for useful critiques of this research work.

A special thanks for my family. My wife Liuba, who inspired me to become a PhD student and supported me uncoditionally throughout my studies. My kids Alon and Uri, who kept me awake at nights so I could think about the research. Finally, I wish to thank my parents for their support and encouragement throughout my study.

Contents

The Effect of Translationese on Statistical Machine Translation

Gennadi Lembersky

Abstract

Much research in Translation Studies indicates that translated texts have unique characteristics that set them apart from original texts. Known as *translationese*, translated texts (in any language) constitute a register of the target language, which reflects both artifacts of the translation process and traces of the original language from which the texts were translated. The goal of this work is utilize these differences in order to improve the quality of statistical machine translation.

First, we investigate the differences between language models compiled from original target-language texts and those compiled from texts manually translated to the target language. Corroborating established observations of Translation Studies, we demonstrate that the latter are significantly better predictors of translated sentences than the former, and hence fit the reference set better. Furthermore, translated texts yield better language models for statistical machine translation than original texts.

Second, we investigate the effect of the translation direction of parallel corpora on the quality of translation models. It has already been shown that phrase tables constructed from parallel corpora translated in the same direction as the translation task outperform those constructed from corpora translated in the opposite direction. We reconfirm that this is indeed the case, but emphasize the importance of using also texts translated in the 'wrong' direction. We take advantage of information pertaining to the direction of translation in constructing phrase tables, by adapting the translation model to the special properties of translationese. We explore two adaptation techniques: First, we create a mixture model by interpolating phrase tables trained on texts translated in the 'right' and the 'wrong' directions. The weights for the interpolation are determined by minimizing perplexity. Second, we define entropy-based measures that estimate the correspondence of target-language phrases to translationese, thereby eliminating the need to annotate the

parallel corpus with information pertaining to the direction of translation. We show that incorporating these measures as features in the phrase tables of statistical machine translation systems results in consistent, statistically significant improvement in the quality of the translation.

List of Tables

List of Figures

Chapter 1

Introduction

Much research in Translation Studies indicates that translated texts have unique characteristics that set them apart from original texts [Toury, 1980, Gellerstam, 1986, Toury, 1995b]. Known as *translationese*, translated texts (in any language) constitute a register of the target language, which reflects both artifacts of the translation process and traces of the original language from which the texts were translated. Registers are commonly defined [Biber and Conrad, 2009] in the literature by their situational context, communicative purposes and linguistic features; this has also to do with their modes or modalities of production, such as speech vs. writing. In a similar manner, translation is constrained situationally both by its mode of production and linguistic features as well as by the intricate relationship between the two. We take it then as a working hypothesis that translation *is* a register in and for itself. Among the better-known properties of translationese are *simplification* and *explicitation* [Blum-Kulka and Levenston, 1983, Blum-Kulka, 1986, Baker, 1993]: translated texts tend to be shorter, to have lower type/token ratio, and to use certain discourse markers more frequently than original texts. Interestingly, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy [Baroni and Bernardini, 2006, van Halteren, 2008, Ilisei et al., 2010, Koppel and Ordan, 2011].

Recently, Kurokawa et al. [2009] applied this observation to statistical machine translation, showing that for an English-to-French MT system, a *translation* model trained on an English-translated-to-French parallel corpus is better than one trained on French-translated-to-English texts. In the first part of this work we investigate whether a *language model* compiled from translated texts may similarly improve the results of machine trans-

lation. We test this hypothesis on several translation tasks, including translation from several languages to English, and two additional tasks where the target language is not English. For each language pair we build two language models from two types of corpora: texts originally written in the target language, and human translations from the source language into the target language. We show that for each language pair, the latter language model better fits a set of reference translations in terms of perplexity. We also demonstrate that the differences between the two LMs are not biased by content, but rather reflect differences in abstract linguistic features.

Research in Translation Studies holds a dual view on *translationese*, the sublangauge of translated texts. On the one hand, there is a claim for so-called *translation universals*, traits of translationese which occur in any translated text irrespective of the source language. Others hold, on the other hand, that each source language 'spills over' to the target text, and therefore creates a sub-translationese, the result of a pair-specific encounter between two specific languages. If both these claims are true then language models based on translations from the source language should best fit target language reference sentences, and language models based on translations from *other* source languages should fit reference sentences to a lesser extent yet outperform originally written texts. To test this hypothesis, we compile additional English LMs, this time using texts translated to English from languages *other* than the source. Again, we use perplexity to assess the fit of these LMs to reference sets of translated-to-English sentences. We show that these LMs depend on the source language and differ from each other. Whereas they outperform O-based LMs, LMs compiled from texts that were translated from the *source* language still fit the reference set best.

Finally, we train phrase-based MT systems [Koehn et al., 2003] for each language pair. We use four types of LMs: original; translated from the source language; translated from other languages; and a mixture of translations from several languages. We show that the translated-from-source-language LMs provide a significant improvement in the quality of the translation output over all other LMs, and that the mixture LMs always outperform the original LMs. This improvement persists even when the original LMs are up to ten times larger than the translated ones. In other words, one has to collect ten times more original material in order to reach the same quality as is provided with translated material.

It is important to emphasize that translated texts abound: in fact, Pym and Chrupała

[2005] show (quantitatively!) that the rate of translations *into* a language is inversely proportional to the number of books published in that language: So whereas in English only around 2% of texts published are translations, in languages such as Albanian, Arabic, Danish, Finnish or Hebrew, translated texts constitute between 20-25 percent of the total publications. Furthermore, such data can be automatically identified (see Section 1.1). The practical impact of our work on MT is therefore potentially dramatic.

The second part of this work focuses on translation models (TMs). Contemporary statistical machine translation (SMT) systems use parallel corpora to train *translation models* that reflect source- and target-language phrase correspondences. Typically, SMT systems ignore the direction of translation of the parallel corpus. Given the unique properties of translationese, which operate asymmetrically from source to target language (and *not* vice versa), it is reasonable to assume that this direction may affect the quality of the translation.

We use the results of Kurokawa et al. [2009] as our departure point, but improve them in two major ways. First, we demonstrate that the other subset of the corpus, reflecting translation in the 'wrong' direction,[1] is also important for the translation task, and must not be ignored; second, we show that explicit information on the direction of translation of the parallel corpus, whether manually-annotated or machine-learned, is not mandatory. This is achieved by casting the problem in the framework of domain adaptation: We use domain-adaptation techniques to direct the SMT system toward producing output that better reflects the properties of translationese. We show that SMT systems adapted to translationese produce better translations than vanilla systems trained on exactly the same resources. We confirm these findings using automatic evaluation metrics, as well as through a qualitative analysis of the results.

Furthermore, we show that the direction of translation used for producing the parallel corpus can be approximated by defining several entropy-based measures that correlate well with translationese, and, consequently, with translation quality. We use the entire corpus, create a single, unified phrase table and then use these measures, and in particular *cross-entropy*, as a clue for selecting phrase pairs from this table. The benefit of this method is that not only does it improve the translation quality, but it also eliminates the need to directly predict the direction of translation of the parallel corpus.

[1] Henceforth, we refer to the 'right' direction as *source-to-target*, or $S \rightarrow T$and to the 'wrong' direction as *target-to-source*, or $T \rightarrow S$.

1.1 Translationese

Numerous studies suggest that translated texts are different from original ones. Gellerstam [1986] compares texts written originally in Swedish and texts translated from English into Swedish. He notes that the differences between them do not indicate poor translation but rather a statistical phenomenon, which he terms *translationese*. He focuses mainly on lexical differences, for example less colloquialism in the translations, or foreign words used in the translations "with new shades of meaning taken from the English lexeme" (p. 91). Toury [1995a] brings many examples of lexical items in Hebrew which have a wider range of functions when translated from Hebrew. The word *na'ara*, for example, is a common translation of the English 'girl'. In Hebrew 'authentic' discourse it refers to a teenager, but when translated from English this word takes up new shades of meaning it has in English, as when it serves in translating 'college girl' and 'cover girl'. The movie 'working girl' was translated into Hebrew as 'na'ara ovedet', where the movie opens with the protagonist's 30th birthday. 'Na'ara' would rarely, if ever' be used for woman of the age of thirty in Hebrew original speech. Later studies consider grammatical differences (see, e.g., Santos [1995]). The features of translationese were theoretically organized under the terms *laws of translation* and *translation universals*.

Toury [1980, 1995b] distinguishes between two laws: the *law of interference* and the *law of growing standardization*. The law of interference pertains to the fingerprints of the source text that are left in the translation product. The law of standardization pertains to the effort to standardize the translation product according to existing norms in the target language (and culture). Interestingly, these two laws are in fact reflected in the architecture of statistical machine translation: Interference in the translation model and standardization in the language model.

The combined effect of these laws creates a hybrid text that partly corresponds to the source text and partly to texts written originally in the target language, but in fact belongs to neither [Frawley, 1984]. Baker [1993, 1995, 1996] suggests several candidates for translation universals, which are claimed to appear in any translated text, regardless of the source language. These include *simplification*, the tendency of translated texts to simplify the language, the message or both; and *explicitation*, their tendency to spell out implicit utterances that occur in the source text.

During the 1990s, corpora have been used extensively to study translationese. For

example, Al-Shabab [1996] shows that translated texts exhibit lower lexical variety (type-to-token ratio) and Laviosa [1998] shows that their mean sentence length is lower, as is their lexical density (ratio of content to non-content words). These studies, although not conclusive, provide some evidence for the simplification hypothesis.

Baroni and Bernardini [2006] use machine learning techniques to distinguish between original and translated Italian texts, reporting 86.7% accuracy. They manage to abstract from content and perform the task using only morpho-syntactic cues. Ilisei et al. [2010] perform the same task for Spanish but enhance it theoretically in order to check the simplification hypothesis. They first use a set of features which seem to capture "general" characteristics of the text (ratio of grammatical words to content words); they then add another set of features, each of which relates to the simplification hypothesis. Finally, they remove each "simplification feature" in turn and evaluate its contribution to the classification task. The most informative features are lexical variety, sentence length and lexical density.

van Halteren [2008] focuses on six languages from Europarl [Koehn, 2005]: Dutch, English, French, German, Italian and Spanish. For each of these languages, a parallel six-lingual sub-corpus is extracted, including an original text and its translations into the other five languages. The task is to identify the source language of translated texts, and the reported results are excellent. This finding is crucial: As Baker [1996] states, translations do resemble each other; however, in accordance with the law of interference, the study of van Halteren [2008] suggests that translations from different source languages constitute different sublanguages.

Kurokawa et al. [2009] were the first to address the direction of translation in the context of SMT. They find that a translation model based on the $S \rightarrow T$ portion of a parallel corpus results in much better translation quality than a translation model based on the $T \rightarrow S$ portion.

1.2 Statistical Machine Translation

The modern age of Statistical Machine Translation began when Brown et al. [1990] proposed the *noisy-channel* approach to machine translation. In this approach, a source sentence is viewed as a target sentence that was passed through a noisy medium and

transformed by it. Since the noisy-channel is a stochastic process, a given source sentence can be the result of millions of target sentences. Thus, in order to translate a given source sentence, s, the most likely target sentence, $\hat{t}$, that could be transformed into s by the noisy-channel has to be identified. This is formally given by:

$$\hat{t} = \arg\max_t P(t|s) = \arg\max_t P(s|t)P(t) \tag{1.1}$$

where $P(s|t)$ is the *translation model*, $P(t)$ is the *language model* and *argmax* represents the search for the target sentence that maximizes their product.

Translation Model: The translation model estimates the *conditional* probability of translating a source sentence s into a target sentence t. Five translation models, commonly known as *IBM Models*, were proposed in Brown et al. [1993] and have become standard in the MT community. The models propose a generative "story" that explains how a target sentence was transformed by the noisy-channel to become a source sentence. The "story" is realized through *alignments* between source and target sentences. An alignment indicates, for each word in the source sentence, the corresponding word in the target sentence from which is was translated. Formally, if the target sentence, $t = t_1\, t_2 \ldots t_l$, has l words, and the source sentence, $s = s_1\, s_2 \ldots s_m$ has m words, then the alignment, a, can be represented by a series, $a_1\, a_2 \ldots a_m$, of m values, each between 0 and l, such that if s_j corresponds to t_i, then $a_j = i$, and if it is not connected to any target word, then $a_j = 0$ [Brown et al., 1990].[2] The length of the source sentence, m, is uniformly distributed over a set of "reasonable" lengths. Thus, the conditional probability $p(s|t)$ is given by Brown et al. [1993]:

$$P(s|t) = \sum_a P(s, a|t) = P(m|t) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, s_1^{j-1}, m, t) P(s_j|a_1^j, s_1^{j-1}, m, t) \tag{1.2}$$

The equation describes a process of generating a source sentence from a target sentence. First, the length of the source sentence, m, is selected given the length of the target sentence. Then, the first position in the source sentence is aligned with the position a_j in the target sentence, based on the knowledge of the target sentence and the length m of the source sentence. Then, the first word in the source sentence is selected, based on the knowledge of the target sentence, the length m and the alignment position a_j. The process continues until the last position in the source sentence is reached.

[2] In the general case, any source word can be connected to several target words, and therefore the maximal range of a_j is between 0 and 2^l, resulting in 2^{lm} possible alignments.

The parameters of the translation model are *trained* on a large bilingual corpus. First, the source and the target sides of bilingual corpora are aligned on a sentence level. Then, each pair of the aligned sentences is aligned on a word level. Last, the translation model probablities are learned using the Expectation Maximization (EM) algorithm [Baum, 1972, Dempster et al., 1977].

Language Model: The language model estimates the *a priori* probability of a target sentence in a target language. The language models used by most of today's MT systems are basic statistical n-grams that model language as a Markov chain of order $n-1$ [Bahl et al., 1983]. These language models were developed for speech recognition applications and were later adopted by the MT community. Statistical language models are independently trained on very large monolingual corpora.

Decoding: The search component is usually refered to as *decoding* [Wang and Waibel, 1998]. Decoding is a very difficult optimization problem due to the huge number of possible translations of a given source sentence. Since an exhaustive search is impractical, weak translations are agressively pruned. Currently, most search algorithms for statistical MT proposed in the literature are based on dynamic-programming (DP) beam search [Tillmann and Ney, 2003]. Decoding searches for the translation that optimizes some scoring function. In the most basic form this is the translation that has the highest probability according to the language model.

1.2.1 Phrase-based Models

An important improvement over the basic SMT model was the transition from single-word models to more sophisticated *phrase-based* models [Koehn et al., 2003, Venugopal et al., 2003, Och and Ney, 2004]. In a phrase-based model, a generalized phrase becomes an atomic element of statistical modeling. A *phrase* in this model is just a consequtive sequence of words with no syntax-based limitations. The advantages of phrase-based models over single-word models were shown by Och and Ney [2000a]: Word context and local reordering are explicitly taken into account, resulting in better translation quality.

Phrase-based statistical MT (PB-SMT) extends the above-mentioned IBM Models by allowing alignments on the phrase-level as well as the word-level. The generative "story"

of this approach can be described as follows [Koehn et al., 2003]: The source sentence, s, is segmented into N sequential phrases($\bar{s}_1 \ldots \bar{s}_N$), assuming uniform distribution over all possible segmentations (each phrase must have at least one word). Each phrase, $\bar{s}_i$, is translated into a target language phrase, $\bar{t}_i$, according to some translation probability $t(\bar{s}_i|\bar{t}_i)$. The target phrases are then reordered, based on some distortion probility distribution $d(pos(\bar{s}_i)|pos(\bar{t}_i))$ that specifies the probability of a source phrase in position $pos(\bar{s}_i)$ to be translated into a target phrase in position $pos(\bar{t}_i)$. Thus, the phrase-based translation model is formally given by:

$$P(s|t) = P(\bar{s}_i|\bar{t}_i) = \prod_{i=1}^{N} t(\bar{s}_i|\bar{t}_i) d(pos(\bar{s}_i)|pos(\bar{t}_i)) \tag{1.3}$$

The parameters of the models t and d can be estimated from a bilingual parallel corpus, as shown in Marcu and Wong [2002].[3] The extracted phrase pairs and their translation probabilities are stored in a special data structure referred to as the *phrase table*, which is then used in decoding.

1.2.2 Statistical Language Modeling (SLM)

A statistical language model (SLM) estimates the *a priori* probability $P(t)$ of a sentence t in a language. In other words, it attempts to give a numeric expression to a likelihood of someone saying sentence t. Most SLMs model the probability $P(t)$ by breaking it to a product of conditional probabilities. If $t = t_1\, t_2, \ldots, t_n$ is a sentence with n words, then $P(t)$ is given by:

$$P(t) \equiv \prod_{i=1}^{n} p(t_i|h_i) \tag{1.4}$$

where $h_i \equiv t_1, t_2, \ldots, t_{i-1}$ is called the *history* [Rosenfeld, 2000].

The language models used by most of today's MT systems are basic statistical n-grams that model language as a Markov chain of order $n-1$ [Bahl et al., 1983], that is the i-th word t_i depends only on $n-1$ preceding words:

$$P(t_i|h_i) \approx p(t_i|t_{i-n+1}, \ldots t_{i-1}) \tag{1.5}$$

The most popular choice for the order of the n-grams is 3. It is a good fit for large training corpora (millions of words). The straightforward way to estimate these conditional

[3] Marcu and Wong [2002] estimate joint-probabilites of the translation and distortion models. Afterwards, joint-probabilites can be marginalized to the conditional probabilities required by the noisy-channel approach.

probabilities is the *maximum likelihood estimation* (MLE):

$$P_{MLE}(t|h) = \frac{count(h,t)}{count(h)} \tag{1.6}$$

where the *count* function provides the number of times the event was observed in the training corpus.

The standard method to assess the quality of a given language model is *perplexity* [Jelinek et al., 1977], which is based on the *cross entropy* measure. Cross entropy makes the connection between the probability distribution of the constructed language model P_M and the true (but unknown) probability distribution $P(D_i)$ of a new data sample D_i, $1 \leq i \leq n$. It is formally given by:

$$H(P, P_M) = -\sum_{i=1}^{n} P(D_i)\, log P_M(D_i) \tag{1.7}$$

Equation (1.7) can be rewritten using the *average log likelihood* of a new random sample, which can be viewed as an empirical estimate of the cross entropy [Rosenfeld, 2000]:

$$H(P, P_M) \approx -\frac{1}{n}\sum_{i=1}^{n} log P_M(D_i) \tag{1.8}$$

The perplexity is given by:

$$PP = 2^{H(P,P_M)} \tag{1.9}$$

Perplexity is used to measure the effect of the language model on the results of a specific language application, such as machine translation (the better the model, the lower the perplexity).

1.2.3 Discriminative Models

Most MT systems use a combination of statistical and linguistic features to discriminate between good and bad hypotheses during decoding. The most commonly used discriminative model is the *log-linear* model [Och and Ney, 2001]. Log-linear modeling is based on the well-founded maximum entropy framework [Berger et al., 1996]. Generally, a log-linear model is defined as a combination of N feature functions $h_i(t,s), 1 \leq i \leq N$, that map input, output or a pair of input and output strings to a numeric value. An example of a feature function might be the logarithm of the probability defined by a translation model $P(s|t)$ or a language model $P(t)$, the number of phrase segmentations in t, etc. For each

feature function there is a model paramater λ_i. These parameters, which are called *feature weights*, determine the contribution of a feature to the overall value of $P(t|s)$ [Lopez, 2008]. Formally, the log-linear model of the translation probability $P(t|s)$ is given by:

$$P(t|s) = \frac{1}{Z} exp\left[\sum_{i=1}^{N} \lambda_i h_i(t,s)\right] \tag{1.10}$$

where Z is a normalizing factor.

Since in SMT, one searches for the target sentence $\hat{t}$ that maximizes the probability $P(t|s)$, equation 1.10 entails:

$$\hat{t} = \arg\max_t P(t|s) = \arg\max_t \left\{\sum_{i=1}^{N} \lambda_i h_i(t,s)\right\} \tag{1.11}$$

1.2.4 Minimum Error-Rate Training (MERT)

Minimum error-rate training (MERT) [Och, 2003] is an algorithm that is typically used to train the discriminative model and determine the feature weights. MERT searches for the vector of feature weights that minimizes a given error function. An error function $E(t,r)$ defines the error in a candidate translation t, by comparing it to a reference translation r. The relation between the error function and the feature weights is formally defined as follows: Given a training corpus of M source sentences s_j, $1 \leq j \leq M$, with a given set of reference translations r_j, $1 \leq j \leq M$ and a set of K candidate translations $t'_{j,k}$ for each source sentence s_j, the feature weights are:

$$\hat{\lambda}_1^N = \arg\min_{\lambda_1^N} \sum_{j=1}^{M} E(\arg\max_{t \in t'_{j,k}} \left\{\sum_{i=1}^{N} \lambda_i h_i(t|s_i)\right\}, r_j) \tag{1.12}$$

The MERT algorithm starts with random values for λ_i, $1 \leq i \leq N$, then it tries to improve each parameter λ_i in turn, holding the others constant. The optimized λ_i, $1 \leq i \leq N$, that most significantly reduce the error at the end of each optimization cycle, are then fed back to the next optimization iteration [Lopez, 2008].

1.3 Evaluation

An automatic MT evaluation metric is a necessary tool for anyone who wants to develop an MT system. Such a tool allows immediate assessment of the contribution of an update to the translation quality. The common approach to MT evaluation uses a set of test

sentences translated by human translators, called *references*. In order to prevent bias towards a specific translation style, several reference translations are used. The translation candidates produced by the MT system are matched against the references. The closer the translation candidate is to the reference, the higher the score it receives. The intuition behind this approach is that MT must be good if it closely resembles human translation [Papineni et al., 2002]. The most widely used metric is the *bilingual evaluation understudy* (BLEU) proposed by Papineni et al. [2002].

BLEU calculates the n-gram *precision* for the set of candidate translations, starting with $n = 1$, up to some maximum n (usually 4). Formally, given a set of hypothesis translations H, the n-gram precision p_n is [Lopez, 2008]:

$$p_n = \frac{\sum_{t \in H} \sum_{g \in ngrams(t)} count(ref(g))}{\sum_{t \in H} \sum_{g' \in ngrams(t)} count(g')} \tag{1.13}$$

where $count(g)$ is the number of n-gram g, in a particular sentence t, and $count(ref(g))$ is the number of n-gram g, in the corresponding reference sentence. The n-gram precisions for different values of n are combined as the geometric average: $\sum_n log\, p_n$.

Based on (1.13), BLEU clearly favors shorter hypotheses over longer ones. In order to correct that, the metric includes a *brevity penalty* that penalizes the hypotheses that are much shorter than the reference. Given h, the total number of words in the entire set of hypothesis translations and r, the overall length of the references, created by summing up the lengths of the closest references to each candidate translation, the brevity penalty is given by:

$$BP = \begin{cases} 1 & if\ h > r \\ e^{1-r/h} & if\ h \leq r \end{cases} \tag{1.14}$$

Thus, BLEU is given by:

$$BLUE = BP \cdot exp(\sum_n log\, p_n) \tag{1.15}$$

Another popular MT evaluation metric is Meteor [Banerjee and Lavie, 2005, Denkowski and Lavie, 2011]. Meteor has several distinctive features that set it apart from the BLEU metric:

1. It takes *recall* into account as well as *precision*, while BLEU focuses only on *precision*.
2. It is well suited for evaluating a single sentence, as well as larger texts, while BLEU is mostly used for text evaluation.

3. Meteor supports several modes of word matching: In the "exact" mode, two words match if they are identical. In the "Porter stem" mode, two words match if they have the same stem. And in the "WN synonymy" mode, two words match if they are synonyms.

4. The latest version of Meteor (1.3) implements improved text normalization, higher-precision paraphrase matching, and discrimination between content and function words [Denkowski and Lavie, 2011].

METEOR starts by creating a word alignment between a translation hypothesis and its reference sentence. An alignment is a mapping between words, such that every word in each sentence maps to at most one word in the other sentence [Banerjee and Lavie, 2005]. Given the number of words mapped between two sentences, m, the number of words in the translation candidate, t, and the number of words in the reference sentence, r, the metric first calculates the precision $P = m/t$ and the recall $R = m/r$. Next, the parameterized harmonic mean of P and R is computed:

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \tag{1.16}$$

where $0 \leq \alpha \leq 1$ determines the significance of precision vs. recall.

In the next stage, METEOR computes a *fragmentation penalty* in order to take into account word-order differences between two sentences. First, the number of *chunks* (sequences of matched adjacent words in the same order) and the number of matched words are used to calculate the *fragmentation fraction*: $frag = ch/m$. Then, the penalty is given by:

$$Pen = \gamma \cdot frag^{\beta} \tag{1.17}$$

where γ determines the maximum penalty ($0 \leq \gamma \leq 1$) and β determines the functional relation between fragmentation and the penalty. Finally, the METEOR score is given by:

$$score = (1 - Pen) \cdot F_{mean} \tag{1.18}$$

Yet another popular metric is Translation Error Rate (TER) [Snover et al., 2006]. Generally speaking, TER measures the amount of human editing required to change a translation hypothesis into a reference translation. Possible editing includes the insertion, deletion, and substitution of single words as well as shifts of word sequences. Formally, TER is defined as the minimum number of edits needed to change a translation hypothesis

so that it exactly matches one of the references, normalized by the average length of the references.

$$TER = \frac{\#\ of\ edits}{average\ \#\ of\ reference\ words} \tag{1.19}$$

1.4 SMT Adaptation

Our work is closely related to research in domain-adaptation. In a typical domain adaptation scenario, a system is trained on a large corpus of "general" (out-of-domain) training material, with a small portion of in-domain training texts. In our case, the translation model is trained on a large parallel corpus, of which some (generally unknown) subset is "in-domain" ($S \rightarrow T$), and some other subset is "out-of-domain" ($T \rightarrow S$). Most existing adaptation methods focus on selecting in-domain data from a general domain corpus. In particular, *perplexity* is used to score the sentences in the general-domain corpus according to an in-domain language model. Gao et al. [2002] and Moore and Lewis [2010] apply this method to language modeling, while Foster et al. [2010] and Axelrod et al. [2011] to translation modeling. Moore and Lewis [2010] suggest a slightly different approach, using cross-entropy *difference* as a ranking function.

Domain adaptation methods are usually applied at the corpus level, while we focus on an adaptation of the *phrase table* used for SMT. In this sense, our work follows Foster et al. [2010], who weigh out-of-domain phrase pairs according to their relevance to the target domain. They use multiple features that help to distinguish between phrase pairs in the general domain and those in the specific domain. We rely on features that are motivated by the findings of Translation Studies, having established their relevance through a comparative analysis of the phrase tables. In particular, we use measures such as *translation model entropy*, inspired by Koehn et al. [2009]. Additionally, we apply the method suggested by Moore and Lewis [2010] using perplexity *ratio* instead of cross-entropy difference.

Koehn and Schroeder [2007] suggest a method for adaptation of translation models. They pass two phrase tables directly to the decoder using multiple decoding paths. As we show in Section 3.4, the application of this method to our scenario does not result in a clear contribution, and we are able the show better results using our proposed method.

Finally, Sennrich [2012] proposes perplexity minimization as a way to set the weights for translation model mixture for domain adaptation. We successfully apply this method

to the problem of adapting translation models to translationese, gaining statistically significant improvements in translation quality.

1.5 Contributions of the Thesis

The main objective of this work is to explore ways to utilize knowledge about the status of texts (original vs. translated) to improve the quality of statistical machine translation.

The focus of our research is two-fold. First, we show that a *language model* compiled from translated texts is better for statistical machine translation than a language model compiled from original texts (chapter 2). The main contributions of the first part of this work are thus a computational corroboration of the following hypotheses:

1. Original and translated texts exhibit significant, measurable differences;

2. LMs compiled from translated texts better fit translated references than LMs compiled from original texts of the same (and much larger) size (and, to a lesser extent, LMs compiled from texts translated from languages other than the source language); and

3. MT systems that use LMs based on manually translated texts significantly outperform LMs based on originally written texts.

Second, we demonstrate that the direction of translation in parallel corpora can be used to improve the quality of translation model and hence the translation quality of SMT systems (chapter 3). The main contribution of this part is a methodology that improves the quality of SMT by building translation models that are adapted to the nature of translationese. We explore two adaptation techniques:

1. A linear interpolation of phrase tables trained on texts translated in the 'right' and the 'wrong' directions. The weights for the interpolation are determined by minimizing perplexity (Section 3.4.2).

2. Enriching phrase tables with entropy-based measures that estimate the correspondence of target-language phrases to translationese. The benifit of this method is that it eliminates the need to annotate the parallel corpus with information pertaining to the direction of translation (Section 3.4.3).

We show that both techniques produce translation models that, when used in SMT systems, significantly outperform baseline, un-adapted models.

The results of the experiments with language models, including the Europarl experiments, abtraction and experiments with LMs of different sizes (Section 2.3.1 and Section 2.3.2) were presented in Lembersky et al. [2011]. The extended version that reports on additional language pairs (translating English into French and German), adds experiments with the Gigaword corpus, presents language model combination techniques (Section 2.3.2) and more detailed evaluation and analysis (Section 2.4) will be published as Lembersky et al. [2012a].

The initial results of translation model experiments, including the baseline Hansard experiments (Section 3.2.2), phrase table analysis, including entropy-base metrics (Section 3.3) and cross-entropy-based adaptation techniques (Section 3.4.3) were presented in Lembersky et al. [2012b]. Experiments with additional language pairs (Europarl experiments: Section 3.2.1), a linear interpolation with perplexity minimization adaptation technique (Section 3.4.2) and more detailed analysis (Section 3.5) were described in a separate paper, which is now under review for a major journal.

The rest of the thesis is organized as follows: Chapter 2 details the LM experiments. The TM experiments are described in chapter 3. We present our attempts to combine the findings of the LM and TM experiments in chapter 4. Finally, chapter 5 discusses the results and their implications, and suggests directions for future research.

Chapter 2

Language Models: Translated vs. Original Texts

This chapter is an adaptation of Lembersky et al. [2012a].

2.1 Overview

We investigate the following three hypotheses:

1. Translated texts differ from original texts;
2. Texts translated from one language differ from texts translated from other languages;
3. LMs compiled from manually translated texts are better for MT than LMs compiled from original texts.

We test our hypotheses by considering translations from several languages to English, and from English to German and French. For each language pair we create a reference set comprising several thousands of sentences written originally in the source language and manually translated to the target language. Section 2.2.3 provides details on the reference sets.

To investigate the first hypothesis, we train two LMs for each language pair, one created from texts originally written in the language (O-based) and the other from texts translated into the target language (T-based). Then, we check which LM better fits the reference set.

Fitness of a language model to a set of sentences is measured in terms of *perplexity* [Jelinek et al., 1977, Bahl et al., 1983]. Given a language model and a test (reference) set, perplexity measures the predictive power of the language model over the test set, by looking at the average probability the model assigns to the test data. Intuitively, a better model assigns higher probability to the test data, and consequently has a *lower* perplexity; it is *less* surprised by the test data. Formally, the perplexity PP of a language model L on a test set $W = w_1\, w_2 \ldots w_N$ is the probability of W normalized by the number of words N [Jurafsky and Martin, 2008, page 96]:

$$PP(L, W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P_L(w_i|w_1 \ldots w_{i-1})}} \tag{2.1}$$

For the second hypothesis, we extend the experiment to LMs created from texts translated from other languages. For example, we test how well a LM trained on French-translated-to-English texts fits the German-translated-to-English reference set; and how well a LM trained on German-translated-to-English texts fits the French-translated-to-English reference set.

Finally, for the third hypothesis, we use these LMs for statistical MT (SMT). For each language pair we build several SMT systems. All systems use a translation model extracted from a parallel corpus which is oblivious to the direction of the translation; and one of the above-mentioned LMs. Then, we compare the translation quality of these systems in terms of the BLEU metric [Papineni et al., 2002] (as we show in Section 2.4.1, other automatic evaluation metrics reveal the same pattern).

2.2 Resources

2.2.1 Language Models

In all the experiments, we use SRILM [Stolcke, 2002], with interpolated modified Kneser-Ney discounting [Chen, 1998] and no cut-off on all n-grams, to train n-gram language models from various corpora. Unless mentioned otherwise, $n = 4$. We limit language models to a fixed vocabulary and map out-of-vocabulary (OOV) tokens to a unique symbol to better control the OOV rates among various corpora. We experimented with two techniques for setting the vocabulary: Use all words that occur more than once in the

evaluation set (see Section 2.2.3); and use the intersection of all words occurring in all corpora used to train the language model. Both techniques produce very similar results, and for brevity we only report the results achieved with the former technique. In addition, we tried various discounting schemes (e.g., Good-Turing smoothing [Chen, 1998]), and also ran experiments with an open vocabulary. The results of all these experiments are consistent with our findings, and therefore we do not elaborate on them here.

Our main corpus is Europarl [Koehn, 2005], specifically portions collected over the years 1996 to 1999 and 2001 to 2009. This is a large multilingual corpus, containing sentences translated from several European languages. However, it is organized as a collection of bilingual corpora rather than as a single multilingual one, and it is hard to identify sentences that are translated into several languages.

We therefore treat each bilingual sub-corpus in isolation; each such sub-corpus contains sentences translated to English from various languages. We rely on the **language** attribute of the **speaker** tag to identify the source language of sentences in the English part of the corpus. Since this tag is rarely used with English-language speakers, we also exploit the **ID** attribute of the **speaker** tag, which we match against the list of British members of the European parliament.[1]

We focus on the following languages: German (DE), French (FR), Italian (IT), and Dutch (NL). For each of these languages, L, we consider the L-English Europarl sub-corpus. In each sub-corpus, we extract chunks of approximately 2.5 million *English* tokens translated from each of these source languages (T-DE, T-FR, T-IT, and T-NL), as well as sentences written originally in English (O-EN). The mixture corpus (MIX), which is designed to represent "general" translated language, is constructed by randomly selecting sentences translated from any language (excluding original sentences). For English-to-German and English-to-French, we use the German-English and French-English Europarl sub-corpora. We extract German (and French) sentences translated from English, French (or German), Italian and Dutch, as well as sentences originally written in German (or French).

Table 2.1 lists the number of sentences, number of tokens and average sentence length, for each English sub-corpus and each original language. Table 2.2 lists the statistics for

[1] We wrote a small script that determines the original language of Europarl utterances in this way. The script is publicly available.

German and French corpora.

German–English			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	82,700	2,325,261	28.1
O-EN	91,100	2,324,745	25.5
T-DE	87,900	2,322,973	26.4
T-FR	77,550	2,325,183	30.0
T-IT	65,199	2,325,996	35.7
T-NL	94,000	2,323,646	24.7

Italian–English			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	87,040	2,534,793	29.1
O-EN	93,520	2,534,892	27.1
T-DE	90,550	2,534,867	28.0
T-FR	82,930	2,534,930	30.6
T-IT	69,270	2,535,225	36.6
T-NL	96,850	2,535,053	26.2

French–English			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	90,700	2,546,274	28.1
O-EN	99,300	2,545,891	25.6
T-DE	94,900	2,546,124	26.8
T-FR	85,750	2,546,085	29.7
T-IT	72,008	2,546,984	35.4
T-NL	103,350	2,545,645	24.6

Dutch–English			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	90,500	2,508,265	27.7
O-EN	97,000	2,475,652	25.5
T-DE	94,200	2,503,354	26.6
T-FR	86,600	2,523,055	29.1
T-IT	73,541	2,518,196	34.2
T-NL	101,950	2,513,769	24.7

Table 2.1: Europarl English-target corpus statistics, translation from **Lang.** to English

English–German			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	81,447	2,215,044	27.2
O-DE	89,739	2,215,036	24.7
T-EN	88,081	2,215,040	25.2
T-FR	77,555	2,215,021	28.6
T-IT	64,374	2,215,030	34.4
T-NL	94,289	2,215,033	23.5

English–French			
Lang.	**Sentences**	**Tokens**	**Len**
MIX	89,660	2,845,071	31.7
O-FR	89,875	2,844,265	31.6
T-EN	96,057	2,847,238	29.6
T-DE	93,468	2,843,730	30.4
T-IT	73,257	2,848,931	38.9
T-NL	102,498	2,835,006	27.7

Table 2.2: Europarl corpus statistics, translation from **Lang.** to German and French

In another set of experiments we address the size of language models, to assess how much more original material is needed compared with translated material (Section 2.3.2). Since Europarl does not have enough training material for this task, we use the Hansard corpus, containing transcripts of the Canadian parliament from 1996–2007. This is a

bilingual French–English corpus comprising about 80% original English texts (EO) and about 20% texts translated from French (FO). We first separate original English texts from texts translated from French and then, for each sub-corpus, we randomly extract portions of texts of different sizes: 1M, 5M and 10M tokens from the FO corpus and 1M, 5M, 10M, 25M, 50M and 100M tokens from the EO corpus; see Table 2.3. For even larger amounts of data, we use the English Gigaword corpus [Graff and Cieri, 2007], from which we randomly extract portions of up to 1G tokens; see Table 2.4. Unfortunately, we do not know how much of this corpus is original; since it includes data from the Xinhua news agency, we suspect that parts of it are indeed translated.

Original French			
Size	**Sent's**	**Tokens**	**Len**
1M	54,851	1,000,076	18.2
5M	276,187	5,009,157	18.1
10M	551,867	10,001,716	18.1

Original English			
Size	**Sent's**	**Tokens**	**Len**
1M	54,216	1,006,275	18.6
5M	268,806	5,006,482	18.6
10M	537,574	10,004,191	18.6
25M	1,344,580	25,001,555	18.6
50M	2,689,332	50,009,861	18.6
100M	5,376,886	100,016,704	18.6

Table 2.3: Hansard corpus statistics

English, various sources			
Size	**Sentences**	**Tokens**	**Len**
100M	4,448,260	107,483,194	24.2
500M	20,797,060	502,380,054	24.2
1000M	41,517,095	1,002,919,581	24.2

Table 2.4: Gigaword corpus statistics

To experiment with a non-European language (and a different genre) we choose Hebrew (HE). We use two English corpora: The *original* (O-EN) corpus comprises articles from the *International Herald Tribune*, downloaded over a period of seven months (from January to July 2009). The articles cover four topics: news (53.4%), business (20.9%), opinion (17.6%) and arts (8.1%). The *translated* (T-HE) corpus consists of articles collected from the Israeli newspaper *HaAretz* over the same period of time. *HaAretz* is published in Hebrew, but portions of it are translated to English. The O-corpus was downsized in

order for both sub-corpora to have approximately the same number of tokens in each topic. Table 2.5 lists basic statistics for this corpus.

Hebrew–English			
Orig. Lang.	**Sentences**	**Tokens**	**Len**
O-EN	135,228	3,561,559	26.3
T-HE	147,227	3,561,556	24.2

Table 2.5: Hebrew-to-English corpus statistics

2.2.2 SMT Training Data

To focus on the effect of the *language* model on translation quality, we design SMT training corpora to be oblivious to the direction of translation. Again, we use Europarl (January 2000 to September 2000) as the main source of our parallel corpora. We also use the Hansard corpus: We randomly extract 50,000 sentences from the French-translated-to-English sub-corpus and another 50,000 sentences from the original English sub-corpus. For Hebrew we use the Hebrew–English parallel corpus [Tsvetkov and Wintner, 2010] which contains sentences translated from Hebrew to English (54%) and from English to Hebrew (46%). The English-to-Hebrew part comprises many short sentences (approximately 6 tokens per sentence) taken from a movie subtitle database. This explains the low average sentence length of this particular corpus. Table 2.6 lists some details on those corpora.

2.2.3 Reference Sets

The reference sets have two uses. First, they are used as the test sets in the experiments that measure the perplexity of the language models. Second, in the MT experiments we use them to randomly extract 1000 sentences for tuning and 1000 (different) sentences for evaluation. All references are of course disjoint from the LM and training materials.

For each language L we use the L-English sub-corpus of Europarl (over the period of October to December 2000). For L-to-English translation tasks we only use sentences originally produced in L, while for English-to-L tasks we use sentences originally written in English. The Hansard reference set comprises only French-translated-to-English sentences. The Hebrew-to-English reference set is an independent (disjoint) part of the Hebrew-to-English parallel corpus. This set mostly comprises literary data (88.6%) and a small

Language pair	Side	Sentences	Tokens	Len
DE-EN	DE	92,901	2,439,370	26.3
	EN	92,901	2,602,376	28.0
FR-EN	FR	93,162	2,610,551	28.0
	EN	93,162	2,869,328	30.8
IT-EN	IT	85,485	2,531,925	29.6
	EN	85,485	2,517,128	29.5
NL-EN	NL	84,811	2,327,601	27.4
	EN	84,811	2,303,846	27.2
Hansard	FR	100,000	2,167,546	21.7
	EN	100,000	1,844,415	18.4
HE-EN	HE	95,912	726,512	7.6
	EN	95,912	856,830	8.9

Table 2.6: Parallel corpora used for SMT training

portion of news (11.4%). All sentences are originally written in Hebrew and are manually translated to English. See Table 2.7 for the figures.

2.3 Experiments and Results

We detail in this section the experiments performed to test the three hypotheses: that translated texts can be distinguished from original ones, and provide better language models for other translated texts; that texts translated from other languages than the source are still better predictors of translations than original texts (Section 2.3.1); and that these differences are important for SMT (Section 2.3.2).

2.3.1 Translated vs. Original texts

Adequacy of O-based and T-based LMs

We begin with English as the target language. We train 1-, 2-, 3- and 4-gram language models for each Europarl sub-corpus, based on the corpora described in Section 2.2.1. For each language L, we compile a LM from texts translated (into English) from L; from texts translated from languages other than L (including a mixture of such languages, MIX); and

Language pair	Side	Sentences	Tokens	Len
DE-EN	DE	6,675	161,889	24.3
	EN	6,675	178,984	26.8
FR-EN	FR	8,494	260,198	30.6
	EN	8,494	271,536	32.0
IT-EN	IT	2,269	82,261	36.3
	EN	2,269	78,258	34.5
NL-EN	NL	4,593	114,272	24.9
	EN	4,593	105,083	22.9
EN-DE	EN	8,358	215,325	25.8
	DE	8,358	214,306	25.6
EN-FR	EN	4,284	108,428	25.3
	FR	4,284	125,590	29.3
Hansard	FR	8,926	193,840	21.7
	EN	8,926	163,448	18.3
HE-EN	HE	7,546	102,085	13.5
	EN	7,546	126,183	16.7

Table 2.7: Reference sets

from texts originally written in English. The LMs are applied to the reference set of texts translated from L, and we compute the perplexity: the fitness of the LM to the reference set. Table 2.8 details the results. The lowest perplexity (reflecting the **best** fit) in each sub-corpus is typeset in boldface, and the highest (*worst* fit) is slanted.

These results overwhelmingly support our hypothesis. For each language L, the perplexity of the language model that was created from L translations is lowest, followed immediately by the MIX LM. Furthermore, the perplexity of the LM created from originally-English texts is highest in all experiments (except the Dutch-to-English translation task, where the perplexity of the 2-gram LM created from texts translated from Italian is slightly higher). The perplexity of LMs constructed from texts translated from languages other than L always lies between these two extremes: It is a better fit of the reference set than original texts, but not as good as texts translated from L (or mixture translations).

This gives rise to yet another hypothesis, namely that translations from typologically

German to English translations				
Orig. Lang.	**1-gram PPL**	**2-gram PPL**	**3-gram PPL**	**4-gram PPL**
Mix	451.50	93.00	69.36	66.47
O-EN	*468.09*	*103.74*	*79.57*	*76.79*
T-DE	**443.14**	**88.48**	**64.99**	**62.07**
T-FR	460.98	99.90	76.23	73.38
T-IT	465.89	102.31	78.50	75.67
T-NL	457.02	97.34	73.54	70.56
French to English translations				
Orig. Lang.	**1-gram PPL**	**2-gram PPL**	**3-gram PPL**	**4-gram PPL**
Mix	472.05	99.04	75.60	72.68
O-EN	*500.56*	*115.48*	*91.14*	*88.31*
T-DE	486.78	108.50	84.39	81.41
T-FR	**463.58**	**94.59**	**71.24**	**68.37**
T-IT	476.05	102.69	79.23	76.36
T-NL	490.09	110.67	86.61	83.55
Italian to English translations				
Orig. Lang.	**1-gram PPL**	**2-gram PPL**	**3-gram PPL**	**4-gram PPL**
Mix	395.99	88.46	67.35	64.40
O-EN	*415.47*	*99.92*	*79.27*	*76.34*
T-DE	404.64	95.22	73.73	70.85
T-FR	395.99	89.44	68.38	65.54
T-IT	**384.55**	**81.90**	**60.85**	**57.91**
T-NL	411.58	98.78	76.98	73.94
Dutch to English translations				
Orig. Lang.	**1-gram PPL**	**2-gram PPL**	**3-gram PPL**	**4-gram PPL**
Mix	434.89	90.73	69.05	66.08
O-EN	*448.11*	100.17	*78.23*	*75.46*
T-DE	437.68	93.67	71.54	68.57
T-FR	445.00	97.32	75.59	72.55
T-IT	448.11	*100.19*	78.06	75.19
T-NL	**423.13**	**83.99**	**62.17**	**59.09**

Table 2.8: Fitness of various LMs to the reference set

related languages form a *similar* 'translationese dialect', whereas translations from more distant source languages form *two different* 'dialects' in the target language (see Koppel and Ordan [2011]).

Linguistic Abstraction

A possible explanation for the different perplexity results among the LMs could be the specific contents of the corpora used to compile the LMs. For example, one would expect texts translated from Dutch to exhibit higher frequencies of words such as "Amsterdam" or even "canal". This, indeed, is reflected by the lower (usually lowest) number of OOV items in language models compiled from texts translated from the source language.

As a specific example, the top five words that occur in the T-FR corpus and the evaluation set, but are absent from the O-EN corpus are: *biarritz, meat-and-bone, armenian, ievoli* and *ivorian*. The top five words that occur in the O-EN corpus, but are absent from the T-FR corpus, are: *duhamel, paciotti, ivoirian, coke* and *spds*. Of those, *Biarritz* seems to be French-specific, but the other items seem more arbitrary.

To rule out the possibility that the perplexity results are due to specific content phenomena, and to further emphasize that the corpora are indeed *structurally* different, we conduct more experiments, in which we gradually abstract away from the domain- and content-specific features of the texts and emphasize their syntactic structure. We focus on French-to-English, but the results are robust and consistent (we repeated the same experiments for all language pairs, with very similar outcomes).

First, we remove all punctuation to eliminate possible bias due to differences in punctuation conventions.[2] Then, we use the Stanford Named Entity Recognizer [Finkel et al., 2005] to identify named entities, which we replace with a unique token ('NE'). Next, we replace all nouns with their part-of-speech (POS) tag; we use the Stanford POS Tagger [Toutanova and Manning, 2000]. Finally, for full lexical abstraction, we replace all words with their POS tags, retaining only abstract syntactic structures devoid of lexical content.

At each step, we train six language models on O- and T-texts and apply them to the reference set (which is adapted to the same level of abstraction, of course). As the abstraction of the text increases, we also increase the order of the LMs: From 4-grams for text without punctuation and NE abstraction, to 5-grams for noun abstraction, to 8-grams for full POS abstraction. In all cases we fix the LM vocabulary to only contain

[2]In fact, there is reason to assume that punctuation constitutes part of the translationese effect. For example, the right parenthesis is much more common in English translated from German, than in original English since it is used as a list item identifier in German. Removing punctuation therefore harms our cause of identifying this effect.

tokens that appear more than once in the "abstracted" reference set. The results, which are depicted in Table 2.9, consistently show that the T-based LM is a better fit to the reference set, albeit to a lesser extent. The rightmost column specifies the improvement, in terms of perplexity, of each language model, compared to the worst-performing model. While we do not show the details here, the same pattern is persistent in all the other Europarl languages we experiment with.

More Language Pairs

To further test the robustness of these phenomena, we repeat these experiments with the Hebrew-to-English corpus and reference set, reflecting a different language family, a smaller corpus and a different domain. We train two 4-gram language models on the O-EN and T-HE corpora. We then apply the two LMs to the reference set and compute the perplexity. The results are presented in Table 2.10. Again, the T-based LM is a better fit to the translated text than the O-based LM: Its perplexity is lower by 12.8%. We also repeat the abstraction experiments on the Hebrew scenario. The results, which are depicted in Table 2.11, consistently show that the T-based LM is a better fit to the reference set.

Clearly, then, translated LMs better fit the references than original ones, and the differences can be traced back not just to (trivial) specific lexical choice, but also to syntactic structure, as evidenced by the POS abstraction experiments.

We further test our findings on other *target* languages, specifically English-German and English-French. We train several 4-gram language models on the corpora specified in Table 2.2. We then compute the perplexity of the German-translated-from-English and French-translated-from-English reference sets (see Section 2.2.3) with respect to these language models. Table 2.12 depicts the results; they are in complete agreement with our hypothesis.

Larger Language Models

Can these phenomena be attributed to the relatively small size of the corpora we use? Will the perplexity of O texts converge to that of T texts when more data become available, or will the differences persist? To address these questions, we use the (much larger) Hansard corpus and the (even larger) Gigaword corpus. We train 4-gram language models for each

No Punctuation		
Orig. Lang.	**Perplexity**	**Improvement (%)**
MIX	105.91	19.73
O-EN	*131.94*	
T-DE	122.50	7.16
T-FR	**99.52**	**24.58**
T-IT	112.71	14.58
T-NL	126.44	4.17
NE Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
MIX	93.88	18.51
O-EN	*115.20*	
T-DE	107.48	6.70
T-FR	**88.96**	**22.77**
T-IT	99.17	13.91
T-NL	110.72	3.89
Noun Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
MIX	36.02	11.34
O-EN	*40.62*	
T-DE	38.67	4.81
T-FR	**34.75**	**14.46**
T-IT	36.85	9.30
T-NL	39.44	2.91
POS Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
MIX	7.99	2.66
O-EN	*8.20*	
T-DE	8.08	1.47
T-FR	**7.89**	**3.77**
T-IT	8.00	2.47
T-NL	8.11	1.11

Table 2.9: Fitness of O- vs. T-based LMs to the reference set (FR-EN), reflecting different abstraction levels

Hansard and Gigaword sub-corpus described in Section 2.2.1. We apply the LMs to the Hansard reference set, but also to the Europarl reference set, to examine the effect on out-of-domain (but similar genre) texts. In both cases we report perplexity (Table 2.13).

Hebrew to English translations		
Orig. Lang.	**Perplexity**	**Improvement(%)**
O-EN	*187.26*	
T-HE	**163.23**	**12.83**

Table 2.10: Fitness of O- vs. T-based LMs to the reference set (HE-EN)

No Punctuation		
Orig. Lang.	**Perplexity**	**Improvement (%)**
O-EN	*401.44*	
T-HE	**335.30**	**16.48**
NE Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
O-EN	*298.16*	
T-HE	**251.39**	**15.69**
Noun Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
O-EN	*81.92*	
T-HE	**72.34**	**11.70**
POS Abstraction		
Orig. Lang.	**Perplexity**	**Improvement (%)**
O-EN	*11.47*	
T-HE	**10.76**	**6.20**

Table 2.11: Fitness of O- vs. T-based LMs to the reference set (HE-EN), reflecting different abstraction levels

The results are fully consistent with our previous findings: In the case of the Hansard reference set, a language model based on original texts must be up to *ten times larger* to retain the low perplexity level of translated texts. For example, whereas a language model compiled from 10 million English-translated-from-French tokens yields a perplexity of 42.70 on the Hansard reference set, a LM compiled from original English texts requires 100 million words to yield a similar perplexity of 43.70 on the same reference set. The Gigaword LMs, which are trained on texts representing completely different domains and genres, produce much higher (i.e., worse) perplexity in this scenario. In the case of the

English to German translations		
Orig. Lang.	**Perplexity**	**Improvement(%)**
Mix	106.37	20.24
O-DE	*133.37*	
T-EN	**99.39**	**25.47**
T-FR	119.21	10.61
T-IT	123.35	7.51
T-NL	119.99	10.03
English to French translations		
Orig. Lang.	**Perplexity**	**Improvement(%)**
Mix	58.71	3.20
O-FR	*60.65*	
T-EN	**49.44**	**18.47**
T-DE	55.41	8.63
T-IT	57.75	4.77
T-NL	54.23	10.57

Table 2.12: Fitness of O- vs. T-based LMs to the reference set (EN-DE and EN-FR)

Europarl reference set, a language model based on original texts must be approximately *five times larger* (and a Gigaword language model approximately *twenty times larger*) than a language model based on original texts to yield similar perplexity.

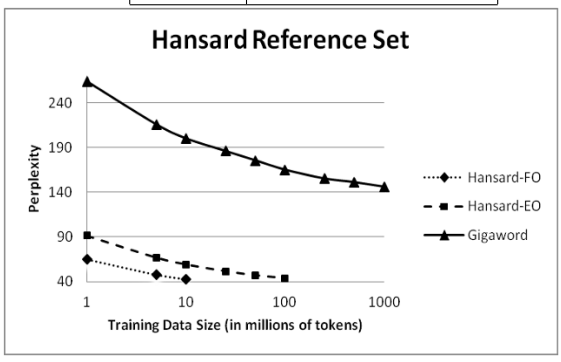
2.3.2 Original vs. Translated LMs for Machine Translation

SMT Experiments

The last hypothesis we test is whether a better fitting language model yields a better machine translation system. In other words, we expect the T-based LMs to outperform the O-based LMs when used as part of machine translation systems. We construct German-to-English, English-to-German, French-to-English, French-to-German, Italian-to-English and Dutch-to-English MT systems using the Moses phrase-based SMT toolkit [Koehn et al., 2007]. The systems are trained on the parallel corpora described in Section 2.2.2. We use the reference sets (Section 2.2.3) as follows: 1,000 sentences are randomly extracted

Hansard Reference Set	
Hansard T-FR	
Size	**Perplexity**
1M	64.68
5M	47.63
10M	42.70
Hansard O-EN	
Size	**Perplexity**
1M	91.40
5M	66.95
10M	59.19
25M	51.59
50M	47.02
100M	43.70
Gigaword	
Size	**Perplexity**
100M	165.03
500M	151.00
1000M	145.88

Europarl Reference Set	
Hansard T-FR	
Size	**Perplexity**
1M	169.66
5M	137.72
10M	128.65
Hansard O-EN	
Size	**Perplexity**
1M	198.93
5M	162.08
10M	150.05
25M	137.31
50M	129.43
100M	123.10
Gigaword	
Size	**Perplexity**
100M	136.72
500M	121.88
1000M	116.55



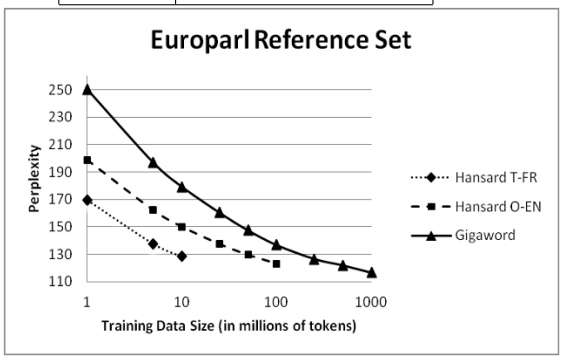


Table 2.13: The effect of LM training corpus size on the fitness of LMs to the reference sets

for minimum error-rate tuning [Och, 2003], and another, disjoint set of 1,000 randomly-selected sentences is used for evaluation. Each system is built and tuned with six different LMs: MIX, O-based and four T-based models (Section 2.2.1). We use BLEU [Papineni et al., 2002] to evaluate translation quality. The results are listed in Tables 2.14 and 2.15.

DE to EN	
LM	**BLEU**
MIX	21.43
O-EN	*21.10*
T-DE	**21.90**
T-FR	21.16
T-IT	21.29
T-NL	21.20

FR to EN	
LM	**BLEU**
MIX	28.67
O-EN	*27.98*
T-DE	28.01
T-FR	**29.14**
T-IT	28.75
T-NL	28.11

IT to EN	
LM	**BLEU**
MIX	25.41
O-EN	24.69
T-DE	*24.62*
T-FR	25.37
T-IT	**25.96**
T-NL	24.77

NL to EN	
LM	**BLEU**
MIX	24.20
O-EN	*23.40*
T-DE	24.26
T-FR	23.56
T-IT	23.87
T-NL	**24.52**

Table 2.14: Machine translation with various LMs; English target language

EN to DE	
LM	**BLEU**
MIX	13.00
O-DE	12.47
T-EN	**13.10**
T-FR	*12.46*
T-IT	12.65
T-NL	12.86

EN to FR	
LM	**BLEU**
MIX	24.83
O-FR	24.70
T-EN	**25.31**
T-DE	*24.58*
T-IT	24.89
T-NL	25.20

Table 2.15: Machine translation with various LMs; non English target language

The results are consistent and fully confirm our hypothesis. Across all language pairs, MT systems using LMs compiled from translated-from-source texts consistently outperform all other systems. Systems that use LMs compiled from texts originally written in the target language always perform worst or second worst. We test the statistical significance of the differences between the results using the bootstrap resampling method [Koehn, 2004]. In all experiments, the best system (translated-from-source LM) is significantly better than the system that uses the O-based LM ($p < 0.01$).

We now repeat the experiment with Hebrew to English translation. We construct a Hebrew-to-English MT system with Moses, using a factored translation model [Koehn and Hoang, 2007]. Every token in the training corpus is represented as two factors: surface form and lemma. The Hebrew input is fully segmented [Itai and Wintner, 2008]. The system is built and tuned with O- and T-based LMs. The O-based LM yields a BLEU score of 11.94, whereas using the T-based LM results in somewhat higher BLEU score, 12.07,

but the difference is not statistically significant ($p = 0.18$). Presumably, the low quality of both systems prevents the better LM from making a significant difference.

LM	**BLEU**	**p-value**
O-based LM	*11.94*	0.18
T-based LM	**12.07**	

Table 2.16: Hebrew-to-English MT results

Larger Language Models

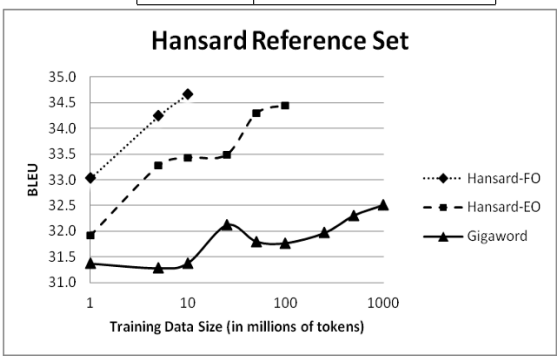
Again, the LMs used in the MT experiments reported above are relatively small. To assess whether the benefits of using translated LMs carry over to scenarios where larger original corpora exist, we build yet another set of French-to-English MT systems. We use the Hansard SMT translation model and Hansard LMs to train nine MT systems, three with varying sizes of translated texts and six with varying sizes of original texts. We train additional MT systems with several subsets of the Gigaword LM. We tune and evaluate on the Hansard reference set. In another set of experiments we use the Europarl French-to-English scenario (using Europarl corpora for the translation model as well as for tuning and evaluation), but we use the Hansard and Gigaword LMs to see whether our findings are consistent also when LMs are trained on out-of-domain material.

Table 2.17 again demonstrates that language models compiled from original texts must be up to *ten times larger* in order to yield translation quality similar to that of LMs compiled from translated texts.[3] In other words, much smaller translated LMs perform better than much larger original ones, and this holds for various LM sizes, both in-domain and out-of-domain. For example, on the Hansard corpus, a 10-million-token T-FR language model yields a BLEU score of 34.67, whereas an O-EN language model of 100 million tokens is required in order to yield a similar BLEU score of 34.44. The systems that use the Gigaword LMs perform much worse in-domain, even with a language model compiled from 1000M tokens. Out-of-domain, the Gigaword systems are better than O-EN, but they require approximately five times more data to match the performance of T-FR systems.

[3]The table only specifies three subsets of the Gigaword corpus, but the graphs show more data points. Note that the X-axis is logarithmic. Incidentally, the graphs show that increases in (Gigaword) corpus size do not monotonically translate to better MT quality.

Hansard TM and Test	
Hansard T-FR	
Size	**BLEU**
1M	33.03
5M	34.25
10M	**34.67**
Hansard O-EN	
Size	**BLEU**
1M	31.91
5M	33.27
10M	33.43
25M	33.49
50M	34.29
100M	34.44
Gigaword	
Size	**BLEU**
100M	31.77
500M	32.31
1000M	32.51

Europarl TM and Test	
Hansard T-FR	
Size	**BLEU**
1M	26.36
5M	27.06
10M	27.22
Hansard O-EN	
Size	**BLEU**
1M	26.06
5M	26.03
10M	26.72
25M	26.72
50M	27.01
100M	27.04
Gigaword	
Size	**BLEU**
100M	27.47
500M	**27.71**
1000M	27.69



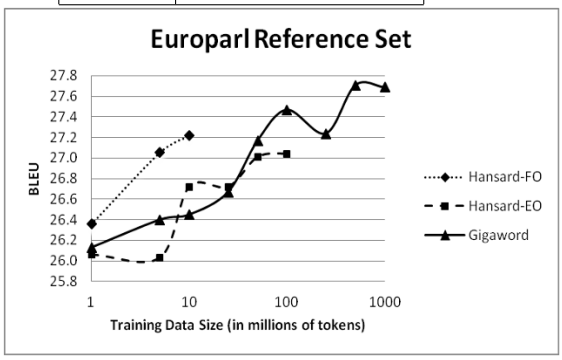


Table 2.17: The effect of LM size on MT performance

Enjoying Both Worlds

The previous section established the fact that language models compiled from translated texts are better for MT than ones compiled from original texts, even when the original LMs are much larger. In many real-world scenarios, however, one has access to texts of both

types. Our results do not imply that original texts are useless, and that only translated ones should be used. In this section we explore various ways to combine original and translated texts, thereby yielding even better language models.

For these experiments we use 10 million English-translated-from-French tokens from the Hansard corpus (T-FR) and another 100 million original-English tokens from the same source (O-EN). We combine them in five different ways: straightforward concatenation of the corpora; a concatenation of the original-English corpus with the translated corpus, upweighted by a factor of 10 and then of 20; log-linear modeling; and an interpolated language model. In each experiment we report both the fitness of the LM to the reference set, in terms of perplexity, and the quality of machine translation that uses this LM, in terms of BLEU.[4] We execute each experiment twice, once (in-domain) with the Hansard reference set and once (out-of-domain) where the translation model, tuning corpus and reference set all come from the Europarl FR-EN sub-corpus, as above. The results are listed in Table 2.18; we now provide a detailed explanation of these experiments.

Hansard TM, LM and Test		
Combination	**PPL**	**BLEU**
O-EN	43.70	34.44
T-FR	42.70	34.67
Concatenation	38.43	34.62
Concatenation x10	41.15	35.09
Concatenation x20	45.07	34.67
Log-Linear LM	–	35.26
Interpolated LM	**36.69**	**35.35**

Europarl TM and Test; Hansard LM		
Combination	**PPL**	**BLEU**
O-EN	123.10	27.04
T-FR	128.65	27.22
Concatenation	116.71	27.14
Concatenation x10	135.09	27.29
Concatenation x20	152.02	27.09
Log-Linear LM	–	27.30
Interpolated LM	**107.82**	**27.48**

Table 2.18: Various combinations of original and translated texts and their effect on perplexity (PPL) and translation quality (BLEU)

Concatenation of O and T texts We train three language models by concatenating the T-FR and O-EN corpora. First, we simply concatenate the corpora obtaining 110 million tokens. Second, we upweight the T-FR corpus by a factor of 10 before the concatenation; and finally, we upweight the T-FR corpus by a factor of 20 before the con-

[4]Except log-linear models, for which we only report the quality of machine translation, since there are two language models in this case and perplexity is harder to compute.

catenation. In the 'in-domain' scenario, the LM trained on a simple concatenation of the corpora reduces the perplexity by more than 10%. The best translation quality is obtained when the T-FR corpus is upweighted by a factor of 10. It improves by 0.42 BLEU points compared to the MT system that uses T-FR ($p = 0.074$), and, more significantly, by 0.65 BLEU points compared to O-EN ($p < 0.05$). In the 'out-of-domain' scenario, there is a small reduction in perplexity (about 5%) with a language model that is trained on a simple concatenation of the corpora. There is also a very small improvement in the translation quality (0.07 BLEU points compared to the T-FR system and 0.25 BLEU points compared to O-EN).

Log-Linear combination of language models The MOSES decoder uses log-linear modeling [Och and Ney, 2001] to discriminate between better and worse hypotheses during decoding. A log-linear model is defined as a combination of N feature functions $h_i(t, s), 1 \leq i \leq N$, that map input ($s$), output ($t$) or a pair of input and output strings to a numeric value. Each feature function is associated with a model parameter λ_i, its *feature weight*, which determines the contribution of the feature to the overall value of $P(t|s)$. Formally, decoding based on a log-linear model is defined by:

$$\hat{t} = \underset{t}{\operatorname{argmax}} P(t|s) = \underset{t}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \lambda_i h_i(t, s) \right\} \tag{2.2}$$

We train two language models, based on T-FR and O-EN. Then, we combine these models by including them as different feature functions. The feature weight of each LM is set by minimum error-rate tuning, optimizing the translation quality; this is the same technique that Koehn and Schroeder [2007] employ for domain adaptation. In-domain, this combination is better by 0.82 BLEU points compared to an MT system that uses O-EN ($p < 0.001$), 0.59 BLEU points compared to the one that uses T-FR ($p < 0.05$). Out of domain, this combination is again not significantly better than using T-FR only (improvement of 0.08 BLEU points, $p = 0.255$).

Interpolated language models In the interpolated scenario, two language models are mixed on a fixed proportion η, according to the following equation [Weintraub et al., 1996]:

$$p(w|h) = (1 - \eta) \cdot p(w|h; LM_1) + \eta \cdot p(w|h; LM_2) \tag{2.3}$$

where w is a word, h is its 'history', and η is the fixed interpolation weight. We use SRILM to train an interpolated language model from LM_1 = O-EN and LM_2 = T-FR.

The interpolation weight is tuned to minimize the perplexity of the combined model with respect to the tuning set; we use the EM algorithm provided as part of the SRILM toolkit to establish the optimized weights. In the in-domain scenario $\eta = 0.46$ and in the out-of-domain scenario $\eta = 0.49$.

The interpolated language model yields additional improvement in perplexity and translation quality compared to all other models. It is significantly better ($p < 0.05$) than the T-FR system on the in-domain scenarios, but the improvement is less significant ($p = 0.075$) out of domain.

In summary, LMs compiled from source-translated-to-target texts are almost as good as much larger LMs that also include large corpora of texts originally written in the target language. Clearly, ignoring the status (original or translated) of monolingual texts and creating a single language model from all of them (the concatenation scenario) is not much better than using *only* translated texts. In order to benefit from (often much larger) original texts, one must consider more creative ways of combining the two sub-corpora. Of the methods we explored here, interpolated LMs provide the greatest advantage. More research is needed in order to find an optimal combination.

2.4 Evaluation and Analysis

One question, however, requires further investigation: Do MT systems based on translated-from-source-language LMs produce better translations, or do they merely generate sentences that are directly adapted to the reference set, thereby only improving a specific evaluation metric, such as BLEU? We address this issue in three ways, showing that the former is indeed the case. First, we use two automated evaluation metrics other than BLEU, and show that the T-based LMs yield better MT systems even with different metrics. Second, we perform a manual evaluation of a portion of the evaluation set. The results show that human evaluators prefer translations produced by an MT system that uses a T-based LM over translations produced by a system built with an O-based LM. Finally, we provide a detailed analysis of the differences between O- and T-based LMs, explaining these differences in terms of insights from Translation Studies.

2.4.1 Automatic Evaluation

First, we use two alternative automatic evaluation metrics, METEOR[5] [Denkowski and Lavie, 2011] and TER [Snover et al., 2006], to assess the quality of the MT systems described in Section 2.3.2. We focus on four translation tasks: From German, French, Italian and Dutch to English.[6] For each task we report the performance of two MT systems: One that uses a language model compiled from original-English texts, and one that uses a language model trained on texts translated from the source language. The results, which are reported in Table 2.19, fully support our previous findings (recall that *lower* TER is better): MT systems that use T-based LMs significantly outperform systems that use O-based LMs.

DE to EN		
Orig. Lang.	**METEOR**	**TER**
O-EN	*28.26*	*64.56*
T-DE	**28.64**	**63.57**

IT to EN		
Orig. Lang.	**METEOR**	**TER**
O-EN	*31.03*	*58.30*
T-IT	**31.16**	**57.63**

FR to EN		
Orig. Lang.	**METEOR**	**TER**
O-EN	*33.05*	*54.45*
T-FR	**33.30**	**53.65**

NL to EN		
Orig. Lang.	**METEOR**	**TER**
O-EN	*29.97*	*60.29*
T-NL	**30.40**	**59.63**

Table 2.19: MT system performance as measured by METEOR and TER

2.4.2 Human Evaluation

To further establish the qualitative difference between translations produced with an English-original language model and translations produced with a LM created from French-translated-to-English texts, we conducted a human evaluation campaign, using Amazon's Mechanical Turk as an inexpensive, reliable and accessible pool of annotators [Callison-Burch and Dredze, 2010]. We created a small evaluation corpus of 100 sentences, selected randomly among all (Europarl) reference sentences whose length is between 15 and 25 words. Each instance of the evaluation task includes two English sentences, obtained from the two MT systems that use the O-EN and the T-FR language models, respectively. An-

[5] More precisely, we use METEOR-RANK, the configuration used for WMT-2011.

[6] All MT systems were tuned using BLEU.

notators are presented with these two translations, and are requested to determine which one is better. The definition given to annotators is: "A better translation is more fluent, reflecting better use of English." Observe that since the only variable that distinguishes between the two MT systems is the different language model, we only have to evaluate the fluency of the target sentence, not its faithfulness to the source. Consequently, we do not present the source or the reference translation to the annotators. We understand that there is always a trade-off between faithfulness and fluency. Increased fluency may come at the cost of reduced faithfulness. However, we believe that on our case this effect is minimal since both language models were trained on a data extracted from the same source. All annotators were located in the US (and, therefore, are presumably English speakers).

As a control set, we added a set of 10 sentences produced with the O-based LM, which were paired with their (manually-created) reference translations, and 10 sentences produced with the T-based LM, again paired with their references. Each of the 120 evaluation instances was assigned to 10 different Mechanical Turk annotators. We report two evaluation metrics: *score* and *majority*. The score of a given sentence pair $\langle e_1, e_2 \rangle$ is i/j, where i is the number of annotators who preferred e_1 over e_2, and $j = 10 - i$ is the number of annotators preferring e_2. For such a sentence pair, the majority is e_1 if $i > j$, e_2 if $i < j$, and undefined otherwise.

The average score of the 10 sentences in the O-vs.-reference control set is 22/78, and the majority is the reference translation in all but one of the instances. As for the T-vs.-reference control set, the average score is 18/82, and the majority is the reference in all of the instances. This indicates that the annotators are reliable, and also that it is unrealistic to expect a clear cut distinction even between human translations and machine-generated output.

As for the actual evaluation set, the average score of O-EN vs. T-FR is 38/62, and the majority is T-FR in 75% of the cases, O-EN in only 25% of the sentence pairs. We take these results as a very strong indication that English sentences generated by an MT system whose language model is compiled from translated texts are perceived by humans as more fluent than ones generated by a system built with an O-based language model. Not only is the improvement reflected in significantly higher BLEU (and METEOR, TER) scores, but it is undoubtedly also perceived as such by human annotators.

2.4.3 Qualitative Analysis

In order to look into the differences between T and O qualitatively, rather than quantitatively, we turn now to study several concrete examples. To do so, we extracted approximately 200 sentences from the French-English Europarl evaluation set; we chose all sentences of length between 15 and 25. In addition, we extracted the 100 most frequent n-grams, for $1 \leq n \leq 5$, from both English-original and English-translated-from-French Europarl corpora. As both corpora include approximately the same number of tokens, we report counts below rather than frequencies.

The differences between O and T texts are consistent with well-established observations of translation scholars. Consider the *explicitation hypothesis* [Blum-Kulka, 1986], which Séguinot [1998, p. 108] spells out thus:

1. "something which was implied or understood through presupposition in the source text is overtly expressed in the translation"
2. "something is expressed in the translation which was not in the original"
3. "an element in the source text is given greater importance in the translation through focus, emphasis, or lexical choice".

Blum-Kulka [1986] uses the term *cohesive markers* to refer to items that are utilized by the translator which cannot be found overtly in the source text. One would expect such markers to be much more prevalent in translationese.

An immediate example of (1) is the case of acronyms: These tend to be spelled out in translated texts. Indeed, the acronym *EU* is ranked 90 among the O-EN unigrams with 3270 occurrences, whereas in T-FR it is ranked 571 with 478 occurrences. On the other hand, the explicit trigram *The European Union* occurs more frequently in T (3678 occurrences) than in O (3526 occurrences).

Other cohesive markers discussed by Blum-Kulka [1986] are over-represented in T compared to O. These include: *therefore* (3,187 occurrences in T, 1,983 in O); *for example* (863 occurrences in T, 701 in O); *in particular* (1336 vs. 1068); *first of all* (601 vs. 266); *in fact* (1014 vs. 441); *in other words* (553 vs. 87); *with regard to* (1137 vs. 310); *in order to* (2,016 vs. 603); *in this respect* (363 vs. 94); *on the one hand* (288 vs. 72); *on the other hand* (428 vs. 76); and *with a view to* (213 vs. 51). A similar list of markers have

been shown to be excellent discriminating features between original and translated texts (from several European languages, including French) in an independent study [Koppel and Ordan, 2011].

Another phenomenon we notice is that the T-based language model does a much better job translating verbs than the O-based language model. In two very large corpora of French and English [Ferraresi et al., 2008], verbs are much more frequent in French than in English (0.124 vs. 0.091). Human translations from French to English, therefore, provide many more examples of verbs from which to model. Indeed, we encounter several examples in which the O-based translation system fails to use a verb at all, or to use one correctly, compared with the T-based system:

Source *Une telle Europe serait un gage de paix et marquerait le refus de tout nationalisme ethnique.*

O *Such a Europe would be a show of peace and would the rejection of any ethnic nationalism.*

T *Such a Europe would be a show of peace and would* ***mark*** *the refusal of all ethnic nationalism.*

Source *Votre rapport, madame Sudre, met l'accent, à juste titre, sur la nécessité d'agir dans la durée.*

O *Your report, Mrs Sudre, its emphasis, quite rightly, on the need to act in the long term.*

T *Your report, Mrs Sudre,* ***places*** *the emphasis, quite rightly, on the need to act in the long term.*

Source *Cette proposition, si elle constitue un pas dans la bonne direction n'en comporte pas moins de nombreuses lacunes auxquelles le rapport evans remédie.*

O *This proposal, if it is a step in the right direction do not least in contains many shortcomings which the evans report resolve.*

T *This proposal, if it is a step in the right direction it* ***contains*** *no less many shortcomings which the evans report resolve.*

Last, there are several cases of *interference*, which Toury [1995b, p. 275] defines as follows: "Phenomena pertaining to the make-up of the source text tend to be transferred to the target text". In the following example, *do not say nothing more* is a literal translation of the French construction *On ne dit rien non plus*. The T-based translation is much more fluent:

Source *On ne dit rien non plus sur la responsabilité des fabricants, notamment en grande-bretagne, qui ont été les premiers responsables.*

O *We* ***do not say nothing more*** *on the responsibility of the manufacturers, particularly in Britain, which were the first responsible.*

T *We* ***do not say anything either*** *on the responsibility of the manufacturers, particularly in great Britain, who were the first responsible.*

Incidentally, there are also some cultural differences between O and T that we deem less important, since they are not part of the 'translationese dialect' but rather indicate differences pertaining to the culture from which the speaker arrives. Most notable is the form *ladies and gentlemen*, which is the tenth most frequent trigram in T, but does not even rank among the top 100 in O. This is already noted by van Halteren [2008], according to whom this form is significantly more frequent in translations from five European languages as opposed to original English.

In terms of (shallow) syntactic structure, we observe that part-of-speech n-grams are distributed somewhat differently in O and in T (we use the POS-tagged Europarl corpus of Section 2.3.1 for the following analysis). For example, *proper nouns* are more frequent in O (ranking 7 among all POS 1-grams) than in T (rank 9). This has influence on longer n-grams: For example, the 3-gram *PRP MD VB* is 20% more frequent in O than in T. The sequence *<S>PRP VBP* is almost twice as frequent in O. The 4-gram *IN DT NN </S>* is 25% more frequent in O. In contrast, the 4-gram *IN DT NNS IN* is 15% more frequent in T than in O. Appendix A provides a long list of examples for statistical differences in POS sequences in original (O-EN) and translated (T-FR) texts. A full analysis of such patterns is beyond the scope of this paper.

Summing up, T-based language models are more fluent and therefore yield better translation results for the following reasons: They are more cohesive, less influenced by structural differences between the languages, such as the under-representation of verbs in origi-

nal English texts, and less prone to interference, i.e., they can break away from the original towards a more coherent model of the target language.

Chapter 3

Translation Models: Utilizing the Direction of the Translation

This chapter is an extended version of Lembersky et al. [2012b], which is currently under review for a journal.

3.1 Overview

We investigate several ways to adapt a translation model to the nature of translationese, thereby making the output of machine translation more similar to actual, human translation. Our departure point is the results of Kurokawa et al. [2009], which we successfully replicate in Section 3.2. First (Section 3.3), we explain *why* translation quality improves when the parallel corpus is translated in the 'right' direction. We do so by showing that the subset of the corpus that was translated in the direction of the translation task (the 'right' direction, henceforth, *source-to-target*, or $S \rightarrow T$) yields *phrase tables* that are better suited for translation of the original language than the subset translated in the reverse direction (the 'wrong' direction, henceforth, *target-to-source*, or $T \rightarrow S$). We use several statistical measures that indicate the better quality of the phrase tables in the former case.

Then (Section 3.4), we explore ways to build a translation model adapted to the unique properties of translationese. We first show that using the entire parallel corpus, including texts that are translated both in the 'right' and in the 'wrong' direction, improves the quality of machine translation. Furthermore, we show methods to overcome the need

to predict the direction of translation used for producing the parallel corpus by defining several entropy-based measures that correlate well with translationese, and, consequently, with the quality of machine translation.

Specifically, we first define a naïve method, akin to *corpus-level adaptation*: We create two phrase tables, one for the $S \rightarrow T$ portion of the corpus, and one for the $T \rightarrow S$ portion, and combine them with a log-linear model. We show that this combination does not perform as well as a simple union of the two portions of the parallel corpus. Better results, however, are obtained by *phrase table adaptation*.

Specifically, we combine two phrase tables, one for the $S \rightarrow T$ portion of the corpus, and one for the $T \rightarrow S$ portion into a mixture model and use perplexity minimization [Sennrich, 2012] to set the model weights. We show that this combination significantly outperforms a simple union of the two portions of the parallel corpus. Additionally, we use the entire corpus, create a single, unified phrase table and then use the statistical measures mentioned above, and in particular *cross-entropy*, as a clue for selecting phrase-pairs from this table. The benefit of this method is that not only does it yield the best results, but it also eliminates the need to directly predict the direction of translation of the parallel corpus.

3.2 Baseline Experiments

3.2.1 Europarl Experiments

The task we focus on in our experiments is translation from French to English (FR-EN) and from English to French (EN-FR). However, to establish the robustness of our approach, we also conduct experiments with other translation tasks, including German–English (DE-EN), English–German (EN-DE), Italian–English (IT-EN) and English–Italian (EN-IT). Our main corpus is Europarl [Koehn, 2005], specifically portions collected over the years 1996 to 1999 and 2001 to 2009. This is a large multilingual corpus, containing sentences translated from several European languages. In most cases the corpus is annotated with the original language and the name of the speaker. For each language pair we extract from the multilingual corpus two subsets, corresponding to the original languages in which the sentences were produced. For example, in the case of FR-EN we extract from our corpus all sentences produced in French and translated into English, and all sentences produced

in English and translated into French. All sentences are lowercased and tokenized using Moses [Koehn et al., 2007]. Sentences longer than 80 words are discarded. Table 3.1 depicts the size of the subsets.

	Original language	#Sentence	#Tokens
FR-EN	French	168,818	4,995,397
	English	134,318	3,441,120
DE-EN	German	200,037	5,571,202
	English	129,309	3,283,298
IT-EN	Italian	69,270	2,535,225
	English	125,640	3,389,736

Table 3.1: Europarl corpus size, in sentences and tokens

We use each subset to train two phrase-based statistical machine translation (PB-SMT) systems [Koehn et al., 2007], translating in both directions between the languages in each language pair. In other words, we train two PB-SMTs for each translation task, each based on a parallel corpus produced and translated in a different direction. We use GIZA++ [Och and Ney, 2000b] with *grow-diag-final* alignment, and extract phrases of length up to 10 words. We prune the resulting phrase tables as in Johnson et al. [2007], using at most 30 translations per source phrase and discarding singleton phrase pairs.

We use all Europarl corpora between the years 1996 to 1999 and 2001 to 2009 to construct English, German, French and Italian 5-gram language models, using interpolated modified Kneser-Ney discounting [Chen, 1998] and no cut-off on all n-grams. We use a specific symbol to mark out-of-vocabulary words (OOVs). We use the portion of Europarl collected over year 2000 for tuning and evaluation. For each translation task we randomly extract 1,000 parallel sentences for the tuning set and another set of 5,000 parallel sentences for evaluation. The sentences are originally written in the translation task's source language and are translated into the translation task's target language. We use the MERT algorithm [Och, 2003] for tuning and BLEU [Papineni et al., 2002] as our evaluation metric. We test the statistical significance of the differences between the results using the bootstrap resampling method [Koehn, 2004].

A word on notation: We use '$S \rightarrow T$' when the translation direction of the parallel corpus corresponds to the translation task and '$T \rightarrow S$' when a corpus is translated in the opposite direction to the translation task. For example, suppose the translation tasks are

English-to-French (E2F) and French-to-English (F2E). We use '$S \rightarrow T$' when the French-original corpus is used for the F2E task or when the English-original corpus is used for the E2F task; and '$T \rightarrow S$' when the French-original corpus is used for the E2F task or when the English-original corpus is used for the F2E task.

Table 3.2 depicts the BLEU scores of the SMT systems. The data are consistent with the findings of Kurokawa et al. [2009]: Systems trained on $S \rightarrow T$ parallel texts always outperform systems trained on $T \rightarrow S$ texts. The difference in BLEU score can be as high as 3 points.

Task	$S \rightarrow T$	$T \rightarrow S$
FR-EN	33.64	30.88
EN-FR	32.11	30.35
DE-EN	26.53	23.67
EN-DE	16.96	16.17
IT-EN	28.70	26.84
EN-IT	23.81	21.28

Table 3.2: BLEU scores of the Europarl baseline systems

3.2.2 Hansard Experiments

The corpora used in these experiments are small (up to 200,000 sentences). Also, the ratio between $S \rightarrow T$ and $T \rightarrow S$ materials varies greatly for different language pairs. To mitigate these issues we use the Hansard corpus, containing transcripts of the Canadian parliament from 1996–2007, as another source of parallel data. The Hansard is a bilingual French–English corpus comprising approximately 80% English-original texts and 20% French-original texts. Crucially, each sentence pair in the corpus is annotated with the direction of translation.

To address the effect of corpus size, we compile six subsets of different sizes (250K, 500K, 750K, 1M, 1.25M and 1.5M parallel sentences) from each portion (English-original and French-original) of the corpus. Additionally, we use the *devtest* section of the Hansard corpus to randomly select French-original and English-original sentences that are used for tuning (1,000 sentences each) and evaluation (5,000 sentences each).

On these corpora we train twelve French-to-English and twelve English-to-French PB-

SMT systems using the MOSES toolkit [Koehn et al., 2007]. We use the same GIZA++ configuration and phrase table pruning as in the Europarl experiments. We also reuse the English and French language models. French-to-English MT systems are tuned and tested on French-original sentences and English-to-French systems on English-original ones.

Table 3.3 depicts the BLEU scores of the Hansard systems. The data are consistent with our previous findings: systems trained on $S \to T$ parallel texts always outperform systems trained on $T \to S$ texts, even when the latter are much larger. For example, a French-to-English SMT system trained on 250,000 $S \to T$ sentences outperforms a system trained on 1,500,000 $T \to S$ sentences.

Task: French-to-English		
Corpus subset	$S \to T$	$T \to S$
250K	34.35	31.33
500K	35.21	32.38
750K	36.12	32.90
1M	35.73	33.07
1.25M	36.24	33.23
1.5M	36.43	33.73

Task: English-to-French		
Corpus subset	$S \to T$	$T \to S$
250K	27.74	26.58
500K	29.15	27.19
750K	29.43	27.63
1M	29.94	27.88
1.25M	30.63	27.84
1.5M	29.89	27.83

Table 3.3: BLEU scores of the Hansard baseline systems

3.3 Phrase Tables Reflect Facets of Translationese

The baseline results suggest that $S \to T$ and $T \to S$ phrase tables differ substantially, presumably due to the different characteristics of original and translated texts. In this section we explain the better translation quality in terms of the better quality of the respective phrase tables, as defined by a number of statistical measures. We first relate these measures to the unique properties of translationese.

Translated texts tend to be simpler than original ones along a number of criteria. Generally, translated texts are not as rich and variable as original ones, and in particular, their type/token ratio is lower. Consequently, we expect $S \to T$ phrase tables (which are based on a parallel corpus whose source is original texts, and whose target is translationese) to have more unique source phrases and a lower number of translations per source phrase. A large number of unique source phrases suggests better coverage of the source text, while

a small number of translations per source phrase means a lower phrase table entropy.

These expectations are confirmed, as the data in Table 3.4 show. We report the total size of the phrase table in tokens ('Total'), the number of unique source phrases ('Source'), and the average number of translations per source phrase ('AvgTran'), computed on the twenty four phrase tables corresponding to our SMT systems.[1] Evidently, $S \rightarrow T$ phrase tables have more unique source phrases, but fewer translation options per source phrase. This holds uniformly for all twenty four tables.

Task: French-to-English						
Set	$S \rightarrow T$			$T \rightarrow S$		
	Total	Source	AvgTran	Total	Source	AvgTran
250K	231K	69K	3.35	199K	55K	3.65
500K	360K	86K	4.21	317K	69K	4.56
750K	461K	96K	4.81	405K	78K	5.19
1M	544K	103K	5.27	479K	85K	5.66
1.25M	619K	109K	5.66	545K	90K	6.07
1.5M	684K	114K	6.01	602K	94K	6.43
Task: English-to-French						
Set	$S \rightarrow T$			$T \rightarrow S$		
	Total	Source	AvgTran	Total	Source	AvgTran
250K	224K	49K	4.52	220K	46K	4.75
500K	346K	61K	5.64	334K	57K	5.82
750K	437K	68K	6.39	421K	64K	6.54
1M	513K	74K	6.95	489K	69K	7.10
1.25M	579K	78K	7.42	550K	73K	7.56
1.5M	635K	81K	7.83	603K	76K	7.92

Table 3.4: Statistic measures computed on the phrase tables: total size, in tokens ('Total'); the number of unique source phrases ('Source'); and the average number of translations per source phrase ('AvgTran')

A well-established tool for assessing the quality of a phrase table involves entropy-based measures. *Phrase table entropy* captures the amount of uncertainty involved in choosing candidate translation phrases [Koehn et al., 2009]. Given a source phrase s and a phrase

[1] The phrase tables were pruned, retaining only phrases that are included in the evaluation set.

table T with translations t of s whose probabilities are $p(t \mid s)$, the entropy H of s is:

$$H(s) = -\sum_{t \in T} p(t \mid s) \times log_2 p(t \mid s) \tag{3.1}$$

To compute the phrase table entropy, Koehn et al. [2009] search through all possible segmentations of the source sentence to find the optimal covering set of test sentences that minimizes the average entropy of the source phrases in the covering set. We refer to this measure as *covering set entropy*, or *CovEnt*.

We also propose a metric that assesses the quality of the *source* side of a phrase table. This metric finds the minimal covering set of a given text in the source language using source phrases from a particular phrase table, and outputs the average length of a phrase in the covering set. This measure is referred to as *covering set average length*, or *CovLen*.

In Chapter 2 we show that *perplexity* distinguishes well between translated and original texts. Moreover, perplexity can reflect the degree of 'relatedness' of a given phrase to original language or to translationese. Motivated by this observation, we design a cross-entropy-based measure that assesses how well each phrase table fits the register of translationese.

We then build language models from translated texts, and compute the cross-entropy of each target phrase in the phrase tables according to these language models.

Given a language model L, the cross-entropy of a text $w = w_1, w_2, \cdots w_N$ is:

$$H(w, L) = -\frac{1}{N} \sum_{i=1}^{N} log_2 L(w_i) \tag{3.2}$$

We build language models of translated texts as follows. For English translationese, we extract 170,000 French-original sentences from the English portion of Europarl, and 3,000 English-translated-from-French sentences from the Hansard corpus (disjoint from the training, development and test sets, of course). We use each corpus to train a trigram language model with interpolated modified Kneser-Ney discounting and no cut-off. All out-of-vocabulary words are mapped to a special token, $\langle unk \rangle$. Then, we interpolate the Hansard and Europarl language models to minimize the perplexity of the target side of the development set ($\lambda = 0.58$). For French translationese, we use 270,000 sentences from Europarl and 3,000 sentences from Hansard, $\lambda = 0.81$.

Similarly to covering set entropy, *covering set cross-entropy (CovCrEnt)* finds the optimal covering set of test sentences that minimizes the *weighted cross-entropy* of the source

phrase in the covering set. Given a phrase table T and a language model L, the weighted cross-entropy W for a source phrase s is:

$$W(s, L) = -\sum_{t \in T} H(t, L) \times p(t \mid s) \tag{3.3}$$

where $H(t, L)$ is the cross-entropy of t according to a language model L.

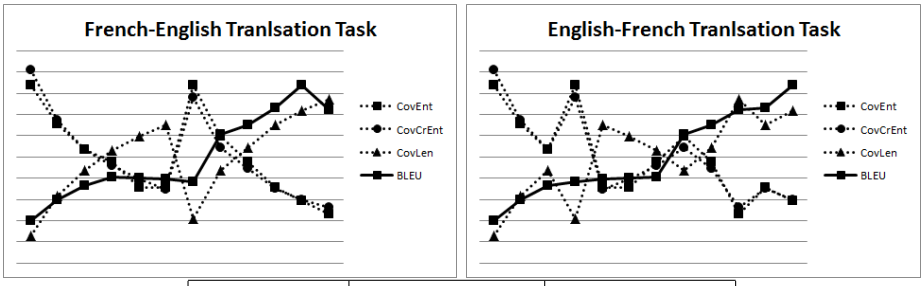
Table 3.5 lists the entropy-based measures computed on our twenty four phrase tables. Again, the data meet our expectations: $S \rightarrow T$ phrase tables uniformly and unexceptionally have lower entropy and cross-entropy, but higher covering set length.

Task: French-to-English						
Set	$S \rightarrow T$			$T \rightarrow S$		
	CovEnt	CovCrEnt	CovLen	CovEnt	CovCrEnt	CovLen
250K	0.36	1.64	2.44	0.45	1.87	2.25
500K	0.35	1.30	2.64	0.43	1.52	2.42
750K	0.35	1.10	2.77	0.43	1.35	2.53
1M	0.34	0.99	2.85	0.42	1.21	2.61
1.25M	0.34	0.91	2.92	0.41	1.12	2.67
1.5M	0.33	0.85	2.97	0.41	1.07	2.71
Task: English-to-French						
Set	$S \rightarrow T$			$T \rightarrow S$		
	CovEnt	CovCrEnt	CovLen	CovEnt	CovCrEnt	CovLen
250K	0.63	1.88	2.08	0.63	2.09	2.02
500K	0.59	1.49	2.25	0.60	1.70	2.16
750K	0.57	1.33	2.33	0.58	1.48	2.25
1M	0.55	1.18	2.41	0.57	1.35	2.32
1.25M	0.54	1.09	2.46	0.55	1.25	2.37
1.5M	0.53	1.03	2.50	0.55	1.17	2.41

Table 3.5: Entropy-based measures computed on the phrase tables: covering set entropy ('CovEnt'); covering set cross-entropy ('CovCrEnt'); and covering set average length ('CovLen')

So far, we have established the hypothesis that $S \rightarrow T$ phrase tables better reflect the register of translationese than $T \rightarrow S$ ones. But does this necessarily affect the quality of the generated translations? To verify that, we measure the correlation between the

quality of the translation, as measured by BLEU (Table 3.3), with each of the entropy-based metrics. We compute the correlation coefficient R^2 (the square of Pearson's product-moment correlation coefficient) by fitting a simple linear regression model. Table 3.6 lists the results; clearly, all three measures are strongly correlated with translation quality. Consequently, we use these measures as indicators of better translations, more similar to translationese. Crucially, these measures are computed directly on the phrase table, and do not require reference translations or meta-information pertaining to the direction of translation of the parallel phrase.



Measure	R^2 (FR–EN)	R^2 (EN-FR)
CovEnt	**0.94**	0.46
CovCrEnt	0.56	0.54
CovLen	0.75	**0.56**

Table 3.6: Correlation of BLEU scores with phrase table statistical measures

3.4 Adaptation of the Translation Model to Translationese

We have thus established the fact that $S \rightarrow T$ phrase tables have an advantage over $T \rightarrow S$ ones that stems directly from the different characteristics of original and translated texts. We have also identified three statistical measures that explain most of the variability in translation quality. We now explore ways for taking advantage of the *entire* parallel corpus, including translations in *both* directions, in light of the above findings. Our goal is to establish the best method to address the issue of different translation direction components in the parallel corpus.

3.4.1 Baseline

As a simple baseline we take the union of the two subsets of the parallel corpus. This gives the decoder an opportunity to select phrases from either subset of the corpus, and

MERT can be expected to optimize this selection process. For each translation task in Section 3.2.1, we concatenate the $S \rightarrow T$ and the $T \rightarrow S$ subsets of the parallel corpora and use the union to train an SMT system (henceforth *UNION*). We use the same language and reordering models, Moses configuration and the same tuning and evaluation sets as in Section 3.2.1. Table 3.7 reports the results. The 'UNION' systems, which use twice as much training data as the $S \rightarrow T$ systems, outperform the $S \rightarrow T$ systems for all language pairs except English-to-Italian. However, only in three cases out of six (German-to-English, English-to-German and Italian-to-English) is the gain statistically significant. Nevertheless, this indicates that the $T \rightarrow S$ subset contains useful material that can (and does) contribute to translation quality.

System	**FR-EN**	**EN-FR**	**DE-EN**	**EN-DE**	**IT-EN**	**EN-IT**
$S \rightarrow T$	33.64	32.11	26.53	16.96	28.70	23.81
UNION	33.79	32.24	26.76	17.36	29.12	23.70
MULTI-PATH	33.81	31.95	26.68	17.39	29.11	23.80
PPLMIN-1	33.86	32.47	**26.83**	**17.80**	29.23	23.86
PPLMIN-2	**33.95**	**32.65**	26.77	17.65	**29.44**	**24.01**

Table 3.7: Evaluation results of various ways for combining phrase tables

We now look at ways to better utilize this portion. First, we train SMT systems with two phrase tables using multiple decoding paths, and combine them in a log-linear model, following Koehn and Schroeder [2007]. The performance of this approach (referred to as *MULTI-PATH*) is either lower or only slightly better that that of the UNION systems (Table 3.7).

3.4.2 Perplexity Minimization

Next, we look at a linear interpolation of the translation models. We need a way to tune the weights of the translation model components, and we use *perplexity minimization*, following Sennrich [2012].

Given n phrase tables, we are looking for a set of n weights $\lambda = \lambda_1, \ldots, \lambda_n$, such that $\sum_{i=1}^{n} \lambda_i = 1$, where λ_i is the interpolation weight of phrase table i. Then, given a phrase

pair (s,t), the linear interpolation of the n models is given by:

$$p(s \mid t; \lambda) = \sum_{i=1}^{n} \lambda_i p(s \mid t) \tag{3.4}$$

To adapt an interpolated translation model to a specific (development) corpus, let $\tilde{p}(s,t)$ be the observed, empirical probability of the pair (s,t) in the development corpus. To obtain the phrase pairs, we process the development set with the same word alignment and phrase extraction tools (with the same configuration, including the maximum phrase size) that we use for training, i.e. GIZA++ and heuristics for phrase extraction. Then the cross entropy H of a translation model with probabilities p to a development corpus with probabilities $\tilde{p}$ is defined as:

$$H = -\sum_{s,t} \tilde{p}(s,t) \times log_2 p(s \mid t) \tag{3.5}$$

To minimize the cross entropy, we look for a weight vector $\hat{\lambda}$ such that:

$$\hat{\lambda} = \arg\min_{\lambda} -\sum_{s,t} \tilde{p}(s,t) \times log_2 \left(\sum_{i=1}^{n} \lambda_i p(s \mid t) \right) \tag{3.6}$$

Each feature of the standard SMT translation model (the phrase translation probabilities $p(t \mid s)$ and $p(s \mid t)$, and the lexical weights $lex(t \mid s)$ and $lex(s \mid t)$) is optimized independently. The lambda values are set using L-BFGS with numerically approximated gradients [Byrd et al., 1995].

This technique is particularly appealing for us due to two reasons: first, in Chapter 2 we show that perplexity is a good differentiator between original and translated texts; second, the perplexity is minimized with respect to some development set. Consequently, if we use a $S \rightarrow T$ corpus for this purpose, we directly adapt the interpolated phrase table to the qualities of the $S \rightarrow T$ translation models as described in Section 3.3. We use this technique to interpolate two pairs of phrase tables: We interpolate the $S \rightarrow T$ and the $T \rightarrow S$ models (we refer to this system as '*PPLMIN-1*') and we also interpolate the $S \rightarrow T$ with the UNION models ('*PPLMIN-2*'), as a simple way of upweighting. Table 3.7 reports the results. In all cases, the interpolated systems yield higher BLEU scores than the simple UNION systems. While the improvements are small (0.2-0.4 BLEU points), they are statistically significant in all cases, except for German-English. Clearly, the interpolated systems outperform the $S \rightarrow T$ systems by 0.2-0.7 BLEU points (statistically significant in all cases). PPLMIN-2 seems to be better than PPLMIN-1 in four out of six systems.

To verify that the improvement in translation quality is due to the adaptation process rather than a quirk resulting from MERT instability, we use *MultEval* [Clark et al., 2011]. This is a script that takes machine translation hypotheses from several runs of an optimizer (MERT) and reports three popular metric scores: BLEU, Meteor [Denkowski and Lavie, 2011] and TER [Snover et al., 2006]. Meteor and BLEU scores are higher for better translations (↑), TER is a lower-is-better measure (↓). In addition, MultEval computes the ratio of output length to reference length (closer to 100% is better), as well as p-values (via approximate randomization). We use MultEval to compare translation hypotheses of the UNION and PPLMIN-2 systems. Table 3.8 depicts the results for French-to-English and English-to-French (other translation tasks produce similar results). The improvement of the adapted systems is clear and robust.

Metric	**System**	**Avg**	p**-value**
French-to-English			
BLEU ↑	UNION	33.7	-
	PPLMIN-2	33.9	0.0001
METEOR ↑	UNION	35.7	-
	PPLMIN-2	35.8	0.0001
TER ↓	UNION	49.7	-
	PPLMIN-2	49.5	0.0001
Length	UNION	99.4	-
	PPLMIN-2	99.5	0.0003
English-to-French			
BLEU ↑	UNION	32.3	-
	PPLMIN-2	32.6	0.0001
METEOR ↑	UNION	53.8	-
	PPLMIN-2	54.0	0.0001
TER ↓	UNION	52.6	-
	PPLMIN-2	52.5	0.004
Length	UNION	98.7	-
	PPLMIN-2	98.9	0.0001

Table 3.8: MultEval scores for UNION and PPLMIN-2 systems

3.4.3 Adaptation without Classification

A prerequisite for interpolating translation models, the method we advocate above, is that the direction of translation of every sentence pair in the parallel corpus be known in advance. When such information is not available, machine learning can automatically classify texts as original or translated [van Halteren, 2008, Baroni and Bernardini, 2006, Ilisei et al., 2010, Koppel and Ordan, 2011]. Naturally, however, the quality of the interpolation of translation models trained on classified (rather than annotated) data is expected to decrease. In this section we establish an adaptation technique that does not rely on explicit information pertaining to the direction of translation, but rather uses perplexity-based measures to evaluate the 'relatedness' of a specific phrase to an original or a translated language "dialect".

For the following experiments we use the Hansard corpus described in Section 3.2.2; *FO* refers to subsets of the parallel corpus that were translated from French to English, *EO* refers to texts translated from English to French. We create three different mixtures of FO and EO: a balanced mix comprising 500K sentences each of FO and EO ('MIX'), an EO-biased mix with 500K sentences of FO and 1M sentences of EO ('MIX-EO'), and an FO-biased mix with 1M sentences of FO and 500K sentences of EO ('MIX-FO'). We use these corpora to train French-to-English and English-to-French MT systems, evaluating their quality on the evaluation sets described in Section 3.2.2. We use the same Moses configuration as well as the same language and reordering models as in Section 3.2.2.

Now, we adapt the translation models by adding to each phrase pair in the phrase tables an additional factor, as a measure of its fitness to the regsiter of translationese. We experiment with two such factors. First, we use the language models described in Section 3.3 to compute the cross-entropy of each translation option according to this model. We add cross-entropy as an additional score of a translation pair that can be tuned by MERT (we refer to this system as *CrEnt*). Since cross-entropy is 'the lower the better' metric, we adjust the range of values used by MERT for this score to be negative.

Second, following Moore and Lewis [2010], we define an adapting feature that not only measures how close phrases are to translated language, but also how far they are from original language, and use it as a factor in a phrase table (this system is referred to as *PplRatio*). We build two additional language models of original texts as follows. For original English, we extract 135,000 English-original sentences from the English portion

of Europarl, and 2,700 English-original sentences from the Hansard corpus. We train a trigram language model with interpolated modified Kneser-Ney discounting on each corpus and we interpolate both models to minimize the perplexity of the source side of the development set for the English-to-French translation task ($\lambda = 0.49$). For original French, we use 110,000 sentences from Europarl and 2,900 sentences from Hansard, $\lambda = 0.61$. Finally, for each target phrase t in the phrase table we compute the ratio of the perplexity of t according to the original language model L_o and the perplexity of t with respect to the translated model L_t (see Section 3.3). In other words, the factor F is computed as follows:

$$F(t) = \frac{H(t, L_o)}{H(t, L_t)} \tag{3.7}$$

We apply these techniques to the French-to-English and English-to-French phrase tables built from the concatenated corpora, and use each phrase table to train an SMT system. We compare the performance of these systems to that of $S \rightarrow T$, UNION and both PPLMIN systems. Table 3.9 summarizes the results.

All systems outperform the corresponding UNION systems. 'CrEnt' systems show significant improvements ($p < 0.05$) on balanced scenarios ('MIX') and on scenarios biased towards the $S \rightarrow T$ component ('MIX-FO' in the French-to-English task, 'MIX-EO' in English-to-French). 'PplRatio' systems exhibit more consistent behavior, showing small, but statistically significant improvement ($p < 0.05$) in all scenarios. Additionally, the new systems perform quite competitively compared to the interpolated systems, wining in four out of six cases. Note again that all systems in the same column (except $S \rightarrow T$) are trained on exactly the same corpus and have exactly the same phrase tables. The only difference is an additional factor in the phrase table that "encourages" the decoder to select translation options that are closer to translated texts than to original ones.

3.5 Analysis

We have demonstrated that SMT systems that are sensitive to the direction of translation perform better. The superior quality of SMT systems that are adapted to translationese is reflected in higher BLEU scores, but also in the scores of other automatic measures for evaluating the quality of machine translation output. In this section we analyze the bet-

Task: French-to-English			
System	**MIX**	**MIX-EO**	**MIX-FO**
$S \rightarrow T$	35.21	35.21	35.73
UNION	35.27	35.36	35.94
PPLMIN-1	35.46	35.59	36.26
PPLMIN-2	**35.75**	35.65	36.20
CrEnt	35.54	35.45	**36.75**
PplRatio	35.59	**35.78**	36.22
Task: English-to-French			
System	**MIX**	**MIX-FO**	**MIX-EO**
$S \rightarrow T$	29.15	29.15	29.94
UNION	29.27	29.44	30.01
PPLMIN-1	29.64	29.94	29.65
PPLMIN-2	29.50	**30.45**	30.12
CrEnt	29.47	29.45	**30.44**
PplRatio	**29.65**	29.62	30.34

Table 3.9: Adaption without classification results

ter performance of translationese-adapted systems, both quantitatively and qualitatively, relating it to established insights in Translation Studies.

3.5.1 Quantitative Analysis

Is the output of translationese-adapted systems indeed more similar to translationese? We begin with a set of properties of translationese that are easy to compute, and evaluate the output of our translationese-adapted SMT systems in terms of these properties.

Type-Token Ratio

Translated texts have been shown to have lower type-to-token ratio (TTR) than original ones [Al-Shabab, 1996]. Figure 3.1 compares the TTR of the translation outputs of $S \rightarrow T$, $T \rightarrow S$, UNION and PPLMIN-2 systems. Obviously, the TTR of $S \rightarrow T$ output is much lower than $T \rightarrow S$ system. Recall that $S \rightarrow T$ systems produce markedly better translations than $T \rightarrow S$ ones, so indeed there is a clear correspondence between the TTR

of the outputs and better translation quality. Figure 3.1 also compares the TTR of the outputs produced from two combination systems, UNION and PPLMIN-2. The UNION outputs are arbitrary: Their TTR is sometimes lower than the corresponding $S \rightarrow T$ system, but sometimes higher than even the corresponding $T \rightarrow S$ system. In contrast, PPLMIN-2 systems (which are the best adapted systems) systematically produce outputs with the lowest TTR, i.e., outputs closest to translationese.

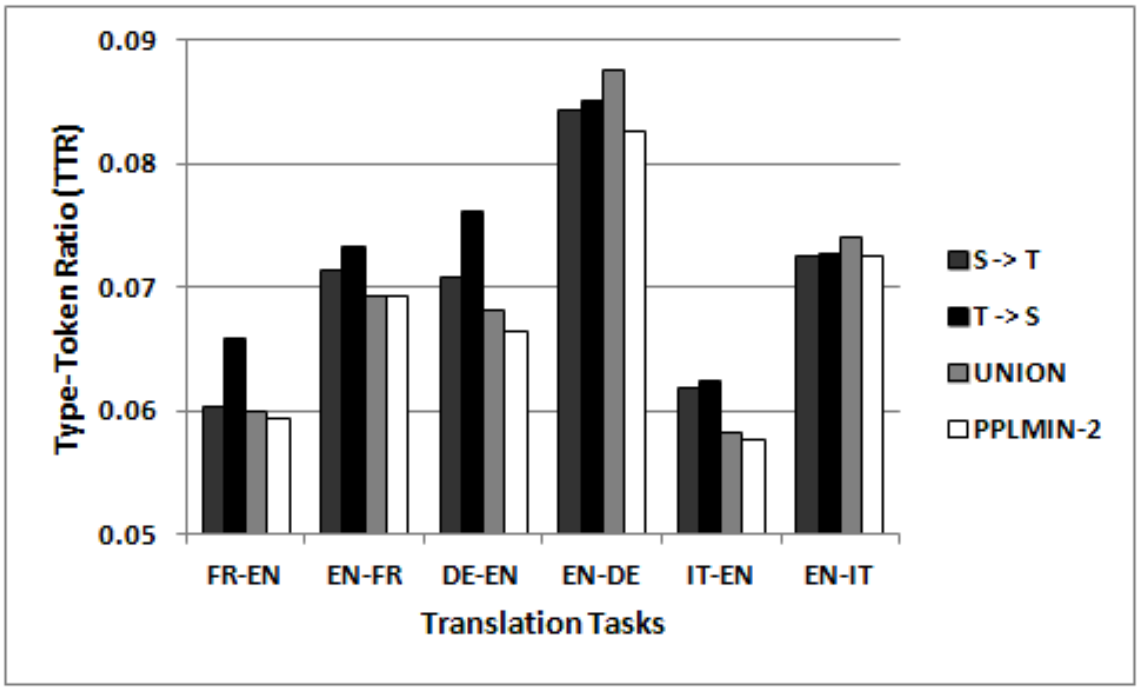


Figure 3.1: Type-token ratio in SMT translation outputs

Hapax Legomena

A related property of translated texts is that they tend to exhibit a much lower rate of *hapax legomena* (words that occur only once in a text) than original texts. We thus count the number of hapaxes in the outputs of each of the SMT systems. The results, which are depicted in Figure 3.2, are not totally conclusive, but are interesting nonetheless. Specifically, in all cases the PPLMIN-2 system exhibits a lower number of hapaxes than the UNION system; and in all systems except the English-Italian one, the number of hapaxes produced by the PPLMIN-2 system is lowest.

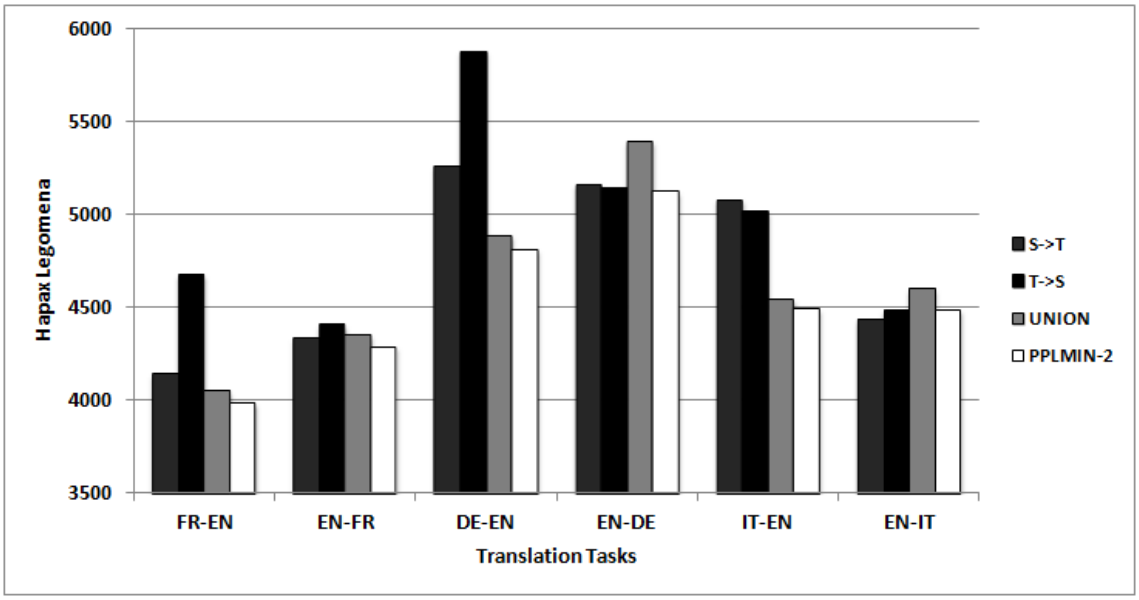


Figure 3.2: Numbers of hapax legomena in SMT translation outputs

Entropy

As another quantitative measure of the contribution of perplexity minimization as a method of adaptation, we list in Table 3.10 the values of the entropy-based measures discussed in Section 3.3, on three types of SMT systems: those compiled from $S \rightarrow T$ texts only, UNION, and PPLMIN-2 ones. Observe that the covering set cross-entropy measure, designed to reflect the fitting of a phrase table's target side to translated texts, is significantly lower in PPLMIN-2 systems than in $S \rightarrow T$ and UNION systems. This indicates that perplexity minimization improves the system's fitness to translationese. Interestingly, the PPLMIN-2 systems have better lexical coverage than the UNION systems. Table 3.10 lists data for French-English and English-French, but other language pairs exhibit similar behavior.

Mean Occurrence Rate

Original texts are known to be lexically richer than translated ones; in particular, translationese uses more frequent and common words [Laviosa, 1998]. To assess the lexical diversity of a given text we define Mean Occurrence Rate (logMOR). logMOR computes the average number of occurrences of tokens in the text with respect to a large reference corpus. Consequently, sentences containing more frequent words have higher logMOR scores. More formally, given a reference corpus R with n word types $r_1 \cdots r_n$, let $C(r_i)$ be

	System	CovEnt	CovCrEnt	CovLen
FR-EN	$S \rightarrow T$	0.43	2.39	2.24
	UNION	0.43	2.20	2.34
	PPLMIN-2	0.43	**2.14**	**2.35**
EN-FR	$S \rightarrow T$	0.64	3.47	2.01
	UNION	0.61	3.09	2.17
	PPLMIN-2	0.61	**2.99**	**2.18**

Table 3.10: Entropy-based measures, computed on phrase tables of baseline and adapted SMT systems

a number of occurrences of the word r_i in the corpus R. Then the logMOR of a sentence $S = s_1 \cdots s_k$ is:

$$logMOR(S) = \frac{1}{k} \sum_{i=1}^{k} log(C(s_i)) \tag{3.8}$$

$C(s_i)$ is calculated from the corpus R if $s_i \in R$. Otherwise, $C(s_i) = \alpha$, where α is a predefined constant depending on the size of the reference corpus. In all our experiments we use $\alpha = 0.5$.

In order to establish the relation between the logMOR measure and translation quality, we compute logMOR scores for each sentence of an SMT system output. Then, we sort the output sentences based on their logMOR scores, split the output into two parts, below and above the median of logMOR, and calculate BLEU score for each portion independently. We perform these calculations on the outputs of UNION and PPLMIN-2 SMT systems for all our translation tasks. We use the Europarl corpus [Koehn, 2005] as a reference for a list of occurrences. Table 3.11 depicts the results. In all cases, the bottom part (below the median) of SMT outputs has significantly lower BLEU scores (up to 5 BLUE points!) than the upper part, indicating that the logMOR measure is a good (post factum) differentiator between poor and good translations.

We now compute the average logMOR score on the outputs of all our SMT systems. Figure 3.3 shows the results. In all cases (except Italian to English), $S \rightarrow T$ is better than $T \rightarrow S$; and in all systems except EN-FR, PPLMIN-2 is best.

Task	UNION		PPLMIN-2	
	Bottom	Upper	Bottom	Upper
DE-EN	24.05	28.72	24.10	28.71
EN-DE	16.06	18.78	16.42	18.82
FR-EN	31.48	35.49	31.85	35.49
EN-FR	28.97	34.83	29.30	35.58
IT-EN	26.07	31.75	26.43	31.97
EN-IT	21.57	25.79	21.97	25.99

Table 3.11: BLEU scores computed on portions of UNION and PPLMIN-2 systems outputs below and above the logMOR median

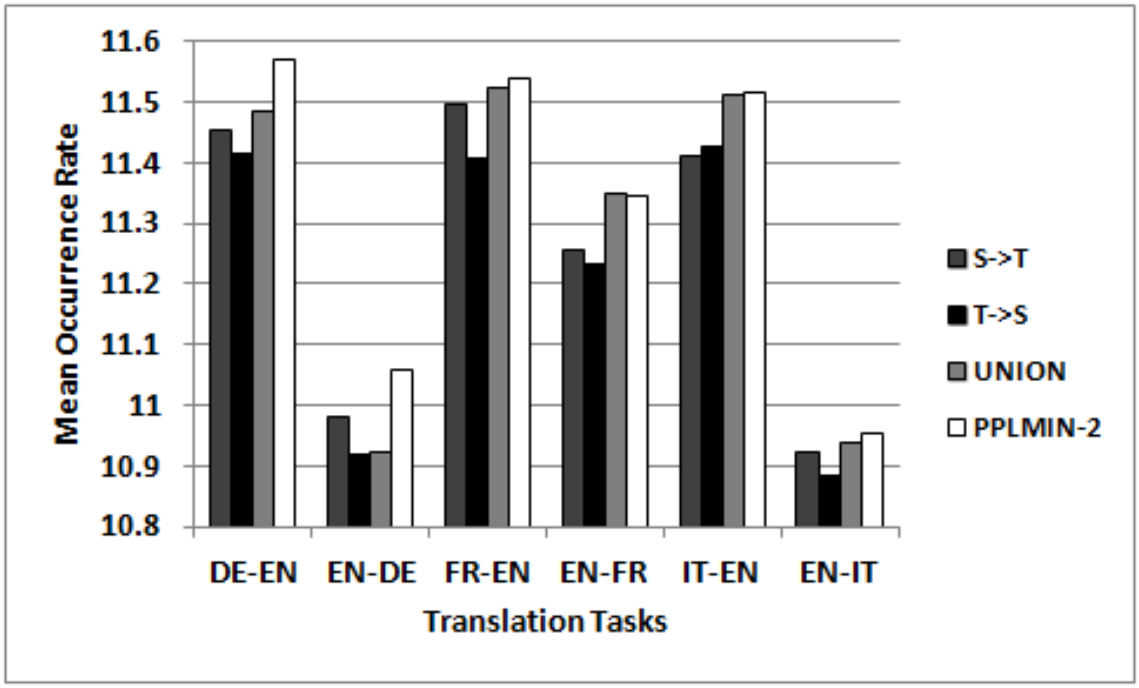


Figure 3.3: Mean Occurrence Rate in SMT translation outputs

3.5.2 Qualitative Analysis

Translation is sometimes described as an attempt to strike a balance between *interference*, the so-called inevitable marks left by the source language on the target text, and *standardization*, the attempt of the translator to *adapt* the translation product to the target language and culture, to break away from the source text towards a more adequate text [Toury, 1995b]. In order to study the effect of the adaptation qualitatively, rather than quantitatively, we focus on several concrete examples. We compare translations pro-

duced by the UNION (henceforth *baseline*) and by the PPLMIN-2 (henceforth *adapted*) French-English Europarl systems. We selected 200 sentences of from the French-English evaluation set for manual inspection, focusing on sentences in which the translations were significantly different from each other. Indeed, we find that the translations are better adapted along several dimensions.

In the following sentences, the baseline follows a more literal translation, whereas the adapted system creates a more adequate, standardized translation.

Source *Monsieur le président, chers collègues, les tempêtes qui ont ravagá la france dans la nuit des 26 et 27 décembre ont fait, on l'a dit, 90 morts, 75 milliards de francs, soit 11 milliards d'euros, de dégâts.*

Baseline *Mr president, ladies and gentlemen, storms that have* ***ravaged*** *france during the night of 26 and 27 december were, as has been said, 90 people dead, 75 billion francs, that is, eur 11 billion, damage.*

Adapted *Mr president, ladies and gentlemen, the storms which have* ***devastated*** *france during the night of 26 and 27 december were, as has been said, 90 people dead, 75 billion francs, or eur 11 billion, damage.*

Source *Tout d'abord, je tiens à saluer tous mes collègues maires, élus locaux, qui, au quotidien, ont dû rassurer la population, organiser la solidarité, coopérer avec les services publics.*

Baseline *First of all, I should like to pay tribute to all my colleagues mayors, local elected representatives, who,* ***in their daily lives****, have had to reassure the population, organise solidarity, cooperate with public services.*

Adapted *First of all, I should like to pay tribute to all my colleagues mayors, local elected representatives, who,* ***on a daily basis****, have had to reassure the population, organise solidarity, cooperate with public services.*

Source *Monsieur le président, je vous remercie de me laisser conclure, et je rappellerai simplement une maxime: "les tueurs en série se font toujours prendre par la police quand ils accélèrent la cadence de leurs crimes".*

Baseline *Mr president, thank you for allowing me to* ***leave conclusion****, and I would like*

to remind you just a maxim: 'the murderers in series are always take by the police when they accélèrent the pace of their crimes'.

Adapted *Mr president, thank you for letting me* ***finish****, and I would like to remind you just a maxim: 'the murderers in series are always take by the police when they accélèrent the pace of their crimes'.*

Note that the baseline is not necessarily incomprehensible, nor even 'impossible' in the target language; in the first example, it is clear what is meant by *storms that have ravaged France*, and moreover, we find such expressions in a 1.5G token-sized corpus [Ferraresi et al., 2008]; it is just half as likely as what is offered by the Adapted System. The second example, on the other hand, misses the point altogether, and the third one is a clear case of interference, where the French *laisser conclure* is transferred verbatim as *leave conclusion*.

Another difference between the two systems is reordering. Sometimes, as in the two examples below, the inability of the Baseline System to reorder the words correctly stems from interference:

Source *Madame la présidente, mes chers collègues, nous croyions, jusqu'à présent, que l'union européenne était, selon les dispositions des traités de rome et de paris qui avaient fondé les communautés, devenues union, une association d'états libres, indépendants et souverains.*

Baseline *Madam president, ladies and gentlemen, we croyions, up to now, that the european union is, according to the provisions of the treaties of rome and paris who had based the communities, become union,* ***an association of states free, independent and sovereign****.*

Adapted *Madam president, ladies and gentlemen, we croyions, up to now, that the european union was, according to the provisions of the treaties of paris and rome who had based communities, become union,* ***an association of free, sovereign and independent states****.*

Source *La convention de lomé bénéficie essentiellement à quelques grands groupes industriels ou financiers qui continuent á piller ces pays et perpétuent leur dépendance économique, notamment des anciennes puissances coloniales.*

Baseline *The lomé convention has mainly to a few* ***large industrial groups or financial*** *which continue to plunder those countries and perpétuent their economic dependence, in particular the former colonial powers.*

Adapted *The lomé convention has mainly to a few* ***large financial and industrial groups*** *which continue to plunder those countries and perpetuate their economic dependence, in particular the former colonial powers.*

Additionally, the Adapted System produces much better collocations. Compare the 'natural' expressions *pay a high price* and *express the concern* with the Baseline System products:

Source *Ces hommes et ces femmes qui bougent à travers l'europe paient leur voiture, leurs taxes nationales, leur pot catalytique, leurs taxes sur les carburants, et paient donc déjà très cher le prix de la magnifique machine et la liberté de circuler.*

Baseline *These men and women who are moving across europe are paying their car, their national taxes, their catalytic converter, their taxes on fuel, and therefore already* ***pay very dearly*** *for the price of the magnificent machine and freedom of movement.*

Adapted *These men and women who are moving across europe pay their car, their national taxes, their catalytic converter, their taxes on fuel, and therefore already* ***pay a high price*** *for the magnificent machine and freedom of movement.*

Source *Je veux dire également le souci que j'ai d'une bonne coopération entre interreg et le fed, notamment pour les caraïbes et l'océan indien.*

Baseline *I would like to* ***say to the concern*** *that I have good cooperation between interreg and the edf, particularly for the caribbean and the indian ocean.*

Adapted *I also wish to* ***express the concern*** *that I have good cooperation between interreg and the edf, particularly for the caribbean and the indian ocean.*

Last, there are a few cases of explicitation. Blum-Kulka [1986] observed the tendency of translations to introduce to the target texts cohesive markers in order to render implicit utterances more explicit. Koppel and Ordan [2011], who used function words to discriminate between translated and non-translated texts, found that cohesive markers, words such as *in fact, however, moreover*, etc., were among the top markers of translationese,

irrespective of source language and domain. And truly we find them also over-represented in the Adapted System:

Source *Nous affirmons au contraire la nécessité politique de rééquilibrer les rapports entre l'afrique et l'union européenne.*

Baseline *We say* ***the opposite*** *the political necessity to rebalance relations between africa and the european union.*

Adapted ***on the contrary****, we maintain the political necessity of rebalancing relations between africa and the european union.*

Source *Cette mention semble alors contredire les explications linguistiques données par l'office et laisse craindre que l'erreur ne revête pas le seul caractère technique que l'on semble vouloir lui donner.*

Baseline *This note seems so contradict the explanations given by the language and leaving office fear that the mistake revête do not only technical nature hat we seems to want to give it.*

Adapted *This note* ***therefore*** *seems to contradict the linguistic explanations given by the office and fear that leaves the mistake revête do not only technical nature that we seems to want to give it.*

In (human) translation circles, translating *out of* one's mother tongue is considered unprofessional, even unethical [Beeby, 2009]. Many professional associations in Europe urge translators to work exclusively into their mother tongue [Pavlović, 2007]. The two kinds of automatic systems built in this paper reflect only partly the human situation, but they do so in a crucial way. The $S \rightarrow T$ systems learn examples from many human translators who follow the decree according to which translation should be made *into* one's native tongue. The $T \rightarrow S$ systems are flipped directions of humans' input and output. The $S \rightarrow T$ direction proved to be more fluent and accurate. This has to do with the fact that the translators 'cover' the source texts more fully, having a better 'translation model'.

Chapter 4

Combining Translation and Language Models

In the previous chapters we focused on either the language or the translation model, training the other model on generic data. When we experimented with language models, we trained our translation models on a parallel corpus which was oblivious to the direction of the translation. When we experimented with translation models, we compiled language models from corpora comprising original and translated texts. In this chapter we examine whether our previous findings have an accumulative effect. In other words, we test if an additional improvement in the translation quality can be gained by combining our findings for translation and language models.

We perform our experiments on French-to-English (FR-EN) and English-to-French (EN-FR) translation tasks. We re-use the Europarl-based translation models from Section 3.2.1. We compile language models from the French-English Hansard-based parallel corpora described in Section 3.2.2. We use 1 million parallel sentences subsets. We train an original French LM on the source side of the $S \rightarrow T$ corpus and we train the translated English LM on the target side of the same corpus. In the same manner we compile the translated French LM and the original English LM from the $T \rightarrow S$ corpus. All language models are 5-grams with an interpolated modified Kneser-Ney discounting [Chen, 1998]. The vocabulary is limited to tokens that appear twice or more in the reference set. All unknown words are mapped to a special token. We tune and evaluate all SMT systems on two kinds of reference sets: Europarl (see Section 3.2.1) and Hansard (see Section 3.2.2).

First, we use all possible combinations of translation and language models to train four SMT systems for each translation task: $T \rightarrow S$ TM with original (O) LM, $T \rightarrow S$ TM with translated (T) LM, $S \rightarrow T$ TM with O LM and $S \rightarrow T$ TM with T LM. All systems are tuned and evaluated on both Europarl and Hansard reference sets. Table 4.1 shows the translation quality of the SMT systems in terms of BLEU. Both translation and language models contribute to the translation quality, but it seems that the contribution of the translation model is more significant. Even in the case of the Hansard reference set, in the English-to-French translation task, the $S \rightarrow T$ TM (compiled from Europarl texts) gains 1.2 BLEU points, while the T LM (compiled from Hansard texts) gains only 0.46 BLEU points.

FR-EN (EUROPARL)			
	LM		
		O	**T**
TM	$T \rightarrow S$	27.06	27.30
	$S \rightarrow T$	30.38	**30.65**

FR-EN (HANSARD)			
	LM		
		O	**T**
TM	$T \rightarrow S$	24.41	25.47
	$S \rightarrow T$	25.46	**26.44**

EN-FR (EUROPARL)			
	LM		
		O	**T**
TM	$T \rightarrow S$	22.33	22.71
	$S \rightarrow T$	**25.11**	24.94

EN-FR (HANSARD)			
	LM		
		O	**T**
TM	$T \rightarrow S$	15.88	16.34
	$S \rightarrow T$	17.08	**17.48**

Table 4.1: Combining TMs and LMs: SMT system evaluation results

Finally, we perform a set of experiments to test whether a combination of the adaptation techniques described in Section 3.4 for translation and language models can further improve the translation quality. First, we build a baseline SMT system with a translation model trained on a concatenation of $S \rightarrow T$ and $T \rightarrow S$ parallel corpora and a language model compiled from a concatentation of translated and original texts. Then, we build two other systems, one with an adapted translation model and one with an adapted language model. Finally, we use the adapted translation and language models to train yet another SMT system. We use the PPLMIN-2 method (Section 3.4) to adapt the translation model and linear interpolation (Section 2.3.2) to adapt the language model. The SMT systems are then tuned and evaluated on the Europarl and the Hansard reference sets. The results, depicted in table 4.2, show that SMT systems with an adapted TM usually outperform

the baseline systems. LM adaptation alone does not improve the translation quality, but if combined with TM adaptation it produces the best results (but not siginificantly better than just TM adaptation).

FR-EN (EUROPARL)			
	LM		
		Concat	**Adapt**
TM	Concat	30.76	30.69
	Adapt	31.06	**31.13**

FR-EN (HANSARD)			
	LM		
		Concat	**Adapt**
TM	Concat	27.65	27.48
	Adapt	**27.76**	27.73

EN-FR (EUROPARL)			
	LM		
		Concat	**Adapt**
TM	Concat	25.55	25.51
	Adapt	25.64	**25.69**

EN-FR (HANSARD)			
	LM		
		Concat	**Adapt**
TM	Concat	**18.69**	18.46
	Adapt	18.65	18.68

Table 4.2: Adapting TMs and LMs: SMT system evaluation results

Chapter 5

Discussion and Future Research

5.1 Language Models

We use language models computed from different types of corpora to investigate whether their fitness to a reference set of translated sentences can differentiate between them (and, hence, between the corpora on which they are based). Our main findings are that LMs compiled from manually translated corpora are much better predictors of translated texts than LMs compiled from original-language corpora of the same size. The results are robust, and are sustainable even when the corpora and the reference sentences are abstracted in ways that retain their syntactic structure but ignore specific word meanings.

Furthermore, we show that translated LMs are better predictors of translated sentences even when the LMs are compiled from texts translated from languages *other* than the source language. However, LMs based on texts translated from the source language still outperform LMs translated from other languages.

We also show that MT systems based on translated-from-source-language LMs outperform MT systems based on originals LMs or LMs translated from other languages. Again, these results are robust and the improvements are statistically significant. This effect seems to be amplified as translation quality improves. Furthermore, our results show that original LMs require five to ten times more data to exhibit the same fitness to the reference set and the same translation quality as translated LMs.

More generally, this study confirms that insights drawn from the field of theoretical translation studies, namely the dual claim according to which translations as such differ

from originals, and translations from different source languages differ from each other, can be verified experimentally and contribute to the performance of machine translation.

5.2 Translation Models

Phrase tables trained on parallel corpora that were translated in the same direction as the translation task perform better than ones trained on corpora translated in the opposite direction. Nonetheless, even 'wrong' phrase tables contribute to the translation quality. We analyze both 'correct' and 'wrong' phrase tables, uncovering a great deal of difference between them. We use insights from Translation Studies to explain these differences; we then adapt the translation model to the nature of translationese.

We investigate several approaches to the adaptation problem. First, we use linear interpolation to create a mixture model of $S \rightarrow T$ and $T \rightarrow S$ translation models. We use perplexity minimization and an $S \rightarrow T$ reference set to determine the weights of each model, thus directly adapting the model to the properties of translationese. We show consistent and statistically significant improvements in translation quality on three different language pairs (six translation tasks) using several automatic evaluation metrics.

Furthermore, we incorporate information-theoretic measures that correlate well with translationese into phrase tables as an additional score that can be tuned by MERT, and show a statistically significant improvement in the translation quality over all baseline systems. We also analyze the results qualitatively, showing that SMT systems adapted to translationese tend to produce more coherent and fluent outputs than the baseline systems. An additional advantage of our approach is that it does not require an annotation of the translation direction of the parallel corpus. It is completely generic and can be applied to any language pair, domain or corpus.

5.3 Combination of Translation and Language Models

Our findings have an accumulative effect. The experiments we performed show that both translation and language models contribute to the quality of translation. However, the contribution of the translation model seems to be more significant. Furthermore, adapting both translation and language models produces the best results in many cases. However,

these results are not significantly better than adapting just the translation model.

5.4 Future Research

Phrase-based statitical machine translation is limited to a very local context, usually disregarding sentence structure and long distance dependencies. In our work we harnessed local differences between original and translated texts to improve the quality of SMT systems. More research is required to uncover structural differences between original and translated texts and investigate their effect on machine translation. It seems that a syntax-based framework for MT is more suited for this kind of research.

One of many possible future research directions that might contribute to our understanding of why phrase tables trained on $S \rightarrow T$ are so much better than phrase tables trained on $T \rightarrow S$ parallel corpora focuses on multi-word expressions (MWEs), in particular, idiomatic expressions. It is reasonable to assume that MWEs are much more common in original texts than in translated ones. If this is the case, then $S \rightarrow T$ phrase tables have much more chances to capture idiomatic expressions than $T \rightarrow S$ phrase tables (since MWEs usually do not transfer across languages.). Consequently, a decoder that uses $T \rightarrow S$ phrase-tables will most likely split an idiom in a source sentence and translate it wrongly.

Another research direction is to integrate the translation universals, such as *simplification* and *explicitation*, into a process of translation. There is evidence that manually translated texts have shorter sentences, use more unmarked words and fewer acronyms. One way to simulate a human translation process is to preprocess source sentences using rules designed based on these observation. Such rules may include splitting of a complex sentence into several smaller parts, replacing marked words with their more common synonyms, expanding acronyms, interpreting idioms, etc .

In many cases, the lack of parallel resources for a pair of languages deprives us from building a reliable SMT system translating between these languages. One possible solution to this situation is to use a pivot language – a third language for which there are enough parallel resources available to train SMT systems translating between this language and each one of the original languages. For example, in order to translate Hebrew to Arabic, one could use English as a pivot, first translating Hebrew to English and then, translating

English to Arabic. Obviously, our findings, regarding the effect of translationese on SMT, can be applied in case of a pivot-based translation. Specifically, the second translation step seems to be more challenging since both languages, the source and the target, are supposed to be translated. We hypothesize that best translation quality will be achieved by an SMT system trained on a parallel corpus where both sides are translated. Such training data can be compiled from a multi-lingual parallel corpus, such as the Europarl.

Appendix A

POS Sequences Statistics

The following tables depict major examples for statistical differences in POS sequences in original (O-EN) and translated (T-FR) texts.

POS Seq.	Orig.	Tran.	Diff.	POS Seq.	Orig.	Tran.	Diff.
WP$	281	605	0.54	**JJR**	6577	5675	0.14
EX	7477	4975	0.33	**WRB**	9621	8307	0.14
FW	443	644	0.31	**PRP**	116019	100951	0.13
PDT	2001	2897	0.31	**JJS**	3192	2794	0.12
WDT	18727	25544	0.27	**RBR**	3414	3860	0.12
VBP	73598	60914	0.17	**VBG**	39519	44455	0.11

Table A.1: Major descrepencies in POS unigrams in original and translated texts

POS Seq.	Orig.	Tran.	Diff.	POS Seq.	Orig.	Tran.	Diff.
<S>EX	2252	1256	0.44	**TO NNP**	2535	1826	0.28
IN </S>	1049	594	0.43	**PRP VBP**	43166	31312	0.27
VBP TO	7957	4531	0.43	**DT JJR**	1468	1069	0.27
NNS VBN	2584	4256	0.39	**RB </S>**	4771	3487	0.27
NNS DT	2464	3928	0.37	**NN WDT**	7641	10432	0.27
NN VBN	2708	4266	0.37	**VBG PRP**	1126	1525	0.26
WDT VBZ	5046	7897	0.36	**VBP JJ**	5773	4279	0.26
IN EX	1883	1222	0.35	**<S>VBG**	772	1040	0.26
EX VBP	2034	1325	0.35	**MD PRP**	1472	1102	0.25
IN WP	1591	1037	0.35	**POS NN**	4186	3142	0.25
<S>NNS	1001	656	0.34	**WRB PRP**	3539	2669	0.25
JJR NN	2376	1583	0.33	**IN PDT**	864	1146	0.25
<S>PRP	31495	21449	0.32	**NNS WDT**	4963	6579	0.25
PDT DT	1932	2818	0.31	**<S>WP**	892	1182	0.25
NN VB	812	1176	0.31	**VBG DT**	8057	10664	0.24
CC VB	3793	2620	0.31	**<S>JJ**	1642	1243	0.24
VBP DT	11110	7683	0.31	**NNP JJ**	1204	912	0.24
DT </S>	1739	1209	0.30	**NNP PRP**	5514	4177	0.24
PRP TO	1964	2799	0.30	**VBP NN**	1533	1165	0.24
EX VBZ	3582	2512	0.30	**WDT MD**	2354	3088	0.24
NNP WDT	1550	2209	0.30	**POS NNS**	1575	1202	0.24
WDT VBP	3176	4507	0.30	**NN POS**	1205	924	0.23
NN DT	6301	8854	0.29	**VB CC**	1864	1435	0.23
IN VBZ	759	1065	0.29	**NNP TO**	3700	2857	0.23
CC NNS	3867	5407	0.28	**JJR IN**	1137	1462	0.22

Table A.2: Major descrepencies in POS 2-grams in original and translated texts

POS Seq.	Orig.	Tran.	Diff.	POS Seq.	Orig.	Tran.	Diff.
NNP NNS CC	409	1640	0.75	**NNP POS NN**	3025	2075	0.31
NNP NNP NNS	694	1585	0.56	**DT NNP CC**	3387	2327	0.31
PRP VBP TO	5841	2711	0.54	**VB DT NNP**	2738	1899	0.31
DT NNS VBN	978	1929	0.49	**PRP VBP PRP**	2369	1646	0.31
WDT VBZ RB	891	1665	0.46	**NN NN NN**	1999	1390	0.30
<S>PRP VBP	15251	8275	0.46	**JJ IN PRP**	2178	1516	0.30
PRP VBP DT	7293	3961	0.46	**DT RB JJ**	3187	2225	0.30
VBP TO VB	7246	3969	0.45	**VBP DT NN**	4673	3276	0.30
DT NNP MD	2479	1381	0.44	**VBZ VBN DT**	1650	1158	0.30
IN NN TO	2866	5136	0.44	**NN DT NN**	2675	3798	0.30
PRP VBP JJ	3209	1882	0.41	**CC PRP VBP**	3943	2790	0.29
DT NN VBN	1187	2005	0.41	**VBG DT JJ**	1945	2746	0.29
NN VBN IN	1857	3130	0.41	**NNS WDT VBP**	2462	3474	0.29
NN DT JJ	1137	1878	0.39	**RB JJ NN**	3723	2650	0.29
PRP VBZ VBN	1932	1172	0.39	**VB IN PRP**	2218	1582	0.29
NNS DT NN	1003	1613	0.38	**<S>PRP VBZ**	5751	4112	0.28
DT NNP </S>	3181	2005	0.37	**<S>IN PRP**	2025	1451	0.28
NNS VBN IN	1839	2915	0.37	**DT IN DT**	2334	1679	0.28
NNP PRP VBP	2547	1620	0.36	**TO DT NNP**	2968	2148	0.28
NNS CC NNS	2764	4344	0.36	**NN MD RB**	1231	1678	0.27
JJ NN WDT	2033	3144	0.35	**NN NNS IN**	3493	2568	0.26
NN WDT VBZ	3518	5393	0.35	**NNP MD VB**	4026	2962	0.26
VB JJ IN	1828	1194	0.35	**JJ NN NNS**	3534	2603	0.26
JJ NN DT	1379	2106	0.35	**DT NN DT**	1573	2131	0.26
<S>PRP VBD	2215	1450	0.35	**PRP VBP VBN**	4786	3535	0.26
DT NNP NNPS	2136	3261	0.34	**WRB PRP VBP**	1690	1250	0.26
NNP NNP PRP	3880	2570	0.34	**DT NNP VBZ**	3617	2678	0.26
JJ NNS DT	1452	2157	0.33	**<S>IN NN**	1880	2537	0.26
IN NN DT	1436	2106	0.32	**NNP VBZ VBN**	2490	1846	0.26
PRP TO VB	1601	2343	0.32	**VBG DT NN**	3777	5031	0.25

Table A.3: Major descrepencies in POS 3-grams in original and translated texts

Bibliography

Omar S. Al-Shabab. *Interpretation and the language of translation: creativity and conventions in translation.* Janus, Edinburgh, 1996.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, July 2011. URL `http://www.aclweb.org/anthology/D11-1033`.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 1993.

Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, September 1995.

Mona Baker. Corpus-based translation studies: The challenges that lie ahead. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam, 1996.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W05/W05-0909`.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September 2006. URL `http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1`.

L. E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8, 1972.

Alison Beeby. Direction of translation (directionality). In Mona Baker and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, pages 84–88. Routledge (Taylor and Francis), New York, 2nd edition, 2009.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): 39–71, 1996. ISSN 0891-2017.

D. Biber and S. Conrad. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2009. ISBN 9780521860604. URL `http://books.google.de/books?id=0HUhombmOJUC`.

Shoshana Blum-Kulka. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and intercultural communication Discourse and cognition in translation and second language acquisition studies*, volume 35, pages 17–35. Gunter Narr Verlag, 1986.

Shoshana Blum-Kulka and Eddie A. Levenston. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*, pages 119–139. Longman, 1983.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990. ISSN 0891-2017.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. ISSN 0891-2017.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208,

September 1995. ISSN 1064-8275. doi: 10.1137/0916069. URL http://dx.doi.org/10.1137/0916069.

Chris Callison-Burch and Mark Dredze. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, June 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0701.

Stanley F. Chen. An empirical study of smoothing techniques for language modeling. Technical report 10-98, Computer Science Group, Harvard University, November 1998.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-2031.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. URL http://dx.doi.org/10.2307/2984875.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91. Association for Computational Linguistics, July 2011. URL http://www.aclweb.org/anthology/W11-2107.

Adriano Ferraresi, Silvia Bernardini, Picci Giovanni, and Marco Baroni. Web corpora for bilingual lexicography, a pilot study of English/French collocation extraction and translation. In *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*, September 2008.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1219840.1219885.

George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1870658.1870702`.

William Frawley. Prolegomenon to a theory of translation. In William Frawley, editor, *Translation. Literary, Linguistic and Philosophical Perspectives*, pages 159–175. University of Delaware Press, Newark, 1984.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1:3–33, March 2002. ISSN 1530-0226. doi: http://doi.acm.org/10.1145/595576.595578. URL `http://doi.acm.org/10.1145/595576.595578`.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

David Graff and Christopher Cieri. *English Gigaword*. Linguistic Data Consortium, Philadelphia, third edition, 2007. LDC Catalog No. LDC2007T07.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL `http://dx.doi.org/10.1007/978-3-642-12116-6`.

Alon Itai and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98, March 2008.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63, November 1977. Supplement 1.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Joint Confer-*

ence on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 967–975. Association for Computational Linguistics, June 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1103.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, February 2008. ISBN 013122798X. URL http://www.worldcat.org/isbn/013122798X.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. MT Summit, 2005.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D07/D07-1091.

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1626355.1626388.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Association for Computational Linguistics, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the*

Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-2045.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for Europe. In *Machine Translation Summit XII*, 2009.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1132.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, 2009.

Sara Laviosa. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta*, 43(4):557–570, December 1998.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1034.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, 2012a.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France, April 2012b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E12-1026.

Adam Lopez. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49, 2008. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1380584.1380586.

Daniel Marcu and William Wong. A phrase-based, joint probability model for statistical machine translation. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 133–139, Morristown, NJ, USA, 2002.

Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118693.1118711.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference, Short Papers*, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1858842.1858883`.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075117.

Franz Josef Och and Hermann Ney. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1086–1090, Morristown, NJ, USA, 2000a. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992730.992810.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Morristown, NJ, USA, 2000b. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075218.1075274.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, Morristown, NJ, USA, 2001. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073133.

Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449, 2004. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/0891201042544884.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Nataša Pavlović. Directionality in translation and interpreting practice. Report on a questionnaire survey in Croatia. *Forum*, 5(2):79–99, 2007.

Anthony Pym and Grzegorz Chrupała. The quantitative analysis of translation flows in the age of an international language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*, pages 27–38. John Benjamins, Amsterdam, 2005.

Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? In *Proceedings of the IEEE*, page 2000, 2000.

Diana Santos. On grammatical translationese. In *In Koskenniemi, Kimmo (comp.), Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics (Helsinki*, pages 29–30, 1995.

Candice Séguinot. Pragmatics and the explicitation hypothesis. *TTR: Traduction, Terminologie, Rédaction*, 11(2):106–114, 1998.

Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France, April 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E12-1055`.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedigns of Association for Machine Translation in the Americas (AMTA-2006)*, 2006.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Procedings of International Conference on Spoken Language Processing*, pages 901–904, 2002. URL `citeseer.ist.psu.edu/stolcke02srilm.html`.

Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, 2003. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120103321337458.

G. Toury. *Descriptive Translation Studies: And Beyond*. Benjamins translation library. John Benjamins, 1995a. ISBN 9789027221452. URL `http://books.google.de/books?id=4sooAQAAIAAJ`.

Gideon Toury. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

Gideon Toury. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam / Philadelphia, 1995b.

Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 63–70, Morristown, NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1117794.1117802.

Yulia Tsvetkov and Shuly Wintner. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3389–3392. European Language Resources Association (ELRA), May 2010. ISBN 2-9517408-6-7.

Hans van Halteren. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.

Ashish Venugopal, Stephan Vogel, and Alex Waibel. Effective phrase translation extraction from alignment models. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 319–326. Association for Computational Linguistics, 2003.

Ye-Yi Wang and Alex Waibel. Fast decoding for statistical machine translation. In *Proceedings of International Conference on Spoken Language Processing*, pages 2775–2778, 1998.

Mitch Weintraub, Yaman Aksu, Satya Dharanipragada, Sanjeev Khudanpur, Herman Ney, John Prange, Andreas Stolcke, Fred Jelinek, and Liz Shriberg. Fast training and portability. LM95 project report, Center for Language and Speech Processing, Johns Hopkins University, April 1996. URL `http://www-speech.sri.com/cgi-bin/run-distill?papers/lm95-report.ps.gz`.