Language Models for Machine Translation: Original vs. Translated Texts

Gennadi Lembersky*
University of Haifa

Noam Ordan*
University of Haifa

Shuly Wintner*
University of Haifa

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.

1. Introduction

Statistical machine translation (MT) uses large target language models (LMs) to improve the fluency of generated texts, and it is commonly assumed that for constructing language models, "more data is better data" (Brants and Xu 2009). Not all data, however, are created the same. In this work we explore the differences between language models compiled from texts originally written in the target language (O) and language models compiled from *translated* texts (T).

This work is motivated by much research in Translation Studies that suggests that original texts are significantly different from translated ones in various aspects (Gellerstam 1986). Recently, corpus-based computational analysis corroborated this observation, and Kurokawa, Goutte, and Isabelle (2009) apply it 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. The main research question we investigate here is whether a *language model* compiled from translated texts may similarly improve the results of machine translation.

We test this hypothesis on several translation tasks, including translation from several languages to English, and two additional tasks where the target language is

* Department of Computer Science, University of Haifa, 31905 Haifa, Israel.
E-mails: glembers@campus.haifa.ac.il, noam.ordan@gmail.com, shuly@cs.haifa.ac.il.

Submission received: 22 August 2011; revised submission received: 25 December 2011; accepted for publication: 31 January 2012

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 on abstract linguistic features.

Research in Translation Studies holds a dual view on **translationese**, the sub-language 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, Och, and Marcu 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% and 25% of the total publications. Furthermore, such data can be automatically identified (see Section 2). The practical impact of our work on MT is therefore potentially dramatic.

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

This article[1] is organized as follows: Section 2 provides background and describes related work. We explain our experimental set-up, research methodology and resources in Section 3 and detail our experiments and results in Section 4. Section 5 discusses the results and their implications, and suggests directions for future research.

2. Background and Related Work

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" (page 91). Only 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, 1995) 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 were 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.

1 Preliminary results were published in Lembersky, Ordan, and Wintner (2011). This is an extended, revised version of that paper, providing fuller data and reporting on more language pairs. Some experiments (in particular, Section 4.2.3) are completely new, as is the bulk of the discussion in Section 5, including the human evaluation.

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 subcorpus 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; in accordance with the law of interference, however, the study of van Halteren (2008) suggests that translation from different source languages constitute different sublanguages. As we show in Section 4.2, LMs based on translations from the source language outperform LMs compiled from non-source translations, in terms of both fitness to the reference set and improving MT.

Kurokawa, Goutte, and Isabelle (2009) show that the direction of translation affects the performance of statistical MT. They train systems to translate between French and English (and vice versa) using a French-translated-to-English parallel corpus, and then an English-translated-to-French one. They find that in translating into French it is better to use the latter parallel corpus, and when translating into English it is better to use the former. Whereas they address the *translation* model, we focus on the *language* model in this work. We show that using a language model trained on a text translated from the source language of the MT system does indeed improve the results of the translation.

3. Methodology and Resources

3.1 Hypotheses

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 3.4 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, Jelinek, and Mercer 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 $W = w_1 w_2 \dots w_N$ is the probability of W normalized by the number of words N (Jurafsky and Martin 2008, page 96):

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

For the second hypothesis, we extend the experiment to LMs created from texts translated from other languages. For example, we test how well an LM trained on French-translated-to-English texts fits the German-translated-to-English reference set; and how well an 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 5.1, other automatic evaluation metrics reveal the same pattern).

3.2 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 3.4); 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 1996–1999 and 2001–2009. This is a large multilingual corpus, containing sentences translated from several European languages. It is organized as a collection of bilingual corpora rather than as a single multilingual one, however, and it is hard to identify sentences that are translated into several languages.

We therefore treat each bilingual subcorpus in isolation; each such subcorpus 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. Because 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.[2]

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

Table 1
Europarl English-target corpus statistics, translation from **Lang.** to English.

German–English				Italian–English			
Lang.	**Sentences**	**Tokens**	**Length**	**Lang.**	**Sentences**	**Tokens**	**Length**
MIX	82,700	2,325,261	28.1	MIX	87,040	2,534,793	29.1
O-EN	91,100	2,324,745	25.5	O-EN	93,520	2,534,892	27.1
T-DE	87,900	2,322,973	26.4	T-DE	90,550	2,534,867	28.0
T-FR	77,550	2,325,183	30.0	T-FR	82,930	2,534,930	30.6
T-IT	65,199	2,325,996	35.7	T-IT	69,270	2,535,225	36.6
T-NL	94,000	2,323,646	24.7	T-NL	96,850	2,535,053	26.2

French–English				Dutch–English			
Lang.	**Sentences**	**Tokens**	**Length**	**Lang.**	**Sentences**	**Tokens**	**Length**
MIX	90,700	2,546,274	28.1	MIX	90,500	2,508,265	27.7
O-EN	99,300	2,545,891	25.6	O-EN	97,000	2,475,652	25.5
T-DE	94,900	2,546,124	26.8	T-DE	94,200	2,503,354	26.6
T-FR	85,750	2,546,085	29.7	T-FR	86,600	2,523,055	29.1
T-IT	72,008	2,546,984	35.4	T-IT	73,541	2,518,196	34.2
T-NL	103,350	2,545,645	24.6	T-NL	101,950	2,513,769	24.7

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 subcorpus. In each subcorpus, 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 1 lists the number of sentences, number of tokens, and average sentence length for each English subcorpus and each original language. Table 2 lists the statistics for German and French corpora.

Table 2
Europarl corpus statistics, translation from **Lang.** to German and French.

English–German				English–French			
Lang.	**Sentences**	**Tokens**	**Length**	**Lang.**	**Sentences**	**Tokens**	**Length**
MIX	81,447	2,215,044	27.2	MIX	89,660	2,845,071	31.7
O-DE	89,739	2,215,036	24.7	O-FR	89,875	2,844,265	31.6
T-EN	88,081	2,215,040	25.2	T-EN	96,057	2,847,238	29.6
T-FR	77,555	2,215,021	28.6	T-DE	93,468	2,843,730	30.4
T-IT	64,374	2,215,030	34.4	T-IT	73,257	2,848,931	38.9
T-NL	94,289	2,215,033	23.5	T-NL	102,498	2,835,006	27.7

Table 3
Hansard corpus statistics.

Original French				Original English			
Size	**Sentences**	**Tokens**	**Length**	**Size**	**Sentences**	**Tokens**	**Length**
1M	54,851	1,000,076	18.2	1M	54,216	1,006,275	18.6
5M	276,187	5,009,157	18.1	5M	268,806	5,006,482	18.6
10M	551,867	10,001,716	18.1	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

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 4.2.2). Because 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 subcorpus, 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 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 4. Unfortunately, we do not know how much of this corpus is original; because it includes data from the Xinhua news agency, we suspect that parts of it are indeed translated.

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 subcorpora to have approximately the same number of tokens in each topic. Table 5 lists basic statistics for this corpus.

3.3 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

Table 4
Gigaword corpus statistics.

English, various sources			
Size	**Sentences**	**Tokens**	**Length**
100M	4,448,260	107,483,194	24.2
500M	20,797,060	502,380,054	24.2
1,000M	41,517,095	1,002,919,581	24.2

Table 5
Hebrew-to-English corpus statistics.

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

(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 subcorpus and another 50,000 sentences from the original English subcorpus. For Hebrew we use the Hebrew–English parallel corpus (Tsvetkov and Wintner 2010) that contains sentences translated from Hebrew to English (54%) and from English to Hebrew (46%). The English-to-Hebrew part comprises many short sentences (approximately six tokens per sentence) taken from a movie subtitle database. This explains the low average sentence length of this particular corpus. Table 6 lists some details on those corpora.

3.4 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 1,000 sentences for tuning and 1,000 (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 subcorpus of Europarl (over the period of October 2000 to December 2000). For *L*-to-English translation tasks we only use sentences originally produced in *L*, and 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 portion of news (11.4%). All sentences are originally written in Hebrew and are manually translated to English. See Table 7 for the figures.

Table 6
Parallel corpora used for SMT training.

Language pair	**Side**	**Sentences**	**Tokens**	**Length**
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

4. 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 4.1); and that these differences are important for SMT (Section 4.2).

4.1 Translated vs. Original texts

4.1.1 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 subcorpus, based on the corpora described in Section 3.2. 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 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 8 details the results. The lowest perplexity (reflecting the **best** fit) in each subcorpus is typeset in boldface, and the highest (*worst* fit) is italicized.

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 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).

Table 7
Reference sets.

Language pair	Side	Sentences	Tokens	Length
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 8
Fitness of various LMs to the reference set.

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**

Boldface = best fit; italics = worst fit.

4.1.2 Linguistic Abstraction. A possible explanation for the different perplexity results among the LMs could be the specific content 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.[3] Then, we use the Stanford Named Entity Recognizer (Finkel, Grenager, and Manning 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 tokens that appear more than once in the "abstracted" reference set. The results, depicted in Table 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 with the worst-performing model. Although we do not show the details here, the same pattern is persistent in all the other Europarl languages we experiment with.

4.1.3 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 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, depicted in Table 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. We then compute the perplexity of the German-translated-from-English and French-translated-from-English reference sets (see Section 3.4) with respect to these language models. Table 12 depicts the results; they are in complete agreement with our hypothesis.

3 In fact, there is reason to assume that punctuation constitutes part of the translationese effect. Removing punctuation therefore harms our cause of identifying this effect.

Table 9
Fitness of O- vs. T-based LMs to the reference set (FR-EN), reflecting different abstraction levels.

	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

Boldface = best fit; italics = worst fit.

4.1.4 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 Hansard and Gigaword subcorpus described in Section 3.2. 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 13).

Table 10
Fitness of O- vs. T-based LMs to the reference set (HE-EN).

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

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 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.

Table 11
Fitness of O- vs. T-based LMs to the reference set (HE-EN), reflecting different abstraction levels.

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**

4.2 Original vs. Translated LMs for Machine Translation

4.2.1 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 3.3. We use the reference sets (Section 3.4) as follows: 1,000 sentences are randomly extracted for minimum error-rate training (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 3.2). We use Bleu (Papineni et al. 2002) to evaluate translation quality. The results are listed in Tables 14 and 15.

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

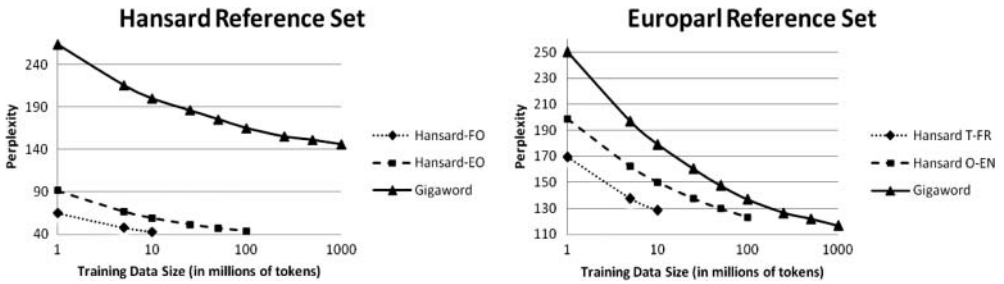
Table 12
Fitness of O- vs. T-based LMs to the reference set (EN-DE and EN-FR).

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 13
The effect of LM training corpus size on the fitness of LMs to the reference sets.

Hansard Reference Set		**Europarl Reference Set**	
Hansard T-FR		**Hansard T-FR**	
Size	**Perplexity**	**Size**	**Perplexity**
1M	64.68	1M	169.66
5M	47.63	5M	137.72
10M	42.70	10M	128.65
Hansard O-EN		**Hansard O-EN**	
Size	**Perplexity**	**Size**	**Perplexity**
1M	91.40	1M	198.93
5M	66.95	5M	162.08
10M	59.19	10M	150.05
25M	51.59	25M	137.31
50M	47.02	50M	129.43
100M	43.70	100M	123.10
Gigaword		**Gigaword**	
Size	**Perplexity**	**Size**	**Perplexity**
100M	165.03	100M	136.72
500M	151.00	500M	121.88
1000M	145.88	1000M	116.55



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.

4.2.2 Larger Language Models. Again, the LMs used in the MT experiments reported here 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

Table 14
Machine translation with various LMs; English target language.

DE to EN		FR to EN		IT to EN		NL to EN	
LM	**Bleu**	**LM**	**Bleu**	**LM**	**Bleu**	**LM**	**Bleu**
MIX	21.43	MIX	28.67	MIX	25.41	MIX	24.20
O-EN	*21.10*	O-EN	*27.98*	O-EN	24.69	O-EN	*23.40*
T-DE	**21.90**	T-DE	28.01	T-DE	*24.62*	T-DE	24.26
T-FR	21.16	T-FR	**29.14**	T-FR	25.37	T-FR	23.56
T-IT	21.29	T-IT	28.75	T-IT	**25.96**	T-IT	23.87
T-NL	21.20	T-NL	28.11	T-NL	24.77	T-NL	**24.52**

Table 15
Machine translation with various LMs; non-English target language.

EN to DE		EN to FR	
LM	**Bleu**	**LM**	**Bleu**
MIX	13.00	MIX	24.83
O-DE	12.47	O-FR	24.70
T-EN	**13.10**	T-EN	**25.31**
T-FR	*12.46*	T-DE	*24.58*
T-IT	12.65	T-IT	24.89
T-NL	12.86	T-NL	25.20

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.

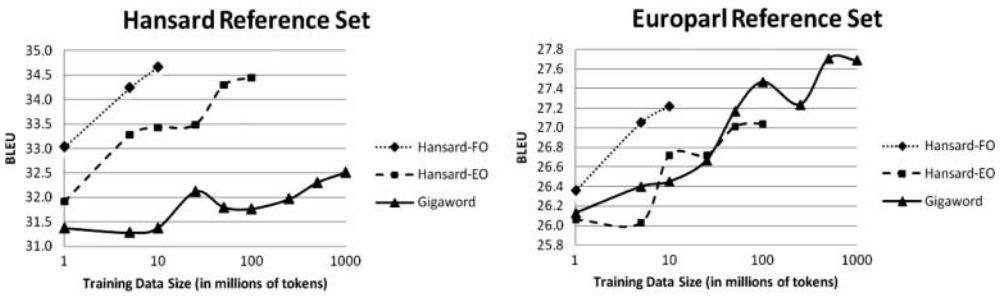
Table 16 again demonstrates that language models compiled from original texts must be up to *ten times larger* in order to yield translation quality similar to that of LMs compiled from translated texts.[4] In other words, much smaller translated LMs perform better than much larger original ones, and 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.

4.2.3 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

4 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.

Table 16
The effect of LM size on MT performance.

Hansard TM and Test		**Europarl TM and Test**	
Hansard T-FR		**Hansard T-FR**	
Size	**Bleu**	**Size**	**Bleu**
1M	33.03	1M	26.36
5M	34.25	5M	27.06
10M	**34.67**	10M	27.22
Hansard O-EN		**Hansard O-EN**	
Size	**Bleu**	**Size**	**Bleu**
1M	31.91	1M	26.06
5M	33.27	5M	26.03
10M	33.43	10M	26.72
25M	33.49	25M	26.72
50M	34.29	50M	27.01
100M	34.44	100M	27.04
Gigaword		**Gigaword**	
Size	**Bleu**	**Size**	**Bleu**
100M	31.77	100M	27.47
500M	32.31	500M	**27.71**
1000M	32.51	1000M	27.69



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

Table 17
Various combinations of original and translated texts and their effect on perplexity (PPL) and translation quality (Bleu).

Hansard TM, LM and Test			Europarl TM and Test; Hansard LM		
Combination	**PPL**	**Bleu**	**Combination**	**PPL**	**Bleu**
O-EN	43.70	34.44	O-EN	123.10	27.04
T-FR	42.70	34.67	T-FR	128.65	27.22
Concatenation	38.43	34.62	Concatenation	116.71	27.14
Concatenation x10	41.15	35.09	Concatenation x10	135.09	27.29
Concatenation x20	45.07	34.67	Concatenation x20	152.02	27.09
Log-Linear LM	–	35.26	Log-Linear LM	–	27.30
Interpolated LM	**36.69**	**35.35**	Interpolated LM	**107.82**	**27.48**

that uses this LM, in terms of Bleu.[5] 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 subcorpus, as above. The results are listed in Table 17; we now provide a detailed explanation of these experiments.

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 concatenation. 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:

$$\hat{t} = \underset{t}{\operatorname{argmax}}\, P(t|s) = \underset{t}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} \lambda_i h_i(t, s) \right\} \tag{2}$$

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

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 with an MT system that uses O-EN ($p < 0.001$), 0.59 Bleu points compared with 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):

$$p(w|h) = (1 - \eta) \cdot p(w|h; LM_1) + \eta \cdot p(w|h; LM_2) \quad (3)$$

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 subcorpora. Of the methods we explored here, interpolated LMs provide the greatest advantage. More research is needed in order to find an optimal combination.

5. Discussion

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. LMs based on texts translated from the source language still outperform LMs translated from other languages, however.

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

Table 18
MT system performance as measured by METEOR and TER.

DE to EN			IT to EN		
Orig. Lang.	**METEOR**	**TER**	**Orig. Lang.**	**METEOR**	**TER**
O-EN	*28.26*	*64.56*	O-EN	*31.03*	*58.30*
T-DE	**28.64**	**63.57**	T-IT	**31.16**	**57.63**

FR to EN			NL to EN		
Orig. Lang.	**METEOR**	**TER**	**Orig. Lang.**	**METEOR**	**TER**
O-EN	*33.05*	*54.45*	O-EN	*29.97*	*60.29*
T-FR	**33.30**	**53.65**	T-NL	**30.40**	**59.63**

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.

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.

5.1 Automatic Evaluation

First, we use two alternative automatic evaluation metrics, METEOR[6] (Denkowski and Lavie 2011) and TER (Snover et al. 2006), to assess the quality of the MT systems described in Section 4.2. We focus on four translation tasks: From German, French, Italian, and Dutch to English.[7] 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 18, 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.

6 More precisely, we use METEOR-RANK, the configuration used for WMT-2011.
7 All MT systems were tuned using Bleu.

5.2 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. Annotators 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 because 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. All annotators were located in the United States (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.

5.3 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 in the following 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, page 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 77 among the O-EN bigrams, whereas in T-FR it does not appear in the top 100. On the other hand, the explicit trigram *The European Union* occurs more frequently in T than in O.

Similarly, an instance of (2) is the cohesive marker *because* in the following example, which appears in T but neither appears in O nor can it be traced back to the original source sentence:

Source *Enfin, ce qui est grave dans le rapport de M. Olivier Tautologie, c'est qu'il propose une constitution tripotage.*

O *Finally, which is serious in the report of Mr Olivier Tautologie, is that it proposes a constitution tripotage.*

T *Finally, and this is serious in the report by Mr olivier Tautologie, it is* ***because*** *it proposes a constitution tripotage.*

Another cohesive marker, *nevertheless*, is correctly generated only in the T-based translation in the following example:

Source *C'est quand même quelque chose de précieux qui a été souligné par tous les membres du conseil européen.*

O *Even when it is something of valuable which has been pointed out by all the members of the European Council.*

T *It is* ***nevertheless*** *something of a valuable which has been pointed out by all the members of the European Council.*

Other cohesive markers discussed by Blum-Kulka (1986) are over-represented in T compared with 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 (1995, page 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, because 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 4.1.2 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. A full analysis of such patterns is beyond the scope of this article.

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 original 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).

5.4 Future Research

This work is among the first to use insights from Translation Studies in order to improve machine translation, and to use computational linguistic methodologies to corroborate Translation Studies hypotheses. We believe that there are still vast opportunities for fertile cross-disciplinary research in these directions. First, we only address the language model in the present work. Kurokawa, Goutte, and Isabelle (2009) investigate the relations between the direction of translation and the quality of the *translation* model used by SMT systems. There are various ways in which the two approaches can be extended and combined, and we are actively pursuing such research directions now (Lembersky, Ordan, and Wintner 2012).

This work also bears on language typology: We conjecture that LMs compiled from texts translated not from the original language, but from a closely related one, can be better than LMs compiled from texts translated from a more distant language. Some of our results support this hypothesis, but more research is needed in order to establish it.

The fact that translations seem to make do with fewer words (cf. also Laviosa 2008) call into question certain norms in comparing corpora in the field of machine translation. Translated and original texts can be expected to either have the same number of sentences or the same number of tokens, but not both. Similarly, they may have the same number of tokens or the same number of types, but not both.

Another interesting question that arises from this study is whether the perplexity of a language model on a reference set is a good predictor of a translation quality measure, such as Bleu. Although our results show a certain correlation between the perplexity and Bleu, we acknowledge the fact that these results need further corroboration. Chen, Beeferman, and Rosenfeld (1998) examine the ability of perplexity to estimate the performance of speech recognition. They find that perplexity often does not correlate well with word-error rates. As it is extremely important to have a reliable measure capable of estimating the effect of language model improvements on translation quality without requiring expensive decoding resources, we believe that finding correspondences between perplexity and the quality of MT is a valuable topic for future research.

Acknowledgments

We are grateful to Cyril Goutte, George Foster, and Pierre Isabelle for providing us with an annotated version of the Hansard corpus. Alon Lavie has been instrumental in stimulating some of the ideas reported in this article, as well as in his long-term support and advice. We benefitted greatly from several constructive suggestions by the three anonymous *Computational Linguistics* referees. This research was supported by the Israel Science Foundation (grant no. 137/06) and by a grant from the Israeli Ministry of Science and Technology.

References

Al-Shabab, Omar S. 1996. *Interpretation and the Language of Translation: Creativity and Conventions in Translation*. Janus, Edinburgh.

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

Baker, Mona. 1993. 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*. John Benjamins, Amsterdam, pages 233–252.

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

Baker, Mona. 1996. 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*. John Benjamins, Amsterdam, pages 175–186.

Baroni, Marco and Silvia Bernardini. 2006. 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.

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Editors Blum-Kulka, editors, *Interlingual and Intercultural Communication Discourse and Cognition in Translation and Second Language Acquisition Studies*, volume 35. Gunter Narr Verlag, Berlin, pages 17–35.

Brants, Thorsten and Peng Xu. 2009. Distributed language models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 3–4, Boulder, CO.

Callison-Burch, Chris and Mark Dredze. 2010. 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, CA.

Chen, Stanley, Douglas Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (BNTUW)*, Landsdowne, PA.

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

Denkowski, Michael and Alon Lavie. 2011. 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, Edinburgh.

Ferraresi, Adriano, Silvia Bernardini, Picci Giovanni, and Marco Baroni. 2008. 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*, Hangzhou.

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

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

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

Graff, David and Christopher Cieri. 2007. *English Gigaword*. Linguistic Data

Consortium, Philadelphia, PA, third edition. LDC Catalog No. LDC2007T07.
Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. 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, Berlin.
Itai, Alon and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.
Jelinek, Frederick, Robert L. Mercer, Lalit R. Bahl, and J. K. Baker. 1977. Perplexity—A measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63.
Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, Upper Saddle River, NJ.
Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona.
Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*, pages 79–86, Phuket.
Koehn, Philipp and Hieu Hoang. 2007. 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.
Koehn, Philipp, 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. 2007. 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.
Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. 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, Edmonton.
Koehn, Philipp and Josh Schroeder. 2007. 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.
Koppel, Moshe and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, OR.
Kurokawa, David, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, Kurokawa.
Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta*, 43(4):557–570.
Laviosa, Sara. 2008. Universals. In Mona Baker and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, 2nd Edition. Routledge (Taylor and Francis), New York, pages 288–292.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2011. Language models for machine translation: Original vs. translated texts. In *Proceedings of EMNLP*, pages 363–374, Edinburgh.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, Avignon.
Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Morristown, NJ.
Och, Franz Josef and Hermann Ney. 2001. Discriminative training and maximum entropy models for statistical machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Morristown, NJ.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ.
Pym, Anthony and Grzegorz Chrupała. 2005. The quantitative analysis of translation flows in the age of an international

language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*. John Benjamins, Amsterdam, pages 27–38.
Santos, Diana. 1995. On grammatical translationese. In *Koskenniemi, Kimmo (comp.), Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics (Helsinki)*, University of Helsinki, pages 29–30.
Séguinot, Candice. 1998. Pragmatics and the explicitation hypothesis. *TTR: Traduction, Terminologie, Rédaction*, 11(2):106–114.
Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas (AMTA-2006)*, pages 223–231, Cambridge, MA.
Stolcke, Andreas. 2002. SRILM—An extensible language modeling toolkit. In *Procedings of International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.
Toury, Gideon. 1980. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv.
Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam / Philadelphia.
Toutanova, Kristina and Christopher D. Manning. 2000. 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.
Tsvetkov, Yulia and Shuly Wintner. 2010. 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, Valleta.
van Halteren, Hans. 2008. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ.
Weintraub, Mitch, Yaman Aksu, Satya Dharanipragada, Sanjeev Khudanpur, Herman Ney, John Prange, Andreas Stolcke, Fred Jelinek, and Liz Shriberg. 1996. Fast training and portability. LM95 project report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.