

# Computational Approaches to Translation Studies

Shuly Wintner<sup>(✉)</sup>

Department of Computer Science, University of Haifa, Haifa, Israel  
shuly@cs.haifa.ac.il

**Abstract.** Translated texts, in any language, have unique characteristics that set them apart from texts originally written in the same language. *Translation studies* is a research field that focuses on investigating these characteristics. Until recently, research in computational linguistics, and specifically in machine translation, has been entirely divorced from translation studies. The main goal of this tutorial is to introduce some of the findings of translation studies to researchers interested mainly in machine translation, and to demonstrate that awareness of these findings can result in better, more accurate machine translation systems (This chapter synthesizes material that has been previously published by the author and colleagues, in particular in Volansky et al. (2015); Rabinovich and Wintner (2015); Lembersky et al. (2011, 2012a, 2012b, 2013); and Twitto et al. (2015)).

## 1 Introduction

Research in *translation studies* reveals that translated texts are ontologically different from original, non-translated ones.<sup>1</sup> Translated texts, in any language, can be considered a “dialect” of that language, known as *translationese*. Several characteristics of translationese have been proposed as universal in a series of hypotheses. Awareness of the special properties of translationese can improve the quality of natural language processing (NLP) applications, in particular machine translation (MT). This chapter provides an introduction to translation studies and its relevance to research in NLP and specifically to MT.

In Sect. 2 we survey some theoretical hypotheses of translation studies. Focusing on the unique properties of translationese, we distinguish between properties resulting from *interference* of the source language (the so-called “fingerprints” of the source language on the translation product) and properties that are source-language-independent, and that are therefore presumably universal. The latter include phenomena resulting from three main processes: simplification, standardization and explicitation. All these phenomena are defined, explained and exemplified.

Section 3 describes several works that use standard (supervised and unsupervised) text classification techniques to distinguish between translations and

---

<sup>1</sup> We use ‘originals’ here as opposed to ‘translations’, although translation are of course also originally created by translators.

originals, in several languages. We focus on the features that best distinguish between the two classes, and how these features corroborate some (but not all) of the hypotheses set forth by translation studies scholars.

Then, we discuss in Sect. 4 several computational works that show that awareness of translationese can improve machine translation. Specifically, we show that language models compiled from translated texts are more fitting to the reference sets than language models compiled from originals. We also show that translation models compiled from texts that were (manually) translated from the source to the target are much better than translation models compiled from texts that were translated in the reverse direction. Finally, in Sect. 5 we touch upon some related issues and current research directions.

## 2 Translationese

Numerous studies suggest that translated texts differ from original ones. Gellerstam (1986) compared texts written originally in Swedish with texts translated from English into Swedish. He noted that the differences between them did not indicate poor translation but rather a statistical phenomenon, which he termed *translationese*.

The features of translationese were theoretically organized under the terms *laws of translation* or *translation universals*. Toury (1980, 1995) distinguished between two laws: the *law of interference* and the *law of growing standardization*. The former pertains to the fingerprints of the source text that are left in the translation product. The latter pertains to the effort to standardize the translation product according to existing norms in the target language and culture. 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 is neither of them (Frawley 1984).

Baker (1993) suggested several candidates for translation universals, which she claimed to appear in any translated text, regardless of the source language: “features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems” (Baker 1993, p. 243). Among the better known universals are *simplification* and *explicitation*, defined and discussed thoroughly by Blum-Kulka and Levenston (1978, 1983) and Blum-Kulka (1986), respectively. A third universal hypothesis is *standardization*, also known as *normalization* (Toury 1995). We now detail these hypotheses.

*Simplification* refers to the process of rendering complex linguistic features in the source text into simpler features in the target text. Strictly speaking, this phenomenon can be studied only vis-à-vis the source text, since ‘simpler’ is defined here in reference to the source text, where, for example, the practice of splitting sentences or refraining from complex subordinations can be observed. And indeed, this is how simplification was first defined and studied in translation studies (Blum-Kulka and Levenston 1983; Vanderauwerea 1985). Baker (1993) suggested that simplification can be studied by comparing translated texts with

non-translated ones, as long as both texts share the same domain, genre, time frame, etc. In a series of corpus-based studies, Laviosa (1998, 2002) confirmed this hypothesis. Ilisei et al. (2010) and Ilisei and Inkpen (2011) provided further evidence for this universal in Romanian and Spanish.

*Explicitation* is the tendency to spell out in the target text utterances that are more implicit in the source. One obvious way in which explicitation is manifested is by overusing cohesive markers such as *because*, *hence*, or *moreover*. Blum-Kulka (1986) exemplified this phenomenon in translations from Hebrew to English, and Øverås (1998) compiled a parallel bidirectional Norwegian-English and English-Norwegian corpus and provided further evidence for explicitation. Koppel and Ordan (2011) found that some of the prominent features in their list of function words were cohesive markers, such as *therefore*, *thus*, and *consequently*.

Translators take great efforts to *standardize* texts (Toury 1995), or, in the words of Baker (1993, p. 244), they have “a strong preference for conventional ‘grammaticality’”. This includes the tendency to avoid repetitions (Ben-Ari 1998), the tendency to use a more formal style manifested in refraining from the use of contractions (Olohan 2003), and the tendency to overuse fixed expressions even when the source text refrains, sometime deliberately, from doing so (Toury 1980; Kenny 2001).

In the last two decades corpora have been used extensively to study translationese. For example, Al-Shabab (1996) showed that translated texts exhibit lower lexical variety than originals; Laviosa (1998) showed that their mean sentence length is lower, as is their lexical density (ratio of content to non-content words). Both these studies provide evidence for the simplification hypothesis. Corpus-based translation studies became a very prolific area of research (Laviosa 2002).

### 3 Identification of Translationese

#### 3.1 Supervised Classification

Since the pioneering work of Baroni and Bernardini (2006), text classification methods, based on standard machine learning techniques, have been extensively used to automatically identify translationese in several languages (van Halteren 2008; Ilisei et al. 2010; Ilisei and Inkpen 2011; Popescu 2011; Koppel and Ordan 2011; Avner et al. 2016). While many of these works were mainly interested in the practical task of distinguishing between originals and translations, Volansky et al. (2015) used the accuracy of classification as a proxy for evaluating the validity of translation studies hypotheses.

In *supervised machine-learning*, a *classifier* is trained on labeled examples the classification of which is known a priori, e.g., *translations* vs. *originals*. Each text has to be *represented*: a set of numeric *features* is extracted from the data (here, chunks of text), and a generic machine-learning algorithm is then trained to distinguish between *feature vectors* representative of one class and those representative of the other. Given enough data for training and given that the

features are indeed relevant, the trained classifier can then be given an ‘unseen’ text, namely a text that is not included in the training set. Such a text is again represented by a feature vector in the same manner, and the classifier can predict the class (variety) it belongs to.

For evaluation, it is customary to use  $k$ -fold cross validation. The ‘unseen’ texts are also labeled, of course, and the prediction of the classifier can be compared to the actual, ‘gold’ label. In  $k$ -fold cross-validation, the training data is divided to  $k$  folds (typically,  $k = 10$ ), and the following procedure is repeated  $k$  times: training on  $k - 1$  folds, then testing on the held-out fold, cyclically. Finally, the accuracy results over the  $k$  folds is averaged and reported.

The experimental setup of Volansky et al. (2015) is as follows: the dataset is taken from Europarl (Koehn 2005), with approximately 4 million tokens in English and the same number of tokens translated from ten source languages: Danish, Dutch, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. The corpus is first tokenized and then partitioned into chunks of approximately 2000 tokens (ending on a sentence boundary). For training and evaluation, each chunk is represented as a *feature vector*. A standard SVM classifier is trained on the feature vectors and is evaluated using ten-fold cross validation. As the task is binary and the dataset is balanced, the baseline accuracy is 50%.

### 3.2 Features

The crux of the method lies in the selection of features. If the features indeed reflect a true characteristic of translationese, one can assume that a classifier based on these features will be accurate. It should be noted that using content words as features is likely to yield good classifiers: one can expect texts translated, say, from French to English, to include more instances of proper names like *Paris* or even more common nouns like *cheese*. Such features, however, reveal nothing about the properties of translationese, and hence to not advance the goal of the investigation. Volansky et al. (2015) experiment with several feature sets which we briefly list below. Features were selected such that they:

- reflect frequent linguistic characteristics one expects to be present in the two types of text;
- be content-independent, indicating formal and stylistic differences between the texts that are not derived from differences in contents, domain, genre, etc.; and
- be easy to interpret, yielding insights regarding the differences between original and translated texts.

Specifically, features were grouped together to reflect the main translation studies hypotheses. The feature types discussed below have varied dimensionality (some are a single value, averaged over the entire chunk, and some can define very long vectors). If not mentioned otherwise, the value of a feature is a simple count.

**Simplification.** The simplification hypothesis was modeled through the following features:

**Lexical variety.** The assumption is that original texts use richer vocabularies than translated ones (Baker 1993; Laviosa 1998). Three different *type-token ratio* (TTR) measures were used, following Grieve (2007), where  $V$  is the number of types and  $N$  is the number of tokens per chunk. All three versions consider punctuation marks as tokens.

1.  $V/N$ , magnified by order of 6.
2.  $\log(V)/\log(N)$ , magnified by order of 6.
3.  $100 \times \log(N)/(1 - v_1/v)$ , where  $V_1$  is the number of types occurring only once in the chunk.

**Mean word length (in characters).** The assumption is that translated texts use simpler words, in particular shorter ones.

**Syllable ratio.** Assuming that simpler words are used in translated texts, one expects fewer syllables per word.

**Lexical density.** The frequency of tokens that are *not* nouns, adjectives, adverbs or verbs (Laviosa 1998).

**Mean sentence length.** Splitting sentences is a common strategy in translation, which is also considered a form of simplification. Baker (1993) renders it one of the universal features of simplification.

**Mean word rank.** The assumption is that less frequent words are used more often in original texts than in translated ones. This is based on the observation of Blum- Kulka and Levenston (1983) that translated texts “make do with less words” and the application of this feature by Laviosa (1998). A theoretical explanation is provided by Halverson (2003): translators use more prototypical language, i.e., they “regress to the mean” (Shlesinger 1989). To compute this, a list of 6000 English most frequent words was used.

**Most frequent words.** The frequencies of the  $N$  most frequent words in the corpus, where  $N = 5, 10, 50$ .

**Explicitation.** Several features were used to model the explicitation hypothesis. The first three feature sets were inspired by an example provided by Baker (1993, pp. 243–4), where the clause *The example of Truman was always present in my mind* was translated into Arabic with a fairly long paragraph, which includes the following: *In my mind there was always the example of the American President Harry Truman, who succeeded Franklin Roosevelt....*

**Explicit naming.** The ratio of personal pronouns to proper nouns.

**Single naming.** The frequency of proper nouns consisting of a single token, not having an additional proper noun as a neighbor.

**Mean multiple naming.** The average length (in tokens) of proper noun sequences.

**Cohesive markers.** Translations are known to excessively use certain *cohesive markers* (Blum-Kulka 1986; Øverås 1998). A list of 40 such markers was used, based on Koppel and Ordan (2011).

**Normalization.** Normalization was modeled through the following features:

**Repetitions.** The number of content words that occur more than once in a chunk.

**Contractions.** The ratio of contracted forms to their counterpart full form(s).

**Average PMI.** Original texts are expected to use more collocations, and in any case to use them differently than translated texts. This hypothesis is based on Toury (1980) and Kenny (2001), who showed that translations overuse highly associated words. To reflect this, the average PMI (Church and Hanks 1990) of all bigrams in the chunk was used.

**Threshold PMI.** The number of bigrams with PMI above 0.

**Interference.** Several features were selected to model the influence of the source language on the translation product:

**POS  $n$ -grams.** The hypothesis is that grammatical structures used in the various source languages interfere with the translations; and that translations have unique grammatical structure. Following Baroni and Bernardini (2006) and Kurokawa et al. (2009), this assumption was modeled by defining as features unigrams, bigrams and trigrams of part-of-speech (POS) tags.

**Character  $n$ -grams.** Unigrams, bigrams and trigrams of characters. This feature was motivated by Popescu (2011); it captures morphological features of the language.

**Prefixes and suffixes.** Character  $n$ -grams are an approximation of morphological structure. In the case of English, the little morphology exhibited by the language is typically manifested as prefixes and suffixes. A more refined variant of the character  $n$ -gram feature, therefore, focuses only on prefixes and suffixes.

**Contextual function words.** This feature is a variant of POS  $n$ -grams, where the  $n$ -grams can be anchored by specific function words. It is defined as the frequencies in the chunk of consecutive triplets  $\langle w_1, w_2, w_3 \rangle$ , where at least two of the elements are function words, and at most one is a POS tag.

**Positional token frequency.** Writers have a relatively limited vocabulary from which to choose words to open or close a sentence, and the choices may be subject to interference (Munday 1998; Gries and Wulff 2012). The value of this feature is the frequency of tokens appearing in the first, second, antepenultimate, penultimate and last positions in a sentence.

**Miscellaneous.** Finally, a number of features that cannot be naturally associated with any of the above hypotheses, but nevertheless shed light on the nature of translationese, were also defined.

**Function words.** Replicating the results of Koppel and Ordan (2011), the same list of function words was used.

**Pronouns.** Pronouns are function words, and Koppel and Ordan (2011) reported that this subset is among the top discriminating features between originals and translations.

**Punctuation.** Punctuation marks organize the information within sentence boundaries and to a great extent reduce ambiguity; according to the explicitation hypothesis, translated texts are less ambiguous (Blum-Kulka 1986). The following punctuation marks were used: ? ! : ; - ( ) [ ] ‘ ’ “ ” / , . Following Grieve (2007), three variants of this feature were defined:

1. The normalized frequency of each punctuation mark in the chunk.
2. A non-normalized notion of frequency:  $n/tokens$ , where  $n$  is the number of occurrences of a punctuation mark; and  $tokens$  is the actual (rather than normalized) number of tokens in the chunk.
3.  $n/p$ , where  $p$  is the total number of punctuations in the chunk; and  $n$  as above.

**Ratio of passive forms to all verbs.** The assumption is that English original texts tend to use the passive form more excessively than translated texts, due to the fact that the passive voice is more frequent in English than in some other languages (cf. Teich (2003) for German-English).

The features defined above are all stylistic features that abstract away from the actual contents of the text. As a “sanity check”, two content-bearing features were used: token unigrams and token bigrams. These features are expected to yield excellent classifiers but not shed any interesting light on translation hypotheses.

### 3.3 Results

Each of the feature types discussed above defines a separate classifier. The accuracy of ten-fold cross-validation evaluation with the various feature sets is reported in Table 1.

As expected, the accuracy of the “sanity” features is perfect. This is not surprising in light of their ability to reflect contents, which is highly related to the source language and culture. However, this provides no interesting insights on the properties of translationese.

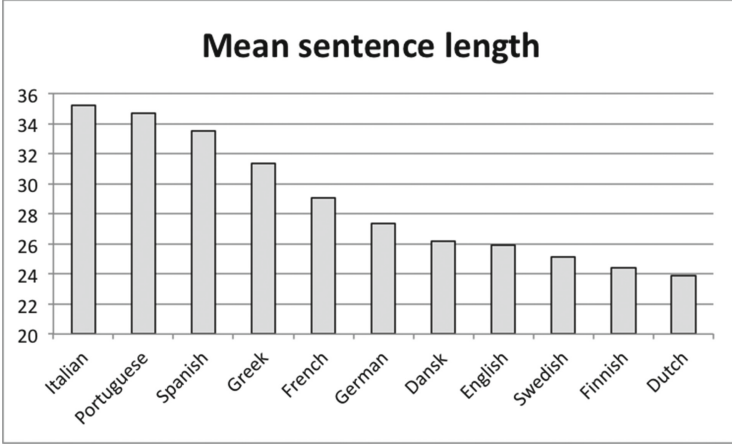
In contrast, the simplification features tell a mixed story. Some of them are reasonably accurate, especially considering the low dimensionality (of, e.g., TTR). Both TTR and mean word rank provide reasonable separation between the two classes. The other features are less discriminating. Most surprising is mean sentence length which, while providing a better-than-baseline classifier, actually behaves conversely to the hypothesis: as it turns out, the mean sentence length of translations in our corpus is actually *higher* than that of originals (Fig. 1).

The explicitation classifiers also yield mixed results. While the various naming features perform almost at chance level, the cohesive markers turn out to be very effective. In contrast, the normalization features do not discriminate well between originals and translations. In particular, the PMI features again behave conversely to the prediction of the hypothesis: English originals turn out to have much more highly collocated bigrams than translations (Fig. 2).

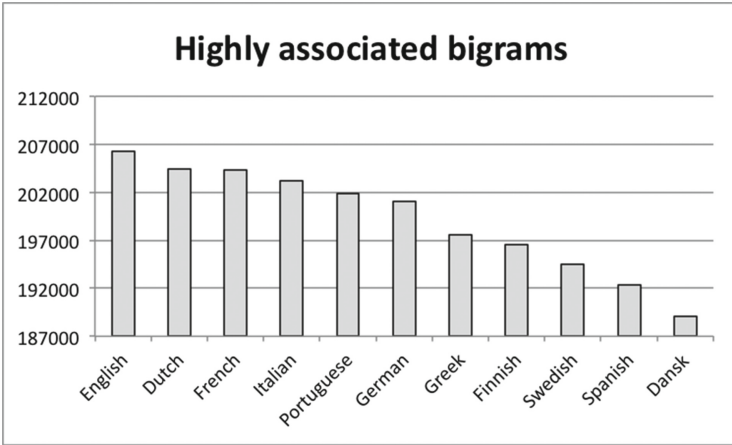
**Table 1.** Classification results

Category	Feature	Accuracy (%)
Sanity	Token unigrams	100
	Token bigrams	100
Simplification	TTR (1)	72
	TTR (2)	72
	TTR (3)	76
	Mean word length	66
	Syllable ratio	61
	Lexical density	53
	Mean sentence length	65
	Mean word rank (1)	69
	Mean word rank (2)	77
	<i>N</i> most frequent words	64
Explicitation	Explicit naming	58
	Single naming	56
	Mean multiple naming	54
	Cohesive markers	81
Normalization	Repetitions	55
	Contractions	50
	Average PMI	52
	Threshold PMI	66
Interference	POS unigrams	90
	POS bigrams	97
	POS trigrams	98
	Character unigrams	85
	Character bigrams	98
	Character trigrams	100
	Prefixes and suffixes	80
	Contextual function words	100
	Positional token frequency	97
Miscellaneous	Function words	96
	Pronouns	77
	Punctuation (1)	81
	Punctuation (2)	85
	Punctuation (3)	80
	Ratio of passive forms to all verbs	65





**Fig. 1.** Mean sentence length of translations from several languages to English vs. original English



**Fig. 2.** Number of bigrams whose PMI is above threshold, by source language

Finally, the interference features are clearly the best discriminators. Part-of-speech n-grams, character n-grams, contextual function words and positional token frequency, all of them features that are highly influenced by the structure (and lexis) of the source language, yield excellent, sometimes even perfect classifiers. This result is robust, and persists even when the dimensionality of the feature vectors is reduced (by limiting vectors to the 300 most frequent features) and when experimenting with originals and translations in languages other than English, including Hebrew (Avner et al. 2016), German, and French (Rabinovich and Wintner 2015).

### 3.4 Unsupervised Classification

Clearly, then, it is possible to automatically distinguish between original and translated texts, with very high accuracy, by employing text classification methods. However, the approaches we surveyed all employed *supervised* machine-learning; they therefore suffer from two main drawbacks: they inherently depend on data annotated as original vs. translated; and they do not scale up to unseen (related or unrelated) domains.<sup>2</sup> These shortcomings undermine the usability of supervised methods for translationese identification in a typical real-life scenario, where no labelled in-domain data are available.

To overcome these issues, Rabinovich and Wintner (2015) proposed to use *unsupervised* machine learning, or *clustering*, as a way to identify translationese. In addition to the Europarl corpus described above, they used the following datasets: (i) the Canadian Hansard, transcripts of the Canadian Parliament; (ii) literary classics written (or translated) mainly in the 19th century; and (iii) transcripts of TED and TEDx talks.

First, they replicated the results of Volansky et al. (2015) on the four datasets, using five of the best-performing features. Accuracy is indeed excellent, as shown in Table 2. The results reflect (supervised) ten-fold cross validation evaluation. In the table, ‘EUR’ stands for Europarl, ‘HAN’ for Hansard, and ‘LIT’ for the literary corpus.

**Table 2.** In-domain (cross-validation) classification accuracy using various feature sets

Feature/corpus	EUR	HAN	LIT	TED
Function words (FW)	96.3	98.1	97.3	97.7
Character trigrams	98.8	97.1	99.5	100.0
POS trigrams	98.5	97.2	98.7	92.0
Contextual FW	95.2	96.8	94.1	86.3
Cohesive markers	83.6	86.9	78.6	81.8

However, when a classifier is trained on one domain and tested on another, the domain-dependence of the supervised method is revealed. This pattern persists even when a classifier is trained on *two* domains and is tested on the third, as shown in Table 3. (The TED corpus was too small to include.)

As these tables clearly show, while in-domain cross-validation evaluation (the rightmost column) shows excellent accuracy, even when the dataset is a mixture of two domains, the classifiers are limited to the domain(s) they were trained on and do not scale up to other datasets.

To remedy the obstacle of domain-dependence, Rabinovich and Wintner (2015) proposed to use unsupervised *clustering* on the entire dataset.

<sup>2</sup> We use “domain” rather freely henceforth to indicate not only the topic of a corpus but also its modality (written vs. spoken), register, genre, date, etc.

**Table 3.** Pairwise (left) and leave-one-out (right) cross-domain classification using function words

Train / Test EUR HAN LIT X-validation				Train / Test EUR HAN LIT X-validation				
EUR		60.8	56.2	96.3	EUR + HAN		63.8	94.0
HAN	59.7		58.7	98.1	EUR + LIT	64.1		92.9
LIT	64.3	61.5		97.3	HAN + LIT	59.8		96.0

More specifically, they employed the KMeans algorithm (Lloyd 1982), using KMeans++ initialization (Arthur and Vassilvitskii 2007) and Principal Component Analysis (Pearson 1901) for dimension reduction. Of course, since the method is unsupervised, the labels of the resulting classes are not known; assuming “gold” labels (that is, judging the class by the majority of the instances in it) the accuracy is surprisingly high, as shown in Table 4.

**Table 4.** Clustering results using various feature sets

Feature/corpus	EUR	HAN	LIT	TED
FW	88.6	88.9	<b>78.8</b>	<b>87.5</b>
Char trigrams	72.1	63.8	70.3	78.6
POS trigrams	<b>96.9</b>	76.0	70.7	76.1
Contextual FW	92.9	<b>93.2</b>	68.2	67.0
Cohesive markers	63.1	81.2	67.1	63.0

Furthermore, Rabinovich and Wintner (2015) suggested a simple yet effective method for determining the correct label of the classes and showed that it was perfectly (100%) accurate on all the datasets they experimented with. As the label can be accurately determined, several classifiers, reflecting different feature sets, can be combined in an ensemble, using voting among classifiers to establish the class of each instance. The results of this ensemble clustering are shown in Table 5, and reveal a fully-unsupervised, highly accurate method for discriminating between originals and translations in a single domain.

Finally, Rabinovich and Wintner (2015) defined two simple methods for clustering in a mixed-domain scenario, *flat* and *hierarchical*. The hierarchical method first clusters a mixture of texts into domains (e.g., using KMeans), and then separates each of the resulting (presumably, domain-coherent) clusters into two sub-clusters, presumably originals and translations. The flat approach assumes that the number of domains,  $k$ , is known, and attempts to divide the data set into  $2 \times k$  clusters, expecting classification by domains and by translationese status, simultaneously. The results, experimenting with a mixture of two and then three different datasets, are shown in Table 6.

**Table 5.** Clustering consensus by voting

Method/corpus	EUR	HAN	LIT	TED
FW	88.6	88.9	78.8	87.5
FW				
Char trigrams	91.1	86.2	78.2	<b>90.9</b>
POS trigrams				
FW				
POS trigrams	<b>95.8</b>	89.8	72.3	86.3
Contextual FW				
FW				
Char trigrams				
POS trigrams	94.1	<b>91.0</b>	<b>79.2</b>	88.6
Contextual FW				
Cohesive markers				

**Table 6.** Flat and hierarchical clustering of domain-mix using function words

Method/corpus	EUR	EUR	HAN	EUR
	HAN	LIT	LIT	HAN
Flat	<b>92.5</b>	60.7	77.5	66.8
Two-phase	91.3	<b>79.4</b>	<b>85.3</b>	<b>67.5</b>

Summing up, it is possible to accurately identify translationese even in mixed-domain scenarios, but the accuracy of the classification deteriorates as the number of different domains increases.

## 4 Applications to Machine Translation

The special properties of translationese have ramifications to NLP applications, and in particular to statistical machine translation (SMT). Until recently, research in SMT was divorced from scholarly work in translation studies. This was changed by a series of works, pioneered by Kurokawa et al. (2009) and further elaborated by Lembersky et al. (2011, 2012a, 2012b, 2013) and Twitto et al. (2015). This section summarizes some of the main results reported in these works.

The standard SMT paradigm (Brown et al. 1990, 1993) is based on the *noisy channel model*, whereby the best translation  $\hat{T}$  of a source-language sentence  $S$  is a target-language sentence  $T$  that maximizes some function combining the *faithfulness* of  $(T, S)$  and the *fluency* of  $T$ . The standard notation assumes that the

task is to translate a *foreign* sentence  $F = f_1, \dots, f_m$  into an *English* sentence  $E = e_1, \dots, e_l$ . Thus, the best translation is:

$$\begin{aligned}\hat{E} &= \arg \max_E P(E \mid F) \\ &= \arg \max_E \frac{P(F|E) \times P(E)}{P(F)} \\ &= \arg \max_E P(F \mid E) \times P(E)\end{aligned}$$

The noisy channel thus requires two components: a *translation model* and a *language model*:

$$\hat{E} = \arg \max_{E \in \text{English}} \underbrace{P(F \mid E)}_{\text{Translation model}} \times \underbrace{P(E)}_{\text{Language model}}$$

The language model is responsible for the fluency of the translation outcome; it estimates  $P(E)$  from a monolingual  $E$  corpus. The translation model is responsible for the faithfulness of the translation, and it estimates  $P(F \mid E)$  from a bilingual parallel corpus. In addition, a *decoder* is used to produce the most probable  $E$  given  $F$ , but it will be ignored here.

#### 4.1 Language Models

As mentioned, language models (LMs) are estimated from monolingual corpora of the target language. The common wisdom in SMT used to be that the larger the corpora, the better the translation quality (Brants and Xu 2009); the research question we discuss here is whether this is indeed the case, and in particular, whether corpora compiled from *translated* texts are better for SMT than those compiled from original texts.

Lembersky et al. (2012b) set out to investigate the fitness of language models compiled from translated texts vs. the fitness of LMs compiled from original texts; and whether these differences carry over to SMT, namely whether language models compiled from translated texts are better for MT than LMs compiled from original texts. The fitness of a language model to a reference corpus is evaluated using *perplexity*: the perplexity  $PP$  of a language model  $LM$  with respect to a sequence of words  $w_1, \dots, w_N$  is defined in terms of the probability  $LM$  assigns to the sequence, as follows:

$$PP(LM, w_1 w_2 \dots w_N) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P_{LM}(w_i \mid w_1 \dots w_{i-1})}}$$

Note that the lower the perplexity, the better the fitness of the LM to the reference set.

Lembersky et al. (2012b) first trained LMs from Europarl corpora and tested them on Europarl reference translations. They experimented with  $n$ -gram language models, where  $n$  ranged between 1 and 4, and with translations from four different languages (French, German, Spanish and Italian) to English. In all

these cases, the results were the same: LMs compiled from translated texts were consistently better than ones compiled from originals; furthermore, LMs compiled from the same source language as the one from which the references were compiled from were the best. Table 7 demonstrates these results for translations from German to English; the results for the other three source languages are very similar. In this and in subsequent tables, ‘O-EN’ is original English, ‘T-XX’ refers to translations to English from some language XX, and ‘Mix’ is a mixture of translated and original texts.

**Table 7.** The perplexity of various language models on a reference set of German translated into English

LM	1-gram	2-gram	3-gram	4-gram
Mix	451.50	93.00	69.36	66.47
O-EN	<i>468.09</i>	<i>103.74</i>	<i>79.57</i>	<i>76.79</i>
T-DE	<b>443.14</b>	<b>88.48</b>	<b>64.99</b>	<b>62.07</b>
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

These results may be attributed to the contents of the various language models, and in particular to specific named entities in them, as mentioned in Sect. 3.2 above. To control this, Lembersky et al. (2012b) further compiled LMs that abstracted away from the actual words in the corpus. They first replaced proper names by a special symbol; then did the same for nouns; and finally, they replaced all words by their parts of speech. In all these cases, the results remained robust, although the differences among the various LMs decreased. In conclusion, LMs compiled from translations, preferably from the same source language as the references, fit the reference set better than LMs compiled from originals.

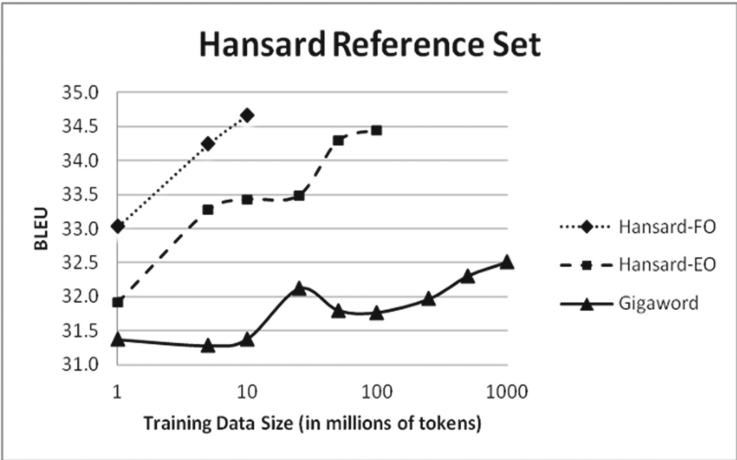
In order to test the second hypothesis, Lembersky et al. (2012b) trained SMT systems (Koehn et al. 2007) using various LMs and evaluated their quality on a reference set. As is common in SMT, the quality was measured in terms of BLEU scores (Papineni et al. 2002), where the higher the score the better the system is. The results are reported in Table 8. Each pair of columns refers to a specific SMT system, translating from some language into English. The rows indicate the corpus used for the language model.

Obviously, in all cases the best SMT systems are those that use LMs compiled from corpora that were translated from the source language. LMs compiled from other translated texts (from different languages) come next. The worst LMs are those that were compiled from originals. The automated evaluation results were corroborated by manual evaluation, in which humans were asked to assess the quality of the translations. Again, human evaluators preferred the SMT outputs that were produced with the LMs that were based on translations.

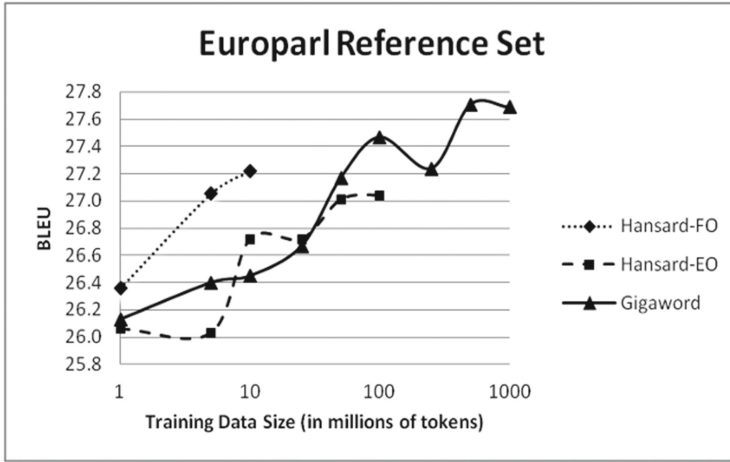
**Table 8.** Quality of SMT using various language models

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	<b>21.90</b>	T-DE	28.01	T-DE	24.62	T-DE	24.26
T-FR	21.16	T-FR	<b>29.14</b>	T-FR	25.37	T-FR	23.56
T-IT	21.29	T-IT	28.75	T-IT	<b>25.96</b>	T-IT	23.87
T-NL	21.20	T-NL	28.11	T-NL	24.77	T-NL	<b>24.52</b>

Going back now to the issue of the size of the training corpus, Fig. 3 plots the quality of French-to-English SMT systems, in terms of BLEU scores, against the size of the French monolingual corpus that was used to train the language model of the systems. The three graphs correspond to a LM compiled from French-translated-to-English texts (FO), a LM compiled from English original texts (EO) and one compiled from a very large corpus which may include originals or translations, in a different domain (Gigaword). Evidently, to reach the same quality obtained using the translated LM, an order of magnitude more original text is needed. The out-of-domain corpus fares much worse. Figure 4 shows a similar plot, where *all* three corpora are out-of-domain, as the reference set used for the evaluation consists of Europarl sentences whereas the LMs are compiled from Hansard or Gigaword materials. Again, much more original data are needed to match the quality of SMT systems built with translated LMs.



**Fig. 3.** SMT quality as a function of the size of the language model



**Fig. 4.** SMT quality as a function of the size of the language model, out-of-domain evaluation

## 4.2 Translation Models

In terms of the *translation* model, Kurokawa et al. (2009) have shown that translation models (TMs) compiled from parallel corpora that were (manually) translated in the same direction as that of the SMT task are better than ones translated in the reverse direction. Lembersky et al. (2013) replicated these results, using more varied datasets (both Europarl and the Canadian Hansards) and more language pairs. Table 9 shows the quality of SMT between three languages (six language pairs in total), using TMs compiled in the same direction as that of the task (source-to-target,  $S \rightarrow T$ ) and in the reverse direction (target-to-source,  $T \rightarrow S$ ).

**Table 9.** Quality of SMT as a function of the direction of the translation of the TM

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

Focusing now on a single language pair, namely French to English, Table 10 shows the massive savings in training materials that is facilitated by using  $S \rightarrow T$  parallel corpora rather than  $T \rightarrow S$  ones. Other language pairs showed very similar patterns.



**Table 10.** Quality of SMT as a function of the direction of the translation of the TM and of various sizes of the training dataset

Task	$S \rightarrow T$	$T \rightarrow S$
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

In a realistic scenario, however, one has access to a large parallel corpus, parts of which were manually translated in the “right” direction, and parts of which in the reverse direction. Lembersky et al. (2013) proposed several methods for adapting the “wrong” subset of the parallel corpus to translationese. The technical details are too complicated to describe here, but the results robustly showed that the SMT systems that resulted from the best adaptation of the translation model were significantly better than ones that used either the entire corpus (in a naïve way) or only its  $S \rightarrow T$  subset.

Of course, to benefit from these results, the parallel corpus has to be annotated with information pertaining to the translation direction; such annotation is typically not available. However, Twitto et al. (2015) showed that this obstacle can be overcome, as the predictions of translationese classifiers are as good as meta-information. First, when a monolingual corpus in the target language is given, to be used for constructing a language model, predicting the translated portions of the corpus, and using only them for the language model, is as good as using the entire corpus. Second, identifying the portions of a parallel corpus that are translated in the direction of the translation task, and using only them for the translation model, is as good as using the entire corpus. Twitto et al. (2015) presented results from several language pairs and various data sets, indicating that these results were robust and general.

## 5 Conclusion

We demonstrated above that awareness of translationese can significantly improve the quality of machine translation. Insights drawn from translation studies can also improve other NLP applications. For example, the task of *native language identification* attempts to identify the mother tongue of non-native writers (typically learners) based on texts they composed in a foreign language (Tetreault et al. 2013). The classifier of Tsvetkov et al. (2013) achieved an accuracy of 80–85% on an 11-way classification task (i.e., texts were authored by native speakers of eleven different languages) using several features that were inspired by the translationese features of Volansky et al. (2015).

The reason that native language identification is similar to the identification of translationese has to do with *interference*. In both of these cases, elements of one linguistic system (the source language in the case of translation, and the native language in the case of non-native speakers) interfere with the production of the target language, and can be traced back by the classifier. In fact, interference is so powerful that it overshadows other, more subtle, properties of translationese. In work in progress, we have been able to demonstrate that translations from related languages (e.g., Spanish and Italian) are closer to each other than translations from more distant languages (e.g., German). This interference is so powerful that it is possible to cluster together related languages based only on their translations to English.

The relations between translationese and non-native language have been explored by Rabinovich et al. (2016), who showed clear similarities but also some significant differences between these two language varieties. In the future, we intend to further explore these relations, focusing not only on advanced, highly fluent non-native speakers but also on learners. We believe that better understanding of the linguistic properties of such language varieties are not only interesting in and of themselves, but may help engineer better NLP systems, as we hope to have shown in this chapter.

**Acknowledgements.** I am grateful to Noam Ordan for his immense help with the research reported here. Thanks are due to all my other collaborators on these works, including Gennadi Lembersky, Vered Volansky, Udi Avner, Naama Twitto and Ella Rabinovich. Special thanks are due to Agata Savary, not least for her continuous encouragement. I am grateful to the three anonymous reviewers whose constructive comments greatly improved the quality of the presentation. This research was supported by a grant from the Israeli Ministry of Science and Technology.

## References

- Al-Shabab, O.S.: Interpretation and the Language of Translation: Creativity and Conventions in Translation. Janus, Edinburgh (1996)
- Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics, pp. 1027–1035 (2007). <http://dl.acm.org/citation.cfm?id=1283383.1283494>. ISBN 978-0-898716-24-5
- Avner, E.A., Ordan, N., Wintner, S.: Identifying translationese at the word and sub-word level. Digit. Scholarsh. Humanit. **31**(1), 30–54 (2016). <http://dx.doi.org/10.1093/llc/fqu047>
- Baker, M.: Corpus linguistics and translation studies: implications and applications. In: Baker, M., Francis, G., Tognini-Bonelli, E. (eds.) Text and Technology: In Honour of John Sinclair, pp. 233–252. John Benjamins, Amsterdam (1993)
- Baroni, M., Bernardini, S.: A new approach to the study of translationese: machine-learning the difference between original and translated text. Lit. Linguist. Comput. **21**(3), 259–274 (2006). <http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1>

- Ben-Ari, N.: The ambivalent case of repetitions in literary translation. Avoiding repetitions: a “universal” of translation? *Meta* **43**(1), 68–78 (1998)
- Blum-Kulka, S.: Shifts of cohesion and coherence in translation. In: House, J., Blum-Kulka, S. (eds.) *Interlingual and Intercultural Communication Discourse and Cognition in Translation and Second Language Acquisition Studies*, vol. 35, pp. 17–35. Gunter Narr Verlag, Tübingen (1986)
- Blum-Kulka, S., Levenston, E.A.: Universals of lexical simplification. *Lang. Learn.* **28**(2), 399–416 (1978)
- Blum-Kulka, S., Levenston, E.A.: Universals of lexical simplification. In: Faerch, C., Kasper, G. (eds.) *Strategies in Interlanguage Communication*, pp. 119–139. Longman, Harlow (1983)
- Brants, T., Xu, P.: 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*, Boulder, Colorado, May 2009, pp. 3–4. Association for Computational Linguistics (2009). <http://www.aclweb.org/anthology/N/N09/N09-4002>
- Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. *Comput. Linguist.* **16**(2), 79–85 (1990). ISSN 0891-2017
- Brown, P.F., Della Pietra, V.J., Della Pietra, S.A., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* **19**(2), 263–311 (1993). ISSN 0891-2017
- Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. *Comput. Linguist.* **16**(1), 22–29 (1990). ISSN 0891-2017
- Frawley, W.: Prolegomenon to a theory of translation. In: Frawley, W. (ed.) *Translation. Literary, Linguistic and Philosophical Perspectives*, pp. 159–175. University of Delaware Press, Newark (1984)
- Gellerstam, M.: Translationese in Swedish novels translated from English. In: Wollin, L., Lindquist, H. (eds.) *Translation Studies in Scandinavia*, pp. 88–95. CWK Gleerup, Lund (1986)
- Gries, S.T., Wulff, S.: Regression analysis in translation studies. In: Oakes, M.P., Ji, M. (eds.) *Quantitative Methods in Corpus-Based Translation Studies. Studies in Corpus Linguistics*, vol. 51, pp. 35–52. John Benjamins, Philadelphia (2012)
- Grieve, J.: Quantitative authorship attribution: an evaluation of techniques. *Lit. Linguis. Comput.* **22**(3), 251–270 (2007)
- Halverson, S.: The cognitive basis of translation universals. *Target* **15**(2), 197–241 (2003)
- Ilisei, I., Inkpen, D.: Translationese traits in Romanian newspapers: a machine learning approach. *Int. J. Comput. Linguist. Appl.* **2**(1–2) (2011)
- Ilisei, I., Inkpen, D., Corpas Pastor, G., Mitkov, R.: Identification of translationese: a machine learning approach. In: Gelbukh, A. (ed.) *CICLing 2010. LNCS*, vol. 6008, pp. 503–511. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-12116-6\\_43](https://doi.org/10.1007/978-3-642-12116-6_43). <http://dx.doi.org/10.1007/978-3-642-12116-6>. ISBN 978-3-642-12115-9
- Kenny, D.: *Lexis and Creativity in Translation: A Corpus-Based Study*. St. Jerome, Northampton (2001). ISBN 9781900650397
- Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: *Proceedings of the Tenth Machine Translation Summit*, pp. 79–86. AAMT (2005). <http://mt-archive.info/MTS-2005-Koehn.pdf>

- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: 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, Prague, Czech Republic, pp. 177–180. Association for Computational Linguistics, June 2007. <http://www.aclweb.org/anthology/P07-2045>
- Koppel, M., Ordan, N.: Translationese and its dialects. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 1318–1326. Association for Computational Linguistics, June 2011. <http://www.aclweb.org/anthology/P11-1132>
- Kurokawa, D., Goutte, C., Isabelle, P.: Automatic detection of translated text and its impact on machine translation. In: Proceedings of MT-Summit XII, pp. 81–88 (2009)
- Laviosa, S.: Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta* **43**(4), 557–570 (1998)
- Laviosa, S.: *Corpus-Based Translation Studies: Theory, Findings, Applications. Approaches to Translation Studies*. Rodopi, Amsterdam (2002). ISBN 9789042014879
- Lembersky, G., Ordan, N., Wintner, S.: Language models for machine translation: original vs. translated texts. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, pp. 363–374. Association for Computational Linguistics, July 2011. <http://www.aclweb.org/anthology/D11-1034>
- Lembersky, G., Ordan, N., Wintner, S.: Adapting translation models to translationese improves SMT. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, pp. 255–265. Association for Computational Linguistics, April 2012a. <http://www.aclweb.org/anthology/E12-1026>
- Lembersky, G., Ordan, N., Wintner, S.: Language models for machine translation: original vs. translated texts. *Comput. Linguist.* **38**(4), 799–825 (2012b). <http://dx.doi.org/10.1162/COLL.a.00111>
- Lembersky, G., Ordan, N., Wintner, S.: Improving statistical machine translation by adapting translation models to translationese. *Comput. Linguist.* **39**(4), 999–1023 (2013). <http://dx.doi.org/10.1162/COLL.a.00159>
- Lloyd, S.: Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**(2), 129–137 (1982). doi:[10.1109/TIT.1982.1056489](https://doi.org/10.1109/TIT.1982.1056489). ISSN 0018-9448
- Munday, J.: A computer-assisted approach to the analysis of translation shifts. *Meta* **43**(4), 542–556 (1998)
- Olohan, M.: How frequent are the contractions? A study of contracted forms in the translational English corpus. *Target* **15**(1), 59–89 (2003)
- Överås, L.: In search of the third code: an investigation of norms in literary translation. *Meta* **43**(4), 557–570 (1998)
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, Morristown, NJ, USA, pp. 311–318. Association for Computational Linguistics (2002). <http://dx.doi.org/10.3115/1073083.1073135>
- Pearson, K.: On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**(6), 559–572 (1901)

- Popescu, M.: Studying translationese at the character level. In: Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N. (eds.) *Proceedings of RANLP 2011*, pp. 634–639 (2011)
- Rabinovich, E., Wintner, S.: Unsupervised identification of translationese. *Trans. Assoc. Comput. Linguist.* **3**, 419–432 (2015). <https://tac12013.cs.columbia.edu/ojs/index.php/tac1/article/view/618>. ISSN 2307-87X
- Rabinovich, E., Nisioi, S., Ordan, N., Wintner, S.: On the similarities between native, non-native and translated texts. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pp. 1870–1881, August 2016. <http://aclweb.org/anthology/P/P16/P16-1176.pdf>
- Shlesinger, M.: Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum. Master's thesis, Tel Aviv University, Faculty of the Humanities, Department of Poetics and Comparative Literature (1989)
- Teich, E.: *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Mouton (2003)
- Tetreault, J., Blanchard, D., Cahill, A.: A report on the first native language identification shared task. In: *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, June 2013
- Toury, G.: *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv (1980)
- Toury, G.: *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam/Philadelphia (1995)
- Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., Dyer, C.: Identifying the L1 of non-native writers: the CMU-Haifa system. In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 279–287. Association for Computational Linguistics, June 2013. <http://www.aclweb.org/anthology/W13-1736>
- Twitto, N., Ordan, N., Wintner, S.: Statistical machine translation with automatic identification of translationese. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, pp. 47–57. Association for Computational Linguistics, September 2015. <http://aclweb.org/anthology/W15-3002>
- van Halteren, H.: Source language markers in EUROPARL translations. In: Scott, D., Uszkoreit, H., (eds.) *Proceedings of the 22nd International Conference on Computational Linguistics, COLING 2008*, Morristown, NJ, USA, pp. 937–944. Association for Computational Linguistics (2008). ISBN 978-1-905593-44-6
- Vanderauwerea, R.: *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature*. Rodopi, Amsterdam (1985)
- Volansky, V., Ordan, N., Wintner, S.: On the features of translationese. *Digit. Scholarsh. Humanit.* **30**(1), 98–118 (2015)

Business Intelligence

6th European Summer School, eBISS 2016, Tours,  
France, July 3-8, 2016, Tutorial Lectures

Marcel, P.; Zimányi, E. (Eds.)

2017, IX, 139 p. 61 illus., Softcover

ISBN: 978-3-319-61163-1