Finding the Origin of a Translated Historical Document

Zahurul Islam
MassineScheffer & Company GmbH
Berlin, Germany
zahurul.islaam@massine.com

Natia Dundia
Department of Modern Languages
Goethe University Frankfurt, Germany
dunduanatia@gmail.com

Abstract

Gospels are one type of translated historical document. There are many versions of the same Gospel that have been translated from the original, or from another Gospel that has already been translated into a different language. Nowadays, it is difficult to determine the language of the original Gospel from where these Gospels were translated. In this paper we use a supervised machine learning technique to determine the origin of a version of the *Georgian* Gospel.

1 Introduction

Translation is a process of rewriting an original text in a different language (Lefevere, 2002). It is one of the oldest text manipulation related processes. Gospels are historical documents that were translated centuries ago. There are many versions of the same Gospel, translated from the original, or from another Gospel that had already been translated into a different language. Nowadays, it is unclear what was the language of the original Gospel from where these were translated. Historians and linguists are uncertain as to the origin of such historical documents. The *Georgian* Gospels are translated from *Armenian* or *Greek* Gospels (Lang, 1957). There are about 300 manuscripts of the four Gospels in *Georgian* that are translated from different languages (Kharanauli, 2000). Linguists are able to narrow down potential origins by looking at different linguistic properties, but skeptical to choose a single origin. We have three such Gospels in *Georgian*, *Armenian* and *Greek*, where linguists believe that *Armenian* or *Greek* are the potential origin. In this paper we use a supervised machine learning technique to find out the correct origin of a version of the *Georgian* Gospel.

One of the challenges of dealing with historical data is the requirement of specific knowledge of languages that are not spoken at present day. If the language is currently spoken, it is likely that many properties have changed due to language evolution. Due to this issue, the available historical data set is very small in size, which proves a challenge for machine learning algorithms.

From the early stage of translation studies research, translation scholars proposed different kinds of properties of source text and translated text. Recently, scholars in this area identified several properties of the translation process with the aid of corpora (Baker, 1993; Baker, 1996; Olohan, 2001; Laviosa, 2002; Hansen, 2003; Pym, 2005; Toury, 1995). These properties are subsumed under four keywords: *explicitation*, *simplification*, *normalization*, *levelling out* and *interference*.

In this paper, we use texts from modern language to train a Support Vector Machine (SVM) that can be used to identify the original source of the *Georgian* Gospel.

The paper is organized as follows: Section 2 introduced the historical documents that we are dealing with here, Section 3 discusses related work, followed by a discussion of the nature of a translated text in Section 4. The methodology is described in Section 5. The corpus of modern languages is described briefly in Section 6 followed by a discussion of different features we used in this paper in Section 7. The experiment and evaluation in Section 8 and finally, we present conclusions in Section 9.

2 The historical documents

Gospels are among the very first documents that were translated into *Georgian* language following the invention of the Georgian alphabet (Lang, 1957). The history begins with the palimpsest manuscripts from the *fifth* or *sixth* centuries and ends with the manuscripts from the *eighteenth* century. There are many open debates on the table about the origin of the Georgian translation of

Language	Sentences	Average Sentence Length	Average Word Length
Georgian	3738	18.96%	4.71%
Armenian	3738	19.15%	4.00%
Greek	3738	20.40%	4.24%

Table 1: Historical corpus statistics

the holy script. According to Blake (1932), many translations were made from the Gospels of *Syrian* and *Armenian*.

However, recent studies show two more sources from where the holy scripture were translated into *Georgian*. The first one is the *Palestinian* and other one is the *Antiochian/Constantinopolian* (Kharanauli, 2000).

The precise date of these translations are unknown, but the earliest translations of the *Georgian* Bible are presented in the lower script of palimpsests, the so-called *Xanmeti* fragments. *Xanmeti* is a term already used by the famous *Georgian* monk, religious writer and translator *George the Athonite*[1]. He denotes the text where the *x-prefix* is employed to mark the second subject and the third object persons in the *Georgian* verb. This prefix has not occurred in the inscriptions since the seventh Century. Based on philological data, these fragments are dated from *fifth* to *seventh* centuries. Codicological study of the folio size reveals that they are fragments of quite large codices, and it can be assumed that these codices included several books of the Bible.

Currently, there are about 300 manuscripts of the four Gospels in *Georgian* (Kharanauli, 2000). Among these, about 40 codices include text version of Georgian Gospels. The Gospel considered for this study is believed to be translated from *Armenian* or *Greek*. These Gospels are digitized and aligned manually. The aligned corpus of the Georgian Gospel manuscripts present the texts in their original form side by side, which means that a) nothing is corrected, not even the mistakes presumably made by copyists; and b) abbreviations remain discernible as they are, with the abbreviated letters being indicated in brackets. Table 1 shows the statistics of the Gospels.

3 Related work

There is no work found that is exactly relevant to the problem we are dealing here. Lang (1957) studied *Georgian* Gospels and their origins. The first *Georgian* Gospels were translated from an *Armenian* version (Lang, 1957). The Gospels that were translated in the late ninth century show signs of revision by reference to the *Greek* Gospels.

Corpus-based translation studies is a recent field of research with a growing interest within the field of computational linguistics. Baroni and Bernardini (2006) started corpus-based translation studies empirically, where they work on a corpus of geo-political journal articles. A SVM was used to distinguish original and translated Italian text using n-gram based features. According to their results, word bigrams play an important role in the classification task.

Van Halteren (2008) uses the *Europarl* corpus for the first time to identify the source language of text for which the source language marker was missing. In their experiments, the support vector regression was the best performing method.

Pastor et al. (2008) and Ilisei et al. (2009; 2010) perform classification of Spanish original and translated text. The focus of their works is to investigate the *simplification* relation that was proposed by (Baker, 1996). In total, 21 quantitative features (e.g. a number of different POS, Average Sentence Length (ASL), the parse-tree depth etc.) were used where, nine (9) of them are able to grasp the simplification translation property.

Koppel and Ordan (2011) have built a classifier that can identify the correct source of the translated text (given different possible source languages). They have built another classifier, which can identify source text and translated text. However, the limitation of this study is that they only used a corpus of English original text and English text translated from various European languages. A list of 300 function words (Pennebaker et al., 2001) was used as feature vector for these classifications.

Popescu (2011) uses *string kernels* (Lodhi et al., 2002) to study translation properties. A classifier was built to classify English original texts and English translated texts from French and German books that were written in the nineteenth century. The *p-spectrum* normalized kernel was used for the experiment. The system performs poorly when the source language of the training corpus is different from the one of the test corpus.

Islam and Hoenen (2013) used a source and translated texts of six European languages in order to classify translated texts according to source languages. As features, they have used the hundred

[1] Wikipage: http://en.wikipedia.org/wiki/George_the_Athonite

most frequent words. It is important to consider the properties of language family when dealing with source and translated texts (Islam and Hoenen, 2013).

Features used by Koppel and Ordan (2011) and Islam and Hoenen (2013) are language dependent. As we use texts from twenty-one European languages to build the training model, we only use features that are language and linguistic tools independent. It is also important to consider different properties of translated and source texts proposed by translation scholars.

4 Translation properties

Recently, translation scholars proposed different translation properties using monolingual or comparable corpus. These properties are described in the following subsections.

4.1 Explicitation

Translators are biased to make translations more *explicit* in order to resolve ambiguities that might be inherited in the translated text. Vinay and Darbelnet (1958) used the term *explicitation* as " a process of introducing information into the target language which is present only implicitly in the source language, but which can be derived from the context or situation"(Vinay and Darbelnet, 1995; Pym, 2005). However, Blum-Kulka (1986) first claimed *explicitation* as a translation universal where she studied translated *French* texts from *English* by professional and non-professional translators. Seguinot (1988) provides an empirical study using two translated texts from *French* to *English*. There is a greater level of *explicitness* in the translated texts as linking words and conversion of subordinate clauses into coordinate clauses.

4.2 Simplification

The *simplification* translation property shows the tendency of a translator to simplify a text in order to improve the readability of a translated text. Blum-Kulka and Levenston (1978) mention the term *simplification* as part of the lexical simplification using a small data set of *English* and *Hebrew* translations. According to them, translators use techniques such as *avoidance* and *approximation* in the translation process to make a translated text simpler for the target readers. Later, Baker (1996) also observed this tendency in the translated texts.

To make a translated text simpler, the translator often breaks up complex sentences into two or more sentences. This tendency can be found in the ASL. That is, the ASL in a translated text will be shorter than a source text.

4.3 Normalization

The *normalization* property shows a translator's effort to meet the normative criteria of the target language. It is a translator's tendency to conform to patterns and practices that are typical of the target language, even to exaggerate their use. This property can be observed in a translated text that contains very little trace of the source language. However, the opposite scenario can be seen as well, where the translation is influenced by the source language. In that case *normalization* will be weakened. The influence of *English* can be visible in many software manuals that are translated from *English*. Hansen (2003) stated that this contrary tendency also can be seen in *interpreting*, where the interpreter tries to finish an unfinished sentence and to render an ungrammatical structure into something grammatical.

4.4 Levelling out

Baker (1996) refers to *levelling out* as "the tendency of translated text to gravitate towards the center of a continuum". That is also known as *convergence* (Laviosa, 2002), where she stated that a "relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features" such as lexical density or sentence length, in contrast to source texts. If we have a sub-corpus of translated texts from different languages to the same language, and have source texts in the same language, then translated texts from different languages will be similar in terms of *lexical density*, TTR, and ASL; but will be different than the source texts. More specifically, translated texts from different languages will be alike but will be different than the source texts.

4.5 Interference

Toury (1995) has a different theory that is different from the translation properties described above. He stated that "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text," That is, some *interference* effects will be observable in translated texts that are carried from source texts. These effects will be in the form of *negative transfer* or in the form of *positive transfer*. As an example, specific properties of the English language are visible in user manuals that have been translated to other languages from English (for instance, word order) (Lzwaini, 2003). We can summarize this translation properties in a way that a translated text from different source languages will be sufficiently different from each other.

5 Methodology

The above section describes the properties of translation. Based on these properties, a translated text is different than the corresponding source text. Properties proposed by translation scholars, focus on texts and the translation process. Our assumption is that even though historical texts were translated many hundreds of years ago, there are some properties that are common to modern texts and the recent translation process.

We model the task as a classification task where we use a SVM implementation to find the correct origin of the *Georgian* Gospel. Linguists believe that the *Georgian* Gospel is a translated document. They narrowed down potential origins by looking at different linguistic properties compared to the *Greek* and *Armenian* Gospel. Before finding the source of the *Georgian* Gospel, it is necessary to check that the Gospel itself is a translated document. If the gospel is classified as a translated document then we can move further to find the source. The gospel that has properties of an original document will be the closest candidate for the origin *Georgian* gospel.

In order to build a training model, we use modern texts from different European languages. We have compiled a suitable corpus for this task from the Europarl corpus (Koehn, 2005). This task requires features that are language independent and do not require any linguistic pre-processing. So, we have explored different features that are quantitative indicators of translation properties mentioned above. Finally, we have collected a list of useful features that are listed in Section 7. We use standard classification *accuracy* and *F-Score* in order to measure usefulness of a feature. At the beginning the feature list contains only ASL. We have added a new feature in the list if and only if the classification *accuracy* and *F-Score* improve by adding the feature with existing feature set. The feature collection process will be continued until the classifier achieves a reasonable *accuracy F-Score*. Figure 1 shows the approach we follow in this paper. Finally, the whole corpus of modern texts will be considered for building the final training model.

The final training model and the collected feature set will be used in order to find the origin of the *Georgian* Gospel. We prepare the Gospels data into two sets similarly as the training data. The first set of data will contain texts from *Armenian* and *Georgian* Gospels and the other one will contain texts from *Greek* and *Georgian* Gospels.

6 Corpus of modern texts

The area of translation studies lack corpora by which scholars can validate their theoretical claims, for example, regarding the scope of characteristics of the translation properties. This scope is obviously affected by the membership of the source and target languages to language families. Though the exploration of universally valid characteristics of translations is an important topic, there are not many resources for testing corresponding hypotheses.

There are many parallel and multilingual corpora available nowadays. Most of them are not useful for translation studies immediately as they require customization. Islam and Mehler (2012) provide a customized resource in which the languages of all source texts and their translations are annotated sufficiently. The resource they provide is a customized version of the well-known *Europarl* corpus (Koehn, 2005). A central feature of this corpus is that it provides information on sentence-related alignments that can be explored for finding characteristics of the translation relation.

The language annotation in the *Europarl* corpus is not reliable because of erroneous annotations introduced by translators. There are many cases where one speaker has multiple speeches in different languages that cause problems for identifying the speaker's native language.

In order to resolve this issue we have collected the name of the member of the *European parliament* and their native language manually. We collected names from the current members list page [2] of the *European parliament*. Names of former members are collected from the correspond-

[2] http://www.europarl.europa.eu/meps/en/full-list.html

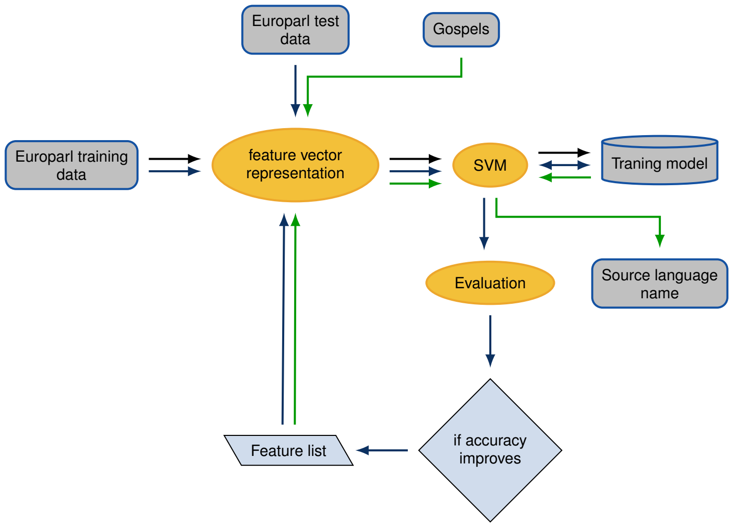

Figure 1: Machine learning approach to find the source of the Georgian Gospel

ing *Wikipedia* pages. The official language of the country of each member is assigned as the native language of a speaker. Members from *Belgium* and *Luxembourg* are not considered as we are not sure about the language spoken by members from these countries in the *European parliament*. Each member from *Finland* is assigned to the *Finnish* language. Finally, the list contains $2,125$ member names and their native language. This list is used to extract source and translated texts from the *Europarl* corpus. The corpus contains $2,646,765$ parallel sentences from 412 language pairs of 21 European languages. We believe that such a corpus is an ideal resource for the problem we are addressing in this paper.

7 Features

As the training corpus contains texts from twenty-one European languages, we only experiment with *lexical* and *information-theoretic* features. Pastor et al. (2008) used various *lexical*, *syntactic* and *discourse* related features. Also, Ilisei et al. (2009; 2010) used similar type of features. The following sub sections describe features that are finally selected for the feature list.

7.1 Lexical features

Different lexical features are being used from the beginning of corpus based translation studies. These features are popular for other NLP applications such as *text readability classification*. The reason behind the popularity is that these are language independent and do not require any linguistic pre-processing.

The ASL is a quantitative measure of syntactic complexity. Generally, the syntax of a longer sentence is more complex than that of a shorter sentence. A translator tries to make a translation *explicit* and also *simple*. Translated texts might become longer due to the *explicitation*. However, opposite can happen when a translator tries to make a translation simpler. Table 2 shows behavior of some features in source and translated texts of four European languages. Translations of German, French and Dutch are more explicit than Spanish. The Average Word Length (AWL) is another useful lexical feature. Most of the cases, the AWL in translated texts is longer than source texts. It would be interesting to see the behavior of AWL in source and translated texts of an agglutinative language.

The *Average number of complex words* feature is related to the AWL. A translated text will be difficult for readers if it contains more complex words. The average length of English written words is 5.5 (Nádas, 1984) letters. We define a *complex word* as any word that contains 10 or more letters.

The Type Token Ratio (TTR), which indicates the lexical density of text, has been considered as useful features by Pastor et al. (2008) and Also, Ilisei et al. (2009; 2010). Low lexical densities involve a great deal of repetition with the same words occurring again and again. Conversely, high lexical density shows the diverseness of a text. A diverse text is supposed to be difficult for readers, generally children (Temnikova, 2012). There are many different version of TTR formulas avail-

	ASL		AWL		Entropy	
	Source	Translation	Source	Translation	Source	Translation
German	26.07	29.34	5.52	5.64	9.95	9.58
French	33.86	34.46	4.65	4.68	9.43	9.12
Spanish	35.99	32.56	4.66	4.74	9.08	9.02
Dutch	25.43	31.13	4.88	5.08	9.30	8.99

Table 2: Observation of different features

able. Carrol (1964) proposed a variation of TTR in order to reduce the sample size effect. Another version of TTR is called Bilogarithmic TTR (Herdan, 1964). Kohler and Galle (1993) also defined a version TTR (see: 1) that consider position of the text. In the Equation 1 x refers to position in the text, $t_x =$ number of types up to position x, $T =$ number of types in the text and N refers to the number of tokes in the whole text. We also used another version of TTR that focuses on document level TTR $\frac{T}{N}$ as well as sentence level TTR $\frac{t}{n}$ (Islam and Mehler, 2013; Islam, 2014; Islam et al., 2014). Lower TTR in sentence level also shows the repetition of the text.

- Köhler–Gale method

$$TTR_x = \frac{t_x + T - \frac{x^T}{N}}{N} \qquad (1)$$

- Root TTR

$$\frac{T}{\sqrt{N}} \qquad (2)$$

- Corrected TTR

$$\frac{T}{\sqrt{2N}} \qquad (3)$$

- Bilogarithmic TTR

$$\frac{\log T}{\log N} \qquad (4)$$

- TTR deviation

$$\sum_{i=0}^{n} \left(\frac{T}{N} - \frac{t_i}{n_i} \right) \qquad (5)$$

7.2 Information-theoretic features

Information theory measures the statistical significance of how documents vary with different types of probability distributions. That is, it determines how much information can be encoded from a document using a certain type of probability distribution. The use of information as a statistical measure of significance is an extension of this process. Information theory allows us to use conditional probabilities. It should be noted that these features are being used for the first time on this kind of problem.

7.2.1 Entropy based features

The most efficient way to send information through a noisy channel is at a constant rate (Genzel and Charniak, 2002; Genzel and Charniak, 2003; **?**). This rule must be retained in any kind of communication to make it efficient. Any text as a medium of communication should satisfy this principle. Genzel and Charniak (2002; 2003) show that the entropy rate is constant in texts. That is, for example, each sentence of a text conveys roughly the same amount of information. In order to utilize this information-theoretic notion, we start from random variables and consider their entropy as indicators of readability.

Shannon (1948) introduced entropy as a measure of information. Entropy, the amount of information in a random variable, can be thought of as the average length of the message needed to have an outcome on that variable. The entropy of a random variable X is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \qquad (6)$$

The more the outcome of X converges towards a uniform distribution, the higher $H(X)$. Our hypothesis is that the higher the entropy, the less readable the text along the feature represented by X. Table 2 shows that translated texts have lower entropy than source texts. This is because translators try to improve the readability of translated texts. In our experiment, we consider the following random variables: *word probability*, *character probability*, *word length probability* and *word frequency probability* (or frequency spectrum, respectively). Note that there is a correlation between the probability distribution of words and the corresponding distribution of word frequencies. As we use SVM for classification, these correlations are taken into consideration.

7.3 Information Transmission-based Features

There is a relation among text difficulty, sentence length, and word length. The usefulness of similar lexical features such as *sentence length* or *number of difficult words in a sentence* is shown in section 7.1. Generally, a longer sentence contains more entities that influence the difficulty level. Similar things happen with longer words. But, a sentence becomes more difficult if it is longer and contains more long words. These kinds of properties can be defined by *joint* and *conditional* probabilities.

In the field of information theory, joint probability measures the likelihood of two events occurring together. That is, two random variables X and Y will be defined in the probability space. The conditional probability gives the probability that the event will occur given the knowledge that another event has already occurred. By considering the joint probability and two random variables X and Y, Shannon's joint entropy can be defined as:

$$H(X,Y) = -\sum_{<x,y>\in XxY} p(x_i,y_i)\log p(x_i,y_i) \tag{7}$$

Two conditional entropies can be defined as:

$$H(X|Y) = -\sum_{y\in Y} P(y_i)\sum_{x\in X} p(x_i|y_i)\log p(x_i|y_i) \tag{8}$$

$$H(Y|X) = -\sum_{x\in X} P(x_i)\sum_{y\in Y} p(y_i|x_i)\log p(y_i|x_i) \tag{9}$$

From the equation 6, 7, 8 and 9, it can be shown that:

$$T_s(X,Y) = H(X) + H(Y) - H(X,Y) \tag{10}$$

The function is called *Information transmission*, and it measures the strength of the relationship between elements of random variables X and Y. Details about this notion can be found in (Klir, 2005). The *sentence length and word length probability* shows the relation between sentence length and word length and *sentence length and difficult word probability* shows the relation between sentence length and the number of difficult words.

8 Experiment

The experiments and evaluations are explained in the following subsections.

8.1 Experiment with modern corpus

The training corpus contains $2,646,765$ parallel sentences from 412 language pairs of 21 European languages. We have divided the corpus into $26,467$ chunks. More specifically, $26,467$ chunks were *source* texts and the same number of chunks were *translations*. It should be noted that a hundred sets of data were randomly generated where 80% of the corpus is used for training and the remaining 20% is used for evaluation. Later, when we get reasonable classification *accuracy* and *F-Score*, the whole corpus will be used to build the final training model. The weighted average of *Accuracy* and *F-score* is computed by considering all sets of data. Note that we have used the SMO (Platt, 1998; Keerthi et al., 2001) classifier model in WEKA (Hall et al., 2009) together with the Pearson VII function-based universal kernel PUK (Üstün et al., 2006).

As we showed in Figure 1, our goal was to build a model using texts from modern European languages and later use that model to identify the source of the *Georgian* Gospel. The challenge was to find features that are language independent and improve the classification accuracy. A feature will be in the feature list if and only if the classification accuracy improves by adding the feature. Many different features were considered, but only useful features are listed in Table 3 and described in Section 7. Additionally, either measure *Accuracy* and *F-score* has to be above average. Individually all features perform reasonably well. However, *information-theoretic* features perform better than lexical features. Table 3 shows evaluation of selected features. Surprisingly *word frequency entropy* is the best performing individual feature. Altogether these features achieve 86.62% of *F-Score*.

8.2 Experiment with target corpus

In order to experiment with the target corpus, we prepare them similarly to the training chunks. Each Gospel was divided into 37 chunks. Each chunk contains 100 verses. Then, these data are divided into two sets. The first set contains chunks from *Armenian* and *Georgian*. The other contains chunks from *Greek* and *Georgian*.

As we stated earlier, the first task is to identify chunks of the *Georgian* Gospel are translations. Table 4 shows the confusion matrix of the first set. In this matrix 36 out of 37 chunks of

Feature	Accuracy	F-Score
ASL	54.01%	53.29%
TTR per document	59.83%	59.18%
TTR per sentence	58.93%	57.42%
Average complex word per document	52.61%	45.74%
Average complex word per sentence	52.52%	45.83%
AWL	56.15%	49.43%
Köhler–Gale TTR	59.58%	58.89%
Root TTR	62.67%	62.67%
Corrected TTR	62.61%	62.61%
Bi-logarithmic TTR	62.23%	62.08%
TTR deviation	60.54%	60.00%
Word entropy	62.02%	61.92%
Word frequency entropy	63.36%	63.39%
Word length entropy	53.81%	50.94%
Character entropy	57.78%	56.58%
Character frequency entropy	57.93%	57.28%
Information transmission of sentence length and word length probability	52.93%	50.26%
Information transmission of sentence length and complex word probability	54.41%	53.86%
All features	86.63%	86.62%

Table 3: Evaluation of lexical features in source and translation identification

	Source	Translation
Armenian	0	37
Georgian	1	36

Table 4: Confusion matrix of *Armenian–Georgian* Gospels

	Source	Translation
Greek	20	17
Georgian	1	36

Table 5: Confusion matrix of *Greek–Georgian* Gospels

the *Georgian* Gospel identified as translated text. So, experimental results show that the *Georgian* Gospel is a translated document. Table 5 shows the same result. All of the *Armenian* chunks are identified as translated documents. However, 20 out of 37 chunks of the *Greek* Gospel are identified as source. Therefore, these two confusion matrices show that *Greek* is most likely the source of the *Georgian* Gospel. It becomes clearer when we have a look on Table 6. Here *Armenian* and *Greek* chunks are labeled as *source* and *Georgian* chunks are labeled as translation. The *accuracy* and *F-Score* of the *Armenian–Georgian* pair is below the baseline 50%. But the *accuracy* and *F-Score* of the *Greek–Georgian* pair is above 75%. So, our experimental results suggest that the *Greek* Gospel is the source of the version of *Georgian* Gospel.

Source-translation	Accuracy	F-Score
Armenian–Georgian	48.64%	32.73%
Greek–Georgian	75.67%	74.48%

Table 6: Classification results of Gospels

9 Conclusion

It is important to identify a document as original or translated from another language. Such a tool is very useful for many NLP applications. Different linguistic features are being explored in recent days for many different NLP applications. However, only simple *lexical* and classical *information-theoretic* features are adequate to build a classifier which is able to identify an original or a translated document. It will be challenging to explore linguistic features for such applications that deal with multilingual data.

There are many versions of the *Georgian* Gospels that are translated from different languages. Linguists are able to narrow down potential origins by looking at different linguistic properties, but skeptical to decide the single origin. We have three such Gospels in *Georgian*, *Armenian* and *Greek*, where linguists believe that *Armenian* or *Greek* are the potential origin. For this paper, we have built a source and translation classifier using modern texts. The classifier is able to identify translated documents that have been translated hundreds of years ago. Based on our experimental evaluation, the *Greek* Gospel is the source of the version of the *Georgian* Gospel.

10 Acknowledgments

We would like to thank Prof. Dr. Alexander Mehler, Prof. Dr. David Scheffer and Armin Hoenen. We also like to thank the anonymous reviewers for their helpful comments. This work was performed when both authors were part of the LOEWE Digital-Humanities project at Goethe-University Frankfurt. Travel was funded by MassineScheffer & Company GmbH.

References

Mona Baker. 1993. Corpus linguistics and translation studies - implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology. In Honour of John Sinclair*, pages 233–354. John Benjamins.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pages 175–186. Amsterdam & Philadelphia: John Benjamins.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of translationese: Machinelearning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Robert Pierpont Blake. 1932. *Khanmeti palimpsest fragments of the Old Georgian version of Jeremiah.* Cambridge Univ Press.

Shoshana Blum and Eddie A Levenston. 1978. Universals of lexical simplification. *Language learning*, 28(2):399–415.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, pages 17–35.

John Bissell Carroll. 1964. *Language and thought.* Prentice-Hall Englewood Cliffs, NJ.

Dimitry Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40st Meeting of the Association for Computational Linguistics (ACL 2002).*

Dimitry Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10–18.

Silvia Hansen. 2003. *The Nature of Translated Text: An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations.* Ph.D. thesis, University of Saarland.

Gustav Herdan. 1964. *Quantitative linguistics.* Butterworths.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2009. Towards simplification: A supervised learning approach. In *Proceedings of Machine Translation 25 Years On, London, United Kingdom, November 21-22.*

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov, 2010. *Identification of translationese: A machine learning approach*, pages 503–511. Springer.

Zahurul Islam and Armin Hoenen. 2013. Source and translation classification using most frequent words. In *6th International Joint Conference on Natural Language Processing (IJCNLP).*

Zahurul Islam and Alexander Mehler. 2012. Customization of the europarl corpus for translation studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC).*

Zahurul Islam and Alexander Mehler. 2013. Automatic readability classification of crowd-sourced data based on linguistic and information-theoretic features. In *14th International Conference on Intelligent Text Processing and Computational Linguistics.*

Md. Zahurul Islam, Md. Rashedur Rahman, and Alexander Mehler. 2014. Readability classification of bangla texts. In *15th International Conference on Intelligent Text Processing and Computational Linguistics (cicLing), Kathmandu, Nepal.*

Zahurul Islam. 2014. Multilingual text classification using information theoretic features. PhD Thesis, Goethe University Frankfurt.

S.S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3):637–649.

Anna Kharanauli, 2000. *Einfhrung in die georgische Psalterbersetzung*, pages 248–308. Vandenhoeck & Ruprecht.

George Jiri Klir. 2005. *Uncertainty and Information.* Wiley-Interscience.

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

Reinhard Köhler and Matthias Galle. 1993. Dynamic aspects of text characteristics. *Quantitative text analysis*, pages 46–53.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *49th Annual Meeting of the Association for Computational Linguistics (ACL).*

David Marshall Lang. 1957. Recent work on the georgian new testament. *Bulletin of the School of Oriental and African Studies*, 19(01):82–93.

Sara Laviosa. 2002. *Corpus-based translation studies. Theory, findings, applications.* Amsterdam/New York: Rodopi.

André Lefevere. 2002. *Translation/history/culture: A sourcebook.* Routledge.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.

Sattar Lzwaini. 2003. Building specialised corpora for translation studies. In *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives, Corpus Linguistics*.

A. Nádas. 1984. Estimation of probabilities in the language model of the ibm speech recognition system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(4):859–861.

Maeve Olohan. 2001. Spelling out the optionals in translation:a corpus study. In *Corpus Linguistics 2001 conference. UCREL Technical Paper number 13. Special issue.*

Gloria Corpas Pastor, Ruslan Mitkov, Naveed Afzal, and Viktor Pekar. 2008. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08).*

Jams W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual.* Erlbaum Publishers.

John C. Platt. 1998. *Fast training of support vector machines using sequential minimal optimization.* MIT Press.

Marius Popescu. 2011. Studying translationese at the character level. In *Recent Advances in Natural Language Processing*.

Anthony Pym. 2005. Explaining explicitation. In *New Trends in Translation Studies. In Honour of Kinga Klaudy*, pages 29–34. Akadmia Kiad.

Candace Séguinot. 1988. Pragmatics and the explicitation hypothesis. *TTR: traduction, terminologie, rédaction*, 1(2):106–113.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423.

Irina Temnikova. 2012. *Text Complexity and Text Simplification in the Crisis Management Domain.* Ph.D. thesis, University of Wolverhampton.

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

B. Üstün, W.J. Melssen, and L.M.C. Buydens. 2006. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81(1):29–40.

Hans Van Halteren. 2008. Source language markers in europarl translations. In *International Conference inComputational Linguistics(COLING)*, pages 937–944.

Jean-Paul Vinay and Jean Darbelnet. 1958. Stylistique comparée de langlais et du français.

Jean-Paul Vinay and Jean Louis Darbelnet. 1995. *Comparative stylistics of French and English: a methodology for translation*, volume 11. John Benjamins.