Shannon's Entropy in Literary Works and Their Translations

Marcin Lawnik
Faculty of Applied Mathematics
Silesian University of Technology
Gliwice, Poland
Marcin.Lawnik@polsl.pl

Abstract— **The aim of the following article is to compare the Shannon's entropy of literary text to its translation into another language. A calculated entropy may provide information about quality of translation. As a space of events was chosen a set of all the various words forming a text. Based on the frequency of their (words) occurrence in a text, the entropy was calculated.**

Keywords: information theory, entropy, literary works, literary translation

I. INTRODUCTION

In literature, one may often find works, which are translated into other languages, e.g. works of William Shakespeare or Johann Wolfgang von Goethe. While analyzing the translation, we might ask ourselves how good is it, in literary sense, compared to the original work. The quality of translation obviously depends on translator's writing skills, knowledge of translated language and lexical resources. From a technical point of view, the answer for this question can be provided by calculating Shannon's entropy of the original, as well as the translation into a foreign language

Shannon's entropy defined in [1], is a measure of average information belonging to single message from given source. It is possible to use this concept to examine the amount of information provided by language [2,3]. However, most of approaches presented in literature calculate entropy on the basis of frequency of individual signs' (letters of alphabet) occurrence in the text. When dealing with language, it seems more important to calculate information establishing as a measure a word, not a letter. The next method, marked [4], allow us to recognize the type of a text, having to choose from variety of texts' topics, e.g. sports, political or economical. It assumes converting the text into corresponding binary code, every letter is given the value of 1, and punctuation marks are given the value of 0. Subsequently, using the definition of Shannon's entropy and its modification, we carry out the analysis of formed binary code.

The method proposed in this article presuppose calculating Shannon's entropy on the basis of the frequency of individual words' occurrence in a given literary text, both in original and translated version. With use of that method, texts by Shakespeare were translated into German and works of Goethe were translated into English. All of the analyzed texts originate from the website of Project Gutenberg [5].

II. METHODOLOGY

All of the texts were edited in order to count words properly – signs were changed into their small equivalents, omitting punctuation marks and others, which are not letters. Footnotes were deleted, as they explain e.g. historical context, and are just an addition added during translation process. Below we may see an example of such editing:

Original text: "A horse, a horse, my kingdom for a horse!"

Edited text: a horse a horse my kingdom for a horse

After preparations presented above, the frequency of occurrence is determined for particular words. Then, on that base, the Shannon's entropy is calculated H(X) given a formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i), \tag{1}$$

where X is a space of possible events (set of various words forming a text), p(x) is the probability of this word's occurrence and *n* is number of different words in a text. H(X) is a positive number and will reach maximum value if every word occurs only once in a text.

In order to compare the amount of information belonging to one event, the normalized entropy NH(X):

$$NH(X) = \frac{H(X)}{L} \tag{2}$$

and the value of maximal entropy MH(X):

$$MH(X) = \log_2 L \tag{3}$$

are computed, where L is a number of words in a text.

III. RESULTS

Table I. and Table II. contains values of entropy H(X) and numbers of words L for texts by Shakespeare and Goethe. Ordinal numbers in those tables (No.) are used as the identifier in a form of horizontal axis in figures presented below. The figures Fig. 1 and Fig. 3 presents correlation between the

work's entropy and maximum entropy for a message of length equal to the number of words in the work. In the figures Fig. 2 and Fig. 4 values NH(X) are presented corresponding to texts written by both authors.

TABLE I. WORKS OF SHAKESPEAR AND THEIR TRANSLATIONS TO GERMAN.

No.	Title	H(X)	L	Author/Translator
1.	Coriolanus	9,276804	29293	W. Shakespeare
2.		9,766302	27526	D. Tieck
3.	Hamlet	9,374843	29704	W. Shakespeare
4.		9,771238	33246	C.M. Wieland
5.	Julius Caesar	9,008386	20822	W. Shakespeare
6.		9,606786	19662	A.W. von Schlegel
7.	King John	9,309804	21703	W. Shakespeare
8.		9,823828	22757	C.M. Wieland
9.	King Lear	9,419567	27526	W. Shakespeare
10.		9,814374	28703	C.M. Wieland
11.	Macbeth	9,35841	18250	W. Shakespeare
12.		9,718761	19043	C.M. Wieland
13.		9,827594	17321	D. Tieck
14.	Othelo	9,116973	27861	W. Shakespeare
15.		9,537414	31520	C.M. Wieland
16.	King Richard II	9,354526	23389	W. Shakespeare
17.		9,810962	23551	C.M. Wieland
18.	King Richard III	9,301778	31310	W. Shakespeare
19.		9,984761	28198	A.W. von Schlegel
20.	Romeo and Juliet	9,353637	25788	W. Shakespeare
21.		9,799269	20877	A.W. von Schlegel
22.		9,812279	26045	C.M. Wieland
23.	Timon of Athen	9,320082	19677	W. Shakespeare
24.		9,692627	21763	C.M. Wieland

TABLE II. WORKS OF GOETHE AND THEIR TRANSLATIONS TO ENGLISH.

No.	Title	H(X)	L	Author/Translator
1.	Egmont	9,777285674	25730	J. H. Goethe
2.		9,358965823	28786	A. Swanwick
3.	Hermann and Dorothea	9,535400771	19830	J. H. Goethe
4.		9,22528389	23598	E. Frothingham
5.	Iphigenia in Tauris	9,662992858	15036	J. H. Goethe
6.		9,41694502	15896	A. Swanwick
7.	The Sorrow of Young Werther	9,790408906	39256	J. H. Goethe
8.		9,280603549	42545	R.D. Boylan
9.	Wilhelm Meister's Apprenticeship and Travels- Book 1	9,837116955	21252	J. H. Goethe
10.		9,27019959	24079	T. Carlyle
11.	Wilhelm Meister's Apprenticeship and Travels- Book 2	9,806078662	21517	J. H. Goethe
12.		9,298190784	24234	T. Carlyle
13.	Wilhelm Meister's Apprenticeship and Travels- Book 3	9,656982998	18212	J. H. Goethe
14.		9,15576823	20590	T. Carlyle
15.	Wilhelm Meister's Apprenticeship and Travels- Book 4	9,85894841	25374	J. H. Goethe
16.		9,312596703	28490	T. Carlyle
17.	Wilhelm Meister's Apprenticeship and Travels- Book 5	9,774658977	23842	J. H. Goethe
18.		9,264506667	26526	T. Carlyle
19.	Wilhelm Meister's Apprenticeship and Travels- Book 6	9,53420899	21095	J. H. Goethe
20.		9,015041488	23019	T. Carlyle
21.	Wilhelm Meister's Apprenticeship and Travels- Book 7	9,570989001	25675	J. H. Goethe
22.		9,135450234	27267	T. Carlyle

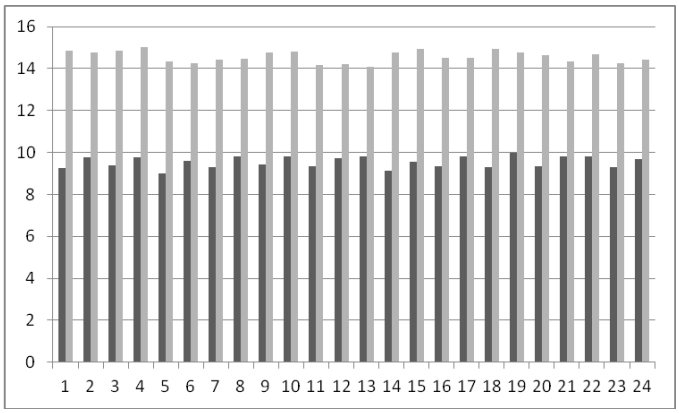

Figure 1. Values of entropy H(X) and their relevant maximum entropy MH(X) of Shakespeare's works.

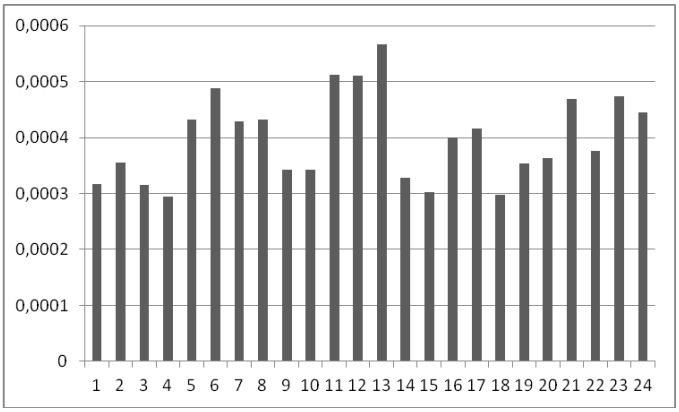

Figure 2. Normalized entropy NH(X) of Shakespeare's literary works and their translations.

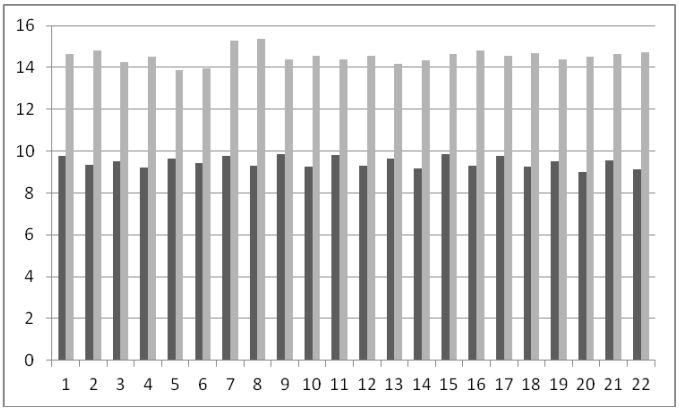

Figure 3. Values of entropy H(X) and their relevant maximum entropy MH(X) of Goethe's works.

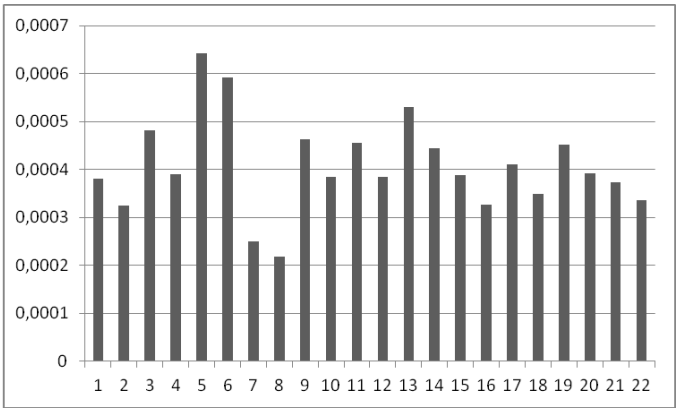

Figure 4. Normalized entropy NH(X) of Goethe's literary works and their translations.

IV. CONCLUSIONS

The results obtained using given method indicates that German has higher entropy than English. Regardless of linguistic correlation , certain translations has entropy higher (lower) than the others. It is perfectly pictured by normalized entropy NH(X), where value for some texts is approximately the same, and for some it significantly differs. Therefore, presented method may be used to measure the quality of translation. What is more, comparing several translations of a literary work, it is possible to determine the one with entropy closest to the original text. When the distinction between entropies are too large, it may signify that a translator were not able to achieve his goal. There might be several reasons for this phenomenon, e.g. neologisms used by the author that need to be properly introduced into the target language.

REFERENCES

[1] Shannon Claude E.: A Mathematical Theory of Communications. Bell System Technical Journal, vol. 27, pp. 379-423, 1948.

[2] Shannon Claude E.:, Prediction and entropy of printed English. Bell Syst. Techn. J., vol. 30, pp. 50–64, 1951.

[3] Moradi H., Grzymala-Busse J.W., Roberts J.A.: Entropy of English text: Experiments with humans and a machine learning system based on rough sets. Information Sciences, vol. 104 , pp. 31-47, 1998.

[4] Papadimitriou C., Karamanos K., Diakonos F.K., Constantoudis V., Papageorgiou H.: Entropy analysis of natural language written texts. Physica A, Volume 389, Issue 16, pp. 3260-3266, 2010.

[5] Project Gutenberg: http://www.gutenberg.org/