Language Complexity in Human and Machine Translation: A Preliminary Study

Gábor Recski¹ and Fanni Kádár²

¹ TU Wien, Vienna, Austria
gabor.recski@tuwien.ac.at
² Deloitte Hungary
fkadar@deloittece.com

Abstract. Systematic comparison between machine translation (MT) and human translation (HT) is mostly limited to the evaluation of MT output with HT as reference, as opposed to a more general study of the properties of MT and HT output texts. We present preliminary experiments investigating differences between MT and HT in terms of readability and language complexity. We perform both quantitative and qualitative comparison of the outputs of machine and human translation, using samples of English text across multiple domains and genres and their Hungarian translations created by humans and by the state-of-the-art machine translation system deep1. Our results show that machine translation produces somewhat simpler text than human translation on 3 out of 4 samples, and on 2 samples this effect is caused primarily by human translators using a higher number of complex words. We release all software used in our experiments to facilitate further studies on larger samples, additional languages and domains, and using alternative MT systems.

Keywords: Machine Translation, Language Complexity, Readability

1 Introduction

Recent advances in machine translation (MT) have fueled the debate about the future role of human translation (HT), yet systematic comparison between MT and HT is still mostly limited to the evaluation of MT output with HT as reference, as opposed to a more general study of the properties of MT and HT output texts. With this preliminary work we propose to investigate differences between MT and HT in terms of the closely related concepts readability and language complexity. We perform both quantitative and qualitative comparison of the outputs of machine and human translation, using samples of English text across multiple domains and genres, each of which have been translated into Hungarian both by a human and by the state-of-the-art MT system deep1. Our results suggest that machine translation produces somewhat simpler text than human translation on 3 out of 4 samples, and that this effect is often caused by human translators using a higher number of complex words.
The remainder of this paper is structured as follows. Section 2 summarizes related work, Section 3 describes our data sources and documents the process of creating the samples. Section 4 describes the metrics used to analyze each sample, Section 5 presents and discusses the results, and Section 6 concludes the paper. All data and software described in this paper is publicly available under an MIT license from GitHub\textsuperscript{3}.

2 Related Work

Most recent work on the evaluation of machine translation (MT), whether it is based on comparison with human reference translations or on direct assessment by human evaluators, is focused on ranking MT systems from best to worst, as dictated by the leaderboard culture dominating the field of natural language processing (NLP) \cite{14,6}. In this context, comparison of MT output to human translation (HT) is usually performed only quantitatively and with the sole purpose of measuring the degree of similarity between MT and HT output. Exceptions include the work of Ahrenberg \cite{1}, who performs a close qualitative comparison of MT and HT on a single newspaper article translated from English into Swedish, focusing on information structure, word order, post-editing required for the MT output, and examples of Translation Procedures \cite{16}. The simple quantitative analysis found that the MT output more closely preserved the number of sentences and tokens of the English original than the human translator, who would often split longer sentences into multiple shorter ones. Another direct comparison of MT and HT is a detailed qualitative comparison of French subtitles of a television show translated from English to French both by deep1 and by professional translators \cite{2}, the quantitative component of this analysis focused on the categorization of errors in the MT output.

Another recent analysis compares HT with the output of post-editing (PE) of MT output \cite{23}, this work involves statistical analysis of samples across 5 languages and multiple domains, focusing on part-of-speech (POS) sequences, text length, lexical diversity, and lexical density, the latter defined as the ratio of content words. These experiments demonstrate that PE output exhibits lower lexical variety and lexical density than MT, that average sentence length in PE output is closer to that in the source text than to that in the HT output, and that POS sequences in PE output are more similar to those typical of the source language. Another quantitative comparison of MT and HT with respect to lexical diversity was performed in \cite{26}, using the Europarl corpus \cite{15} and showing for two language pairs and a wide variety of MT systems that MT produces text with lower lexical richness than HT. \cite{23} extends this study to include measures of morphological variation, focusing on inflectional paradigms of lemmas and showing that MT reduces their diversity compared to HT across all language pairs.

The evaluation of text readability has a long tradition in both applied linguistics and NLP \cite{13,11}. Commonly used quantitative measures of readability such as

\begin{verbatim}
https://github.com/recski/comp-trans
\end{verbatim}
as the Flesch-Kincaid test [5] and Gunning’s Fog Index [7] rely on simple
statistics about the average length of words and sentences, such common measures differ only in the exact methods for quantifying these two properties and in the parameters for combining them to obtain a single dimension of readability. Studies on language complexity also use measures such as type-token ratio and word entropy [27], but their relevance to readability remains unclear.

3 Data

We extracted 4 samples of English text from various domains and genres, each taken from publicly available sources. The TED sample is the text of three short talks given at various TED conferences, transcripts of which are available online [4] with translations in multiple languages, including Hungarian. For creating the sample we used a preprocessed dataset [5] that has been previously used for improving neural machine translation systems [21]. Our sample contains the complete transcript of two presentations. The first was given by Barat Ali Ba-
toor, the Hungarian translation was created by Zsuzsanna Dan and reviewed by Orsi Mance. The second talk is by Kees Moliker, the Hungarian translator of the transcript was Mária Ruzsáné Cseresnýés, the reviewer of the translation was Anna Patai.

Our second sample, FGM, is an excerpt from the script of the movie A few good men, both English and Hungarian texts were extracted from SRT-formatted subtitles downloaded from OpenSubtitles [6]. The English text is an exact transcript of the movie dialogue, the Hungarian subtitles were extracted from the DVD edition, the translator could not be identified. The data was converted to raw text using the srt library [7]. The 1984 sample is an excerpt from the novel 1984 [8] by George Orwell. The original text was downloaded from Project Gutenberg [8], the Hungarian version (translated by László Szögyértő) was extracted from the Szeged Treebank [3]. DC567 is a single document from the multilingual JRC-ACQUIS corpus [22], a 2006 communication of the European Commission titled A Contribution to the EU’s Growth and Jobs Strategy.

Basic statistics describing the original English and the two Hungarian translations of each of the four samples is presented in Table 1. The four samples were selected to be comparable in length, as measured by the number of words in the English original text. Differences in average sentence length (number of words per sentence) and relative vocabulary size (number of unique lemmas compared to the number of words, also known as the type-token ratio, as measured by Herdan’s C, see Section 4 for details) are characteristic of each genre. When comparing the English originals, the novel 1984 contains the longest sentences and has the largest vocabulary. The DC567 text, a policy statement on matters

---

4 https://www.ted.com/participate/translate
5 https://github.com/neulab/word-embeddings-for-nmt
6 https://www.opensubtitles.org/
7 https://pypi.org/project/srt/
8 http://gutenberg.net.au/ebooks01/0100021.txt
of the economy, is a close second in average sentence length but has a smaller vocabulary, comparable to the presentations in TED3. The movie dialogues in FGM have the shortest sentences and the smallest vocabulary.

Table 1. Basic statistics of the four samples. sens and words are the total numbers of sentences and words in each sample, respectively. w/s is the average number of words per sentence (including punctuation), |V| is the vocabulary size measured as the number of unique lemmas, C_l is Herdan’s C on the set of lemmas (see Section 4 for details).

|          | sens | words | w/s | |V| | C_l |
|----------|------|-------|-----|---|---|-----|
| 1984     |      |       |     |   |   |     |
| EN       | 258  | 6181  | 23.96| 1428 | .83 |
| HU (human) | 251  | 5040  | 20.08| 1641 | .87 |
| HU (deepl)| 257  | 5077  | 19.75| 1630 | .87 |
| TED3     |      |       |     |   |   |     |
| EN       | 320  | 4940  | 15.52| 990  | .81 |
| HU (human) | 298  | 4086  | 13.71| 1223 | .85 |
| HU (deepl)| 303  | 4247  | 14.02| 1178 | .85 |
| FGM      |      |       |     |   |   |     |
| EN       | 753  | 7241  | 9.62 | 1106 | .79 |
| HU (human) | 562  | 3680  | 6.55 | 1095 | .85 |
| HU (deepl)| 1056 | 6244  | 5.91 | 1444 | .83 |
| DC567    |      |       |     |   |   |     |
| EN       | 277  | 6175  | 22.29| 1059 | .80 |
| HU (human) | 260  | 5845  | 22.48| 1460 | .84 |
| HU (deepl)| 262  | 5798  | 22.13| 1338 | .83 |

In addition to the existing human translations we translated all four samples from English into Hungarian using deepL. Each sample was passed to deepL in a single API call, with the entire text as a single text parameter, using default values for all other parameters. The API was accessed using the deepL-python client[9]. Each MT output as well as the human translations and the English originals were tokenized and lemmatized using the default models of stanza[11], accessed via a wrapper[12] provided by the tuw-nlp library.

4 Methods

We compute for each of our samples simple statistical measures of language complexity. A common measure of lexical diversity is the type-to-token ratio (TTR) of a document, measuring the ratio between the number of distinct lemmas (types) and the total number of words in the text (tokens). A common version

---

9 [https://www.deepl.com/docs-api/translate-text/markup/](https://www.deepl.com/docs-api/translate-text/markup/)
11 [https://stanfordnlp.github.io/stanza/](https://stanfordnlp.github.io/stanza/)
of this measure is Herdan’s \( C \) \cite{herdan1926quantitative}, also used to measure language complexity by \cite{herdan1964language}, calculated as
\[
\log \frac{\text{#types}}{\text{#tokens}}
\]
For each document we calculate Herdan’s \( C \) both on the set of words and on the set of lemmas. Since the type-token ratio is highly sensitive to text length, we also compute the Measure of Textual Lexical Diversity (MTLD) \cite{yamagata2011measure} for each sample, which measures the mean length of word sequences that maintain a given TTR value (0.72 by default). We calculate MTLD on both tokens and lemmas, using the implementation available from the \texttt{lexical-diversity} python package\cite{lexical_diversity}

Finally, we also calculate Gunning’s Fog Index \cite{gunning1955reading}, a common measure of text readability based on average sentence length and the ratio of complex words, the latter defined as words containing 3 or more syllables. The formula for calculating the Fog Index is the following
\[
0.4 \times \left( \frac{\text{#words}}{\text{#sentences}} + 100 \times \frac{\text{#complex words}}{\text{#total words}} \right)
\]
While this metric is widely used for measuring the readability of English text, its applicability to Hungarian is yet to be investigated (see \cite{balog2019evaluation} for a first discussion). For this study we used two large text corpora of English and Hungarian to measure the distribution of word length (as measured by the number of syllables) and approximate a more realistic language-specific parameter of the FOG index for Hungarian. For English, we used word counts extracted from the UMBC-Webbase\cite{umbc_webbase} corpus \cite{umbc_webbase}. Syllables of each word were counted by lemmatizing them using \texttt{stanza} (see Section 3) and retrieving the number of syllables for each word from the CMU pronunciation dictionary\cite{cmu_dict} provided by the Python library \texttt{pronouncing}\cite{pronouncing}. Since the dictionary’s coverage is limited, but the full vocabulary of the corpus contains 6.7 million word types, we only lemmatized words with at least 100 occurrences in the UMBC corpus (175 000 types) and then proceeded to further filter out those lemmas that are not present in the CMU dictionary. The remaining set of types still covers 2.89 billion of the 3.34 billion tokens in the UMBC corpus. For Hungarian we used a version of the Hungarian Webcorpus\cite{hlt_bme_hu_webcorpus} that has been lemmatized using \texttt{emtsv}\cite{emtsv} \cite{nytud}. Since the number of syllables in a Hungarian word can be determined by simply counting the number of vowel characters, we could efficiently process all 8.8 billion tokens in the corpus.

The distribution of the syllable counts in the two corpora is presented in Figure 1, the raw data is available from the project repository. We observe that by considering all English words with 3 or more syllables as complex the Gunning

13 \url{https://pypi.org/project/lexical-diversity/}
14 \url{https://ebiquity.umbc.edu/resource/html/id/351/UMBC-webbase-corpus}
15 \url{http://www.speech.cs.cmu.edu/cgi-bin/cmudict}
16 \url{https://pypi.org/project/pronouncing/}
17 \url{https://hlt.bme.hu/en/resources/webcorpus2}
18 \url{https://github.com/nytud/emtsv}
The Gunning Fog Index treats about 17% of all tokens in a large web corpus as complex, while for Hungarian a similar ratio (about 15%) can be achieved by considering words with 4 or more syllables as complex. Therefore in the present study we shall calculate the Gunning Fog Index for Hungarian texts using this adjusted parameter, noting that further quantitative and qualitative studies would be required to determine whether this metric can be considered a reliable measure of the readability of Hungarian text.

![Fig. 1. Distribution of the number of syllables in words of the UMBC-Webbase corpus (left) and the Hungarian Webcorpus (right)](image)

### 5 Results

The complexity measures described in Section 4 are shown in Table 2. We observe that while Herdan’s C values show little or no difference between human and machine translations of the same text, the MTDL measure is considerably lower for MT outputs across all samples, suggesting that MT produces lexically less diverse text. Additionally, the Gunning Fog Index is lower for texts produced by deep1 than for those produced by human translators on 3 of the 4 samples, suggesting that machine translation generates somewhat simpler text across a variety of genres and domains. The two factors and how they contribute to Gunning’s F for each sample are visualized in Figure 2, showing that while on the 1984 and DC567 samples the difference is mainly due to the MT output using fewer complex words, on FGM it is primarily caused by shorter average sentences.

We also inspected each sample manually to understand possible sources of quantitative differences between each pair of human and machine translation. First, we observe that in the FGM sample of movie subtitles the human translator considerably altered and shortened the contents of the original text, most likely to reduce the number of characters that viewers have to read over a given period.
Table 2. Complexity measures for the four samples. \( w/s \) is the average number of words per sentence, \( L_c/L \) is the ratio of complex lemmas, \( C \) and \( M \) are Herdan’s \( C \) and MTLD, each measured on both the set of all words (\( C_w, M_w \)) and all lemmas (\( C_l, M_l \)). \( F \) is Gunning’s Fog Index. For details of each of these measures see Section 4.

<table>
<thead>
<tr>
<th></th>
<th>( w/s )</th>
<th>( L_c/L )</th>
<th>( C_w )</th>
<th>( C_l )</th>
<th>( M_w )</th>
<th>( M_l )</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>EN 23.96</td>
<td>6.76%</td>
<td>.85</td>
<td>.83</td>
<td>77.7</td>
<td>57.9</td>
<td>12.29</td>
</tr>
<tr>
<td></td>
<td>HT 20.08</td>
<td>6.77%</td>
<td>.90</td>
<td>.87</td>
<td>105.6</td>
<td>72.6</td>
<td>10.74</td>
</tr>
<tr>
<td></td>
<td>MT 19.75</td>
<td>6.36%</td>
<td>.89</td>
<td>.87</td>
<td>82.1</td>
<td>58.6</td>
<td>10.45</td>
</tr>
<tr>
<td>TED3</td>
<td>EN 15.52</td>
<td>5.98%</td>
<td>.83</td>
<td>.81</td>
<td>57.0</td>
<td>39.5</td>
<td>8.60</td>
</tr>
<tr>
<td></td>
<td>HT 13.71</td>
<td>5.58%</td>
<td>.89</td>
<td>.85</td>
<td>67.1</td>
<td>46.5</td>
<td>7.72</td>
</tr>
<tr>
<td></td>
<td>MT 14.02</td>
<td>6.19%</td>
<td>.89</td>
<td>.85</td>
<td>55.2</td>
<td>38.2</td>
<td>8.08</td>
</tr>
<tr>
<td>FGM</td>
<td>EN 9.62</td>
<td>4.17%</td>
<td>.82</td>
<td>.79</td>
<td>62.6</td>
<td>43.8</td>
<td>5.51</td>
</tr>
<tr>
<td></td>
<td>HT 6.55</td>
<td>4.24%</td>
<td>.89</td>
<td>.85</td>
<td>54.4</td>
<td>40.0</td>
<td>4.31</td>
</tr>
<tr>
<td></td>
<td>MT 5.91</td>
<td>3.91%</td>
<td>.87</td>
<td>.83</td>
<td>38.9</td>
<td>30.9</td>
<td>3.93</td>
</tr>
<tr>
<td>DC567</td>
<td>EN 22.29</td>
<td>18.38%</td>
<td>.83</td>
<td>.80</td>
<td>89.0</td>
<td>68.5</td>
<td>16.27</td>
</tr>
<tr>
<td></td>
<td>HT 22.48</td>
<td>19.26%</td>
<td>.89</td>
<td>.84</td>
<td>109.7</td>
<td>71.4</td>
<td>16.70</td>
</tr>
<tr>
<td></td>
<td>MT 22.13</td>
<td>18.32%</td>
<td>.88</td>
<td>.83</td>
<td>93.9</td>
<td>59.2</td>
<td>16.18</td>
</tr>
</tbody>
</table>

Fig. 2. Contributing factors of Gunning’s Fog Index. On each sample, \( E, H, \) and \( M \) show values for the English original, the human translation, and the machine translation, respectively.
of time, a standard practice in the production of movie subtitles. This practice is illustrated by the excerpt in Figure 3. As a result, the Hungarian subtitles contain only about half as many sentences and words as the machine translation of the English subtitles (see Table [1] in Section 3), calling into question whether the two texts should be considered comparable.

Next we observed that the differences in average sentence length in 1984 are primarily due to the fact that the human translator is more likely to merge multiple English sentences into a single Hungarian sentence. In the 1984 sample 258 English sentences are translated into 257 and 251 sentences by deep1 and the human translator, respectively, a difference of 2.3%, while the difference in the number of tokens (5077 and 5040) is only 0.7%. A closer inspection of the samples revealed that most of this difference was caused by a single decision of the human translator to translate the five sentences in Figure 4 as a single sentence. Finally we looked at examples of MT using fewer complex words than the human translator. Recall that we consider a word complex if it has 4 or more syllables, based on the calculation presented in Section 4. Figure 5 shows an example sentence from the DC567 sample in English and in the two Hungarian translations, illustrating the human translators’ preference for longer words.

6 Conclusion

We presented a preliminary study comparing human and machine translations of the same texts from the point of view of language complexity and readability. Samples of English text from four domains were extracted with existing Hungarian translations written by humans, the machine translation system deep1 was used to create machine translations. HT and MT output was compared quantitatively using multiple measures of lexical diversity and Gunning’s Fog Index of readability. For Hungarian we have estimated a language-specific parameter for calculating Gunning’s F on the basis of the distribution of word length in large corpora of English and Hungarian. Qualitative analysis of the samples was also performed to uncover reasons for the observed quantitative differences. We find that on 3 out of 4 samples human translations exhibit slightly higher complexity levels than machine translations. On the two samples containing fiction and legal text this effect is the result of the human translator choosing more complex words, while in the case of the sample containing movie subtitles it appears to a byproduct of the human translator shortening the text to about half its length, most likely in compliance with external requirements, a finding that questions the suitability of movie subtitles for this type of study.

In light of these preliminary results we believe that further study is required to determine whether MT and HT output exhibit systematic differences in readability and whether such effects are specific to certain domains, genres, or languages. When interpreting measures of readability, their applicability to a given language should also be investigated. We release all code used in the paper as
Original subtitles | Human translation
---|---
Your father’s Lionel Kaffee... former Navy judge advocate and attorney general of the United States. Died 1985. | Apja Lionel Kaffee, korábbi igazságügyminiszter. 'Your father was Lionel Kaffee, former Attorney General'
You went to Harvard law. Then you joined the Navy... probably because that’s what your father wanted you to do. | A Harvardra járt, és az apja kívánságára jött ide. 'You went to Harvard, and came here at the request of your father'
And now you’re just treading water for three years in the JAG Corps. Just laying low till you can get out and get a real job. | Csak tölti itt az idejét, míg be nem jön a jól fizető állás. 'You are just killing time until a well-paying job comes along'
If that’s the situation, that’s fine. I won’t tell anyone. | Am legyen! Nem árulom el senkinek. 'So be it! I won’t tell anyone'
But it’s my feeling that if this case is handled in the same... fast-food, slick-ass, Persian bazaar manner... with which you seem to handle everything else... | De ha a szokásos bazar mexikói módon kezeli ezt az ügyet... 'But if you’re going to handle this case in your usual bazaar manner..'
then something’s gonna get missed. | ...akkor épp a lényeg sikkad el. '... you will miss the point'
And I wouldn’t be doing my job if I allowed Dawson and Downey... to spend any more time in prison than absolutely necessary... | De nem hagyonom, hogy a bíróinkönkem azokat őket... 'But I cannot let them rot in prison..'
because their attorney had predetermined the path of least resistance. | ...mert a védelmük a legkisebb ellenállást kezdtek. '... just because their attorney prefers the least resistance'

Fig. 3. Excerpt from the FGM sample illustrating that the human translator considerably shortened the subtitles, most likely to comply with requirements on the maximum amount of text that can be displayed over a certain period of time. In this sample, 185 tokens in the original text were reduced to just 91 tokens in the translation.
They were the homes of the four Ministries between which the entire apparatus of
government was divided. The Ministry of Truth, which concerned itself with news, en-
tertainment, education, and the fine arts. The Ministry of Peace, which concerned itself
with war. The Ministry of Love, which maintained law and order. And the Ministry of
Plenty, which was responsible for economic affairs.

Fig. 4. A series of five sentences from the 1984 sample that was mapped to a single
Hungarian sentence by the human translator

open-source software\textsuperscript{19} to facilitate similar studies on additional languages and
domains, on larger samples, or using alternative MT systems.

Acknowledgements We thank the two anonymous reviewers for their many
useful suggestions, for pointing out additional references, and for recommending
MTDL as an additional measure of lexical diversity.

\textsuperscript{19} https://github.com/recski/comp-trans
The Doha Development Agenda remains our first priority.

and the Commission is working intensively to restart the Doha negotiation after its suspension in July 2006.

Fig. 5. Sample sentence from the DC567 data, illustrating the human translator’s preference for longer words. Bold indicates Hungarian words with 4 or more syllables in their lemma, considered as complex when calculating Gunning’s F (see Section 4)
References


