Measuring the Impact of Neural Machine Translation on Easy-to-Read Texts: An Exploratory Study

KAPLAN, Abigail, RODRIGUEZ VAZQUEZ, Silvia, BOUILLON, Pierrette

Abstract

Over the last decade, translation research has experienced an increased interest in the study of controlled languages (CLs). The vast majority of studies carried out during that time have focused on technical documentation, with a strong emphasis on the implications of combining CL approaches with machine translation (MT) for the quality of the final translation product. However, few research efforts have been devoted to exploring the impact of MT on texts that follow the guidelines of other forms of controlled language, such as plain language (PL) or Easy to Read (EtR). The most recent example can be found in the work by Rossetti (2018), who measured the machine translatability of PL summaries of health-related complex systematic reviews through the estimation of number and severity of errors in the MT output. Yet, the focus was still on target text fluency and adequacy rather than usability of the resulting PL text. Similarly, target users have rarely included people with special needs, as most studies of this nature are aimed at non-native speakers of a given language. By shifting the focal point towards the end user, [...]

Available at:
http://archive-ouverte.unige.ch/unige:123648

Disclaimer: layout of this document may differ from the published version.
Measuring the Impact of Neural Machine Translation on Easy-to-Read Texts

An Exploratory Study

Abigail Kaplan, Silvia Rodríguez Vázquez, Pierrette Bouillon

Klaara 2019, Helsinki – September 2019
Research Context (I)

- National Project “Proposal and Implementation of a Swiss Centre for Barrier-free Communication” (P-16)
  - 2017-2020, approved by swissuniversities, funded by federal contributions, ZHAW and UNIGE.

- **Goal:** Develop practical guidance and technological resources to facilitate BFC among three stakeholders in higher education institutions in Switzerland:
  - Students (with and without disabilities)
  - Teaching and research staff members
  - Administrative staff members

- Specific focus on inclusive tertiary education for students with visual, hearing and cognitive impairments
Research Context (II)

10 Research Areas

- RA4: Easy to Read and Plain Language
- RA6: Accessibility of Multilingual Digital Documents

More info at https://bfc.unige.ch
1. **Practical perspective:** Switzerland is a multilingual country (*Multilingual Language Policy at federal/cantonal level*)
   - All official documents should be available in French, German and Italian
   - Producing EtR versions in all three languages can be costly (time, resources, money, expertise)

2. **Academic perspective:** Growing interest in Machine Translation and Controlled Languages over time
   - Focus usually placed on technical documentation (PL and EtR often omitted)
     - Recent work by Rossetti (2018, 2019) on MT of PL summaries of health-related content
   - Observations often made on text fluency, adequacy and/or translatability, rather than usability/actionability of the resulting text
     - HOCL (text comprehension by humans) vs MOCL (text comprehension by computers)
   - Target users in MT-related research were rarely people with disabilities
Research Motivation (II)

**General assumption:** If E2R are more comprehensible for humans, then it should also be the case for machines!

**BUT:**

![Translation example](https://translate.google.com/)

To succeed in reform, *all people* must who are in charge of disability insurance participate.
3. **Hype of Neural Machine Translation (NMT)**

- **Statistical Machine Translation (SMT)**
  - No representation of meaning: *cat* – *cats*, just two text strings
  - Learn word associations with probabilities (*cat* – *chat* more frequent than *cat* – *chien*)
  - With training data, we could expect good results for EtR texts

- **Neural Machine Translation (NMT)**
  - Deep Learning; meaning of the word is represented using embeddings

As opposed to SMT, **NMT output is less predictable** (grammar and vocabulary changes can be introduced by the system) (Neubig et al. 2015)
An Exploratory Study

Project Goal

Explore the impact of neural machine translation (NMT) on the usability of EtR texts by people with intellectual disabilities (ID)

Pre-study steps:

- **Comparison** of popular EtR guidelines (FR-EN)
  1. **PL** (Federal Plain Language Guidelines, 2011)
  2. **E2R** (Inclusion Europe, European Commission, 2009)
  3. **FALC** (Inclusion Europe, European Commission, 2009)
  4. **Rédaction administrative** (Comité d’Orientation pour la Simplification du Langage Administratif, 2002)

- **Selection** of E2R rules

Our first study

Are EtR guidelines followed in the source text respected in the resulting NMT output for the French-English language combination?
**Methodology**

**Materials (I)**

**European Easy-to-Read Guidelines**

- 25 rules selected
- 4 ad-hoc categories created (General, Word Level, Sentence Level, Structure)

**DQF-MQM Error Typology**

- 8 categories (Accuracy, Fluency, Terminology, Style, Design, Locale Convention, Verity, Other)
- 2-7 subcategories for each category (e.g. omission, grammar, unidiomatic)
- 3 main severity levels used (critical, major, minor)

<table>
<thead>
<tr>
<th>Category</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Use the right language for your audience</td>
</tr>
<tr>
<td></td>
<td>Explain the topic and any difficult words</td>
</tr>
<tr>
<td>Word Level</td>
<td>Use common words that are easy to understand</td>
</tr>
<tr>
<td></td>
<td>Do not use difficult words, and explain them if you do</td>
</tr>
<tr>
<td></td>
<td>Use the same word to describe the same thing and be consistent throughout the document</td>
</tr>
<tr>
<td></td>
<td>Do not use figures of speech like metaphors</td>
</tr>
<tr>
<td></td>
<td>Do not use words from other languages</td>
</tr>
<tr>
<td></td>
<td>Do not use initials or abbreviations, and explain them if you do</td>
</tr>
<tr>
<td></td>
<td>Do not use percentages or big numbers</td>
</tr>
<tr>
<td></td>
<td>Make sure pronoun use is clear, and repeat the proper noun if it is not</td>
</tr>
<tr>
<td></td>
<td>Do not use special characters</td>
</tr>
<tr>
<td></td>
<td>Do not use numbers with ordinal indicators or suffixes</td>
</tr>
<tr>
<td></td>
<td>Write numbers as digits (instead of words, do not use Roman numerals)</td>
</tr>
<tr>
<td></td>
<td>Do not use contractions</td>
</tr>
<tr>
<td></td>
<td>Write dates in full</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>Speak directly to the audience using the second person “you”</td>
</tr>
<tr>
<td></td>
<td>Use positive instead of negative sentences</td>
</tr>
<tr>
<td></td>
<td>Use active instead of passive constructions</td>
</tr>
<tr>
<td></td>
<td>Use simple punctuation</td>
</tr>
<tr>
<td></td>
<td>Use the present tense whenever possible</td>
</tr>
<tr>
<td>Structure</td>
<td>Start new sentences on a new line</td>
</tr>
<tr>
<td></td>
<td>Do not split a word over two lines with a hyphen</td>
</tr>
<tr>
<td></td>
<td>Keep sentences to one line, or break the line where a person would naturally pause in their speech</td>
</tr>
<tr>
<td></td>
<td>Use clear and descriptive headings</td>
</tr>
<tr>
<td></td>
<td>Use bullet points instead of lists separated by commas</td>
</tr>
</tbody>
</table>
Text selected:

- **Administrative text**
  (originally published in FR, Swiss Federal Administration)

- **7-page FALC summary**
  (published December 7, 2015) of a 160-page report on a Swiss Disability Insurance law reform
**Methodology**

**Procedure**

**Translation error annotation**

- 414 segments x 3 evaluators = 1242 reviewed segments

**Annotation of E2R rule violations**

- 414 segments x 3 evaluators = 1242 reviewed segments

---

**Annotators’ profile:**

- 3 graduate students in FR>EN translation
- All native English speakers (1 UK, 2 US)
- 2/3 had previous experience with the DQF-MQM Error Typology
- 1/3 had previous experience with E2R
  - All were given a copy of the error typology documentation and a copy of the E2R guidelines, with examples
Preliminary Results

Human Evaluation (I)

DQF-MQM Error Typology

Comparison of Major & Minor Translation Errors Across 3 MT Systems

- Yandex
- Google
- DeepL

Average number of errors
DQF-MQM Error Typology

**Minor Errors**

- **Accuracy**
  
  “elle” (refers to “personne”) → “her” / “she”

- **Fluency**
  
  “Disability insurance also gives money to businesses / who give work to disabled people.”

- **Terminology**
  
  “personnes invalides” → “disabled people” vs. “the disabled” vs. “disabled persons” vs. “people with disabilities” (preferred)

- **Style**
  
  “pour aider les personnes invalides à retravailler” → “to help people with disabilities rework”

**Major Errors**

- **Accuracy**
  
  “Situation” → “Location”

- **Fluency**
  
  “To succeed in reform, all people must / who deal with disability insurance are participating.”

- **Terminology**
  
  “invalide veut dire...” → “invalid means”

- **Style**
  
  “pendant un moment” → “during a moment”

---

**EXAMPLES**

- **DQF-MQM Error Typology**

  - **Minor Errors**
    - **Accuracy**
      - “elle” (refers to “personne”) → “her” / “she”
    - **Fluency**
      - “Disability insurance also gives money to businesses / who give work to disabled people.”
    - **Terminology**
      - “personnes invalides” → “disabled people” vs. “the disabled” vs. “disabled persons” vs. “people with disabilities” (preferred)
    - **Style**
      - “pour aider les personnes invalides à retravailler” → “to help people with disabilities rework”
  
  - **Major Errors**
    - **Accuracy**
      - “Situation” → “Location”
    - **Fluency**
      - “To succeed in reform, all people must / who deal with disability insurance are participating.”
    - **Terminology**
      - “invalide veut dire...” → “invalid means”
    - **Style**
      - “pendant un moment” → “during a moment”
DQF-MQM Error Typology

Preliminary Results
Human Evaluation (II)

Comparison of Major & Minor Translation Errors Across 3 MT Systems

- Yandex
- Google
- DeepL

Black bars represent E2R violations because we classified them under the style category (as company style errors)
Preliminary Results

European Easy-to-Read Guidelines

Human Evaluation (II)

Comparison of Easy-to-Read Guideline Violations Across 3 MT Systems

- Yandex
- Google
- DeepL

Average number of violations:

- Total
- Word Level
- Sentence Level
- General
- Structure
European Easy-to-Read Guidelines

- **Word Level**

  “Par exemple : Qui reçoit de l’argent” → “For example : who receives money”

  “un apprentissage” → “an apprenticeship”

  “Donner de l’argent aux employeurs / pour qu’ils leur apprennent un métier” →
  “Give money to employers / to teach them a job”

- **Sentence Level**

  “Pour ces personnes, on doit améliorer l’assurance-invalidité.” →
  “For these people, disability insurance must be improved.”

  “ou quand elles ont besoin d’un objet pour les aider” →
  “when they need something to help them”
## Preliminary Results

### Human vs Automatic Evaluation

<table>
<thead>
<tr>
<th>Category</th>
<th>E2R Guideline</th>
<th>Linguistic Feature(s)</th>
<th>Yandex SMT Score</th>
<th>Google NMT Score</th>
<th>DeepL NMT Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>Always use the right language for the people your information is for.</td>
<td>Flesch Reading Ease</td>
<td>63.069</td>
<td>66.12</td>
<td>64.066</td>
</tr>
<tr>
<td>Word Level</td>
<td>Use easy to understand words that people will know well.</td>
<td>Average word length (syllables)</td>
<td>1.618</td>
<td>1.575</td>
<td>1.597</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Word Frequency</td>
<td>2.45</td>
<td>2.515</td>
<td>2.507</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Age of Acquisition</td>
<td>353.953</td>
<td>345.882</td>
<td>350.241</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Familiarity</td>
<td>580.903</td>
<td>583.949</td>
<td>582.226</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>Keep your sentences short.</td>
<td>Average sentence length (words)</td>
<td>6.781</td>
<td>7.36</td>
<td>7.55</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>Speak to people directly.</td>
<td>2nd Person Pronoun Incidence</td>
<td>5.175</td>
<td>1.224</td>
<td>2.387</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>Use active language rather than passive language.</td>
<td>Agentless Passive Voice Density</td>
<td>5.175</td>
<td>6.12</td>
<td>5.967</td>
</tr>
<tr>
<td>Sentence Level</td>
<td>Use positive sentences rather than negative ones.</td>
<td>Negation Density</td>
<td>6.468</td>
<td>4.896</td>
<td>5.967</td>
</tr>
</tbody>
</table>

### Selection of Coh-Metrix 3.0
Indices that could quantify adherence to E2R guidelines

http://cohmetrix.com/
Project Goal

Explore the impact of neural machine translation (NMT) on the usability of E2R texts by people with intellectual disabilities

Preliminary conclusions

DeepL (NMT) seems to have performed best in terms of both E2R and translation, while Yandex (SMT) performed worst (human evaluation).

New E2R violations can be introduced in the MT process, but E2R violations in the ST are sometimes resolved.

Texts that are easy for humans are not necessarily easy for machines (pre-editing might still be needed).
Limitations and Future Work

1. Limitations
   - Scope
     - One language combination, one direction
     - One text, one field
   - Annotators
     - Limited number, inexperience

2. Future Work
   - User testing
   - Source text evaluation
   - Measure post-editing effort
   - Eye tracking
BFC 2020 – 5-6 June 2020
Winterthur, Switzerland

- **1st Call for Papers is out!**
- **Easy to Read** is one of the main topics of the conference
- **Submission deadline:** 29<sup>th</sup> Nov 2019
- **More info at:**
  [https://www.zhaw.ch/en/linguistics/research/conference-on-barrier-free-communication/](https://www.zhaw.ch/en/linguistics/research/conference-on-barrier-free-communication/)


Merci
Danke
Grazie
Thanks
Kiitos!