SPANISH TRANSLATIONS OF COCHRANE PLAIN LANGUAGE SUMMARIES:

Assessing the impact of a controlled language checker on machine translation quality

Alessandra Rossetti
Dublin City University
alessandra.rossetti2@mail.dcu.ie

DCU
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn
INTERACT
International Network
on Crisis Translation
Horizon 2020
European Union funding
for Research & Innovation

Funding sources and lack of conflict of interest

- This research has received funding from the Irish Research Council, and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 734211.

- I have no actual or potential conflict of interest in relation to this presentation.

Background and motivation (1)

- Cochrane aims to make high-quality health information accessible for users worldwide (Chandler et al. 2017): about 75% of the world population do not speak English (Cochrane Translations 2018)

- Cochrane multi-language strategy: translation of cochrane.org, abstracts and **plain language summaries (PLS)** from English into a variety of languages (e.g. Croatian, Spanish, French, Thai). On average, the **Spanish** version of the Cochrane Library was searched around 4,000,000 per year between 2012 and 2014

- Translations:
  - Mainly conducted and/or revised for accuracy by volunteer health domain experts
  - Time-consuming and onerous tasks, especially for volunteers with no linguistic training

Background and motivation (2)

- **Controlled language checkers** and **machine translation (MT)** to streamline the translation process and make it more sustainable (Birch 2017; Von Elm et al. 2013)

- A controlled language is a **set of rules** adopted to make a text more comprehensible and **translatable**. A controlled language checker is the software that checks for adherence to those rules (O'Brien 2010)
  - Automatically and consistently flagging translatability issues in a text
  - Providing suggestions on how to solve them

Research question

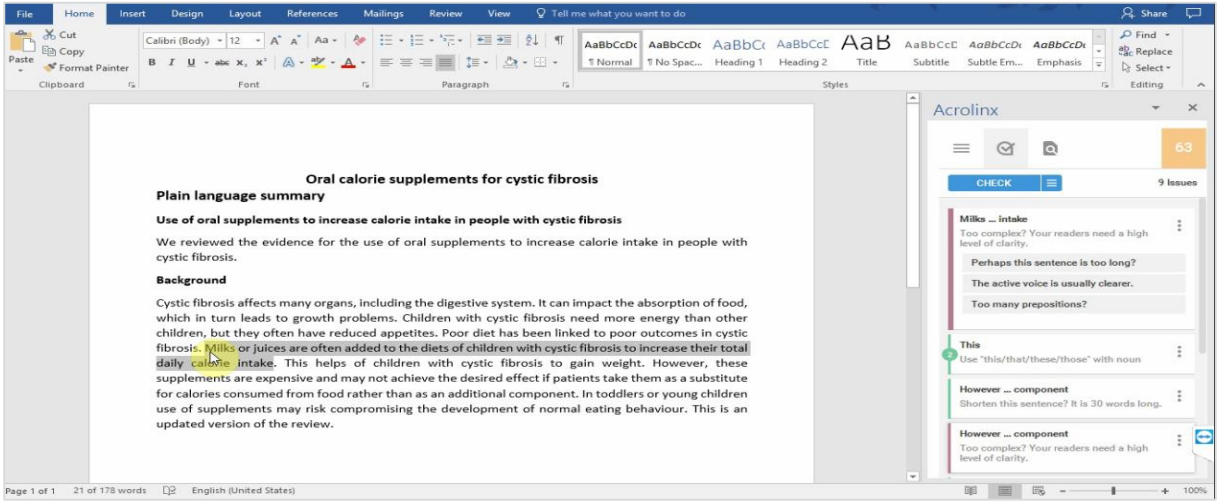
- Does using the **Acrolinx controlled language checker** to revise Cochrane PLS increase their machine translatability* from English into Spanish?

*Machine translatability as "a measure that indicates how well a given sentence can be translated by a particular MT system" (Izumi, Uchimoto and Isahara 2006)

The freely available MT system Google Translate was adopted

Experimental set up - stage 1

- 12 Cochrane authors asked to use Acrolinx
  - To check for translatability issues in PLS previously produced
  - To revise the PLS accordingly
- Acrolinx as a plugin in Microsoft Word



Experimental set up - stage 2 (1)

- 12 **pre-Acrolinx PLS** and corresponding Spanish MT outputs
- 12 **post-Acrolinx PLS** and corresponding Spanish MT outputs

- 41 evaluators (native speakers of Spanish and health domain experts)

- Asked to rate each sentence in the MT outputs for adequacy and fluency (4-point Likert scale) (Linguistic Data Consortium 2002)
  - **Adequacy**: How much of the information contained in the English source sentence (SS) appears in the Spanish target sentence (TS)? From 1 (none of it) to 4 (all of it)
  - **Fluency**: Indicate the extent to which the Spanish target sentence (TS) is in grammatically well-formed and fluent Spanish. From 1 (incorrect and disfluent) to 4 (correct and fluent)

Experimental set up - stage 2 (2)

- Within-subject design: 2 PLS and corresponding outputs per evaluator
- Participants divided into 12 groups, different PLS per each group — between 3 and 4 evaluators per group
- Presentation of pre-Acrolinx and post-Acrolinx PLS counterbalanced to avoid order effects
- MT evaluation conducted remotely (on Google Forms) and independently — no influence between evaluators
- Follow-up questions on the quality of the MT outputs

Findings on machine translatability (1)

- Differences in adequacy and fluency scores between pre-Acrolinx and post-Acrolinx PLS were **slight** — the use of this controlled language checker did not increase the machine translatability of PLS

Grand means and SD (across all evaluators) of…	Pre-Acrolinx PLS	Post-Acrolinx PLS
Adequacy (1-4 scale)	3.72 (0.33)	3.78 (0.27)
Fluency (1-4 scale)	3.27 (0.6)	3.33 (0.55)

Findings on machine translatability (2)

- Fluency and adequacy scores were **relatively high,** suggesting that Google Translate produced reasonably good translations
- **Adequacy** (content) rated as **higher** than fluency (style)
- Both findings were supported by the evaluators' follow-up comments

Comments on overall good quality of MT output

I consider that the translation has a good quality, and it is completely understandable

I think most of the translated text goes from acceptable to high quality

I think the overall machine translated texts were pretty great

The translations from English to Spanish accomplished the required level of understanding for clinical purposes

Comments on fluency/style issues

In all the sentences in Spanish I could find all the components that the sentence in English has. But sometimes the word that the machine used wasn't the right word or not the most used word in Spanish, then the sentence in Spanish sounds weird.

When you read it, you know that is a machine translator and not a native speaker.

My impression of the translators is that they make the translation very literally, and that makes lose the fluency in reading the translated, however the general context of what the translated document is about is well understood, which is very favourable.

When I was trying to think of how you would say things in Spanish, because I already had this Spanish sounding translation it was actually really hard to think differently and more naturally.

Inter-rater agreement

Groups	ICC (adequacy)	ICC (fluency)
1	0.68	0.73
2	0.34	0.51
3	0.29	0.49
4	0.8	0.72
5	0.14	0.55
6	0.54	0.36
7	0.62	0.49
8	-0.05	0.31
9	0.36	0.62
10	0.67	0.4
11	0.43	0.12
12	0.48	0.04

The intraclass correlation coefficient (ICC) showed that the **agreement** between raters was between **poor** and **moderate**

Possible explanation: lack of training of participants on evaluation of MT output

Implications (for healthcare consumers) and future work

- Encouraging results regarding the adoption of freely available MT systems in the translation workflow at Cochrane, as a faster and cheaper alternative to human translation
  - Might encourage volunteer health domain experts to take part in the tasks of correcting and/or validating the MT output
  - Might boost the number of translations (of PLS) made available online, thus increasing the accessibility of Cochrane content for linguistically diverse health consumers
- Fluency/style errors (more common than adequacy/content errors) are easier to correct and less detrimental than content errors for health consumers (Koponen 2010; Stymne 2013)
- Future work:
  - Testing the impact of other controlled language checkers on machine translatability of PLS
  - Repeating the same study after training participants on evaluation of MT output
  - Repeating the same study with evaluation at the document (rather than the sentence) level (Läubli, Sennrich and Volk 2018)
  - Assessing machine translatability into other (e.g. Asian) languages

Thank you for your attention!

Alessandra Rossetti

Dublin City University

alessandra.rossetti2@mail.dcu.ie

DCU
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn
INTERACT
International Network
on Crisis Translation
Horizon 2020
European Union funding
for Research & Innovation

References

- Birch, A. 2017. *D5.4: Report on Second Year's MT Evaluation* [Online].
- Chandler, J., Higgins, J. P. T., Deeks, J. J., Davenport, C. and Clarke, M. J. Chapter 1: Introduction. IN: Higgins, J. P. T., Churchill, R., Chandler, J. and Cumpston, M. S. (eds.) *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.2.0* [Online].
- Cochrane Translation 2018. *Infographic* [Online].
- Izumi, E., Uchimoto, K. and Isahara, H. 2006. Measuring intelligibility of Japanese learner English. *IN:* Salakoski, T., Ginter, F., Pyysalo, S. and Pahikkala, T. (eds.) *Advances in Natural Language Processing*. Berlin and Heidelberg: Springer, pp. 476-487.
- Koponen, M. 2010. Assessing machine translation quality with error analysis. *IN: Electronic Proceedings of the KäTu Symposium on Translation and Interpreting Studies* [Online].
- Läubli, S., Sennrich, R. and Volk, M. 2018. Has machine translation achieved human parity? A case for document-level evaluation [Online].
- Linguistic Data Consortium 2002. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations* [Online].
- O'Brien, S. 2010. Controlled language and readability. *Translation and Cognition - ATA Scholarly Monograph XV Series*. Amsterdam and Philadelphia: John Benjamins, pp. 143-165.
- Stymne, S. 2013. Using a grammar checker and its error typology for annotation of statistical machine translation errors. *IN: Proceedings of the 24th Scandinavian Conference of Linguistics (24SCL)*. Joensuu, Finland, 25-27 August, pp 332-344.
- Von Elm, E., Ravaud, P., MacLehose, H., Mbuagbaw, L., Garner, P., Ried, J. and Bonfill, X. 2013. Translating Cochrane reviews to ensure that healthcare decision-making is informed by high-quality research evidence. *PLOS Medicine*, 10(9), e1001516.