The reception of intralingual and interlingual automatic subtitling: an exploratory study within the HBB4ALL project

Anna Matamala, Andreu Oliver (UAB)
Aitor Álvarez, Andoni Azpeitia (Vicomtech)
anna.matamala/andreu.oliver@uab.cat
aalvarez/aazpeitia@vicomtech.org

Translating and the Computer 37 (TC37-Asling 2015), London 26-27/11/15.

Anna Matamala, Andreu Oliver, Aitor Álvarez, Andoni Aizpetia. 2015. "The reception of intralingual and interlingual automatic subtitling: an exploratory study within the HBB4ALL project".

As presented at the *37th Conference Translating and the Computer*, London, UK, November 26-27, 2015.

Table of Contents

1. Aims

Test whether automatic interlingual subtitles (English into Spanish) and intralingual subtitles (English) help to improve understanding of news content originally broadcast in English.

2. Technological components

- Technology provided by Vicomtech-IK4

a) Automatic Subtitling Component, composed by:

  - LVCSR engine built with KALDI (Povey *et al.*, 2011) operating in real-time.
  - An HMM-GMM acoustic model and 3-gram language Model estimated through KenLM (Heafield, 2011) toolkit.
  - Automatic punctuation and capitalization.
  - EBU-TT-D format subtitles generation.

2. Technological components

- Technology provided by Vicomtech-IK4

b) Moses SMT component (Koehn et al., 2007):

  - Corpora from OPUS repository: news and general domain.
  - Data selection using Bilingual Cross-Entropy Difference (Axelrod et al., 2011).
  - Two phrase-based models combined through perplexity minimization (Sennrich, 2012).
  - Final combined model tuned using 5-gram language model.

3. Testing

- **Materials**: 3 comparable short clips from Reuters.
- **Viewing conditions**: no subtitles/ intralingual /interlingual.
- **Methods**: comprehension questionnaires (improved in main test).

3. Testing

- **Participants: preliminary testing**

	#Participants	English level	Subtitles
Group 1	10	Low	Interlingual
Group 2	20	Low	Intralingual
Group 3	26	High	No subtitles

3. Testing

- **Participants:**

  main experiment

English levels	#Participants
A1	0
A2	2
B1	8
B2	7
C1	8
C2	5
Total	**30**

4. Results: comprehension levels

- Preliminary testing

English skills	Subtitles in…	Clip 1	Clip 2	Clip 3	Total
Lower	Spanish	29.5%	35.5%	41.9%	35.73%
	English	30%	37.75%	41.25%	35.73%
Higher	No subtitles	42.85%	30.03%	47.80%	41.66%

- Low level of English: no significant differences.
- Understanding increases from clip 1 to 3 (methodological limitations).

4. Results: comprehension levels

- Main test

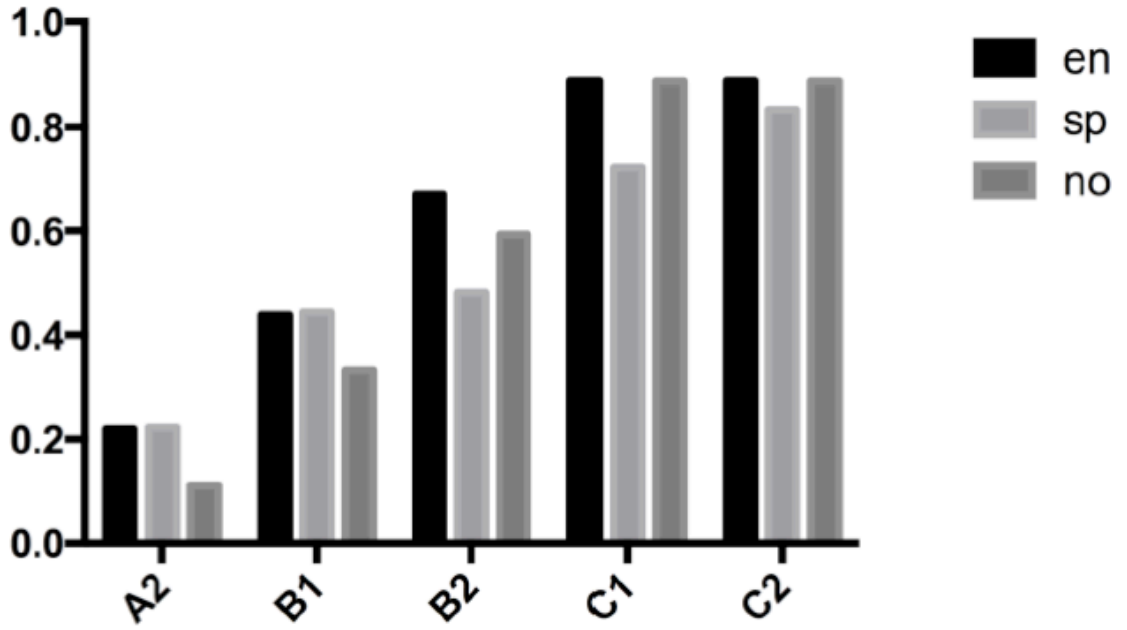

Figure 1. Percentage of correct replies (en: English intralingual subtitles, sp: Spanish interlingual subtitles, no: without subtitles)

4. Results: comprehension levels

- Less proficient: improvement in comprehension for both intralingual and interlingual subtitles, but comprehension is low.
- Most proficient: no improvement in intralingual, comprehension decreases with interlingual.
- Medium-level of English: improvement in intralingual, comprehension decreases with interlingual.

5. Conclusions

- Automatic subtitles, useful for participants with a middle-range level of English, but only if intralingual.
- Distracting effect in highly proficient participants?

Acknowledgements

Anna Matamala and Andreu Oliver are members of TransMedia Catalonia, a research group funded by the Catalan government (2014SGR027). This research is part of the HBB4ALL project, which was co-funded by the European Commission under the Competitiveness and Innovation Framework Program (CIP) and by 12 partners from several fields.

References

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* December 2011.

Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics. Edinburgh, UK. 189-197.

Koehn, Philipp, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions.* Association for Computational Linguistics, June 2007. 177-180.

Axelrod, A., He, X. & Gao, J. 2011. Domain Adaptation Via Pseudo In-Domain Data Selection. In *Proceedings of Empirical Methods in Natural Language Processing*. Edinburgh, UK. 355-362.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Avignon, France, 539-549.

Glossary

- IK4: Alliance of R&D centres in the Basque Country http://www.ik4.es/es/default.asp
- LVCSR: Large Vocabulary Continuous Speech Recognition.
- KALDI toolkit: http://kaldi.sourceforge.net/about.html
- HMM-GMM: Hidden Markov Model – Gaussian Mixture Model
- KenLM toolkit: Kenneth Heafield Language Model: https://kheafield.com/code/kenlm/
- EBU-TT-D format: European Broadcasting Union Timed Text part 'D': https://tech.ebu.ch/ebu-tt
- N-gram: probabilistic models which exploit the ordering of words predicting the next word from the previous N-1 words. In a bit of terminological ambiguity, the term N-gram is usually used to refer to either the word sequence or the predictive model.
- HBB4ALL: "Hybrid Broadcast Broadband for All" European project. http://www.hbb4all.eu/

The reception of intralingual and interlingual automatic subtitling: an exploratory study within the HBB4ALL project

Anna Matamala, Andreu Oliver (UAB),
Aitor Álvarez, Andoni Azpeitia (Vicomtech)
anna.matamala/andreu.oliver@uab.cat
aalvarez/aazpeitia@vicomtech.org

Translating and the Computer 37 (TC37-Asling 2015), London 26-27/11/15.