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

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. © AsLing, The International Association for Advancement in Language Technology, 2015. Distribution without the authorisation from ASLING is not allowed. AsLing asking to be informed of such postings, including URLs or URIs where available, to the email address: <presentations@asling.org>.”
Table of Contents

1. Aims
2. Technological components
3. Testing
4. Results: comprehension levels
5. Conclusions
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

<table>
<thead>
<tr>
<th></th>
<th>#Participants</th>
<th>English level</th>
<th>Subtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>10</td>
<td>Low</td>
<td>Interlingual</td>
</tr>
<tr>
<td>Group 2</td>
<td>20</td>
<td>Low</td>
<td>Intralingual</td>
</tr>
<tr>
<td>Group 3</td>
<td>26</td>
<td>High</td>
<td>No subtitles</td>
</tr>
</tbody>
</table>
3. Testing

- Participants:
  main experiment

<table>
<thead>
<tr>
<th>English levels</th>
<th>#Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0</td>
</tr>
<tr>
<td>A2</td>
<td>2</td>
</tr>
<tr>
<td>B1</td>
<td>8</td>
</tr>
<tr>
<td>B2</td>
<td>7</td>
</tr>
<tr>
<td>C1</td>
<td>8</td>
</tr>
<tr>
<td>C2</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>
4. Results: comprehension levels

• Preliminary testing

<table>
<thead>
<tr>
<th>English skills</th>
<th>Subtitles in...</th>
<th>Clip 1</th>
<th>Clip 2</th>
<th>Clip 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>Spanish</td>
<td>29.5%</td>
<td>35.5%</td>
<td>41.9%</td>
<td>35.73%</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>30%</td>
<td>37.75%</td>
<td>41.25%</td>
<td>35.73%</td>
</tr>
<tr>
<td>Higher</td>
<td>No subtitles</td>
<td>42.85%</td>
<td>30.03%</td>
<td>47.80%</td>
<td>41.66%</td>
</tr>
</tbody>
</table>

• Low level of English: no significant differences.

• Understanding increases from clip 1 to 3 (methodological limitations).
4. Results: comprehension levels

• Main test
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.


• LVCSR: Large Vocabulary Continuous Speech Recognition.
• KALDI toolkit: [http://kaldi.sourceforge.net/about.html](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/](https://kheafield.com/code/kenlm/)
• EBU-TT-D format: European Broadcasting Union Timed Text part ‘D’: [https://tech.ebu.ch/ebu-tt](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/](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)

annamatamala/andreu.oliver@uab.cat
aalvarez/aazpeitia@vicomtech.org