The Reception of Intralingual and Interlingual Automatic Subtitling: An Exploratory Study within The HBB4ALL Project

Anna Matamala
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental, Universitat Autònoma de Barcelona
Anna.matamala@uab.cat

Andreu Oliver
Departament de Psicologia Bàsica, evolutiva i de l'Educació, Universitat Autònoma de Barcelona
Andreu.oliver@uab.cat

Aitor Álvarez
Human Speech and Language Technologies Department, Vicomtech-IK4
aalvarez@vicomtech.org

Andoni Azpeitia
Human Speech and Language Technologies Department, Vicomtech-IK4
aazpeitia@vicomtech.org

Abstract

This paper presents the results of a preliminary experiment and a main test within the HBB4ALL project that aimed to determine whether automatic interlingual and intralingual subtitling help to better understand news content. Results tend to indicate that the usefulness of automatic subtitling correlates with the participants' English level, enhancing comprehension only in certain groups.

1 Introduction

HBB4All[1] is an EC-funded project that builds on HbbTV, the European standard for broadcast and broadband multimedia converged services, and looks at how HbbTV technologies can enhance access services such as subtitling. Within the project, user testing related to automatic subtitling has been carried out by Universitat Autònoma de Barcelona (UAB) and Vicomtech-IK4 research centre. Automatic subtitles were generated through two main components based on Large Vocabulary Continuous Speech Recognition (LVCSR) and Statistical Machine Translation (SMT) technologies. The component based on LVCSR technology generated intralingual subtitles, whilst the one using SMT technology created interlingual subtitles. This study presents the results of user testing on automatic subtitling. The goal was to determine whether automatic interlingual subtitling (English to Spanish) and/or automatic intralingual subtitling (English) help to improve understanding of news content originally broadcast in English.

The paper is structured as follows. Section 2 looks at the technological components used to generate the intralingual and interlingual subtitles. Section 3 presents the preliminary experiment, and Section 4 describes the main test. Section 5 draws conclusions and describes future work.

[1] http://www.hbb4all.eu/

Proceedings of the 37th Conference Translating and the Computer, pages 12–17,
London, UK, November 26-27, 2015.

2 Technological Components

Vicomtech-IK4 provided technology to automatically generate and translate EBU-TT-D subtitles from audiovisual content. Intralingual subtitles were generated through the Automatic Subtitling Component, which was composed by a LVCSR engine. It was responsible for transcribing audio input stream according to an acoustic model, vocabulary and language model. The recognition engine was based on an HMM-GMM (hidden Markov model – Gaussian mixture model) acoustic model with context-dependent phone states and it was trained using KALDI (Povey *et al.*, 2011). The language model was a trigram language model and it was estimated through KenLM (Heafield, 2011) toolkit. The transcription was then automatically punctuated and capitalized, and EBU-TT-D format subtitles were generated.

Interlingual subtitles were created through the SMT Component, which allows the automatic translation of subtitles from English to Spanish in EBU-TT-D format. The SMT technology was built using the Moses SMT system (Koehn *et al.*, 2007). The English into Spanish SMT model was trained over parallel corpora that were collected from the OPUS[2] repository. A balanced adaptation to the news domain and a general language coverage were reached through data selection technique, which was performed using a bilingual cross-entropy difference approach (Axelrod *et al.*, 2011). The resulting data were then prepared using in-house tokenization and true casing models, and used to train two separate phrase-based models, which were finally combined through perplexity minimization on a selected in-domain development test, following Sennrich (2012). The final combined model was tuned using a 5-gram language model created from the entire selected monolingual data.

3 Preliminary Experiment

This section describes the preliminary testing, including its methods, materials, and results.

3.1 Methods and Materials

56 Political Science students volunteered to take part in the experiment. They were categorised by expert lecturers in two levels of English: lower and higher, as it was deemed that English proficiency would affect the results.

Eight short clips from the Reuters[3] video service were initially prepared with intralingual and interlingual subtitles. The clips were about breaking news on business, finance and markets, and lasted around three minutes each. After an analysis of the content, three clips were selected, aiming to reach a balance in terms of number of speakers, content, topic and length.

Following Day and Park (2005), comprehension questionnaires were developed for each clip (20 questions per clip, mostly multiple-choice), and an analysis of the clips allowed to control the information provided visually (Cross, 2011).

3.2 Procedure

Three viewing conditions were prepared: no subtitles, intralingual English subtitles, and interlingual Spanish subtitles. For practical reasons, a randomized viewing was not possible. Table 1 presents the number of participants per group, their English level and the viewing condition.

[2] http://opus.lingfil.uu.se/
[3] http://www.reuters.com/

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

Table 1. Groups in the preliminary test

Participants replied to the questionnaires once they had watched the clips. The data gathered allowed the comprehension of students with low English (Group 1 and Group 2) consuming intralingual and interlingual subtitles to be compared. It was also possible to compare results of students with low English level using subtitles, either intra- or interlinguistic (Group 1 and Group 2), against students with better level of English without subtitles (Group 3). These preliminary experiments were the perfect ground for testing the methodology.

3.3 Results

Table 2 presents the comprehension levels of students with low English level using intralingual and interlingual automatic subtitles. The percentages refer to the number of correct replies to the questions for each clip.

Subtitle language	Clip 1	Clip 2	Clip 3	Total
Spanish (interlingual)	29.5%	35.5%	41.9%	35.73%
English (intralingual)	30%	37.75%	41.25%	35.73%

Table 2. Percentage of correct replies

The difference is not significant between groups, although higher comprehension levels were expected for intralingual subtitling, where quality levels are higher. Besides, the percentage of correct replies is very low (below 40%), and understanding seems to increase from clip 1 to 3.

On the other hand, when comparing the comprehension of participants with a low level using subtitles (Group 1 and 2) with that of participants with a high level not using subtitles (Group 3), results show no major differences (Table 3).

English skills	Subtitle language	Clip 1	Clip 2	Clip 3	Total
Lower	Spanish (interlingual)	29.5%	35.5%	41.9%	35.73%
	English (intralingual)	30%	37.75%%	41.25%	35.73%
Higher	No subtitles	42.85%	30.03%	47.80%	41.55%

Table 3. Comparison of correct replies

These preliminary results left many open questions. First, students with lower English skills who watched clips with either type of subtitles presented almost identical percentages in comprehension. It remained to be seen what would happen if the same clips were shown without subtitles. Secondly, students with higher English skills presented slightly higher comprehension percentages when watching the original content without subtitles, although the difference was minimum. Because of the experiment design, it was not possible to see whether the difference was due to their English proficiency or to the fact that the absence of subtitles may avoid split attention and actually increase comprehension in certain groups.

4 Main Experiment

The main experiment also included three conditions: automatic intralingual subtitles (English), automatic interlingual subtitles (Spanish), and English content without subtitles. The hypotheses were that both intralingual and interlingual automatic subtitles should increase comprehension compared to clips with no subtitles, whilst interlingual subtitling would not increase comprehension compared to clips with intralingual subtitles. Also, it was expected that subtitles would be more useful as English proficiency decreased.

4.1 Methods and Materials

Tests were carried out with 30 students (13 male, 17 female, mean age: 25.2). Materials included the three same news stories selected for the preliminary tests (see 3.1), in the three conditions described above. Automatic subtitles were the same as those produced for the preliminary test, but comprehension questionnaires were adapted based on the preliminary test results.

English skills were controlled through an on-line test[4] that lasted a maximum of 20 minutes and allowed us to classify participants in six levels (Table 4).

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

Table 4. English levels and number of participants

Very few participants were included in the lowest levels (A1 and A2), whilst the number of participants between B1 and C1 provides a more balanced sample. This is why a qualitative descriptive approach was taken in the data analysis.

4.2 Procedure

Participants were welcomed individually in a lab and were instructed that they would watch three clips on the news domain in English (one without subtitles, one with English subtitles, and one with Spanish subtitles). Clips were played twice. After the first viewing, participants could read the questions. After the second, they had to reply to the questionnaire. The viewing order was randomized.

4.3 Results

Figure 1 summarises the results obtained, namely the percentage of correct replies per English level and condition.

In the less proficient participants (A2), both automatic interlingual and intralingual subtitles increase the comprehension from 11% to 22%, although comprehension is very low (below 22%). This pattern is exactly the same for B1 participants, although comprehension levels increase: 33% with no subtitles, and up to 44% with subtitles.

[4] www.examenglish.com/leveltest/listening_level_test.htm

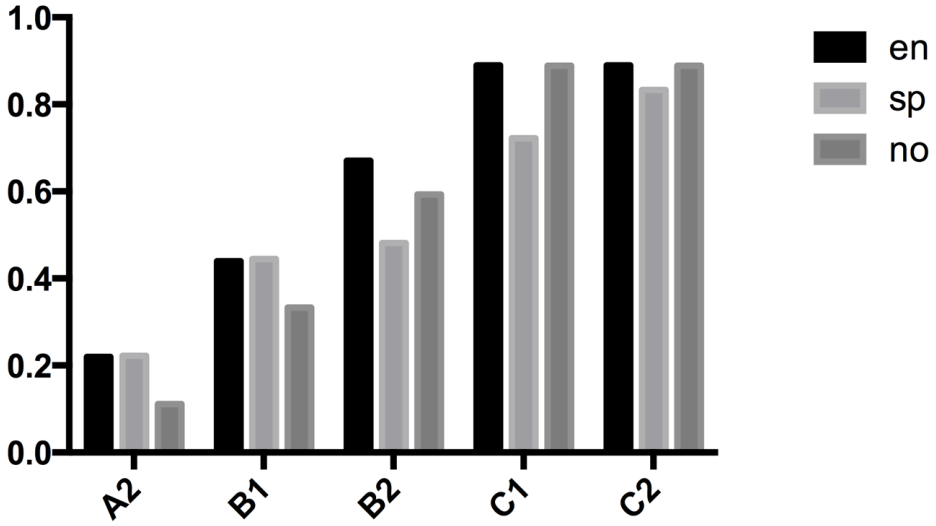

Figure 1. Percentage of correct replies (en: English intralingual subtitles, sp: Spanish interlingual subtitles, no: without subtitles)

In the most proficient participants (C1, C2), no improvement is observed with and without subtitles. Comprehension levels without subtitles are 89% for both C1 and C2. The same value is obtained for automatic intralingual subtitles. However, comprehension decreases for C1 and C2 participants when interlingual subtitles are used, with a more striking decrease in C1 (72%). Subject to further testing, this may indicate that automatic interlingual subtitles may detract the viewers' attention and affect comprehension negatively.

Finally, a different trend is observed in B2 participants: comprehension with automatic intralingual subtitles (67%) is better than without subtitles (59%), but comprehension decreases considerably with interlingual subtitles (48%).

If 50% of correct replies is viewed as a threshold to consider that the news has been understood, this is only achieved in the following conditions:

- B2: intralingual (67%), no subtitles (59%)
- C1: intralingual (89%), interlingual (72%), no subtitles (89%)
- C2: intralingual (89%), interlingual (83%), no subtitles (89%).

5 Conclusions

Results from the preliminary test pointed to some methodological weaknesses which were addressed in the main test, in which the following conclusions were reached.

Regarding the hypothesis that intralingual automatic subtitling increases comprehension as compared to clips with no subtitles, it has been confirmed for participants whose English level is between A2 and B2, but comprehension stays the same for intralingual automatic subtitling and no subtitles for C1 and C2.

Concerning the hypothesis that interlingual automatic subtitling increases comprehension compared to clips with no subtitles, it has been confirmed for the less proficient participants (A2, B1), although comprehension levels are low. As for B2, C1 and C2, comprehension is better without subtitles than with interlingual subtitles, which could prove a distracting effect of these subtitles.

Regarding the hypothesis that interlingual automatic subtitling does not increase comprehension compared to clips with intralingual subtitles, it has been confirmed for all

participants. Comprehension stays the same (A2, B1) or improves (B2, C1, and C2) with intralingual subtitles in English compared to interlingual subtitles.

A general conclusion is that automatic subtitles are useful for participants with a middle-range level of English (B2) but only if intralingual, at least in the current stage of development. In participants with low English proficiency, both intralingual and interlingual automatic subtitling increase comprehension but levels remain very low, so no substantial effect is observed. In highly proficient participants, subtitles do not increase comprehension; on the contrary, interlingual subtitles may affect comprehension negatively, possibly due to a distracting effect. Despite the trends observed, further testing is still needed with wider samples, more clips, other language pairs, and improved technologies.

Acknowledgments

This research is part of the project Hybrid Broadcast Broadband for all, funded by the EC (FP7 CIP-ICT-PSP.2013.5.1. 621014). A. Matamala and A. Oliver are TransMedia Catalonia members, a research group funded by Generalitat de Catalunya (2014SGR0027).

References

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation Via Pseudo In-Domain Data Selection. *Proceedings of Empirical Methods in Natural Language Processing*. Edinburgh, UK, 355-362.

Cross, Jeremy. 2011. Comprehending news videotexts: the influence of visual content. *Language Learning & Technology*, *15*(2): 44–68.

Day, Richard R., and Jeong-suk Park. 2005. Developing reading comprehension questions. *Reading in Foreign Language*, *17*(1): 60–73.

Guichon, Nicolas, and Sinead Mclornan. 2008. The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, 36(1): 85-93.

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007, 177-180.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, George Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. December 2011.

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