Annotating racism: A corpus-based study of cross-linguistic racist discourse annotation and analysis

Technological advancements in Corpus Linguistics and tools for processing and compiling linguistic corpora open new ways on how we utilize corpora. Annotation is being widely used in descriptive linguistic studies [1] [2] [3] facilitating systematic lexico-grammatical analysis of linguistic resources.

This holds increasingly true also for translation corpora, with a particular focus on the examination of translation strategies and norms [2] [4]. This paper presents an ongoing PhD research to examine, from a descriptive viewpoint, and hence annotate, the translational norms of the socio-culturally marked discourse of racism, and the shifts remarked during the discourse transfer from a source language (EN) to two target-languages (EL&ES). Our aim is to present problems and impediments that arise during the annotation process applied in this study. Our work so far reveals issues related both to the annotation methodology and schemata and the implementation of the annotation process in the software utilised.

We have compiled a representative audiovisual corpus (five feature films with a total playtime of 09:05 hours) comprising transcribed (time-aligned) dialogues and their subtitles, i.e. a special trilingual parallel audiovisual corpus. The aim of the project is to facilitate comparison between source and target texts and allow conclusions on translational norms and behaviours [5] [11] with regard to subtitling practices in Greece and Spain. Racism, as manifested in discourse, had been under-researched until recently [6] [7]. However, the issue of racism and racist discourse gathers new and focused research interest in view of the European current social, political and economic backdrop, Greece being an example, where immigration flows and a sharp economic crisis changed the prevalent attitudes towards non-native Greeks. (See, for example the sharp rise of the extreme right-wing party of “Golden Dawn” [8]; and the recent opinion polls attesting this rise [9]). Realistic films on the subject are representative of discourses emanating from racist stances, while cinema, as a medium widely accessible to the public, apart from reflecting society, communicates ideas. On the other hand, subtitles are among the most read translations and text types in countries with a subtitling tradition [10].

In order to ensure conformity with standards for audio-visual material, video segmentation and transcription were performed using ELAN [12]. Each SL utterance is assigned a time slot and a speaker and is aligned to its respective TL utterances. The final output is a TEI-conformant [13] .xml document. Further linguistic annotation was considered vital for our research in order to isolate the instances relevant to the specific type of discourse. We found sentiment/subjectivity analysis [14] [15] [16] highly relevant to the analysis of racist stances, since they are expressed through emotions and/or opinions. We have annotated texts [19] on the clause level, i.e. on the level of extended units of meaning [17] and isolated negative instances, related to persons of ethnicities other than the speaker's, as candidates of racist stance.

The next step was to compare the annotated instances to the respective TL utterances on the basis of register shifts that occur through the translation process and alter the strength of the utterance or add nuances of racist stance. This presupposes the use of a second annotation schema, i.e. one annotation schema is used for each modality (SL oral text/TL subtitles). The comparison of instances of emotion/opinion and the shifts in their strength reveal the translation strategies followed by the subtitlers.

Thus, while annotation begun in ELAN (mostly for para-linguistic information), and although ELAN could be theoretically used for every kind of annotation, complex annotation schemas, as the ones used here, could not be used through its interface in an effective way. Therefore, further linguistic annotation was performed using the GATE platform [18] [19]. The tool was selected for its user-friendliness and versatility in fulfilling the requirements of our classification model.

The option of a trilingual representation of texts, along with their original audiovisual text (currently possible through ELAN), that could also visualize the annotations made in GATE (currently viewed only through the GATE interface) would prove valuable for this research. The data from our research so far will be utilised statistically, as part of our ongoing research (PhD) project. The paper will also present our preliminary findings from the utterances examined.

References:

[1] McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

[2] Zanettin, F. (2012). *Translation-driven Corpora*. Manchester: St. Jerome.

[3] Sinclair, J. M. (2004). *Trust the text. Language, corpus and discourse*. London: Routledge, 190-191.

[4] Laviosa, S. (1998). The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta* 43(4), 474-479.

[5] Toury G. (1995). *Descriptive translation studies and beyond*. Amsterdam: John Benjamins, 258-279.

[6] Dijk T. van (2005). *Racism and Discourse in Spain and Latin America.* Amsterdam: John Benjamins.

[7] Reisigl, M. & Wodak, R. (2001). *Discourse and Discrimination. Rhetorics of Racism and Antisemitism.* London: Routledge.

[8] See the “MP Zaroulia's Racist Delirium in the Parliament” youtube video, <http://www.youtube.com/watch?v=7QzB9burSts>, accessed on 24 October 2012.

[9] Hmerisia Online. (2012, September 9). Poll, first comes SYRIZA, third the Golden Dawn. *Hmerisia Online.* Accessed 24 October 2012 <http://www.imerisia.gr/article.aspcatid=26509&subid=2&pubid=112918932> [in Greek].

[10] Gottlieb, H. (1997). *Subtitles, Translation & Idioms*. Copenhagen: University of Copenhagen, 153. In: Pedersen, J. (2011). *Subtitling Norms for Television. An exploration focusing on extralinguistic cultural references*. Amsterdam: John Benjamins, 125.

[11] Saridakis, I.E. (2010). *Corpora and Translation. Theories and Applications*. Athens: Papazisi Publications, 157-164 [in Greek].

[12] Brugman, H. & Russell, A. (2004). Annotating Multimedia / Multi-modal resources with ELAN . In: *Proceedings of LREC 2004, 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

[13] TEI Consortium (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 1.3.0 of May 2011 <tei-c.org>.

[14] Wiebe, J., Wilson, T., Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 165-210.

[15] Asher, N., Benamara, F., Mathieu, Y. Y. (2009). Appraisal of opinion expressions in discourse.

Lingvisticae Investigationes, 32(2).

[16] Martin, J.R. & White, P.R.R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave, London, UK.

[17] Sinclair, J. (1996). The search for units of meaning. *Textus*, 9.1, 75–106.

[18] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia.

[19] Mouka, E., Giouli, P., Fotopoulou, A., Saridakis, I.E. (2012). Opinion and Emotion in Movies: a Modular Perspective to Annotation. *LREC 2012: ES^3 2012, 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals. Proc.*, 104-109.