SubCo: A Learner Translation Corpus of Human and Machine Subtitles

José Manuel Martínez Martínez, Mihaela Vela
Universität des Saarlandes
Campus, 66123 Saarbrücken, Germany
{j.martinez,m.vela}@mx.uni-saarland.de

Abstract

In this paper, we present a freely available corpus of human and automatic translations of subtitles. The corpus comprises the original English subtitles (SRC), both human (HT) and machine translations (MT) into German, as well as post-editions (PE) of the MT output. HT and MT are annotated with errors. Moreover, human evaluation is included in HT, MT, and PE. Such a corpus is a valuable resource for both human and machine translation communities, enabling the direct comparison – in terms of errors and evaluation – between human and machine translations and post-edited machine translations.

Keywords: subtitling; quality assessment; MT evaluation

1. Introduction

This paper describes a freely available corpus[1] consisting of original English subtitles (SRC) translated into German. The translations are produced by human translators and by SUMAT (Volk, 2008; Müller and Volk, 2013; Etchegoyhen et al., 2014), a MT system developed and trained specifically for the translation of subtitles. Both human (HT) and machine translations (MT) are annotated with translation errors (HT_ERROR, MT_ERROR) as well as evaluation scores at subtitle level (HT_EVAL, MT_EVAL). Additionally, machine translations are completed with post-editions (PE) which are also evaluated (PE_EVAL).

The corpus was compiled as part of a course on subtitling targeted at students enrolled in the BA Translation Studies programme at Saarland University. The students carried out the human translations and the error and evaluation annotation. Though novice translators, this kind of evaluation is already an improvement, since human evaluation of MT datasets is often carried out by lay translators (computational linguists or computer scientists).

To the best of our knowledge, this is the first attempt to apply human error annotation and evaluation to subtitles produced by humans and a MT system, taking into account the distinctive features of subtitling such as temporal and spatial constrains.

Such a resource can be useful for both human and machine translation communities. For (human) translation scholars a corpus of translated subtitles can be very helpful from a pedagogical point of view. In the field of MT, the possibility of comparing between human and machine translations, especially using manual metrics like error analysis and evaluation scores, can enhance MT development, in particular, applied to the specific task of subtitling.

2. Subtitling and Quality Assessment in Human and Machine Translation

Subtitling implies taking into consideration that translated text has to be displayed synchronously with the image. This means that certain constrains on space (number of lines on screen, number of characters per line) and time (synchrony, time displayed on screen) have to be respected when performing subtitling.

According to Díaz Cintas and Remael (2007) subtitling is defined as

> a translation practice that consists of presenting a written text, generally on the lower part of the screen, that endeavours to recount the original dialogue of the speakers as well as the discursive elements that appear in the image (letters, inserts, graffiti, inscriptions, placards, and the like), and the information that is contained on the soundtrack (songs, voices off).

These restrictions have an impact in the translation leading to text reduction phenomena. According to Díaz Cintas and Remael (2007) the text reduction for subtitling can be either:

(i) partial (implying condensation and reformulation of the information) or

(ii) total (implying omission of information).

2.1. Subtitling and Machine Translation

MT systems developed specifically for the translation of subtitles have to take into account these aspects. The improvement of MT technology in the last years has lead to its successful deployment by increasing the speed and amount of text translated. MT has also been applied to subtitling taking into account its distinctive features (Volk, 2008; Müller and Volk, 2013; Etchegoyhen et al., 2014). Nevertheless, the usage of MT to translate subtitles still implies post-editing the MT output, a process which has been proven to be faster and more productive than translating from scratch (Guerberof, 2009; Zampieri and Vela, 2014).

2.2. Quality Assessment in Human Translation

Manual evaluation and analysis of translation quality (HT and MT) has proven to be a challenging and demanding task. According to Waddington (2001), there are two main

[1] http://hdl.handle.net/11858/00-246C-0000-0023-8D18-7

approaches to human evaluation in translation: a) analytic, and b) holistic.
Analytic approaches (Vela et al., 2014a; Vela et al., 2014b) tend to focus on the description of the translational phenomena observed following a given error taxonomy and, sometimes, by considering the impact of the errors. Holistic approaches (House, 1981; Halliday, 2001) tend to focus on how the translation as a whole is perceived according to a set of criteria established in advance.
Pioneering proposals (House, 1981; Larose, 1987) focus on the errors made in combination with linguistic analysis at textual level (Vela and Hansen-Schirra, 2006; Hansen-Schirra et al., 2006; Vela et al., 2007; Hansen-Schirra et al., 2012; Lapshinova-Koltunski and Vela, 2015).
Williams (1989) goes a step further proposing a combined method to measure the quality of human translations by taking into consideration the severity of the errors. According to him, there are two types of errors:

- major, which is likely to result in failure in the communication, or to reduce the usability of the translation for its intended purpose
- minor, which is not likely to reduce the usability of the translation for its intended purpose; although, it is a departure from established standards having little bearing on the effective use of the translation

Depending on the number and impact of the errors, he proposes a four-tier scale to holistically evaluate human translations:

(i) superior: no error at all, no modifications needed;

(ii) fully acceptable: some minor errors, but no major error, it can be used without modifications;

(iii) revisable: one major error or several minor ones, requiring a cost-effective revision; and

(iv) unacceptable: the amount of revision to fix the translation is not worth the effort, re-translation is required

2.3. Quality Assessment in Machine Translation

The evaluation of machine translation is usually performed by lexical-based automatic metrics such as BLEU (Papineni et al., 2002) or NIST (Doddington, 2002). Evaluation metrics such as Meteor (Denkowski and Lavie, 2014), Asiya (Gonzàlez et al., 2014), and VERTa (Comelles and Atserias, 2014), incorporate lexical, syntactic and semantic information into their scores. More recent evaluation methods are using machine learning approaches (Stanojević and Sima'an, 2014; Gupta et al., 2015; Vela and Tan, 2015; Vela and Lapshinova-Koltunski, 2015) to determine the quality of machine translation. The automatically produced scores have been correlated with human evaluation judgements, usually carried out by ranking the output of the MT system (Bojar et al., 2015; Vela and van Genabith, 2015) or by performing post-editing (Gupta et al., 2015; Scarton et al., 2015; Zampieri and Vela, 2014) on the MT system's output.
Error analysis of MT output is a valuable source of information complementing quality assessment, because it discloses the possible weaknesses of MT systems. Vilar et al. (2006), Farrús et al. (2010) and more recently Lommel et al. (2013)[2], suggest the usage of error typologies to evaluate MT output, furthermore Stymne (2011) introduces a modular MT evaluation tool to handle this kind of error typologies.
The evaluation of machine translated subtitles was performed by Etchegoyhen et al. (2014) on the SUMAT output, by letting professional subtitlers post-edit and rate the post-editing effort. Moreover, the same authors measure the MT output quality by running BLEU on the MT output and hBLEU by taking the post-edited MT output as reference translation, showing that the metrics correlate with human ratings.

2.4. Learner Corpora

Learner corpora annotated with errors have been built before, either by collecting and annotating translations of trainee translators as described by Castagnoli et al. (2006) or by training professional translators as specified by Lommel et al. (2013). Castagnoli et al. (2006) collected human translations of trainee translators and got them annotated by translation lecturers, based on a previously established error annotation scheme. The goal was mainly pedagogical: to learn from translation errors made by students enrolled in European translation studies programmes.
A learner corpus similar to ours has been reported by Wisniewski et al. (2014). Their corpus contains English to French machine translations which were post-edited by students enrolled in a Master's programme in specialised translation. Parts of the MT output have undergone error annotation based on an error typology including lexical, morphological, syntactic, semantic, and format errors, and errors that could not be explained.

3. Building the Corpus

In this section, we describe the corpus: the participants, the materials, the annotation schema, and the tasks carried out to compile it.

3.1. Participants

The participants of the experiments were undergraduate students enrolled in the BA Translation Studies programme at Saarland University. The subtitling corpus is the output of a series of tasks fulfilled during a subtitling course. Metadata about the participants were collected documenting: command of source (English) and target (German) languages, knowledge of other languages, professional experience as translator, and other demographic information such as nationality, gender and birth year.
We collected metadata from 50 students who carried out 52 English to German translations. Table 1 gives an overview over the information collected, showing that most of the students are native or near-native speaker of German (C2[3]

[2]In fact, Lommel et al. (2013) propose an error typology for both HT and MT.

[3]Linguistic competence categories as in the Common European Framework
https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages#Common_reference_levels.

Source language command	C2 (6) C1 (34) B2 (10)
Target language command	C2 (44) C1 (2) B2 (1) B1 (3)
Other languages	Yes (46) No (4)
Professional experience	Yes (10) No (40)
Nationality	German (41) Other (9)
Gender	Female (40) Male (10)
Birth year	1983 – 1994

Table 1: Participants overview.

level), with no professional experience, but good to very good knowledge of English (B2 and C1 level).

3.2. Materials

The source text used for this experiment is the documentary film *Joining the Dots* by Pablo Romero Fresco (Romero Fresco, 2013). A master template of the original subtitles in English was provided by the film maker. Therefore, no spotting was required. Using the same master subtitle template for all translators eased comparability of subtitles across translations and translators. The source text contained 132 subtitles amounting to 1557 words. The students produced the human translations from English into German. The machine translations were produced with SUMAT's online demo (Del Pozo, 2014). The MT output was finally post-edited by translation students.

We arranged the quality assessment annotation as two different tasks: 1) error analysis and 2) evaluation. Each task had its own annotation schema. We were interested in obtaining an analytical description of the translational phenomena observed, and a general idea of the impact of such errors on the translation units taken as a whole.

Annotators were provided with guidelines illustrating where to mark the errors, how to mark the appropriate text spans, and typical cases for each error category. In addition, all annotators practised in class both the translation, the post-editing of the MT output, as well as the error annotation and evaluation of the human and machine translation.

3.3. Error Analysis

We developed an error annotation schema and an evaluation instrument based partly on MQM and Mellange TLC taxonomies. The error annotation schema consists of 4 dimensions: 1) content, 2) language, 3) format, and 4) semiotics.

The first two categories correspond to classical error types described in the literature:

- content: omission, addition, content shift, untranslated, terminology; and
- language: syntax, morphology, function words, orthography.

The last two categories are our contribution aimed at describing specific features of subtitling:

- format: punctuation, font-style, capitalisation, number of characters per line, number of lines per subtitle, number of seconds per subtitle, line breaks, positioning of subtitle, colour of subtitle, audio synchronisation, video synchronisation; and
- semiotics: cases where there is a contradiction between other channels contributing to the meaning of the text and the translation.

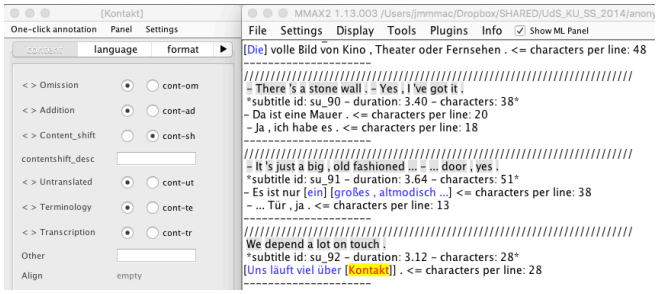

Figure 1: Error analysis with MMAX2.

Table 2 provides an overview on the error annotation scheme used. An additional field labeled *other* was provided for each category, in case annotators found a phenomenon not already listed. Moreover, it was possible to add a description of the annotation to provide a more insightful feedback.

3.4. Evaluation

An approach similar to SICAL ratings described in Williams (1989) was adopted for the evaluation task of both HT and MT. The quality of a translation is measured in four levels in light of its acceptability.

In our case, the labels were:

- perfect
- acceptable
- revisable
- unacceptable

The evaluation was carried out for each dimension of analysis: 1) content, 2) language, 3) format and 4) semiotics.

3.5. Tasks

The students were asked to carry out the following tasks:

- a human translation of the source text (HT)
- a post-edition of the machine translation produced by SUMAT (PE)
- evaluation of:
  - at least one human translation (HT error analysis and evaluation)

Phenomenon	Positive Marker	Negative Marker	To be marked in ST	To be marked in TT	Explanation
CONTENT	lexical semantic level	relation between ST and TT			
Omission	x	x	x		Omission of a word or phrase in TT which was present in ST.
Addition	x	x		x	Addition of a word or phrase in TT, which was not present in ST.
Content Shift	x	x	x	x	The solution in TT is not the same as the element in ST; not the same sense (*contrasense*, *faux sense*, nonsense), condensation and reformulation.
Untranslated	x	x	x	x	Element that should be translated that is not translated; depends on the function, TT, community, target audience.
Terminology	x	x	x	x	Term is used appropriate/inappropriate concerning function, audience and domain of TT or inconsistent within TT.
LANGUAGE	morphosyntactic level	just about TT			
Syntax		x		x	False word order in sentence, syntactic properties do not conform to target language system.
Morphology		x		x	Mistake in agreement, tense, mode, aspect.
Function words		x		x	Wrong usage of prepositions, articles and particles.
Orthography		x		x	Words are not spelled properly.
FORMAL					
Punctuation		x		x	Inappropriate use of punctuation.
Font-style		x		x	Inappropriate font-style.
Capitalisation		x		x	Inappropriate capitalisation.
Nr. of characters/line		x		x	Too many or too few characters/line; ideal: 35 characters/line.
Nr. of lines/subtitle		x		x	Too many or too few lines/subtitle; ideal: 1-2 lines/subtitle .
Nr. of seconds/subtitle		x		x	Too many or too few seconds/subtitle; ideal: 1 second/word, 2 seconds/line, 4 seconds for 2 lines.
Line break		x		x	Inappropriate splitting of lines within subtitle in TT.
Subtitle break		x		x	Inappropriate sentence splitting in TT across subtitles.
Positioning of subtitle		x		x	Inappropriate position of the subtitle on screen; should be centered.
Colour of subtitle		x		x	Subtitle has an inappropriate colour; we do not work with colours.
Audio synchronisation		x		x	Audio stream does not synchronise with the subtitle.
Video synchronisation		x		x	Subtitles do not synchronise with the video; be aware of change of scene, cut.
SEMIOTICS					
Inter-semiotic coherence		x		x	Contradiction/conflict between the translation and any of the other channels that construct meaning together with the subtitles.

Table 2: SubCo error annotation scheme.

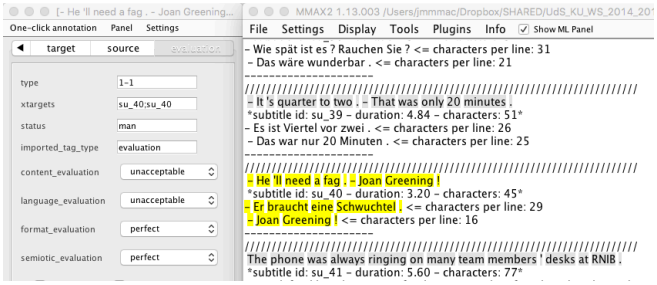

Figure 2: Evaluation with MMAX2.

- the machine translation (MT error analysis and evaluation)
- a post-edition (PE evaluation)

The human translation was done as a homework assignment. It was not a timed translation, students were allowed to make use of any documentation resource, and they employed Aegisub[4] as subtitle editor.
The post-edition of the machine translation produced by SUMAT was performed with PET (Aziz et al., 2012) as a class assignment at the computer labs. Students had 60 minutes to revise the content and linguistic aspects of 132 subtitles. In a second phase, they produced a version where subtitling formal conventions were considered.

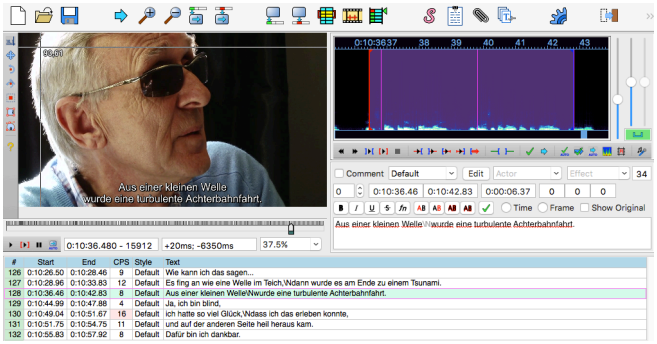

Figure 3: Human translation of subtitles with Aegisub.

The quality assessment of the translations was carried out with MMAX2 (Müller and Strube, 2003). The assessment consisted of two tasks: error analysis and evaluation at subtitle level. Both of them were homework assignments and without any restrictions regarding time and documentation available.
Short supervised training was provided before the quality assessment assignments in order to get the students acquainted with the usage of the annotation schemas and tools.

4. Corpus Details and Statistics

The corpus described here was built over two semesters during a subtitling course offered to students enrolled in the BA Translation Studies programme at Saarland University. Figure 4 depicts the structure of the corpus and Table 1 gives an overview of the students taking part in the courses during the summer and winter terms in 2014.

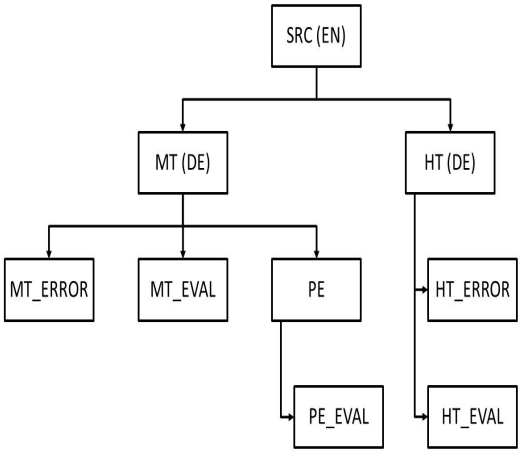

Figure 4: SubCo corpus design.

Table 3 lists the number of tasks collected during the classes.

Task	SS2014	WS2014	Total
Metadata	25	25	50
HT	25	27	52
PE		21	21
HT_ERROR	24		24
HT_EVAL	24	44	68
MT_ERROR	25		25
MT_EVAL	25	21	46
PE_EVAL		21	21

Table 3: Statistics over corpus structure

We collected 50 human and SUMAT produced machine translations. 21 machine translations were post-edited, 25 machine translations and 24 human translations were annotated with errors. The human translations were evaluated 68 times, meaning that some students evaluated more than one translation, but never their own. Machine translations were evaluated 46 times.
The English original subtitles were translated into German with SUMAT and were post-edited by 21 students and annotated in terms of errors by 25 students. These post-edited machine translations were evaluated 21 times.
A preliminary analysis has been carried out on the subset made up of: HT_ERROR, HT_EVAL, MT_ERROR, MT_EVAL for the summer term 2014.
We use box plots to visualise a summary of the distribution underlying the samples and to compare central measure values and spread of the data across groups. Moreover, notched box plots help to check if the differences observed are significant: if the notches of two box plots overlap, there is no evidence that their medians differ (Chambers et al., 1983).
Box plots in Figure 5 illustrate the amount of subtitles per text considered perfect, acceptable, revisable and unacceptable.
HT (white) typically shows a much higher number of perfect subtitles per text than MT (grey). By contrast, MT

[4] http://www.aegisub.org

shows more unacceptable subtitles per text than HT. Subtitles qualified as acceptable exhibit no clear differences, whereas revisable MT segments outnumber HT. Human output shows wider IQR (higher spread of variation) probably because each evaluation is performed on a different target rendering of the source text, while MT evaluation is based on the same target version. The semiotic dimension reveals a different behaviour in contrast, where differences between HT and MT are negligible, intersemiotic coherence errors seem to be quite infrequent.

Humans tend to produce mostly perfect and acceptable subtitles. SUMAT however shows a more homogeneous distribution. All in all, MT output seems to require more revision effort than human translations for all four dimensions.

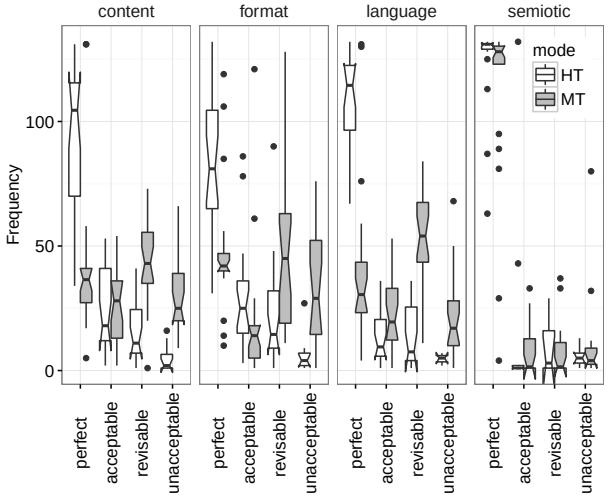

Figure 5: Quality evaluation.

Error Analyses complete our understanding of typical pitfalls for each dimension.

Most frequent content errors (as Figure 6 points up) are content shifts, omissions and additions no matter the mode of translation. By and large, the machine makes more errors than the humans.

Figure 7 shows that the most frequent format error is number of characters per line, specially for machine translations, followed by seconds per subtitle (which captures the ratio characters/second, often been too high if lines are too long). Punctuation and capitalisation errors also have a small share, probably due to the peculiar semiotics of capital letters and some punctuation marks in subtitling, differences are negligible though.

Box plots for language error analysis (see Figure 8) show that SUMAT have in general more difficulties in this area than human beings, but for orthography. The greatest pitfall for MT is syntax, followed by morphology and function words. In all three categories, the differences with human performance are significant.

Figure 9 discloses a very similar behaviour regarding semiotic errors for both modes. In absolute terms, they are negligible.

We conclude this analysis of errors with a methodological remark – *other* is a category barely used. This might indicate that only a few cases were not taken into account by our error taxonomy.

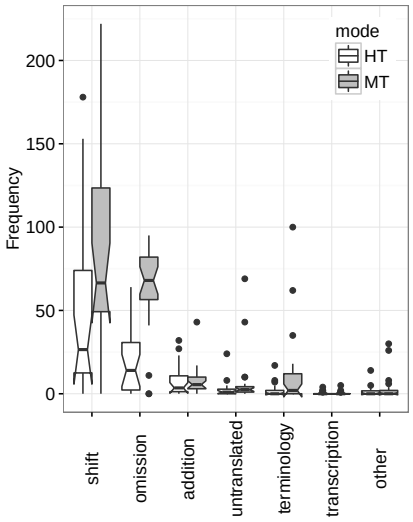

Figure 6: Content Error Analysis.

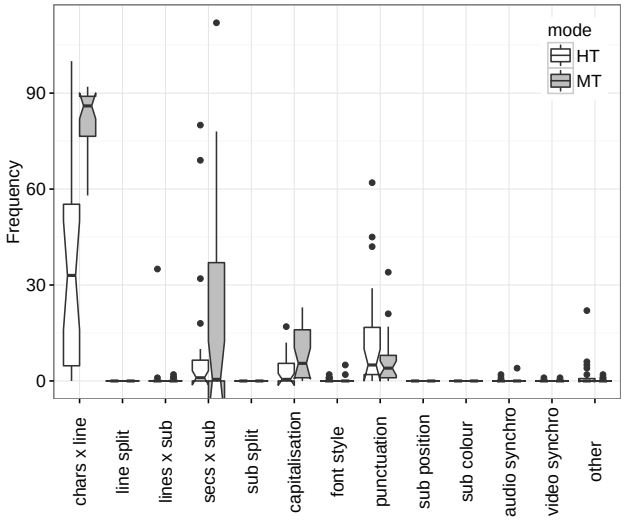

Figure 7: Format Error Analysis.

5. Conclusion

We presented here SubCo, a corpus of human and machine translation subtitles. The machine translation was carried out with SUMAT and post-edited thereafter. Both human and machine translations were annotated in terms of errors. Moreover, human and machine translations as well as the post-editions of SUMAT were manually evaluated. Although human error annotation and evaluation are very time-consuming tasks, we have shown in Section 4. that this kind of data can provide interesting insights on the nature of human and machine translation in general, and subtitling in particular. Therefore, this resource is a valuable contribution for automatic error detection and MT systems developers, who can benefit from this freely available resource.

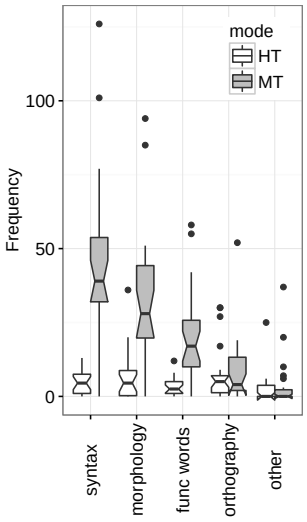

Figure 8: Language Error Analysis.

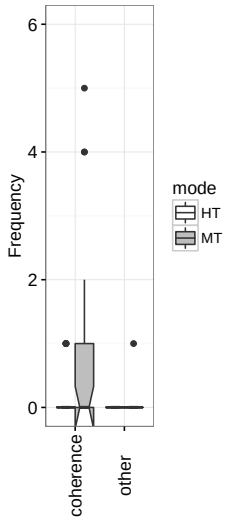

Figure 9: Semiotic Error Analysis.

Future work will involve a thorough evaluation of inter-annotator agreement, and a more fine grained study at subtitle level with a two-fold goal: 1) to identify the most difficult subtitles, and 2) to obtain a more detailed knowledge of the relationship between some types of errors and their impact in the quality of translations.

6. Acknowledgments

This work has partially been supported by the CLARIN-D[5] (Common Language Resources and Technology Infrastructure) project.

7. Bibliographical References

Aziz, W., Castilho, S., and Specia, L. (2012). PET: a Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–2987, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N., and Volanschi, A. (2006). Designing a learner translator corpus for training purposes. In *Proceedings of Teaching and Language Corpora Conference TaLC 2006*, pages 1–19, Université Paris VII.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth.

Comelles, E. and Atserias, J. (2014). VERTa Participation in the WMT14 Metrics Task. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 368–375, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Del Pozo, A. (2014). SUMAT Final Report. Technical report, VICOMTECH.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Díaz Cintas, J. and Remael, A. (2007). *Audiovisual Translation: Subtitling*. Translation Practices Explained. St. Jerome, Manchester, UK.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technologies (HLT)*, pages 138–145.

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Loenhout, G. V., del Pozo, A., Maucec, M. S., Turner, A., and Volk, M. (2014). Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 46–53, Reykjavik, May. European Language Resources Association (ELRA).

Farrús, C. M., Ruiz Costa-Jussà, M., Mariño Acebal, J. B., and Rodríguez Fonollosa, J. A. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *14th Annual Conference of the European Association for Machine Translation*, pages 167–173.

Gonzàlez, M., Barrón-Cedeño, A., and Màrquez, L. (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. In *Proceedings of the 9th Workshop on Statistical Machine Translation (WMT)*, pages 394–401, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Guerberof, A. (2009). Productivity and Quality in the Post-editing of Outputs from Translation Memories and Ma-

[5] http://de.clarin.eu

chine Translation. *International Journal of Localization*, 7(1).

Gupta, R., Orăsan, C., and van Genabith, J. (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Halliday, M. (2001). Towards a theory of good translation. In Erich Steiner et al., editors, *xploring Translation and Multilingual Text Production: Beyond Content*, pages 13–18. Berlin and New York: Mouton de Gruyter.

Hansen-Schirra, S., Neumann, S., and Vela, M. (2006). Multi-dimensional Annotation and Alignment in an English-German Translation Corpus. In *Proceedings of the Workshop on Multi-dimensional Markup in Natural Language Processing (NLPXML-2006) held at EACL*, April.

Hansen-Schirra, S., Neumann, S., and Steiner, E. (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. de Gruyter, Berlin, New York.

House, J. (1981). *A Model for Translation Quality Assessment*. Tübinger Beiträge zur Linguistik. Gunter Narr Verlag.

Lapshinova-Koltunski, E. and Vela, M. (2015). Measuring 'Registerness' in Human and Machine Translation: A Text Classification Approach. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation (DiscoMT)*, pages 122–131, Lisbon, Portugal, September. Association for Computational Linguistics.

Larose, R. (1987). *Théories contemporaines de la traduction*. Presses de l'Université du Québec.

Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2013). Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35 (ASLIB)*, London, UK, November.

Müller, C. and Strube, M. (2003). Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, Japan.

Müller, M. and Volk, M. (2013). Statistical Machine Translation of Subtitles: From OpenSubtitles to TED. In Iryna Gurevych, et al., editors, *Language Processing and Knowledge in the Web SE - 14*, volume 8105 of *Lecture Notes in Computer Science*, pages 132–138. Springer, Berlin/Heidelberg.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Romero Fresco, P. (2013). Accessible filmmaking: Joining the dots between audiovisual translation, accessibility and filmmaking. Number 20, pages 201–223.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015). Searching for context: a study on document-level labels for translation quality estimation. *The 18th Annual Conference of the European Association for Machine Translation (EAMT 2015)*.

Stanojević, M. and Sima'an, K. (2014). BEER: BEtter Evaluation as Ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Stymne, S. (2011). Blast: A tool for error analysis of machine translation output. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 56–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vela, M. and Hansen-Schirra, S. (2006). The Use of Multi-level Annotation and Alignment for the Translator. In *Proceedings of Translating and the Computer 28 (ASLIB)*, November.

Vela, M. and Lapshinova-Koltunski, E. (2015). Register-Based Machine Translation Evaluation with Text Classification Techniques. In *Proceedings of the 15th Machine Translation Summit*, pages 215–228, Miami, Florida, November. Association for Machine Translations in the Americas.

Vela, M. and Tan, L. (2015). Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*, pages 402–410, Lisbon, Portugal, September. Association for Computational Linguistics.

Vela, M. and van Genabith, J. (2015). Re-assessing the WMT2013 Human Evaluation with Professional Translators Trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 161–168, May.

Vela, M., Neumann, S., and Hansen-Schirra, S. (2007). Querying Multi-layer Annotation and Alignment in Translation Corpora. In *Proceedings of the Corpus Liguistics Conference (CL)*, July.

Vela, M., Schumann, A.-K., and Wurm, A. (2014a). Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 47–56, April.

Vela, M., Schumann, A.-K., and Wurm, A. (2014b). Human Translation Evaluation and its Coverage by Automatic Scores. In *Proceedings of the LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation (MTE)*, pages 20–30, May.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. pages 697–702.

Volk, M. (2008). The Automatic Translation of Film Subtitles. A Machine Translation Success Story? In Joakim Nivre, et al., editors, *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, number 7 in Studia Linguistica Upsaliensia, pages 202–214. Acta Universitatis Upsaliensis.

Waddington, C. (2001). Different Methods of Evaluating Student Translations: The Question of Validity. *Meta*,

46(2):311–325.

Williams, M. (1989). The Assessment of Professional Translation Quality: Creating Credibility out of Chaos. *TTR : traduction, terminologie, rédaction*, 2:13–33.

Wisniewski, G., Kübler, N., and Yvon, F. (2014). A Corpus of Machine Translation Errors Extracted from Translation Students Exercises. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3585–3588, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Zampieri, M. and Vela, M. (2014). Quantifying the Influence of MT Output in the Translators' Performance: A Case Study in Technical Translation. In *Proceedings of the EACL Workshop on Humans and Computer-assisted Translation (HaCat)*, pages 93–98, April.